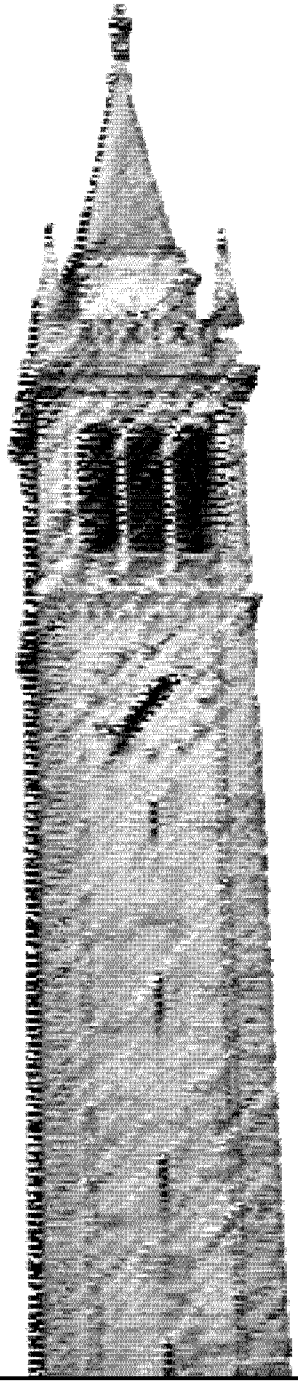
Magnetic Resonance Image Reconstruction with Greater Fidelity and Efficiency



Ke Wang

Electrical Engineering and Computer Sciences
University of California, Berkeley

Technical Report No. UCB/EECS-2023-178
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2023/EECS-2023-178.html

May 16, 2023

Magnetic Resonance Image Reconstruction with Greater Fidelity and Efficiency

By

Ke Wang

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering - Electrical Engineering and Computer Sciences

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Michael Lustig, Chair
Professor Stella Yu, Co-chair
Professor Alexei Efros
Associate Professor Moriel Vandsburger

Spring 2023

Magnetic Resonance Image Reconstruction with Greater Fidelity and Efficiency

Abstract

Magnetic Resonance Image Reconstruction with Greater Fidelity and Efficiency

By

Ke Wang

Doctor of Philosophy in Engineering - Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor Michael Lustig, Chair

Professor Stella Yu, Co-chair

Magnetic resonance imaging (MRI) is an effective imaging modality offering tremendous benefits to both science and medicine. It provides exceptional contrast for visualizing soft tissue, can capture images from any orientation, and does not involve any ionizing radiation. Its remarkable versatility enables a wide range of applications, including assessing blood flow, imaging brain activity with functional MRI (fMRI), and quantifying susceptibility mapping, ushering in a new era of clinical diagnosis and brain research.

However, due to its physics limitations, acquiring MRI data is inherently time-consuming, which significantly extends scan times and limits throughput in hospitals. As a result, there is great interest in reconstructing diagnostic-quality images from limited measurements (k-space data) to shorten scan times. For instance, parallel imaging (PI) capitalizes on spatially sensitive receive coil arrays to simultaneously acquire multiple MRI measurements. Compressed Sensing (CS) techniques have been employed to iteratively reconstruct undersampled data into high-quality images by utilizing sparse priors. More recently, end-to-end deep learning (DL) based reconstruction techniques have been introduced, leveraging deep neural networks to learn the reconstruction pipeline directly from extensive training datasets, rather than relying on hand-crafted prior knowledge.

Although DL-based methods have demonstrated significant success surpassing PI and CS capabilities, several challenges persist that limit the fidelity and efficiency, for example: 1) Loss functions used in DL-based reconstruction are mostly hand-crafted, either pixel-wise (*e.g.,* ℓ_1,ℓ_2 losses) or based on local statistics (*e.g.* SSIM loss), inadequately capture perceptual information, leading to compromised image quality and blurring; 2) Memory constraints during network training restrict the applicability of DL reconstruction for high-dimensional MRI (e.g., 2D+time, 3D, 3D+time); 3) The confidence or reliability of reconstructed structures remains insufficiently investigated, posing a challenge for DL-based approaches in clinical

applications. 4) Unlike natural images, MRI data is inherently complex-valued and faces challenges due to the limited availability of fully-sampled ground truth. This constraint inevitably restricts the applications of deep learning-based MRI techniques to tasks without access to adequate ground truth.

In this dissertation, we introduce a series of projects aimed at overcoming existing obstacles and achieving enhanced fidelity and efficiency in Magnetic Resonance (MR) image reconstruction.

Chapter 3 begins by reconstructing high-fidelity contrast-weighted images from highly undersampled Magnetic Resonance Fingerprinting (MRF) scan. It introduces a supervised learning method that directly synthesizes contrast-weighted images (T1-weighted, T2-weighted, and FLAIR) from an MRF scan. This technique generates multi-contrast images with significantly reduced scan times, as detailed in the [paper link].

Chapters 4-6 feature physics-informed DL-based reconstruction from undersampled k-space data. First, A novel patch-based Unsupervised Feature Loss (UFLoss) is proposed as a novel perceptual loss function and incorporated into the training of DL-based reconstruction frameworks in order to preserve perceptual similarity and high-order statistics ([paper link]). Next, I employ our previously proposed memory-efficient learning framework to minimize the memory required for backpropagation, facilitating the training of DL-based unrolled reconstructions for large-scale 3D MRI and 2D+time cardiac cine MRI ([paper link]). Then, I present our uncertainty estimation framework to identify when and where a reconstruction model is producing potentially misleading results. Our framework produces confidence intervals at each pixel of a reconstruction image with a rigorous finite-sample statistical guarantee. Our in-vivo knee and brain results probe the quality of our uncertainty estimation model, which allows us to identify specific regions where the model performs poorly ([abstract link]).

A distinctive aspect of MRI lies in its inherently complex-valued data. The final section of this dissertation concentrates on the representation learning of complex-valued data. In contrast to deep learning applied to natural images, MRI faces the challenge of a scarcity of fully-sampled ground truth and well-annotated data. To tackle this challenge, I introduce Complex-valued Scattering Representation (CSR) as a universal complex-valued representation, which so far demonstrates superior performance in both real-valued (*e.g.,* RGB image) and complex-valued (*e.g.,* MRI) image classification tasks compared to its counterparts, particularly when training samples are limited. Although this dissertation has not applied CSR to DL-based reconstruction, it represents a promising direction for future research.

Collectively, these approaches embody the central theme and progress toward MR image reconstruction with high fidelity, high efficiency, and high reliability.

To my parents, my advisors, Yuhan, and my friends who always support me!

Contents

List of Figures

List of Tables

Acknowledgments

Every time I open a dissertation, my initial instinct is to skip directly to the acknowledgments. Delving into these heartfelt expressions allows me to catch a glimpse of the author's Ph.D. journey, an aspect that is not fully captured by their research publications. It is difficult to believe that after five incredible years, I have finally reached this milestone and am writing my own acknowledgments.

I would like to first express my deepest gratitude to my advisors Miki and Stella, for their continuous support and guidance throughout my Ph.D. journey. Miki, you have been my guiding light throughout my Ph.D. journey. As the best advisor on this planet (and in my heart), you possess an abundance of brilliant ideas and invaluable insights. You have provided us with ample freedom to explore our research directions while consistently offering your support and invaluable hands-on advice. You have set high standards for your students, which at times made me feel disappointed in myself. Nevertheless, in the end, I truly appreciate your push to take us out of our comfort zones and help us grow into better researchers. I started collaborating with Stella during my first year and then became your student. Having you as my advisor has been one of the most fortunate aspects of my Ph.D. journey. You are always responsive and knowledgeable about every detail of the projects. I wish you all the best in Michigan; you are a hero in my heart.

I would like to extend my heartfelt gratitude to my qualifying exam and dissertation committee members: Chunlei Liu, Moriel Vandsburger, and Alexei Efros. Chunlei, as the first professor I met at Berkeley, you welcomed me with open arms and set the stage for a fantastic journey. Your expertise in MRI and brain imaging has been an invaluable resource throughout my studies. Moriel, I will always cherish our memorable running times in Wisconsin. If I were to retake your class, I promise I would remain fully engaged and never fall asleep for even a single second. I would also like to thank Shirley Salanio, for always being there for us and showing such genuine care and concern. I would love to thank my undergraduate advisor, Kui Ying for introducing me to the amazing field of MRI.

Throughout my Ph.D., I have been fortunate to work and collaborate with numerous outstanding researchers. Prof. Ana Arias, I enjoyed teaching with you, and thank you for inviting me to your group trips! Prof. Shreyas Vasanawala, your insights, and expertise in clinical MRI have truly guided my research direction. I greatly appreciate our conversations and the invaluable knowledge you have shared with me! I would like to thank Professors Peder Larson and Kevin Johnson for our discussions in the Lung MRI meetings. I would like to thank my collaborators from Berkeley: Utkarsh Singhal, Anastasios Angelopoulos, Amit Kohli, my collaborators from Philips: Mariya Doneva, Jakob Meineke, and Thomas Amthor, my collaborators from GE: Uri Wollner, Rafi Brada, Anja Brau, Xucheng Zhu, Marc Lebel, and Graeme McKinnon, and my collaborators from Stanford: Christopher Sandino, Arjun Desai. My Ph.D. work could not have been accomplished without your invaluable help and discussions.

Upon joining Miki's group (MikGroup), I immediately felt embraced by an incredibly welcoming and collaborative environment. To my great mentor and collaborator, Jon, you

are so knowledgeable, friendly, and always excited about new ideas. When you left Berkeley, I felt a mix of emotions - sad to see you go, yet happy and excited for your new chapter. I can't wait to see more amazing works from your group, and I will always be rooting for you! To Frank, the person who introduced me to computational MRI, I sincerely appreciate your guidance and help through my first year. I miss so much the times we went to Aki's together, and wish you all the best! Zhiyong, you brought so much joy and positivity to the group; I wish you great success in SJTU. To my big brother Karthik, unfortunately, I wasn't able to dunk before my graduation; that's a bummer, maybe I have already passed the golden time. However, the times we spent playing basketball, climbing, and hanging out are some of the most unforgettable memories of my Ph.D. journey. To Xucheng, even though we are not in the same research group, we have spent a lot of time together. You have always supported and encouraged me, and I am grateful for your friendship. To the other senior members of MikGroup: Anita Flynn, Michael Kellman, Alan Dong, Wenwen Jiang, Volkert Roeloffs, Efrat Shimron, and Gopal Nataraj, I would like to express my sincere appreciation for the interactions and discussions we have had throughout the years.

During my years at Berkeley, I have been so fortunate to meet so many wonderful friends and peers. I would like to thank MikGroup members: Ekin, Suma, Shreya, Alfredo, Rebekah, Julian, Kaite, Daniel, Jolene, Celine, Catherine, Ana, and others; Friends from Stella's group: Yubei Chen, Frank Wang, Daniel, Peter Ren, and Peter Wang; Friends from Chunlei's group: Jingjia, Victor, Tanya, and Zoe. and my cherished friends: Efe, and Xu Shen.

Dear Ekin, meeting you has been one of the greatest blessings of my past few years. Your kindness, patience, and willingness to help have made a significant impact on my life. We have shared so many great memories together, from seeing the Warriors lose by 30+ points to catching ISMRM deadlines and indulging in amazing food. I wish you all the happiness in life and a successful career ahead!

Dear Suma, it has been a great fortune to start our Ph.D. journey together. We have passed many milestones together and witnessed each other's growth. I was thrilled to see your excitement and passion when you began your BPT projects. I was so proud of being your colleague when you received the well-deserved recognition. You are one of the most self-disciplined people I know, and I wish you all the best in your future endeavors. Please know that I will always be there for you if you need me!

Dear Shreya, I still vividly remember the first time we met - you were giving a talk at our group meeting. At that time, I never could have imagined that we would become such close friends over the next five years. Every time I interact with you, I feel incredibly comfortable. You have a warm and welcoming personality, like sunshine, that brightens the mood of those around you.

Dear Alfredo, your professionalism has had a profound impact on me. It has been a great pleasure working with you over the past few years. We have collaborated on many projects, each of which has become a wonderful memory. Your dedication, hard work, and enthusiasm are truly inspiring. Thank you for being an excellent collaborator and friend!

To Jingjia, both of us did our undergrads in China, which makes us quickly become close friends! I always enjoyed chatting with you and your positive attitude never failed to brighten

up my day. I wish you all the best in your future endeavors in New York! Remember, Qiang Ren Suo Nan - strong people can always tackle challenges!

To my friend Rebekah, I have always been in awe of your incredible musical talents. Perhaps, one day after retiring from your amazing work as a scientist and engineer, you could consider pursuing a career as a musician. To Ruiming, I always enjoy spending time with you, whether it's playing basketball, grabbing a meal, or just hanging out. Let's continue to make great memories together! To Julian, you brought so much energy and positivity to the group. I can't wait to see what amazing things you'll achieve with your golden hands in the future! To Alisha, chatting with you has always been a truly pleasant experience! Your warm and positive personality always enlightens my days, let's catch up in Houston!

To Banghua and Yu, it's hard to believe that we have known each other for more than seven years, from Tsinghua to Berkeley. We have shared so many great memories together, and I will never forget our Alaska trip and the amazing northern lights we witnessed!

During my Ph.D., I had the opportunity to complete two internships at Adobe, working in the EPG team and Adobe Research. These internships were some of the most rewarding experiences of my Ph.D. journey. I would like to express my sincere gratitude to Michaël Gharbi, my mentor at Adobe Research. Working with you has been one of the most memorable experiences of my Ph.D. journey. As a remarkable mentor and cherished friend, you possess a unique ability to uplift and inspire everyone and everything around you. I'm very happy to see our paper published and I believe our paths will cross again in the future. I would love to thank my other mentors and collaborators at Adobe: Zichuan Liu, Xin Lu, Zhihao Xia, He Zhang, and Eli Shechtman. Working with all of you has been a fantastic learning experience!

I would like to express my deepest love and gratitude to my parents, Yuhong Liu and Xiusheng Wang. Without your endless support and love, I would not be the person I am today. You have always been there for me, encouraging me to pursue my dreams! Though I rarely say it, I want you to know how much I love you, Mom and Dad. I would like to extend a heartfelt thank you to my wonderful extended family in the US, my Aunt Hengrui, my Uncle Jun, and my three lovely cousins! Thank you for hosting me during Thanksgiving and engaging in countless conversations over the past few years. Your warmth and support have truly made me feel at home in this country.

Lastly, I want to dedicate this spot to my partner Yuhan. Thank you for being my constant support and source of strength throughout the past two years. Being with you eliminates my negative energies and infuses me with positivity to take on any challenges ahead. You are intelligent, caring, and dedicated. I am always proud of you for every single step you have taken and will take in the future. I'm excited and looking forward to our life together!

Ke Wang
May 2, 2023

Chapter 1

Introduction

1.1 Background

Magnetic resonance imaging (MRI) is a highly effective imaging modality that offers immense advantages to both scientific and medical fields. It delivers outstanding contrast for visualizing soft tissue structures, enables image acquisition from any orientation, and unlike X-Ray, Computed Tomography (CT) and Positron Emission Tomography (PET), operates without the use of ionizing radiation. MRI's extraordinary versatility supports an extensive array of applications. These include evaluating blood flow dynamics, monitoring brain activity using functional MRI (fMRI), and quantifying susceptibility mapping. Consequently, MRI has paved the way for a new era in clinical diagnosis and brain research, revolutionizing our understanding of human anatomy and physiology while enhancing patient care.

Despite its benefits, the inherent time-consuming nature of MRI data acquisition, owing to physics limitations, considerably prolongs scan times and restricts throughput in healthcare settings. A standard MRI scan may require 20-60 minutes, whereas a full-body CT scan can be completed in just a matter of seconds.

As a result, there is considerable interest in reconstructing diagnostic-quality images from a limited number of measurements with the aim of decreasing scan times and improving the overall efficiency of MRI procedures. It is important to note that MRI data is acquired in the frequency domain, commonly referred to as k-space. For instance, parallel imaging (PI) [107, 90, 36] capitalizes on spatially sensitive receive coil arrays to simultaneously acquire multiple MRI measurements. Compressed Sensing (CS) [64] techniques have been employed to iteratively reconstruct under-sampled data into high-quality images by utilizing sparse priors. More recently, end-to-end deep learning (DL)-based reconstruction methods [16, 68, 92, 99, 37, 2, 110] have been proposed to learn the regularization terms directly from a large training dataset.

Although DL-based methods have demonstrated significant success surpassing PI and CS capabilities, several challenges remain that limit the image fidelity and efficiency, for example: 1) Loss functions used in DL-based reconstruction are mostly hand-crafted, either

pixel-wise (*e.g.,* ℓ_1,ℓ_2 losses) or based on local statistics (*e.g.* SSIM loss [130]), inadequately capture perceptual information, leading to compromised image quality and blurring [44, 2, 37, 125]; 2) Memory constraints during network training restrict the applicability of DL reconstruction for high-dimensional MRI (*e.g.*, 2D+time, 3D, 3D+time) [95, 127]; 3) The confidence or reliability of reconstructed structures remains insufficiently investigated, posing a challenge for DL-based approaches in clinical applications [129]. 4) Unlike natural images, MRI data is inherently complex-valued and faces challenges due to the limited availability of fully-sampled ground truth. This constraint inevitably restricts the applications of deep learning-based MRI techniques to tasks without access to adequate ground truth.

1.2 Contribution

This dissertation introduces a series of projects aimed at overcoming existing obstacles and achieving enhanced fidelity and efficiency in MR image reconstruction. Chapter 3 highlights DL-based reconstruction from highly under-sampled MRF scans, introducing a supervised learning-based method that directly synthesizes contrast-weighted images (T1-weighted, T2-weighted, and FLAIR) from a single MRF scan. This approach produces multi-contrast images with substantially reduced scan time ([paper link] [125]).

In Chapter 4, a novel patch-based Unsupervised Feature Loss (UFLoss) is introduced as a perceptual loss function, incorporated into the training of DL-based reconstruction frameworks to preserve perceptual similarity and high-order statistics ([paper link] [126]).

Chapter 5 employs a previously proposed memory-efficient learning framework to reduce backpropagation memory requirements, enabling DL-based unrolled reconstructions for large-scale 3D MRI and 2D+time cardiac cine MRI ([paper link] [127]).

In Chapter 6, an uncertainty estimation framework is presented, which identifies when and where a reconstruction model produces potentially misleading results. The framework generates confidence intervals for each pixel in a reconstructed image, providing rigorous finite-sample statistical guarantees. In-vivo knee and brain results assess the quality of the uncertainty estimation model, highlighting specific regions with poor performance ([abstract link] [129]).

One of MRI's unique characteristics is its inherently complex-valued data. The final section (Chapter 7) of this dissertation focuses on complex-valued data representation learning. Unlike deep learning in natural images, MRI faces challenges due to limited fully-sampled ground truth and well-annotated data. To address this, I introduce Complex-valued Scattering Representation (CSR) as a universal complex-valued representation. It has demonstrated superior performance in both real-valued (*e.g.*, RGB image) and complex-valued (*e.g.*, MRI) image classification tasks compared to alternatives, especially when training samples are limited. While CSR has not been applied to DL-based reconstruction in this dissertation, it offers a promising direction for future research.

These methods collectively form the central theme, progressing toward advanced MR image reconstruction, characterized by high fidelity, high efficiency, and high reliability.

1.3 Outline

The organization of this dissertation is presented as follows:

Chapter 2: Overview of MR imaging and reconstruction

This chapter offers a comprehensive overview of MRI imaging and reconstruction, serving as the foundation for subsequent chapters. It begins with the fundamental physics principles and signal equations of MRI. Next, MRI reconstruction is introduced as a generalized inverse problem, delving into several representative reconstruction techniques: Parallel Imaging, Compressed Sensing, and ultimately, the focus of this thesis, DL-based MRI reconstruction.

Chapter 3: Direct Contrast Synthesis from MR Fingerprinting

This chapter showcases DL-based reconstruction from highly under-sampled MRF scans. Here, I introduce a supervised learning-based method (N-DCSNet) that directly synthesizes contrast-weighted images (T1-weighted, T2-weighted, and FLAIR) from a single, short MRF scan. N-DCSNet not only significantly reduces scan time but also has the ability to inherently mitigate slice in-flow artifacts and spiral off-resonance blurring.

Chapter 4: Unsupervised Feature Loss for DL-based MRI reconstruction

This chapter presents Unsupervised Feature Loss (UFLoss), a novel patch-based unsupervised learning-based feature loss, designed to overcome the limitations of existing handcrafted loss functions (*i.e.*, their inability to capture high-level perceptual information). UFLoss improves the training of DL-based reconstruction methods by enabling them to capture more detailed textures, finer features, and sharper edges. This results in a higher overall image quality within the context of DL-based reconstruction frameworks.

Chapter 5: Memory-efficient learning for high-dimensional MRI reconstruction

In this chapter, I utilize our previously proposed Memory-Efficient Learning (MEL) framework [51] to substantially decrease the Graphic Processing Unit (GPU) memory consumption during the backpropagation of unrolled networks. MEL facilitates the training of high-dimensional MRI reconstruction (*e.g.*, 3D MRI, 2D+time cardiac cine MRI) on a 12 GB GPU, significantly alleviating the computational burden associated with high-dimensional DL-based MRI reconstruction.

Chapter 6: Rigorous uncertainty estimation for MRI reconstruction

In this chapter, I propose a rigorous uncertainty estimation framework to identify when and where a reconstruction model is producing potentially misleading results. Specifically, our framework produces confidence intervals at each pixel of a reconstruction image such

that $1 - \alpha$ of these intervals contain the true pixel value with high probability (typically, $\alpha = 0.05$). Without any constraints on the reconstruction model, our framework acts as a plug-and-play module, and may significantly improve the accuracy of the diagnosis and clinical interpretation of DL-based reconstructions.

Chapter 7: Complex-valued Scattering Representations

In contrast to deep learning applied to real-valued natural images, MRI deals with inherently complex-valued images and faces challenges arising from the limited availability of fully-sampled ground truth and well-annotated data. To address this, Chapter 7 introduces Complex-valued Scattering Representations (CSR) as a universal complex-valued representation. CSR has exhibited superior performance in both real-valued (*e.g.*, RGB images) and complex-valued (*e.g.*, MRI) image classification tasks compared to its counterparts, especially when the number of training samples is limited.

Chapter 8: Summary and future work

This chapter provides a summary of the approaches introduced throughout this dissertation and delineates potential avenues for future research.

Chapter 2

MR Imaging and Reconstruction

In this chapter, the objective is to provide a comprehensive overview of MR imaging and reconstruction, which will facilitate a more thorough understanding of the subsequent chapters. We begin by delving into the fundamental MRI signal equation and the process of reconstruction. Following that, we introduce two conventional image reconstruction approaches: parallel imaging and compressed sensing. Lastly, we present a review of DL-based MRI reconstruction approaches and their remaining challenges.

For readers who are interested in diving deeper into the subject matter, we recommend the following resources:

- Prof. Nishimura's book, Principles of Magnetic Resonance Imaging [75], offers a thorough understanding of MRI physics and basic reconstruction techniques.
- Prof. Zhi-pei Liang and Prof. Paul C. Lauterbur's book, Principles of Magnetic Resonance Imaging: *A Signal Processing Perspective* [60], provides a more mathematical exploration of MRI from a signal processing standpoint.
- Dr. Sandino's review paper, Compressed Sensing: From Research to Clinical Practice with Deep Neural Networks [96], presents a comprehensive introduction to deep learning-based reconstruction methods in the context of compressed sensing.

These resources collectively provide a well-rounded understanding of the field, covering various aspects of MRI reconstruction, from physics and basic principles to mathematical perspectives and cutting-edge deep learning approaches.

2.1 MR Imaging

MR physics

MRI relies on the principles of nuclear magnetic resonance (NMR) to generate detailed images of the internal structures of the human body.

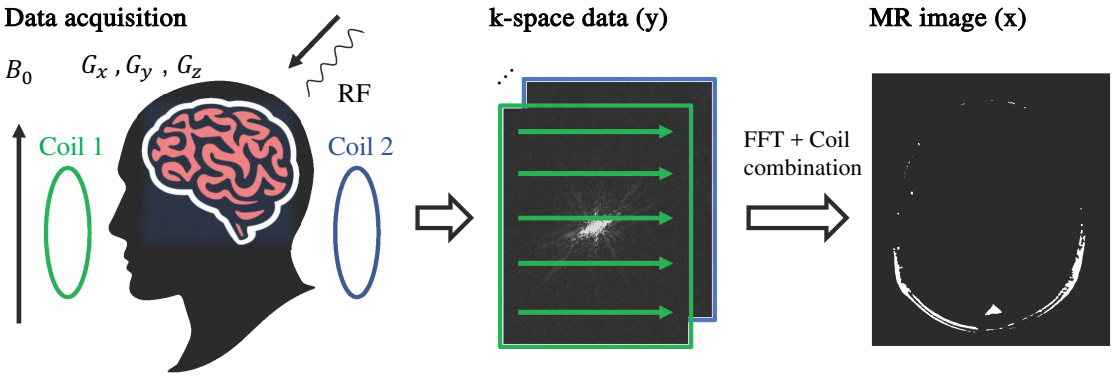


Figure 2.1: **Overview of MR physics and image reconstruction.** An MRI scanner uses a strong magnetic field B_0 and radiofrequency (RF) signals to excite protons in the body. These protons emit signals that are encoded by gradient fields (G_x, G_y, G_z) and detected by coils placed outside the body. These signals reside in the spatial frequency domain (*i.e.*, k-space), representing the image's Fourier coefficients. They are measured, digitized, and sent to a computer for reconstruction using a discrete Fourier transform and coil combination steps, resulting in a reconstructed MR image.

At its core, MR physics revolves around the interaction between the magnetic moments of atomic nuclei, predominantly hydrogen nuclei in water molecules, and the externally applied magnetic field. As shown in Figure 2.1, when placed in a strong static magnetic field (B_0), the nuclear spins of hydrogen atoms align either parallel or anti-parallel to the direction of the field, creating a net magnetization.

Upon the application of a radiofrequency (RF) pulse, the magnetization vector is perturbed, effectively mixing the longitudinal and transverse components. Following this excitation, the precession generates a varying magnetic field in the transverse direction. The changes in the magnetic field are then detected by the receiving coils through induction.

3D Magnetic field gradients (G_x, G_y, G_z) that vary linearly in space are employed to establish a spatial relationship with precession frequency. Consequently, magnetization is spatially distinguished by associating frequency with the position in a linear manner within the received signal. Hence, the received signal measures the spatial Fourier transform of the object being imaged. In the realm of MRI, we call the spatial frequency domain, where the MR signal resides, k-space.

Once we acquire the k-space data, and assume the data is fully-sampled, we can apply an inverse Fourier transform and coil combination to reconstruct MR images.

Image contrast

One of the key features of MRI is its ability to generate images with varying contrast by programming the scanner using different sequence designs, which allows for better differentiation of various tissues and structures within the body.

Image contrast in MRI is primarily governed by biophysical tissue properties, such as proton density (PD), longitudinal/transverse relaxation (T1/T2), magnetic susceptibility, and diffusion. These parameters offer valuable insights into tissue composition and microstructure, serving as excellent biomarkers for the diagnosis and assessment of various diseases.

By adjusting the imaging sequence, MRI can emphasize specific tissue properties to generate a range of contrast types, including T1-weighted, T2-weighted, proton density-weighted, Fluid-attenuated inversion recovery (FLAIR), diffusion-weighted, and susceptibility-weighted images. Each of these contrast types provides unique information about the underlying tissue characteristics, enabling clinicians and researchers to identify pathological changes and monitor disease progression effectively. Figure 2.2 presents representative image contrasts from a 2D brain slice.

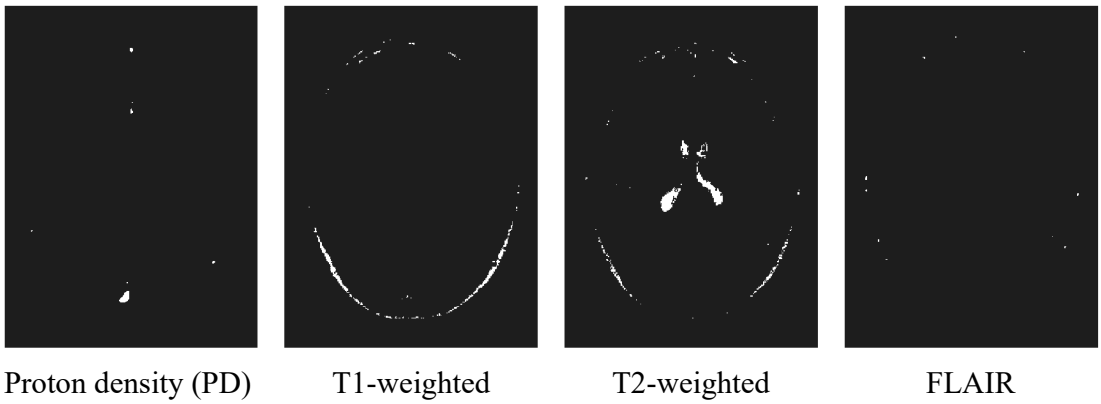


Figure 2.2: **Example of MRI contrasts for a 2D brain slice.** From left to right, we showcase four distinct image contrasts: Proton Density (PD), T1-weighted, T2-weighted, and FLAIR. Each of these contrasts offers unique information about the underlying tissue properties for clinical diagnosis.

In clinical settings, multiple imaging contrasts are often acquired to gather diverse and complementary information about the patient's anatomy and potential abnormalities. By combining various contrasts, healthcare professionals can obtain a more comprehensive understanding of the underlying tissue properties, which helps them make more accurate diagnoses, assessments, and treatment plans.

However, acquiring multiple imaging contrasts significantly prolongs the scan time, requiring patients to remain still for tens of minutes and therefore hindering the scanner throughput.

In Chapter 3 of this dissertation, we introduce a novel deep learning-based reconstruction approach to generate multiple MR imaging contrasts from a single acquisition, significantly reducing scan time while preserving image quality.

MRI signal equations

Derived from MR physics and taking into account multiple receiving coils, the linear relationship between the k-space signal $s_i(t)$, obtained from the i-th coil, and the image space can be described as:

$$s_i(t) = \int_{\mathbf{r}} M_{xy}\mathbf{S}_i(\mathbf{r})(\mathbf{r}, t)e^{-j2\pi\mathbf{k}(t)\cdot\mathbf{r}}d\mathbf{r} + w(t), \tag{2.1}$$

where in our formulation, $s_i(t) \in \mathbb{C}$ represents the complex-valued acquired signal from the i-th coil at time t, M_{xy} denotes the transverse component of the magnetization at position $\mathbf{r}$ and time t, $\mathbf{S}_i(\mathbf{r})$ represents the coil sensitivity from the i-th coil at position $\mathbf{r}$, $j = \sqrt{-1}$, $\mathbf{k}(t)$ corresponds to the k-space encoding, and $w(t)$ is complex-valued Gaussian noise.

By discretizing the above equation, the MR signal equation can be written using a matrix form:

$$\mathbf{y} = \mathbf{PFSx} + \mathbf{w}, \tag{2.2}$$

where $\mathbf{x}$ denotes the image, $\mathbf{y}$ refers to the acquired k-space data, $\mathbf{F}$ symbolizes the Fourier transform operator, $\mathbf{S}$ corresponds to the sensitivity maps, and $\mathbf{w}$ is the noise component. We also introduce $\mathbf{P}$ as the operator that selects the acquired k-space samples. For simplicity, we define the forward model $\mathbf{E} = \mathbf{PFS}$.

It's worth noting that, in our formulation, $\mathbf{P}$ is shared across channels, and $\mathbf{w}$ is white Gaussian and uncorrelated across channels.

2.2 MR reconstruction

Given the forward model $\mathbf{y} = \mathbf{Ex}$, the objective of MRI reconstruction is to obtain an image, denoted by $\hat{\mathbf{x}}$, that closely approximates the underlying image $\mathbf{x}$, using the acquired k-space measurements $\mathbf{y}$. This process is an inverse problem.

Without relying on additional prior knowledge, the system encoding operator $\mathbf{E}$ needs to be critical or overdetermined to form a well-posed inverse problem. For fully-sampled k-space and normalized sensitivity maps, we have $\mathbf{E^H E} = \mathbf{S^H F^H P^H PFS} = \mathbf{I}$, where $^\mathbf{H}$ represents the conjugate transpose. Since $\mathbf{P} = \mathbf{I}$ for fully-sampled k-space, an image can be reconstructed by:

$$\hat{\mathbf{x}} = \mathbf{E^H y} = \mathbf{S^H F^H P^H y} = \mathbf{S^H F^H y}. \tag{2.3}$$

In this formulation, $\mathbf{F^H}$ performs the orthonormal discrete inverse Fourier transform and can be efficiently computed using the Fast Fourier Transform (FFT) algorithm.

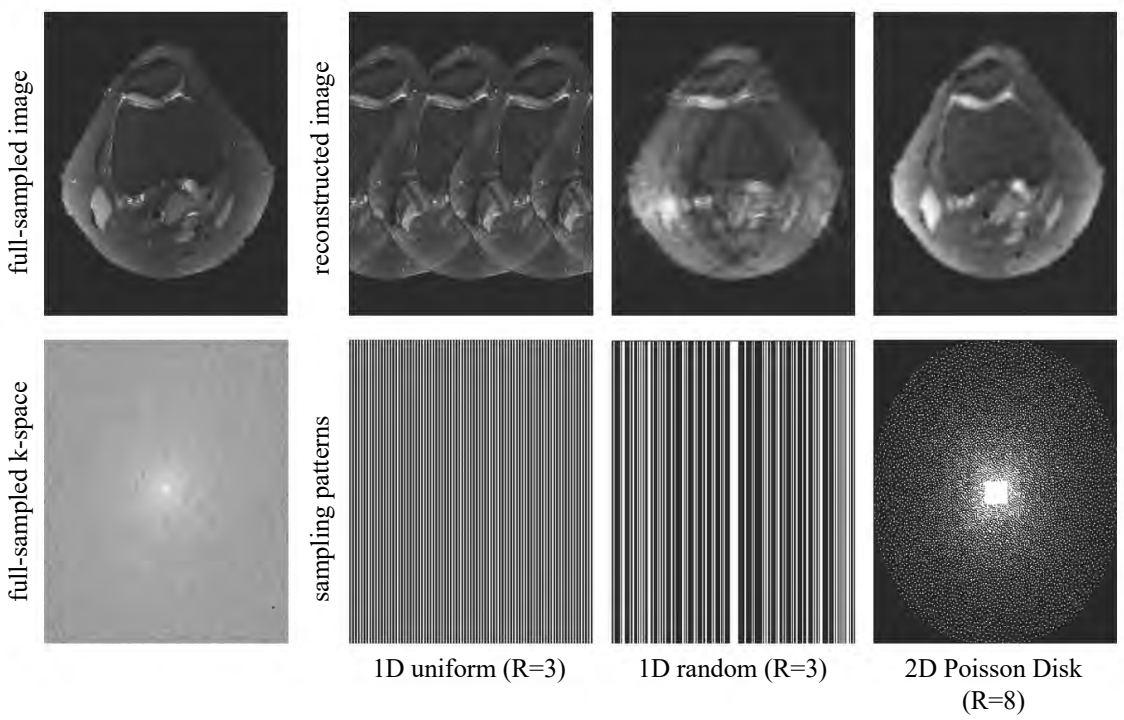


Figure 2.3: **Direct inverse FFT reconstruction from zero-filled under-sampled k-space** Applying inverse FFT to zero-filled under-sampled k-space data can result in aliasing or blurring artifacts. We visualize three undersampling patterns, from left to right: 1)1D uniform undersampling with undersampling rate R=3; 2)1D random undersampling with R=3; 3)2D Poisson Disk undersampling with R=8. The k-space visualization is in log scale.

In practice, to reduce scan time, under-sampled k-space is often acquired. However, when the k-space is under-sampled, the inverse problem becomes ill-posed.

According to the Nyquist–Shannon sampling theorem, directly applying inverse FFT to zero-filled under-sampled k-space results in aliasing or blurring. Figure 2.3 displays artifacts from three different undersampling patterns (R: undersampling rate): 1) 1D uniform undersampling (R=3); 2) 1D random undersampling (R=3); 3) 2D Poisson Disk undersampling (R=8). Different undersampling patterns lead to varying artifacts based on their Point Spread Functions (PSFs). The following subsections introduce representative reconstruction approaches from under-sampled k-space.

Parallel imaging

Parallel imaging has been widely used in clinical settings to reduce scan time by acquiring less k-space data using multiple coils.

As visualized in Figure 2.1, the k-space data is acquired using an array of coils positioned outside the body. Each coil captures complementary signals characterized by coil sensitivities. Parallel imaging (PI) approaches leverage redundant spatial information to reconstruct alias-free images from under-sampled k-space data. Representative PI approaches include SMASH [107], SENSE [90] in image space, GRAPPA [36] and SPIRiT [65] in k-space. In this section, we will provide a brief overview of SENSE and refer readers to relevant literature for other approaches.

Sensitivity encoding (SENSE) formulates the image reconstruction as a linear inverse problem. Given the forward model (Equation 2.2), a SENSE reconstruction is obtained by solving the following least square equation:

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \|\mathbf{E}\mathbf{x} - \mathbf{y}\|_2^2. \tag{2.4}$$

This equation has a closed-form solution:

$$\hat{\mathbf{x}} = (\mathbf{E}^{\mathbf{H}}\mathbf{E})^{-1}\mathbf{E}^{\mathbf{H}}\mathbf{y}. \tag{2.5}$$

In practical situations, the substantial computational expense associated with matrix inversion typically leads to the adoption of the conjugate gradient algorithm as a solution method [101]. With modern hardware design, SENSE has been widely used in clinical routines. The commonly employed undersampling factors for SENSE are typically 2x or 3x.

When the undersampling rate increases, the least square problem shown in Equation 2.4 becomes ll-posed, leading to noise amplification and reduced image quality. At the same time, the quality of reconstructed images relies on the accuracy of the sensitivity maps. These maps can be obtained and estimated from a separate scan or directly estimated from the under-sampled k-space data (*e.g.*, J-SENSE [137], Nonlinear Inversion [117], or ESPIRiT [116]).

Compressed sensing

In addition to utilizing multi-coil information for improved reconstruction, compressed sensing (CS) presents an alternative approach by harnessing prior knowledge on the image itself. This technique enables the reconstruction of high-quality images from highly under-sampled k-space, further enhancing the efficiency of the image reconstruction process.

In CS, the k-space data is acquired in a (pseudo)-random manner. The reconstruction process takes advantage of the sparse structure prior within a specific domain. To accomplish this, a regularization term is integrated into Equation 2.4, resulting in a refined regularized least square problem:

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \|\mathbf{E}\mathbf{x} - \mathbf{y}\|_2^2 + \lambda R(\mathbf{x}), \tag{2.6}$$

where R represents a regularization term that encourages sparsity in the chosen transform domain, while λ serves as a weighting parameter. In particular, R is usually designed as the

ℓ_1 norm of the wavelet coefficients of $\mathbf{x}$. Let $\mathbf{\Psi}$ be the wavelet transform operator and the optimization problem becomes:

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \|\mathbf{E}\mathbf{x} - \mathbf{y}\|_2^2 + \lambda \|\mathbf{\Psi}\mathbf{x}\|_1, \tag{2.7}$$

which can be solved using iterative algorithms, for example, Fast Iterative Soft Thresholding Algorithm (FISTA) [9]. For dynamic MRI reconstruction, R imposes low-rank prior knowledge on the inverse problem [82, 83].

During the past decade, PI and CS have successfully enabled a broad range of clinical applications, and all major MRI vendors have implemented products based on them.

Despite the advancements of PI and CS, several challenges persist:

- 1) The regularization functions employed in CS are either hand-crafted (such as sparse transformation) or rely on relatively simple learned features (for example, dictionary learning as described in [93]). These approaches are known to be less effective in accurately modeling the underlying data distribution, as discussed in [37].
- 2) CS reconstruction is highly sensitive to the tuning parameters, which can impact the overall performance and accuracy of the reconstructed image. The optimal selection of these parameters remains a challenging task.
- 3) The reconstruction time for CS can be relatively long due to the necessity of iterative optimization. This may result in a slower reconstruction process, hindering its practical applicability in real-time scenarios or clinical settings where timely image reconstruction is crucial.

Next, we will introduce DL-based reconstructions that aim to revolutionize MR reconstruction.

2.3 DL-based MRI reconstruction

Deep Learning (DL) is a subset of Machine Learning that leverages artificial neural networks to model complex patterns and structures within data. By employing advanced algorithms and hierarchical feature learning, DL has demonstrated remarkable success in various fields, including computer vision [121], natural language processing [62], and speech recognition [50]. It is this versatility and adaptability that make DL an ideal candidate for MRI reconstruction.

In the context of MRI, DL can be employed to learn a learnable non-linear function f_θ with network weights θ that effectively maps the acquired k-space data $\mathbf{y}$ to an estimated image $\hat{\mathbf{x}}$:

$$\hat{\mathbf{x}} = f_\theta(\mathbf{y}). \tag{2.8}$$

f_θ is trained and optimized to achieve that the estimated image $\hat{\mathbf{x}}$ closely approximates the ground truth $\mathbf{x}$. Once trained, during inference time, f_θ facilitates efficient image reconstruction without relying on lengthy iterative methods, which often require numerous

iterations to converge. This results in a significant reduction in reconstruction time, enhancing the overall efficiency of the process. By utilizing sophisticated neural network architectures such as Convolutional Neural Networks (CNNs) [3] and Recurrent Neural Networks (RNNs) [139], DL-based reconstruction can learn to model more complex relationships between under-sampled k-space data and the images, enabling more accurate reconstruction.

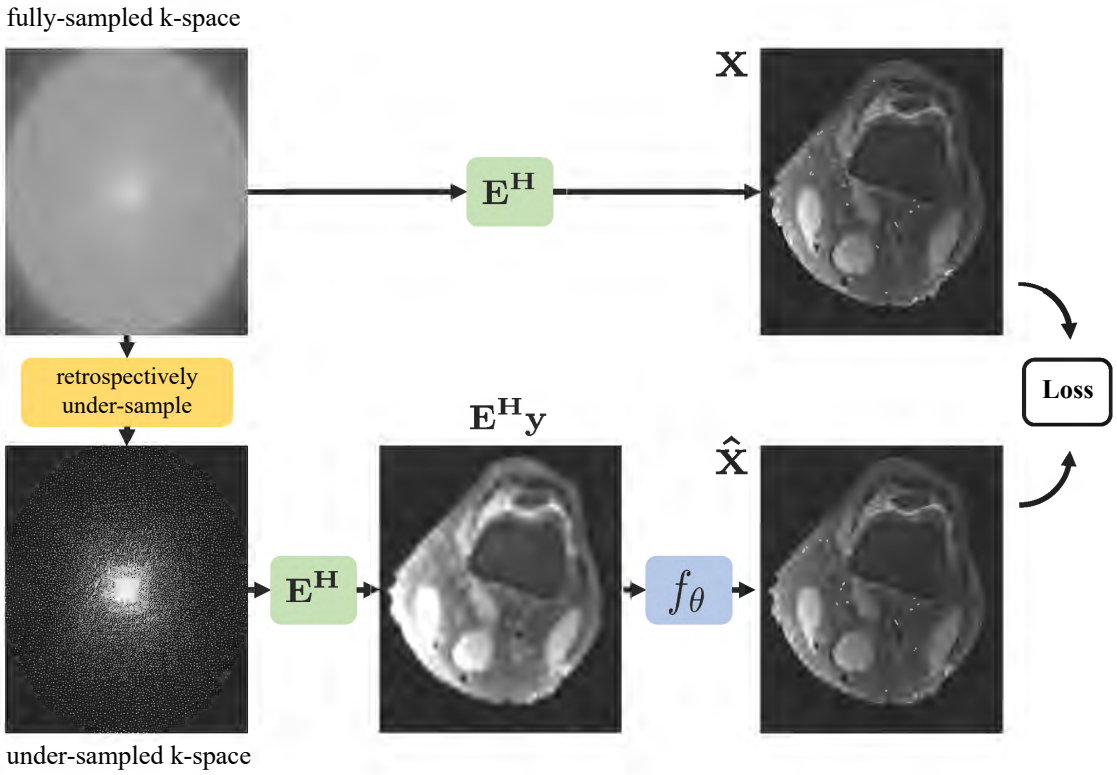


Figure 2.4: **Supervised learning for DL-based MRI reconstruction** Given fully-sampled k-space data, corresponding under-sampled data can be simulated by retrospectively undersampling from the fully-sampled data. The input low-quality image is then reconstructed using the adjoint system operator. Ultimately, the network weight f_θ is optimized by minimizing the loss between the network outputs and the ground truth.

Supervised learning for DL-based image reconstruction

Supervised learning is the predominant method employed in DL-based reconstruction, utilizing vast quantities of fully-sampled training data to learn the reconstruction. Figure 2.5 summarizes the supervised learning framework for DL-based reconstruction.

Given a fully-sampled k-space (usually from a 2D slice), the first step is to under-sample it retrospectively, creating a simulated under-sampled k-space represented as $\mathbf{y}$. The function f_θ is typically constructed using convolutional neural networks (CNNs), which take advantage of spatial relationships in the image domain. Consequently, it is a standard approach in deep learning-based reconstruction to initially convert the input k-space measurements into an aliased image through the conjugate MR signal model $\mathbf{E^h}$, denoted by $\mathbf{E^H y}$, before processing with the CNN. The ground truth image, $\mathbf{x}$, is derived from converting the fully-sampled k-space to image space, while the reconstructed image, $\hat{\mathbf{x}}$, results from applying f_θ: $\hat{\mathbf{x}} = f_\theta(\mathbf{E^H y})$.

In supervised learning, a loss function, represented as $L(\cdot)$, is commonly used to evaluate the difference between the network output, $\hat{\mathbf{x}}$, and the ground truth, $\mathbf{x}$. Frequently utilized loss functions include ℓ_1 loss, ℓ_2 loss (also called MSELoss), and Structural Similarity (SSIM) loss. For ℓ_1 loss, $L(\cdot)$ can be written as:

$$L(\hat{\mathbf{x}}, \mathbf{x}) = \|\hat{\mathbf{x}} - \mathbf{x}\|_1. \tag{2.9}$$

Chapter 4 will provide a more comprehensive overview of the loss function for DL-based reconstruction. During the training process, the network weights θ are optimized to minimize the average loss over the entire training dataset through iterative gradient descent algorithms [52].

The architecture or design of f_θ holds significant importance in deep learning-based reconstruction. A straightforward solution involves implementing an end-to-end feed-forward network, such as U-Net [94, 141], ResNet [39], or Vision Transformer [27].

Nevertheless, feed-forward networks possess inherent drawbacks. Firstly, they typically lack robustness to alterations in sampling patterns, meaning that a network trained on one type of undersampling pattern and tested on another may fail to yield high-quality reconstruction results. Secondly, within the realm of MRI, training a feed-forward model demands a sizable model and an extensive training dataset to ensure its generalizability.

These limitations considerably hinder the performance, efficiency, and clinical adoption of feed-forward networks in MRI reconstruction applications.

In order to overcome these limitations, physics-informed unrolled reconstructions have been proposed, which integrate the underlying physics and signal model directly into the network architecture [1, 37, 110].

Physics-informed unrolled reconstruction

Unrolled networks incorporate physics and signal modeling (signal model $\mathbf{E}$) into the network architecture and have emerged as a powerful approach for DL-based MRI reconstruction, offering significant improvements in image quality and acquisition efficiency [124, 110, 37, 2].

These networks bridge the gap between traditional iterative optimization algorithms and deep learning techniques, providing a robust and efficient method for reconstructing high-

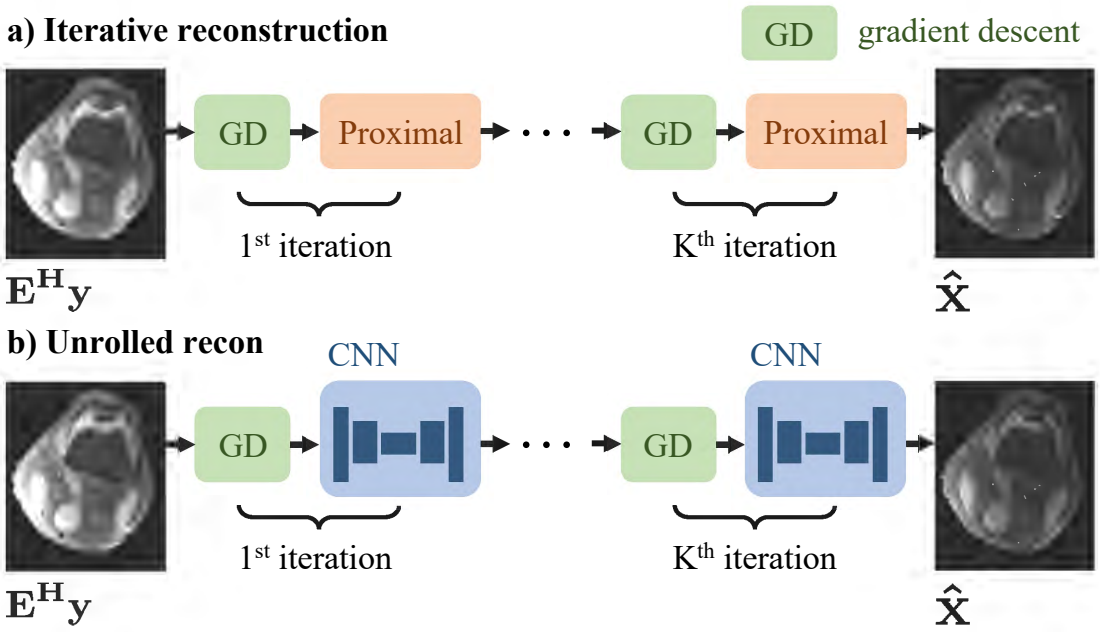


Figure 2.5: **Diagram of iterative reconstruction and unrolled reconstruction.** a) Conventional iterative reconstruction alternates between the gradient descent step and proximal step, with the proximal step being determined by the hand-crafted regularization term, R. b) Physics-informed unrolled reconstruction learns the regularization function by replacing the proximal step with a learnable CNN.

quality MR images from under-sampled k-space data. The unrolled network framework is inspired by the structure of iterative optimization algorithms to solve the inverse problem equation 2.6, with the aim of learning R directly from the data.

To solve Equation 2.6, as shown in Figure 2.5 a, the problem is divided into two alternating steps that are executed repeatedly, which we called proximal gradient descent. During the k-th iteration, a gradient update is carried out as follows:

$$\mathbf{x}^{(k+)} = \mathbf{x}^{(k)} - 2t\mathbf{E}^{\mathbf{H}}(\mathbf{E}\mathbf{x}^{(k)} - \mathbf{y}), \tag{2.10}$$

where t is the gradient step size, $\mathbf{x}^{(k+)}$ is the intermediate result. Subsequently, the proximal problem associated with regularization R is solved:

$$\mathbf{x}^{(k+1)} = \mathbf{prox}_{\lambda R}(\mathbf{x}^{(k+)}) = \arg\min_{u} R(u) + \frac{1}{2t\lambda}\|u - \mathbf{x}^{(k+)}\|_2^2, \tag{2.11}$$

where u is a helper variable. The updated $\mathbf{x}^{(k+1)}$ is then passed to the next iteration. It's worth mentioning that the solution for the proximal step of ℓ_1 regularization is a simple soft-thresholding step.

Unrolled networks essentially learn the proximal step in equation 2.11, thereby implicitly learning the regularization function R instead of relying on hand-crafted alternatives.

As illustrated in Figure 2.5 b, one of the most widely used unrolled network [96] replaces the proximal step in equation 2.11 to an image-to-image CNN-based denoiser (*e.g.*, U-Net [94]), where we denote as D_θ. Now, the proximal step for the unrolled network can be rewritten as:

$$\mathbf{x}^{(k+1)} = D_\theta(\mathbf{x}^{(k+)}), \tag{2.12}$$

where θ are learnable parameters. Unrolled networks usually fix the number of iterations K. The exact number depends on the specific tasks.

By incorporating both signal modeling (Equation 2.10) and data-driven deep learning (Equation 2.12), unrolled networks offer a more robust and efficient alternative to feed-forward networks.

In recent years, numerous unrolled networks [2, 37, 110, 127] have been proposed for MRI reconstruction, showcasing superior image quality and robustness to sampling patterns in comparison to their counterparts. More importantly, unrolled networks utilize the underlying physics and signal modeling, resulting in a reduced need for large model sizes and extensive training data, which offers significant advantages in terms of efficiency and practicality.

Two prominent unrolled networks include Model-based Deep Learning (MoDL)[2] and Variational Networks[37]. We will delve into MoDL in detail in Chapters 4 and 5.

Challenges for DL-based MRI reconstructions

Despite the significant success of deep learning-based MRI reconstructions, particularly unrolled reconstructions, several challenges persist in the field, which limits the fidelity and efficiency. Some of these challenges include:

- Loss functions employed in DL-based reconstruction methods are predominantly hand-crafted. These can be pixel-wise (*e.g.*, ℓ_1 and ℓ_2 losses) or based on local statistics (*e.g.*, SSIM loss [130]). However, such loss functions may inadequately capture perceptual information, which can result in compromised image quality and blurring [44, 2, 37, 125].

- DL-based unrolled reconstruction comprises of multiple learnable networks (one for each iteration), which demand substantial GPU memory resources for back-propagation. Consequently, memory constraints during network training can limit the applicability of deep learning reconstruction techniques for high-dimensional MRI data, such as 2D+time, 3D, and 3D+time MRI [95, 127].

- The confidence or reliability of reconstructed structures remains insufficiently investigated, posing a challenge for DL-based approaches in clinical applications [129].

- Unlike natural images, MRI data is inherently complex-valued and faces challenges due to the limited availability of fully-sampled ground truth. This constraint inevitably

restricts the applications of deep learning-based MRI techniques to tasks without access to adequate ground truth (*e.g.*, 3D MRI, 2D/3D dynamic MRI).

This dissertation presents a series of projects that aims to tackle these challenges and unlock the potential of deep learning for large-scale MRI reconstructions.

Chapter 3

Direct Contrast Synthesis from MR Fingerprinting

3.1 Introduction

In the first chapter of this dissertation, emphasis is placed on the reconstruction of multiple contrast-weighted images from a single highly under-sampled acquisition. In particular, it explores the generation of diagnostic contrast-weighted images from a single, short MR Fingerprinting scan.

As introduced in Chapter 2, image contrast in MRI is dominated by biophysical tissue properties, such as proton density (PD), longitudinal/transverse relaxation (T1/T2), magnetic susceptibility, and diffusion. These parameters provide information on the tissue composition and its microstructure and are excellent biomarkers for diagnosing and assessing disease. Measuring the quantitative value of tissue parameters, *i.e.*, through quantitative MRI (qMRI), is desirable, because it provides a standardized metric for tissue properties [89]. However, qMRI has been notoriously challenging to implement and standardize in clinical practice. Traditional mapping sequences require many lengthy scans to map a single parameter and thus are unsuitable for rapid imaging.

Consequently, current diagnostic examinations are composed of a series of several scans, each *qualitatively* emphasizing one of the physical parameters above. For example, routine brain MRI includes PD-weighted scans, wherein brighter pixel intensities indicate a higher density of protons; T1-weighted (T1w) scans, wherein brighter intensities indicate shorter T1 recovery; T2-weighted (T2w) scans, wherein brightness indicates longer T2 relaxation; fluid-attenuated inversion recovery (T1/T2-FLAIR), wherein fluid signals are suppressed; and diffusion scans, wherein brighter intensities indicate less diffusivity. The relative contrast differences within and across these scans can aid in the assessment of disease.

Owing to the need for multiple scans to obtain multiple contrasts, the typical MRI protocol is lengthy, requiring patients to remain still for tens of minutes and hindering scanner throughput.

In recent years, notable research efforts have focused on acquiring or synthesizing multi-contrast images from single scans or fewer scans to shorten the total examination time [113, 111, 128, 66, 67, 122]. These techniques have shown early success in clinical practice [43, 118]. For example, synthetic MR methods [10, 35, 113, 131, 43] acquire multiple short scans and use parameter fitting and physical models to simulate a variety of contrast-weighted images. T2 shuffling [111, 112] reconstructs multiple contrast-weighted images along the transverse relaxation curve by using a single volumetric fast spin echo acquisition, through randomly shuffling the phase encoding view ordering and performing subspace modeling. Similarly, multitasking [18, 14] approaches use tensor low-rank constraints to reconstruct multiple contrast-weighted images from a single rapid acquisition. The above approaches all require scan parameters to be carefully chosen to limit confounding factors and isolate a small number of qMRI parameters contributing to the overall image contrast.

Instead of decreasing confounding factors, an alternative approach, known as magnetic resonance fingerprinting (MRF) [67, 66, 43] was proposed to mix many quantitative parameters by using a short acquisition with randomized scan parameters. MRF has accelerated the pace of clinical qMRI by demonstrating the ability to rapidly and reliably generate multiple quantitative parameter maps from a single scan. MRF acquisition is usually based on gradient echo sequences and consists of rapid repetition times (TR) with under-sampled spiral readouts, in which the flip angle is modified for every TR, such that the steady state of spin dynamics is never achieved. MRF produces a sequence of images in which tissues with different relaxation and field properties (T1, T2, PD, B0, and B1) produce a unique time series or "fingerprint." The quantitative parameters of the tissue are then extracted by matching the resulting time series of each pixel to the closest signal in a precomputed dictionary constructed by simulating the Bloch equation for parameter combinations within a realistic range.

The fact that quantitative parameters can be extracted from MRF also indicates that the information embedded should be sufficient to synthesize contrast-weighted images. Although quantitative parameter maps provide meaningful physical tissue parameters, clinicians still rely primarily on contrast-weighted images for clinical diagnosis. Therefore, an opportunity exists for MRF to enable both parameter maps and synthetic contrast MRI to be provided by a single sequence.

One approach to synthesizing contrast-weighted images from MRF is to first fit the quantitative parameters and then simulate the contrast-weighted images [10]. Figure 3.1 and Figure 3.2a show the spin-dynamic simulation pipeline, which uses quantitative parameter maps to synthesize different contrast-weighted images by using the Bloch equation or extended phase graphs (EPG) [133]. Unfortunately, contrast-weighted images generated in this manner often exhibit artifacts because of many sources of error. Errors can arise from discrepancies between the MRF sequence and the dictionary simulation, for example, when flow, diffusion, magnetization transfer, excitation slice profile, or partial volume is not modeled appropriately. This limitation is most pronounced in FLAIR contrast, in which errors are seen along the boundaries of cerebrospinal fluid [118].

An alternative, and relatively more straightforward, pipeline avoids explicit modeling

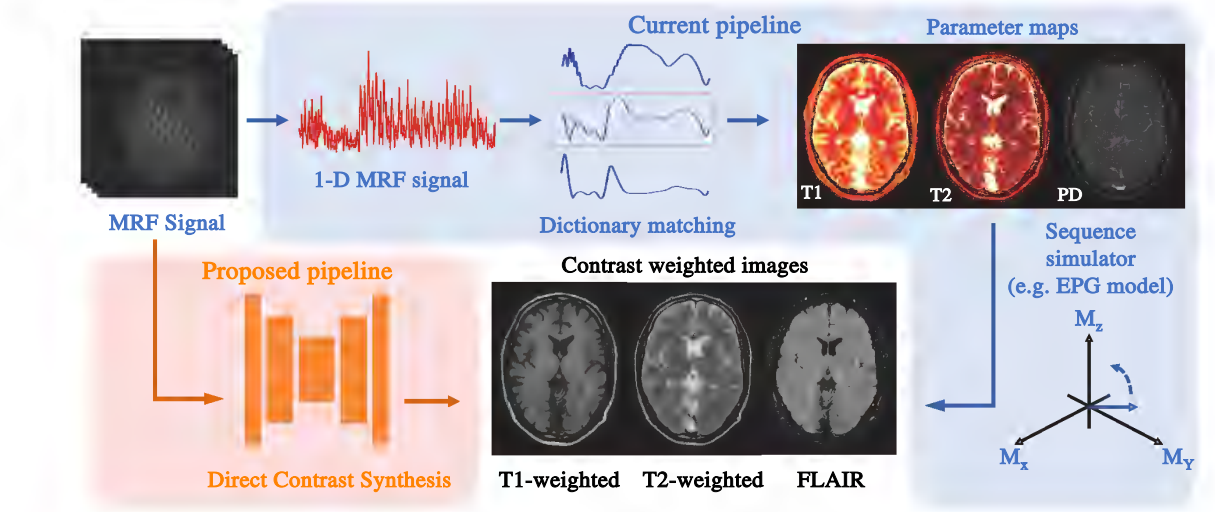


Figure 3.1: **Contrast synthesis from MRF via a current simulation-based pipeline and proposed direct contrast synthesis (DCS) pipeline.** The simulation-based method takes the predicted quantitative parameter maps from MRF and synthesizes different contrast-weighted images by simulating the MRI physics. Our proposed DCS uses a spatial CNN to transform the MRF time series directly into different contrast-weighted images. DCS bypasses dictionary matching and contrast simulation steps, avoids modeling and acquisition imperfections, and produces high-fidelity contrast-weighted images.

and instead directly learns how to synthesize contrast-weighted images from the MRF data through neural networks. We refer to this approach as Direct Contrast Synthesis (DCS). Previous work [120] has proposed a supervised DCS method in which a network was trained to take a single voxel MRF time series and map it to a specific contrast weighting (*e.g.*, T1w, T2w, or FLAIR). This approach, which we refer to as **PixelNet**, is illustrated in Figure 3.2b. By training on many pairs of MRF and contrast-weighted images, PixelNet can achieve better results than dictionary mapping and simulation-based contrast synthesis. However, by processing each pixel independently, PixelNet does not leverage the spatial structure in the data and thus can suffer from noise and spatial inconsistency. To address this issue, we propose to implement DCS as an image sequence-to-image translation task to leverage structural information. In the field of computer vision, image-to-image translation is an established problem that aims to translate an image from a source domain to a target domain (e.g., reconstructing objects from edge maps [44] and colorizing images [144]). Recent studies have shown promising results through image-to-image convolutional neural networks (CNNs) and generative adversarial networks (GANs) [34, 44]. The seminal work of pix2pix [44] investigated conditional adversarial networks as a general-purpose solution to image-to-image translation problems. CycleGAN [148] improved upon the technique of learning image-to-image translation in the absence of paired examples. Image-to-image translation has

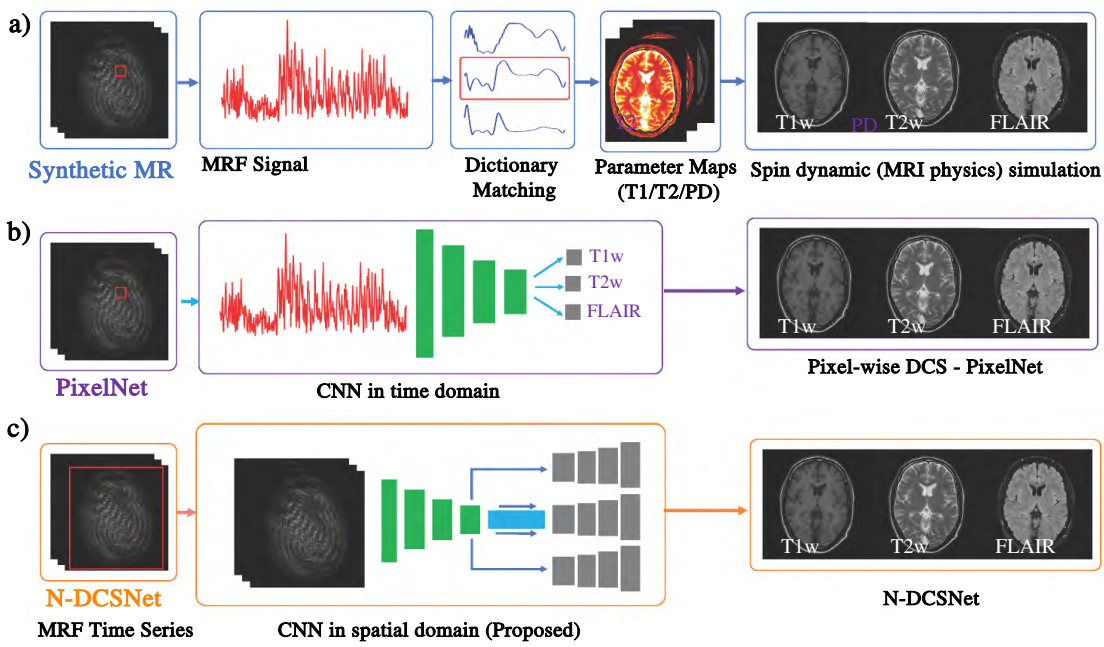


Figure 3.2: **Three possible pipelines to generate contrast-weighted images from MRF.** a) **Synthetic MR** generates multi-contrast images through dictionary matching and sequence simulation (e.g., Bloch equation, EPG). b) **PixelNet** uses a 1D pixel-wise time-domain CNN to output a qualitative contrast weighting for each voxel. c) Our proposed **N-DCSNet** leverages a GAN-based architecture and spatial-convolutional network to synthesize multi-contrast images.

also been applied in the fields of medical imaging and MRI. For example, references [69, 38, 134] learned cross-modality image synthesis between MRI and CT images; references [138, 22] synthesized T2w images from T1w images; reference [91] synthesized 7T high-resolution, high-SNR images from 3T input images; and reference [123] introduced a multi-task deep learning model to synthesize multi-contrast MRI images from multi-echo sequences. Recently, reference [21] proposed a residual transformer-based deep learning model for multi-modal cross-contrast MR image synthesis.

Inspired by previous works, we propose to use a conditional GAN-based architecture for DCS from MRF that enables substantial improvements in image quality and computation efficiency over simulation-based contrast synthesis and PixelNet. We refer to our approach as **N-DCSNet**, first described in reference [124], where **N** represents N different contrasts that can be synthesized by our network (here $N = 3$). Figure 3.2 summarizes the three pipelines of producing synthetic, multi-contrast images.

As illustrated in Figures 3.1 and 3.2c, **N-DCSNet** directly synthesizes different contrast weighted images, i.e., T1w, T2w, or FLAIR, from the MRF time series data through a

spatial CNN. Our generator is designed as a U-Net with a single encoder and multi-branch decoders [94]. We implement a multi-layer CNN (PatchGAN) [44] as the discriminator. The generator is based on spatial convolutions, thus allowing the network to learn and exploit spatial structural information. Different contrast-weighted outputs share the same encoder to exploit the shared information across contrasts. Separate decoders are designed to learn the unique features of each contrast. During the training procedure, we leverage a conditional GAN framework, wherein the time average of the MRF time series is also used as an input to the discriminator to constrain the GAN training.

In vivo experiments on healthy volunteers show that **N-DCSNet** can generate high-fidelity, multi-contrast images from MRF time-series. Our approach outperforms contrast synthesis from parameter maps and PixelNet both qualitatively and quantitatively. Furthermore, we demonstrate that **N-DCSNet** can inherently mitigate some artifacts that appear in MRF, such as slice in-flow artifacts and spiral off-resonance blurring. Our main contributions can be summarized as follows:

- We introduce a spatial CNN-based method to learn the mapping between MRF time series and contrast-weighted images (*i.e.,* T1w, T2w, and FLAIR). Our approach can avoid the simulation errors typically seen in Synthetic MR.
- We use a conditional GAN-based framework to encourage finer textures and produce more faithful contrasts. Additionally, our **N-DCSNet** can inherently mitigate slice in-flow artifacts as well as spiral off-resonance blurring.
- **N-DCSNet** outperforms simulation-based contrast synthesis from parameter maps and PixelNet qualitatively and according to quantitative metrics. It also has significant computation advances. During inference, our approach is significantly faster than simulation-based contrast synthesis and PixelNet, thus improving the potential for clinical adoption.

3.2 Data acquisition and formulation of N-DCSNet

In this section, we first describe the data acquisition protocols and the simulation-based contrast synthesis via parameters used as our baseline for comparisons (§ 3.2). Then, we introduce our GAN-based framework design for **N-DCSNet** (§ 3.2). Next, we detail the loss functions (§ 3.2) and the training process. Finally, we compare our method with previous approaches (§ 7.6).

Data acquisition and contrast synthesis via parameters

Data acquisition: After obtaining IRB approval, we scanned 21 men, ranging from 29 to 61 years of age, with a 1.5 T Philips Ingenia scanner using a 15-channel head coil. A total of 13 channels were selected by using automatic coil selection. To avoid conducting so-called "data

crimes," [103] we report our data preparation pipeline as follows. Four consecutive axial brain scans were acquired for each examination session. The participants were instructed to remain still throughout the examination so that data across scans remained registered. The scans were as follows:

- A spoiled gradient echo [47] MRF sequence with 500 time points, constant TE=3.3 ms, TR=20 ms (Each TR consisted of a spiral-out readout. The spirals between two consecutive TR were rotated by 9°.). The readout time is 12 ms and the undersampling factor is 20.
- T1w spin echo with TE=15 ms, TR=450 ms, flip angle=69°, and two averages.
- T2w turbo spin echo (TSE) with TE=110 ms, TR=1990-2215 ms, ETL=16, flip angle=90° and two averages.
- FLAIR inversion recovery TSE with TE=120 ms, TR=8500 ms, TI=2500 ms, ETL=41, flip angle=90° and two averages.

All scans were acquired with an in-plane resolution of 0.72×0.72 mm (FOV 230×230 mm, matrix size 320×320) and nine to ten slices with a thickness of 5 mm. Of the 21 participants, 17 were scanned twice (on different days), thus resulting in a total of 38 examinations. FLAIR sequences were acquired for only 26 of the 38 examinations. Only the 26 examinations with all four sequences were used in this study, of which 21 were used for training, two were used for validation, and three were used for testing. The data from participants used for testing were not included in any of the training sets.

To further minimize residual motion or misalignment between scans, we employ a 2D rigid in-plane registration per slice, aligning the ground truth contrast-weighted images with the time-averaged MRF image. Moreover, we manually inspect the images and discard those exhibiting significant in-plane and through-plane movements.

Pre-processing: Each of the three contrast-weighted image data was normalized with respect to the 95th percentile of the intensity values for each image. MRF time series images were reconstructed from each TR by using gridding with density compensation [79, 45] followed by coil combination with Philips' CLEAR. The MRF data were then normalized as follows. For each dataset, an averaged image from the 500 time points was computed. The 95th percentile of the magnitude values from the average MRF image was then used to normalize the time series.

Parameter maps and contrast simulation: The dictionary for MRF parameter mapping was simulated by using EPG [133]. The dictionary consisted of 22,031 MRF signals with T1 parameters ranging from 4 to ms to 3,000 ms and T2 parameters ranging from 2 ms to 2,000 ms. Each simulated signal in the dictionary was scaled to have a Euclidean norm equal to one. We used cosine similarity [67] to match the acquired MRF signal to the nearest neighbor in the simulated dictionary (Figure 3.2 a). Additional factors, such as B1 inhomogeneity and slice profile, were not included in the simulated dictionary.

The parameter maps (T1, T2) obtained from dictionary matching were then used to simulate the contrast-weighted images. The T1w spin echo (SE) has a closed form for specific TE and TR, and PD parameters:

$$\mathrm{SE}(\mathrm{PD}, \mathrm{T1}, \mathrm{T2}, \mathrm{TE}, \mathrm{TR}) = \mathrm{PD} \cdot (1 - e^{-\frac{\mathrm{TR}-\mathrm{TE}}{\mathrm{T1}}}) \cdot e^{-\frac{\mathrm{TE}}{\mathrm{T2}}}. \tag{3.1}$$

PD was computed by taking the magnitude of the inner product between the acquired MRF signal and the nearest neighbor in the simulated dictionary. The T2w and FLAIR sequences are based on TSE and do not have closed forms. For these, we used EPG [133] to simulate the contrast-weighted images.

N-DCSNet framework

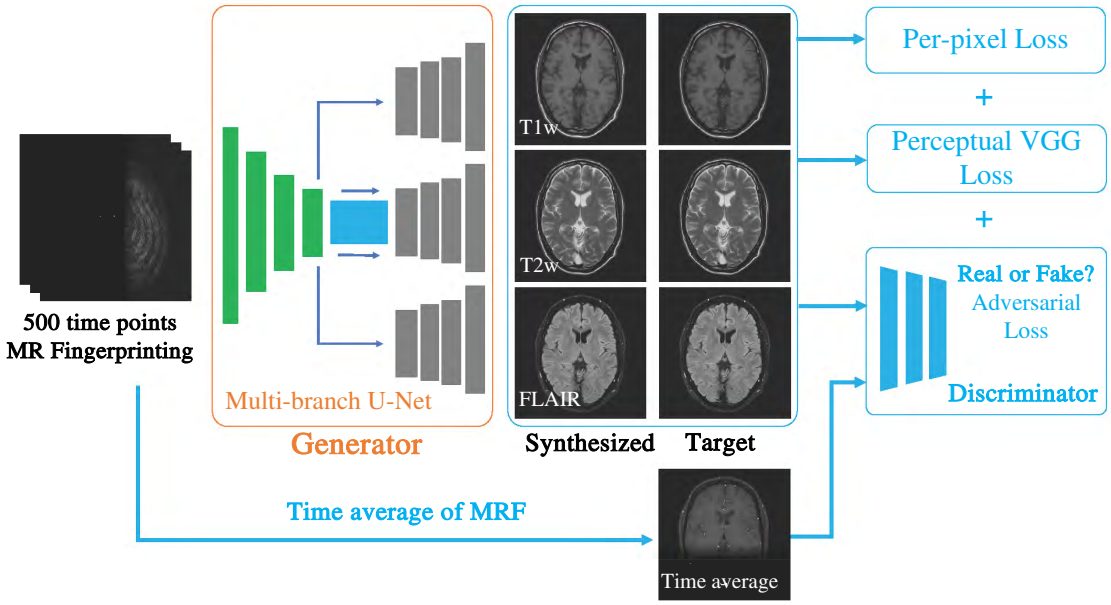


Figure 3.3: **Illustration of our proposed N-DCSNet framework.** Given a complex-valued MRF time series $\mathbf{MRF}_{in} \in \mathbb{C}^{t \times h \times w}$, with number of time points $t \in \mathbb{N}$ and image dimensions $h, w \in \mathbb{N}$, **N-DCSNet** synthesizes three contrast-weighted images (T1w, T2w, and FLAIR) with a single network. We designed a multi-branch U-Net as the generator and a multi-layer CNN as the discriminator by following the conditional GAN training strategy. To constrain the GAN training, we additionally input the time average of MRF to the discriminator. A combination of per-pixel ℓ_1 loss, perceptual VGG loss, and adversarial loss is imposed on the network. **N-DCSNet** generates high-fidelity contrast-weighted images with sharper edges, finer textures, and more faithful contrasts than simulation-based contrast synthesis and PixelNet.

Figure 3.3 illustrates the overall pipeline of our proposed **N-DCSNet**. Our network expects the complex-valued MRF time series $\mathbf{MRF}_{in} \in \mathbb{C}^{t\times h\times w}$ as input, where t, h, and w correspond to the number of time points, image height, and image width, respectively $(t, h, w \in \mathbb{N})$. The network outputs are real-positive (magnitude) contrast weighted images $\hat{\mathbf{T1w}}, \hat{\mathbf{T2w}}, \hat{\mathbf{FLAIR}} \in \mathbb{R}^{h\times w}$. In our experiments, $t = 500, h = w = 320$.

We designed a conditional GAN-based framework for **N-DCSNet**, the standard framework in references [44, 148], consisting of a generator (G) and a discriminator (D).

First, for the input complex-valued MRF data with dimensions 500×320×320, we concatenate the real and imaginary parts along with the time dimension as channels to the network. This results in a real-valued input with dimensions 1000×320×320.

Our generator is a modified U-Net [94], which consists of one shared encoder and multiple independent decoders. The shared encoder exploits structural similarities across the multi-contrast images, whereas the independent decoders learn the unique features of the different contrasts. At test time, **N-DCSNet** produces multi-contrast images with a single network. The discriminator (D) is a multi-layer CNN (patchGAN) [44] that penalizes structure at a patch scale. D aims to classify whether each $N \times N$ patch in an image is real or fake. We run this discriminator convolutionally across the image, averaging all responses to provide the final output of D. To constrain the GAN training, we follow reference [44] and further input the magnitude of the MRF time-averaged image to the discriminator to provide structural guidance. This image has mixed contrast, because of averaging, and significantly fewer spiral undersampling artifacts than the MRF time-series images. We denote it $\mathbf{MRF}_{avg}$.

During training, the generator G learns to predict high-quality contrast-weighted images that cannot be distinguished from the real acquired images (ground truth) by an adversarially trained discriminator D. Meanwhile, D is simultaneously trained to distinguish the generated images (labeled as "fake") from the ground truth images (labeled as "real").

Loss functions

Our proposed **N-DCSNet** is fully supervised, with the purpose of generating high-fidelity contrast-weighted images that are close to the ground truth real acquisitions. The loss function of our generator G is a combination of three components: 1) ℓ_1 reconstruction loss, 2) perceptual loss, and 3) adversarial loss. Given our generator G and the input MRF signal $\mathbf{MRF}_{in}$, G outputs the synthesized contrast-weighted images (T1w, T2w, and FLAIR):

$$\hat{\mathbf{T1w}}, \hat{\mathbf{T2w}}, \hat{\mathbf{FLAIR}} = G(\mathbf{MRF}_{in}). \tag{3.2}$$

Then the cumulative ℓ_1 loss is formulated as:

$$L_{\ell_1} = \mathbb{E}_{\mathbf{MRF}_{in}}(\|\hat{\mathbf{T1w}} - \mathbf{T1w}\|_1 + \|\hat{\mathbf{T2w}} - \mathbf{T2w}\|_1 + \|\hat{\mathbf{FLAIR}} - \mathbf{FLAIR}\|_1), \tag{3.3}$$

where $\mathbf{T1w}, \mathbf{T2w}$, and $\mathbf{FLAIR}$ represent the real, ground-truth acquisitions of the three contrast-weighted images (§ 3.2). Per-pixel losses such as the ℓ_1 loss are known to exhibit

image blurring [44, 49, 59, 125]. Therefore, we incorporate additional perceptual and adversarial losses to encourage detailed reconstructions.

Perceptual losses [49, 125] have been used successfully in super-resolution and image synthesis [123] tasks to improve image quality and encourage delicate structures. The underlying idea is that layer features of task-based networks, such as image classification networks, can capture high-level perceptual information in the image. Therefore, minimizing the loss in the feature space can preserve such perceptual information [49]. In this work, the perceptual loss is implemented as the ℓ_2 distance between **relu2-2** layer features of an ImageNet [24] pretrained VGG Network [104]. We denote the function used to extract these features as $\phi(\cdot)$, where $\phi(\mathbf{x})$ extracts the **relu2-2** layer features of a specific image $\mathbf{x}$. Each contrast-weighted image is scaled to $[0, 1]$, duplicated three times, and concatenated along the channel dimension (to simulate RGB channels) before feeding into $\phi(\cdot)$. Then, the overall VGG perceptual loss term can be written as:

$$L_{\text{vgg}} = \mathbb{E}_{\mathbf{MRF}_{in}}(\|\phi(\mathbf{T\hat{1}w})-\phi(\mathbf{T1w})\|_2+\|\phi(\mathbf{T\hat{2}w})-\phi(\mathbf{T2w})\|_2+\|\phi(\mathbf{FL\hat{A}IR})-\phi(\mathbf{FLAIR})\|_2). \tag{3.4}$$

The third component of our loss function is an adversarial loss. This term is used to further encourage high-frequency details and achieve more realistic synthesized outputs [44]. The generator G is trained to produce outputs that cannot be distinguished from "real" images. We concatenate the acquired images $[\mathbf{MRF}_{avg}, \mathbf{T1w}, \mathbf{T2w}, \mathbf{FLAIR}]$ along the channel dimension, and treat it as the "real" sample $\mathbf{S}_{real} = [\mathbf{MRF}_{avg}, \mathbf{T1w}, \mathbf{T2w}, \mathbf{FLAIR}]$. Meanwhile, we create $\mathbf{S}_{fake} = [\mathbf{MRF}_{avg}, \mathbf{T\hat{1}w}, \mathbf{T\hat{2}w}, \mathbf{FL\hat{A}IR}]$ as the "fake" sample $\mathbf{S}_{fake}$. Then, the adversarial loss for our generator is given by:

$$L_{\text{adv}} = -\mathbb{E}_{\mathbf{S}_{fake}}[\log(D(\mathbf{S}_{fake}))]. \tag{3.5}$$

The overall objective function for the generator becomes:

$$L_{\text{G}} = L_{\ell_1} + \lambda_{\text{vgg}} L_{\text{vgg}} + \lambda_{\text{adv}} L_{\text{adv}}, \tag{3.6}$$

where λ_{vgg} and λ_{adv} are the weights of the perceptual loss and adversarial loss, respectively. In our experiments, we empirically set $\lambda_{\text{vgg}} = 0.03$ and $\lambda_{\text{adv}} = 0.015$.

Our discriminator is adversarially trained to detect the generators' outputs as "fake" images. According to reference [34], the objective function for our discriminator L_{D} is given by:

$$L_{\text{D}} = -\mathbb{E}_{\mathbf{S}_{real}}[\log(D(\mathbf{S}_{real}))] - \mathbb{E}_{\mathbf{S}_{fake}}[\log(1 - D(\mathbf{S}_{fake}))]. \tag{3.7}$$

We update the parameter weights of G and D by alternatively minimizing the objectives L_{G} and L_{D}.

Experiments

To demonstrate its effectiveness, we evaluate our **N-DCSNet** against simulation-based contrast synthesis (synthesis via parameters) and PixelNet on the same testing dataset (detailed in §3.2). The EPG simulation using the dictionary-matched parameters was run for all voxels in parallel by using the joblib package [48] on 24 CPUs. On the basis of the architecture introduced in reference [120], we implemented PixelNet as a 1D temporal CNN to map the MRF time series at every voxel to the corresponding three contrast-weighted scans. The PixelNet network consists of three convolutional layers followed by three fully connected layers and is trained with an ℓ_2 loss. The inference time for the different approaches is calculated by computing the average runtime of 20 separate runs of a single MRF slice. Ablation studies were also conducted to analyze the impacts of the different loss functions on the synthesized contrast-weighted images.

Evaluation metrics

To quantitatively compare our results to the ground truth, we report the following evaluation metrics: normalized root mean square error (nRMSE), peak signal-to-noise ratio (PSNR), structural similarity (SSIM) [130], learned perceptual image patch similarity (LPIPS) [145] with AlexNet [56], and Fréchet inception distance (FID) score [40]. When computing LPIPS and FID, the output images were scaled to the range $[0, 255]$ and saved as png files.

Implementation details

All the proposed algorithms and networks were implemented with PyTorch 1.8 [86] on 24 GB NVIDIA 3090 graphics processing units (GPUs). Our generator and discriminator were trained by using Adam optimizer [52], with a batch size of 4 and a learning rate of 1×10^{-4}.

We supervise the direct contrast synthesis with magnitude contrast weighted images. However, the MRF time series is inherently complex-valued. To reduce the sensitivity to phase, during training, we augment the phase of the MRF data on the fly by multiplying each time-series with random constant phase $e^{j\theta}$, where $j = \sqrt{-1}$ and θ is uniformly distributed between $[0, 2\pi]$.

The ablation study evaluating the loss functions' contributions was performed by comparison of the proposed combined loss function (Equation 3.6) against L_{ℓ_1}, $L_{\ell_1} + \lambda_{\text{vgg}} L_{\text{vgg}}$, and $L_{\ell_1} + \lambda_{\text{adv}} L_{\text{adv}}$ losses.

3.3 Comparisons with contrast synthesis via parameters and PixelNet

Figure 3.4 summarizes the results of the different contrast synthesis methods applied to a representative 2D brain slice. Compared with EPG simulation-based synthesis (synthesis via parameters) [133], and PixelNet [120], **N-DCSNet** produces finer and cleaner structural

details, sharper edges, and better perceptual agreement with the true acquisition (ground truth).

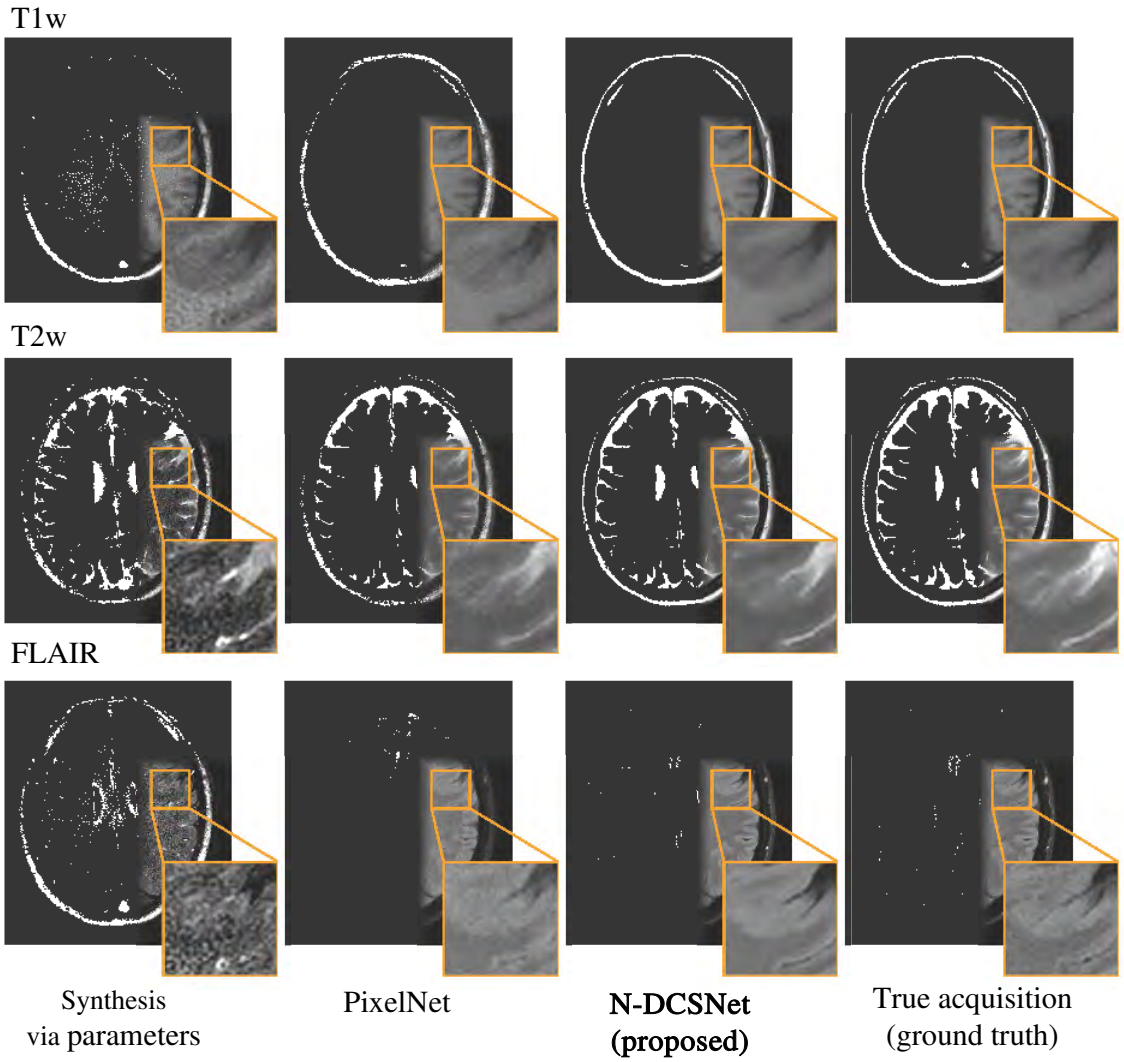


Figure 3.4: **Representative contrast synthesis results of different methods (upper brain).** From left to the right, we compare our proposed **N-DCSNet** with simulation-based contrast synthesis via parameters [133], PixelNet [120], and the true acquisition. **N-DCSNet** shows better visual agreement with the true acquisition, producing finer textures and higher overall image quality than the other approaches. Zoomed-in details are displayed next to each image.

The EPG simulation-based results (synthesis via parameters) exhibit incorrect contrast and noise artifacts due to the modeling and acquisition imperfections (as expected in §3.1).

PixelNet significantly improves the synthesized image quality, but the noise artifacts persist (as shown in T1w and T2w). In comparison, **N-DCSNet** leverages both temporal and spatial information, producing more faithful contrast, preserving finer details, and showing better agreement with the ground truth images.

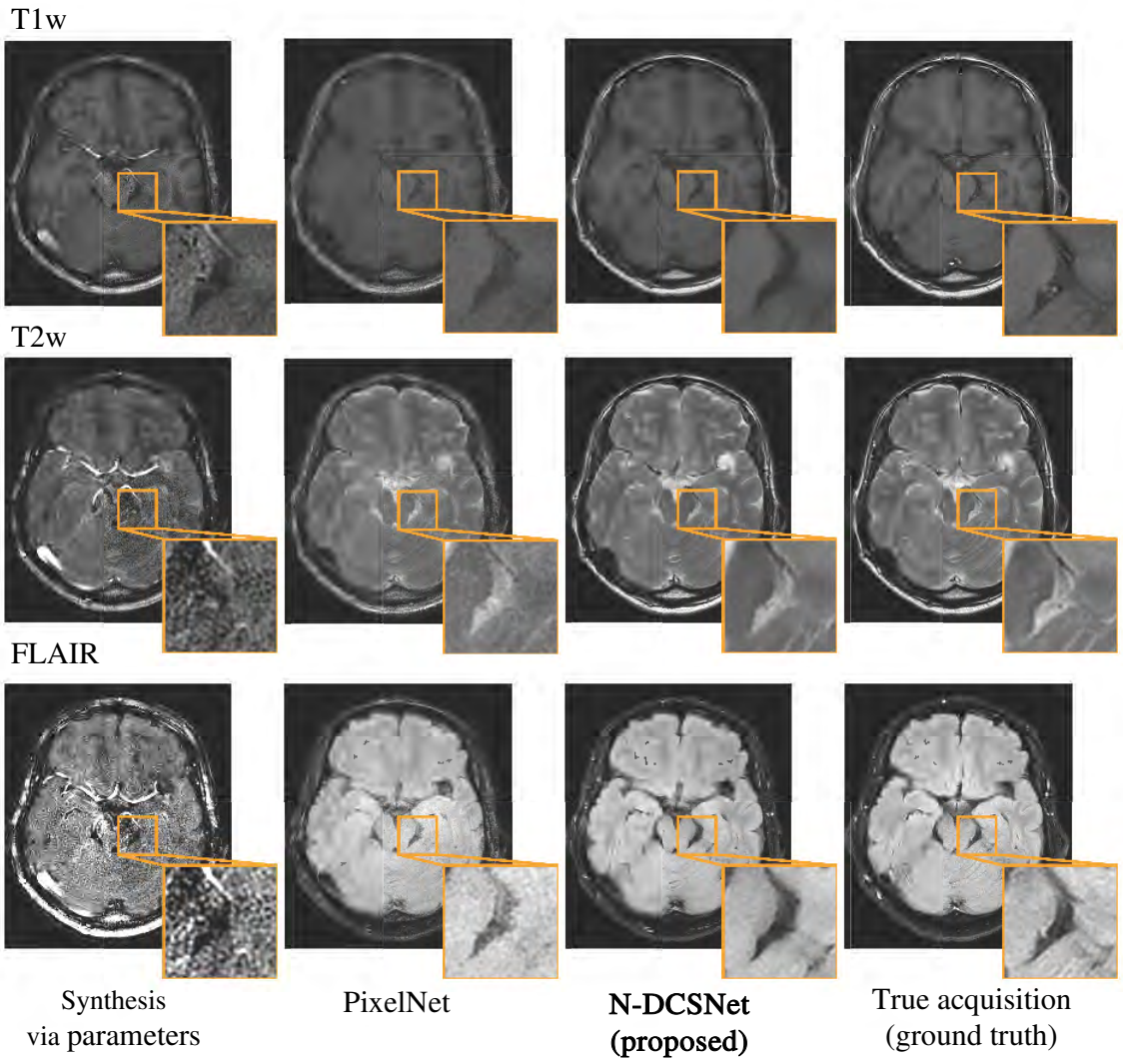


Figure 3.5: **Representative contrast synthesis results of different methods (lower brain).** From left to the right, we compare our proposed **N-DCSNet** with simulation-based synthesis via parameters [133], PixelNet [120], and the true acquisition. Zoomed-in images show the inflow (vasculature) regions where parameter-based synthesis (left column) fails to deliver correct contrast, owing to the moving blood flow. In comparison, **N-DCSNet** successfully reconstructs delicate textures and produces high-quality contrast-weighted images.

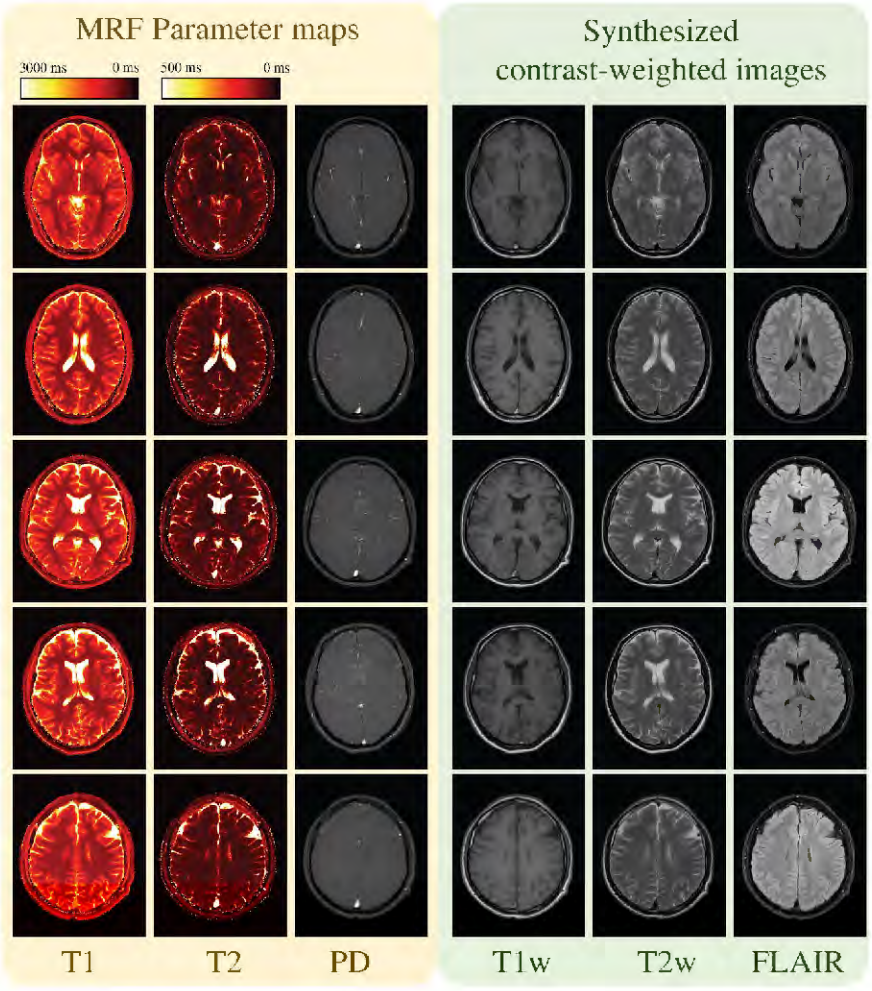


Figure 3.6: **Gallery of N-DCSNet synthesized contrast-weighted images alongside parameter maps.** N-DCSNet synthesizes high-fidelity contrast-weighted images (right three columns) from MRF data. Concurrently, the parameter maps (*i.e.*, PD, T1, T2) can be obtained through dictionary matching (left three columns). Our approach showcases the feasibility of generating complementary parameter maps and contrast-weighted images from a single scan.

Figure 3.5 compares the results of another representative 2D slice from the lower brain. Regions of the vasculature are zoomed in and expanded at the bottom right corners. Because of the blood flow, MRF cannot retrieve accurate parameter maps by dictionary matching [30]. Therefore, synthesis via parameters fail to deliver precise contrasts in the vasculature regions (as shown in T2w images). In comparison, **N-DCSNet** produces accurate contrast and can successfully reconstruct the delicate vessel structures (as shown in T2w results). From the

synthesized FLAIR images, we observe that PixelNet produces noisier images with flattened contrasts in the back of the brain. Instead, **N-DCSNet** successfully depicts the detailed textures and produces high-quality, sharper images.

Figure 3.6 displays an extensive collection of **N-DCSNet** synthesized images, accompanied by the corresponding parameter maps (*i.e.*, PD, T1, T2). These parameter maps are obtained through dictionary matching. Our approach highlights the capability to produce complementary parameter maps and contrast-weighted images from a single scan.

Table 3.1 compiles the quantitative evaluation metrics (nRMSE, PSNR, SSIM, LPIPS, and FID) of different methods (synthesis via parameters [133], PixelNet [120] and **N-DCSNet**) for each contrast. We compute the metrics across the testing dataset and report the mean and standard deviation (std). As indicated in the table, for all three contrasts (T1w, T2w, and FLAIR), our method consistently outperforms other methods in all five evaluation metrics. Of note, LPIPS and FID use learned features to measure perceptual similarity between two images or two distributions, thus resulting in better matching with human judgment than pixel-wise (nRMSE) or patch-wise (SSIM) metrics [145, 40]. **N-DCSNet**, compared with PixelNet, significantly reduces the LPIPS by more than 30% and the FID by more than 50% for all three contrasts, thus demonstrating the superiority of our proposed method in terms of perceptual image quality.

Contrasts	Methods	nRMSE (%) ↓	PSNR (dB) ↑	SSIM ↑	LPIPS ↓ ($\times 10^{-2}$)	FID ↓
T1w	Synthesis via parameters [133]	6.44 ± 1.25	24.0 ± 1.93	0.786 ± 0.030	20.1 ± 1.14	130.8
	PixelNet [120]	4.58 ± 0.83	26.9 ± 1.71	0.880 ± 0.026	11.3 ± 1.81	109.6
	N-DCSNet (ours)	**3.57 ± 0.64**	**29.1 ± 1.63**	**0.923 ± 0.019**	**6.33 ± 1.87**	**57.32**
T2w	Synthesis via parameters [133]	13.4 ± 1.68	17.5 ± 1.11	0.671 ± 0.032	21.1 ± 1.60	148.1
	PixelNet [120]	5.24 ± 0.64	25.7 ± 1.11	0.853 ± 0.027	12.6 ± 1.92	114.1
	N-DCSNet (ours)	**3.76 ± 0.59**	**28.6 ± 1.35**	**0.921 ± 0.017**	**5.77 ± 1.02**	**57.01**
FLAIR	Synthesis via parameters [133]	19.4 ± 2.75	14.3 ± 1.25	0.576 ± 0.028	20.6 ± 2.50	185.4
	PixelNet [120]	4.69 ± 0.67	26.7 ± 1.30	0.797 ± 0.025	11.3 ± 1.35	126.9
	N-DCSNet (ours)	**3.64 ± 0.65**	**29.0 ± 1.75**	**0.883 ± 0.018**	**8.63 ± 0.839**	**63.17**

Table 3.1: **Quantitative comparisons (nRMSE, PSNR, SSIM, LPIPS, and FID) among different contrast synthesis methods (mean ± standard deviation).** We calculate the metrics for each contrast (T1w, T2w, and FLAIR) separately. **N-DCSNet** is compared with contrast synthesis via parameters [133] and PixelNet [120]. Our proposed method consistently outperforms other approaches in all five metrics for each contrast. **Bold** corresponds to the best results. ↑ means that higher is better, ↓ means that lower is better.

Table 3.2 summarizes the inference times of the different approaches. As indicated in the table, simulation-based synthesis (synthesis via parameters) requires an average of 24.37 seconds because of the time-consuming dictionary matching and contrast simulation proce-

	Synthesis via parameters	PixelNet	N-DCSNet (ours)
Inference time (s) ↓	24.37	0.3421	**0.01617**

Table 3.2: **Inference times of different methods for contrast synthesis from a 2D MRF time series. N-DCSNet** reduces the inference time by more than 20 fold with respect to that of PixelNet, demonstrating superior computation efficiency and the potential for clinical adoption. All experiments are implemented on a single NVIDIA 3090 GPU for fair comparison. **Bold** corresponds to the best result.

dures that are repeated for each voxel across the entire image. PixelNet is more efficient, and averages 0.3421 seconds by leveraging parallel GPU computing (on a single NVDIA 3090). In comparison, our **N-DCSNet** has 20 times faster inference time than PixelNet. **N-DCSNet** requires an average of 0.01617 seconds to synthesize three contrast-weighted images from a single 2D MRF time series, demonstrating superior computation efficiency and a potential for clinical translation. All experiments were run on a single NVIDIA 3090 GPU.

3.4 Ablation study of different loss functions

To investigate and better understand the effects of loss functions (§ 3.2) on the resulting image quality, we conducted an ablation study by comparing our overall loss function L_{G} (Equation 3.6) to L_{ℓ_1}, $L_{\ell_1} + \lambda_{\mathrm{vgg}} L_{\mathrm{vgg}}$ and $L_{\ell_1} + \lambda_{\mathrm{adv}} L_{\mathrm{adv}}$ losses. We trained separate models with different objective functions and used the same training setup and datasets (i.e., training set, learning rate, epochs, etc.). Figure 3.7 shows the results on a representative 2D brain slice. The model trained with pure L_{ℓ_1} (left column) suffers from degraded perceptual image quality and exhibits some blurring, in agreement with the findings in literature [125, 44, 123]. Adding perceptual VGG loss (second column) encourages finer details and sharper edges. However, blurring artifacts remain (as seen in T2w and FLAIR). Adding adversarial loss on top of L_{ℓ_1} (third column) encourages even finer structures but suffers from residual blurring (T2w) and recurrent checkerboard artifacts (FLAIR). By incorporating both perceptual loss and adversarial loss, the model trained with our proposed objective (fourth column, Equation 3.6) further improves the synthesized image quality by reconstructing more delicate textures (T2w example) and producing more faithful contrast (FLAIR example).

Table 3.3 summarizes the five evaluation metrics for **N-DCSNet** trained with the different loss functions. Because the model trained with pure L_{ℓ_1} loss optimizes the pixel distances, it produces the best nRMSE and PSNR results. However, nRMSE and PSNR are known not to match human perception [145]. For perception-representative metrics (SSIM, LPIPS, and FID), **N-DCSNet** trained with our proposed full objective outperforms the other loss

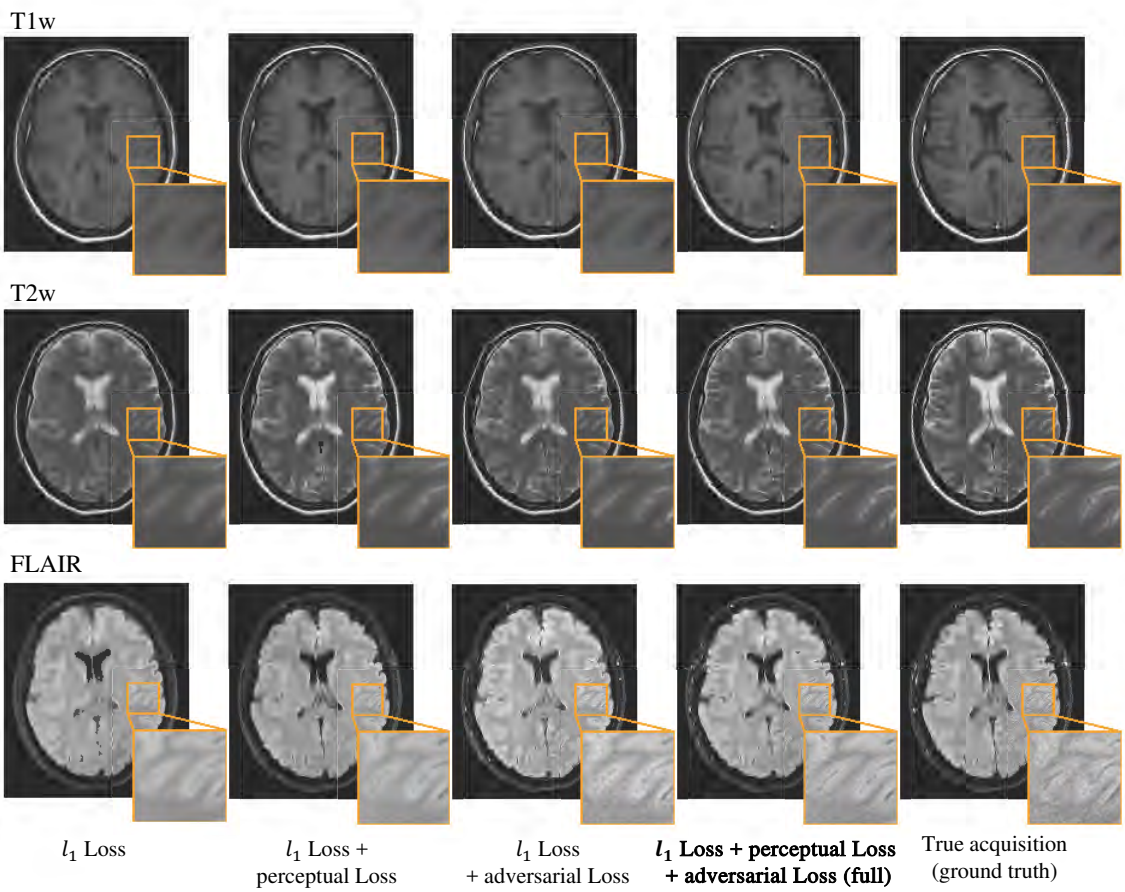


Figure 3.7: **Representative visual comparison of N-DCSNet with different loss functions.** From left to right, our full objective (fourth column; Equation 3.6) is compared with L_{ℓ_1}, $L_{\ell_1}+\lambda_{\text{vgg}}L_{\text{vgg}}$, $L_{\ell_1}+\lambda_{\text{adv}}L_{\text{adv}}$ and the ground truth. Perceptual VGG loss encourages sharper edges than pure L_{ℓ_1}, whereas adversarial loss further improves the image quality. The model trained with our full objective is able to recover subtle structures and show better visual agreement with the ground truth.

functions for all three contrasts (except SSIM for T1w), thus demonstrating the effectiveness of our loss functions in producing high-fidelity contrast-weighted images.

3.5 Mitigation of spiral off-resonance artifacts

Beyond the aforementioned superior performance, we also demonstrate cases in which **N-DCSNet** effectively mitigates the off-resonance artifacts within the MRF time series caused by B0 inhomogeneity and the long readout time of spiral acquisitions. Previous studies have

Contrasts	Methods	nRMSE (%) ↓	PSNR (dB) ↑	SSIM ↑	LPIPS ↓ ($\times 10^{-2}$)	FID ↓
T1w	L_{ℓ_1}	**3.34 ± 0.63**	**29.7 ± 1.69**	0.918 ± 0.018	8.02 ± 2.40	67.39
	$L_{\ell_1} + \lambda_{\text{vgg}} L_{\text{vgg}}$	3.43 ± 0.82	29.5 ± 2.16	**0.926 ± 0.022**	9.14 ± 2.53	66.66
	$L_{\ell_1} + \lambda_{\text{adv}} L_{\text{adv}}$	3.68 ± 0.93	28.7 ± 1.72	0.921 ± 0.020	8.04 ± 2.18	62.94
	$L_{\ell_1} + \lambda_{\text{vgg}} L_{\text{vgg}} + \lambda_{\text{adv}} L_{\text{adv}}$	3.57 ± 0.64	29.1 ± 1.63	0.923 ± 0.019	**6.33 ± 1.87**	**57.32**
T2w	L_{ℓ_1}	**3.57 ± 0.67**	**29.2 ± 1.64**	0.914 ± 0.018	10.08 ± 1.73	71.44
	$L_{\ell_1} + \lambda_{\text{vgg}} L_{\text{vgg}}$	3.67 ± 0.61	28.8 ± 1.45	0.918 ± 0.018	8.67 ± 1.46	64.55
	$L_{\ell_1} + \lambda_{\text{adv}} L_{\text{adv}}$	3.79 ± 0.72	28.4 ± 1.55	0.919 ± 0.024	7.57 ± 1.18	60.30
	$L_{\ell_1} + \lambda_{\text{vgg}} L_{\text{vgg}} + \lambda_{\text{adv}} L_{\text{adv}}$	3.76 ± 0.59	28.6 ± 1.35	**0.921 ± 0.017**	**5.77 ± 1.02**	**57.01**
FLAIR	L_{ℓ_1}	**3.44 ± 0.66**	**29.4 ± 1.72**	0.879 ± 0.017	11.1 ± 0.98	93.01
	$L_{\ell_1} + \lambda_{\text{vgg}} L_{\text{vgg}}$	3.73 ± 0.61	28.7 ± 1.55	0.878 ± 0.019	10.7 ± 1.02	96.01
	$L_{\ell_1} + \lambda_{\text{adv}} L_{\text{adv}}$	3.68 ± 0.93	28.1 ± 1.71	0.869 ± 0.021	9.62 ± 1.08	78.71
	$L_{\ell_1} + \lambda_{\text{vgg}} L_{\text{vgg}} + \lambda_{\text{adv}} L_{\text{adv}}$	3.64 ± 0.65	29.0 ± 1.75	**0.883 ± 0.018**	**8.63 ± 0.839**	**63.17**

Table 3.3: **Quantitative comparisons (nRMSE, PSNR, SSIM, LPIPS, and FID) of N-DCSNet with different loss function designs (mean ± standard deviation).** The model trained with pure L_{ℓ_1} optimizes the per-pixel distances, producing the lowest nRMSE and highest PSNR. The model trained with our full objective outperforms other loss function designs in perceptual metrics SSIM, LPIPS, and FID. **Bold** corresponds to the best results. ↑ indicates that higher is better, ↓ indicates that lower is better.

demonstrated the feasibility and potential of deep learning in off-resonance corrections [142, 23]. As shown in Figure 3.4, 3.5, parameter-based synthesis and PixelNet present blurry scalp fat signals in boundary regions of the brain, because of the MRF off-resonance effects (seen in T1w). In comparison, benefiting from spatial convolutions, **N-DCSNet** reconstructs a clean and sharp scalp fat signal, overcomes the off-resonance artifacts, and agrees well with the ground truth.

Figure 3.8 shows a representative example in which the MRF time-averaged image and PixelNet exhibit off-resonance signal loss artifacts in the regions close to the skull (indicated by the zoomed-in details). **N-DCSNet** significantly reduces the artifacts and recovers the correct contrasts and structures. Some residual artifacts can be observed, as indicated by the red arrows. Figure 3.9 presents another example in which the MRF time-averaged image and PixelNet exhibit several off-resonance signal loss artifacts and geometric distortion near the nasal region. Most brain structures are blurred out, primarily because of the considerable B0 homogeneity and the long readout time for the spiral acquisition. As visualized in the figure, **N-DCSNet** accurately recovers most of the delicate brain structures near the nasal region. The red arrows indicate the residual artifacts.

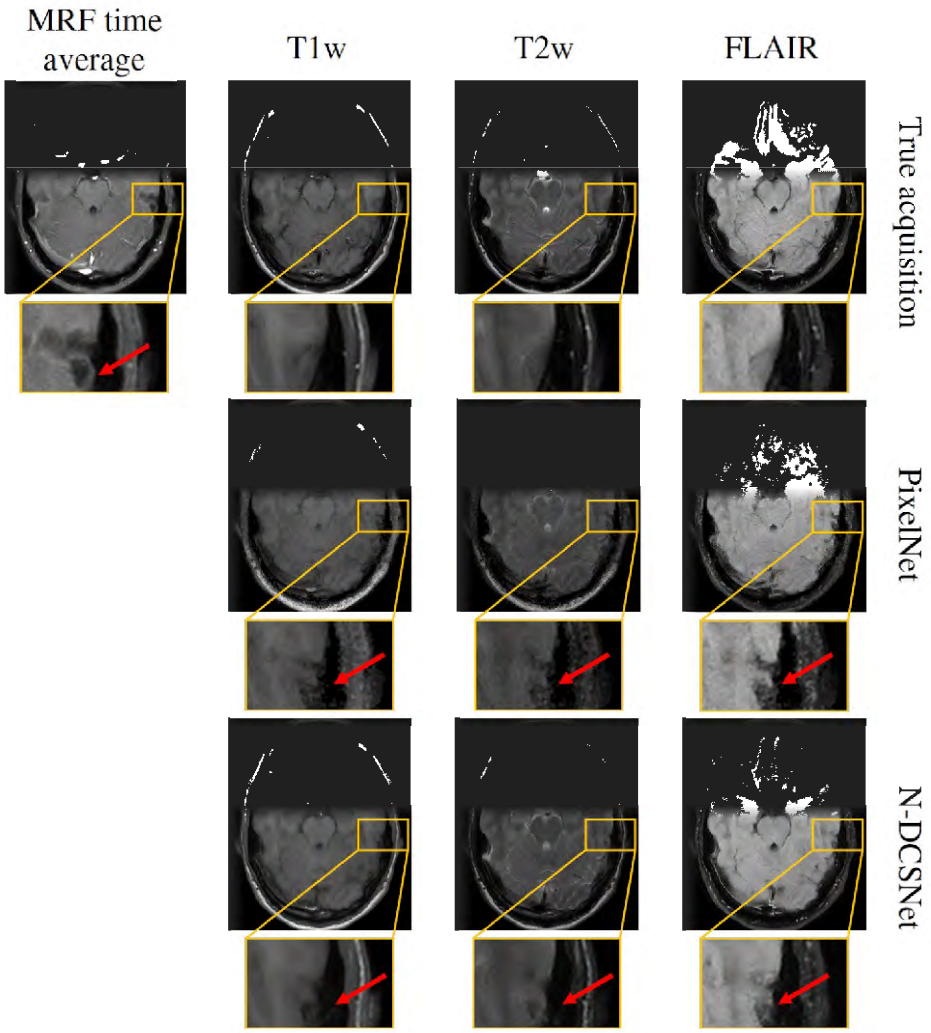


Figure 3.8: **Representative N-DCSNet results in mitigating spiral off-resonance artifacts in an MRF time series near the skull region.** The MRF time-averaged image and PixelNet results exhibit spiral off-resonance artifacts near the skull region (zoomed-in images) because of B0 inhomogeneity and the long readout time. **N-DCSNet** recovers the structure and produces contrast-weighted images with few residual artifacts. True acquisitions are displayed as references. Red arrows point to the regions with residual artifacts.

3.6 Discussion

In this work, we present a novel high-fidelity direct contrast synthesis framework, **N-DCSNet**, for synthesizing multi-contrast images from a single MRF scan. **N-DCSNet** directly learns a mapping between the MRF time series and the desired contrast weighted images (*i.e.*,

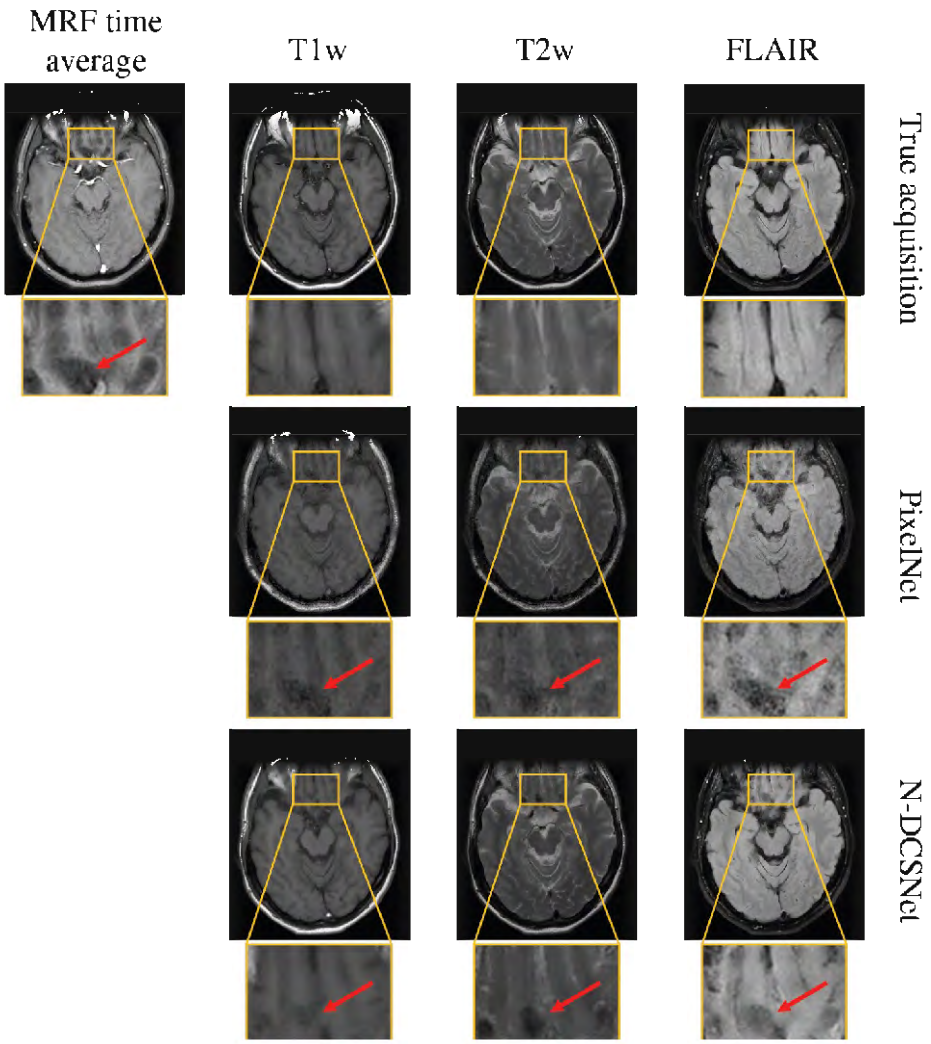


Figure 3.9: **Representative N-DCSNet results in mitigating off-resonance artifacts near the nasal region.** MRF time-averaged images display spiral off-resonance artifacts near the nasal region (as seen in zoomed-in images) due to the lengthy readout time. PixelNet also struggles to restore the structures and exhibits significant noise and distortions. **N-DCSNet** successfully mitigates the artifacts and produces contrast-weighted images with few residual artifacts. True acquisitions are displayed as references. Red arrows point to regions with residual artifacts.

T1w, T2w, and FLAIR) and thus bypasses the mapping and simulation steps required for contrast synthesis from parameter maps.

As briefly introduced in § 3.1, the sources of error contrast synthesis via parameter maps are attributed mainly to 1) factors that are not included in the dictionary simulation (e.g.,

B0/B1 homogeneity, slice profile, and flow effects), 2) approximation and error propagation in the contrast synthesis simulation (EPG algorithm) [133], and 3) artifacts (noise and aliasing) from highly undersampled MRF scans (example shown in Figure 3.1). As indicated by the visual comparison results (Figure 3.4, 3.5), the parameter-based contrast synthesis method does not deliver the correct contrast and produces noisier outputs (particularly for T2w and FLAIR results). One possible way to improve the results is modeling more parameters during the dictionary simulation procedure, such as B1 inhomogeneity [13], flow [30], and partial volume [25]. Unfortunately, including more simulation parameters forces the dictionary to grow in size, thereby prolonging the dictionary matching time (Table 3.2), or severely sacrificing parameter resolution and range.

Direct contrast synthesis leverages paired training data to learn a mapping from MRF signals to contrast-weighted images without explicitly modeling the aforementioned conditions. The previous DCS method PixelNet [120] proposed a 1D temporal CNN that maps the MRF time series at each pixel to the contrast weightings for that pixel and improves synthesized image quality and inference time (Table 3.2). However, because PixelNet treats each pixel independently, it does not leverage the unique spatial structural information within the MRF data. *In vivo* results (Figure 3.4, 3.5) indicate that PixelNet exhibits severe noise artifacts and diminished fine textures, particularly in FLAIR scans.

Our **N-DCSNet** shows significant improvements by introducing a conditional GAN-based framework with a spatial convolution network as the generator. **N-DCSNet** produces more faithful contrasts and is able to recover finer structures with overall better image quality than the other methods examined (Figure 3.4 and 3.5). Moreover, as described in section § 3.5 and shown in Figure 3.8 and 3.9, we demonstrate cases in which **N-DCSNet** effectively mitigates spiral off-resonance artifacts.

In our approach, we directly input the MRF time series to the network without performing pre-reconstruction on the MRF data. With the current MRF undersampling factor of 20, our method generates high-fidelity synthesized contrast-weighted images. For even higher undersampling factors (or for improved quality), incorporating pre-reconstruction techniques (e.g., subspace reconstruction [**zhao2018improved**]) could be a promising direction as it may yield less-aliased inputs. However, depending on the constraints, it could also result in the removal of some information and substantially lengthen the inference time.

Despite significant improvements over previous approaches, we observe that there remains some residual oversmoothing in our results compared to the ground truth (Figure 3.4 and 3.5). This could be attributed to the following reasons: 1) Limited training data constrains the GAN training potential, diminishes robustness against data outliers, and may potentially lead to oversmoothing. 2) MRF and ground truth contrast-weighted scans were obtained at different times. Despite careful experimental design and in-plane registration, small-scale through-plane motion and misalignment can cause oversmoothing. Improving the experimental setup (*e.g.*, hardware setups) to manage motion could mitigate this issue. 3) The MRF input images are relatively noisy due to the high undersampling rate and high resolution. The network is trained to reduce noise. However, this process (training on noisy inputs) can result in oversmoothing. We believe that some of these limitations can be

mitigated through improved experimental design and a larger training dataset.

Another limitation of this work is that the DCS frameworks (PixelNet and **N-DCSNet**) can generate only contrast-weighted images with fixed sequence parameters (e.g., TE or TR) and are therefore less flexible than simulation-based contrast synthesis from parameter maps. Separate networks must be trained for different MRF parameters or contrast acquisitions. Additionally, our **N-DCSNet** requires paired data; however, our approach allows each decoder branch to be trained independently, potentially relaxing this constraint, although further investigation is required. In this work, we trained **N-DCSNet** on a limited number of healthy volunteer data (21 examinations, 203 slices). To facilitate future clinical adoption, larger and more diverse clinical training data (*e.g.*, with pathology) are necessary.

In the future, we plan to extend our framework to more diverse contrast synthesis, including but not limited to gradient echo imaging, diffusion-weighted imaging, and susceptibility-weighted imaging.

3.7 Conclusion

In this work, we propose **N-DCSNet** to directly synthesize multi-contrast MR images from a single MRF acquisition. This method significantly reduces examination time. By directly training a network to generate contrast-weighted images from MRF, our method does not require any model-based simulation and therefore avoids reconstruction errors due to simulation. *In vivo* experiments demonstrate that **N-DCSNet** produces high-fidelity contrast-weighted images with sharper contrast and minimal artifacts (in-flow and spiral off-resonance artifacts), and significantly outperforms simulation-based contrast synthesis and PixelNet, both visually and according to metrics. Additionally, our proposed method can inherently mitigate some off-resonance artifacts within MRF data, thereby producing high-quality contrast-weighted images with minimal residual artifacts.

Chapter 4

Unsupervised Feature Loss for DL-based MRI reconstruction

4.1 Introduction

As we introduced in Chapter 2, DL-based MRI reconstruction methods, such as those presented in [16, 68, 92, 99, 37, 2, 110], have achieved remarkable success by learning regularization terms directly from extensive training datasets, surpassing the capabilities of PI and CS techniques. However, it is well-established that the performance of DL-based methods relies heavily on the loss functions employed during training. Commonly used loss functions for training include pixel-wise ℓ_1, ℓ_2, and patch-wise structural similarity index (SSIM) [130] losses [2, 37, 16]. These loss functions, though, are often hand-crafted or based on local statistics and may not accurately capture the perceptual information of fine structures. As a result, reconstructed images may exhibit degraded perceptual quality and blurring compared to un-accelerated scans [68, 136].

To address these issues, Generative Adversarial Networks (GANs) [33, 44, 71] with adversarial losses have been proposed to exploit the implicit feature information by incorporating discriminators into the reconstruction pipeline [61, 68, 136]. Unfortunately, GANs are notoriously hard to train, easily fall into mode collapse, and are sensitive to hyperparameter selections. Additionally, the adversarial loss is a less-constrained instance-to-set loss function, where improper training parameters may result in unexpected instabilities during the training and artifacts in the reconstructions [96, 72].

Aside from the adversarial loss, recent works in computer vision have shown that CNN-based perceptual losses can be used to learn high-level image feature representations [49, 145]. These perceptual loss functions are based on feature layers of classification networks (such as the VGG Net [104]). They are typically designed to work for natural images with a fixed channel number (RGB) and are usually trained in a supervised manner with human-annotated labels, *e.g.*, from ImageNet [24]. Therefore, simply using perceptual VGG losses may not be ideal for MRI reconstruction tasks. For MR data sets, the dimensionality of

the data can vary from application to application (*e.g.*, 2D/3D complex-valued data, 2D/3D dynamic data), while at the same time, human-annotated labels for MR images are much harder to obtain. More importantly, it is also unclear what kind of human annotations would be best for comparing the image quality for MR images.

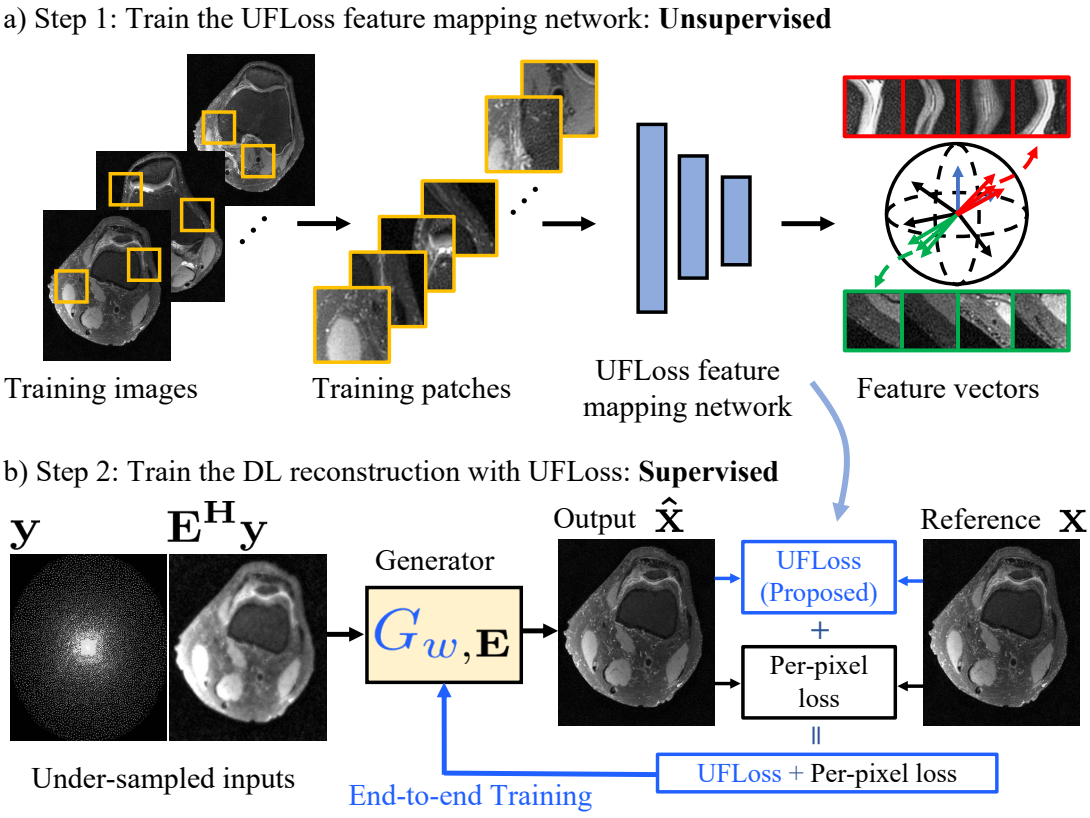


Figure 4.1: **Overview of training the DL-based reconstruction with UFLoss.** We split the pipeline into two steps. a) Step 1: We pre-train the UFLoss feature mapping network on fully-sampled image patches without human annotations, where the aim of the training is to maximally separate out all the patches in the feature space. b) Step 2: For the training of the DL-based reconstruction, $G_{w,\mathbf{E}}$ represents a reconstruction network with learnable parameters w, and given system encoding operator $\mathbf{E}$. The inputs of $G_{w,\mathbf{E}}$ are under-sampled k-space $\mathbf{y}$, and zero-filled reconstruction $\mathbf{E^H y}$. We feed-forward $\mathbf{E^H y}$ through $G_{w,\mathbf{E}}$ to obtain the output reconstruction results. We adopt the pre-trained UFLoss network from (a) to compute the UFLoss in the feature space. Then, end-to-end training is performed with respect to the combination of UFLoss and per-pixel loss. Note that the training of DL-based reconstruction with UFLoss is still supervised.

In this Chapter, we propose a novel unsupervised learned feature loss (Figure 4.1) to capture the perceptual and high-order statistical difference within MR images, which we call Unsupervised Feature Loss (UFLoss). The UFLoss is a large-patch-wise loss function that provides instance-level discrimination by mapping similar patches to similar low-dimensional feature vectors using a pre-trained mapping network (which we refer to as UFLoss feature mapping network or UFLoss network) [126]. The rationale of using features from large-patches (typically 40×40 pixels for a 300×300 pixels image) is that we want our UFLoss to capture mid-level structural and semantic features instead of using small patches (typically around 10×10 pixels), which only contain local edge information. On the other hand, we avoid using global features due to the fact that our training set (typically around 5000 slices) is usually not large enough to capture common and general features at a large-image scale.

Different from adversarial loss, UFLoss is a more-constrained instance-to-instance loss function, which leads to more stable training with clear and straightforward stop criterion. Meanwhile, unlike the VGG perceptual loss, pre-training the UFLoss network requires no supervision, and thus is able to capture high-level structural information specifically for MR images without any human annotations. Similar to the VGG perceptual loss, UFLoss can also be easily incorporated into the training of DL-based reconstruction networks without modifying the network architecture. Figure 4.1 shows the overall pipeline for using our UFLoss to train a DL-based reconstruction. We first pre-train the UFLoss network on fully sampled image patches without accompanying annotated labels (Figure 4.1a). This step maps patches to a lower-dimensional space while attempting to maximally separate them in the feature space. The outcome is that similar patches end up being close together in the feature space while dissimilar ones end up further apart. This pre-trained feature mapping network is then adopted to compute the UFLoss during the training of the DL-based reconstruction (Figure 4.1b), which corresponds to the ℓ_2 distance in the feature space summed across all images patches. End-to-end training is performed with respect to a combination of UFLoss and per-pixel ℓ_1/ℓ_2 or SSIM losses.

To demonstrate the power of UFLoss, we focus on a representative unrolled DL-based reconstruction framework: MoDL [2]. We conduct experiments to show that UFLoss is a valid loss function sensitive to increasing low-level intensity deformation. Our results for patch retrieval and patch correlation in MR images demonstrate that *visually similar* patches are indeed close in the feature space.

In terms of computation costs, our UFLoss is added during training as an additional loss function without modifying the reconstruction network architecture. This imposes about 50% increase in training time and memory requirements during training. However, in inference time, the UFLoss has no impact at all on the reconstruction time as well as the memory requirements. Our experiments on 2D and 3D in-vivo data show that the addition of the UFLoss encourages more realistic reconstructions with more subtle details and improved overall image quality compared to conventional and learning-based methods with other losses (pure ℓ_2 loss and ℓ_2+VGG perceptual loss).

4.2 Unrolled reconstruction for under-sampled MRI

In conventional under-sampled MRI, the PICS inverse problem can be formulated as [64]:

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \frac{1}{2} \|\mathbf{E}\mathbf{x} - \mathbf{y}\|_2^2 + \lambda Q(\mathbf{x}), \tag{4.1}$$

where $\mathbf{x}$ is the image to be reconstructed, and $\mathbf{y}$ is the measured data in k-space. $\mathbf{E}$ describes the system encoding matrix, which can be further expanded to: $\mathbf{E} = \mathbf{UFS}$, where $\mathbf{F}$ is the Fourier transform operator, $\mathbf{S}$ represents the multiple sensitivity maps, and $\mathbf{U}$ corresponds to the k-space sampling operator. For the Cartesian case, $\mathbf{U}$ is a diagonal matrix with 1's corresponding to collected k-space and 0's to un-acquired k-space. For non-Cartesian, $\mathbf{U}$ is a k-space re-sampling operator from a Cartesian grid to the acquired non-Cartesian trajectory. The goal of this problem is to reconstruct the image which has the lowest error compared to the measured k-space data in the least-squares sense. However, when the sampling rate is below the Nyquist rate, Equation 4.1 becomes ill-posed. Therefore, a regularization term $Q(\mathbf{x})$ with a weighting parameter λ, which incorporates prior knowledge about the image, is added to constrain the optimization problem. For conventional CS MRI, $Q(\mathbf{x})$ is often chosen to promote sparsity in a certain transform domain such as wavelets or finite spatial differences.

A number of first-order iterative methods have been developed for efficiently solving the minimization problem in Equation 4.1 for the case where $Q(\mathbf{x})$ is convex [9, 11]. To further develop fast and high-fidelity reconstructions, recent methods have attempted to directly learn the proximal function Q and the corresponding parameters from a large set of fully-sampled training data in an unrolled fashion [16, 68, 92, 99, 37, 2, 110].

A widely used unrolled reconstruction framework is MoDL [2], where the reconstruction is formulated as:

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \|\mathbf{E}\mathbf{x} - \mathbf{y}\|_2^2 + \lambda \|\mathbf{x} - D_w(\mathbf{x})\|_2^2 . \tag{4.2}$$

In this formulation, D_w is a learned CNN denoiser/artifact removal network and w are the learned weighting parameters. The CNN-based prior $\|\mathbf{x} - D_w(\mathbf{x})\|_2^2$ results in high values when $\mathbf{x}$ is corrupted by noise and aliasing. Similar to ADMM [11], we can solve the optimization problem in the following half-quadratic splitting steps:

$$\mathbf{z}^k = D_w(\mathbf{x}^k) \tag{4.3}$$

$$\begin{aligned} \mathbf{x}^{(k+1)} &= \arg\min_{\mathbf{x}} \|\mathbf{E}\mathbf{x} - \mathbf{y}\|_2^2 + \lambda \left\|\mathbf{x} - \mathbf{z}^k\right\|_2^2 \\ &= (\mathbf{E}^{\mathbf{H}}\mathbf{E} + \lambda \mathbf{I})^{-1}(\mathbf{E}^{\mathbf{H}}\mathbf{y} + \lambda \mathbf{z}^k) \end{aligned} \tag{4.4}$$

Equation 4.4 can be solved using the Conjugate Gradient (CG) Method while Equation 4.3 is viewed as a CNN-based forward-pass step. MoDL is formulated as an unrolled network, where in each iteration, a CG layer is followed by a CNN-based proximal step. The unrolled reconstruction can be denoted as $\hat{\mathbf{x}} = G_w(\mathbf{y}, \mathbf{E})$, where $\mathbf{y}$, $\mathbf{E}$ and w correspond to the

under-sampled k-space measurements, the encoding matrix, and the learnable weights of the reconstruction network, respectively. Training the unrolled model becomes supervised learning with a pre-defined loss function:

$$\min_{w} \sum_{i} \mathcal{L}(G_w(\mathbf{y}_i, \mathbf{E}_i), \mathbf{x}_i), \tag{4.5}$$

where $\mathbf{x}_i$ is the i^{th} fully-sampled ground truth image, and $\mathbf{y}_i$ is the retrospectively under-sampled k-space computed by applying the encoding matrix $\mathbf{E}_i$ to generate $\mathbf{y}_i = \mathbf{E}_i\mathbf{x}_i$. The loss function $\mathcal{L}(\cdot)$ can be combinations of ℓ_1, ℓ_2, SSIM, and other losses. Once trained, a new under-sampled scan denoted by $\mathbf{y}$ with the encoding operator $\mathbf{E}$ is reconstructed as:

$$\hat{\mathbf{x}} = G_w(\mathbf{y}, \mathbf{E}). \tag{4.6}$$

4.3 UFLoss feature mapping network

As shown in Figure 4.2a), a patch-wise mapping network (UFLoss feature mapping network) is trained to map patches from image-space to a low-dimensional unit-norm feature space, aiming to capture high-level structural differences. The UFLoss network can then be used for training a DL-based reconstruction. In contrast to conventional supervised computer vision tasks, the UFLoss network is trained from fully sampled image patches in an unsupervised fashion. In other words, the training does not use any human annotation, which has been challenging to obtain in large-scale MRI datasets. The training is motivated by contrastive learning [135], where a feature mapping function f_θ is learned such that each patch is maximally separated from other patches in a lower-dimensional hypersphere feature space.

Mathematically, we formulate our unsupervised feature mapping using the softmax criterion. Suppose we have N patches $\{\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_N\}$ cropped from the fully sampled images from the training set, with their corresponding unit-norm features $\{\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_N\}$ with $\mathbf{v}_i = f_\theta(\mathbf{p}_i) \in \mathbb{R}^d$. For a certain patch $\mathbf{p}$ with feature $\mathbf{v} = f_\theta(\mathbf{p})$, the probability of it being identified as the i^{th} patch under a linear classifier is:

$$P(i|\mathbf{v}) = \frac{\exp\left(\mathbf{w}_i^{\mathbf{T}}\mathbf{v}\right)}{\sum_{j=1}^{N} \exp\left(\mathbf{w}_j^{\mathbf{T}}\mathbf{v}\right)}, \tag{4.7}$$

where $\mathbf{w}_j$ is the weight vector of class j (or patch j), and $\mathbf{w}_j^{\mathbf{T}}\mathbf{v}$ shows how well the feature vector $\mathbf{v}$ matches the j^{th} patch. However, the above formulation Equation 4.7 requires a class prototype $\mathbf{w}$ in addition to the patch feature itself, making direct comparison between patches infeasible. To address this problem, we follow the approach in [135] to turn the instance-wise classification into a metric learning problem, where $\mathbf{w}_j^{\mathbf{T}}\mathbf{v}$ in Equation 4.7 is replaced with $\mathbf{v}_j^{\mathbf{T}}\mathbf{v}$. That is, the j^{th} patch feature is its class prototype itself. The probability then becomes:

$$P(i|\mathbf{v}) = \frac{\exp\left(\mathbf{v}_i^{\mathbf{T}}\mathbf{v}\right)/\tau}{\sum_{j=1}^{N} \exp\left(\mathbf{v}_j^{\mathbf{T}}\mathbf{v}\right)/\tau}, \tag{4.8}$$

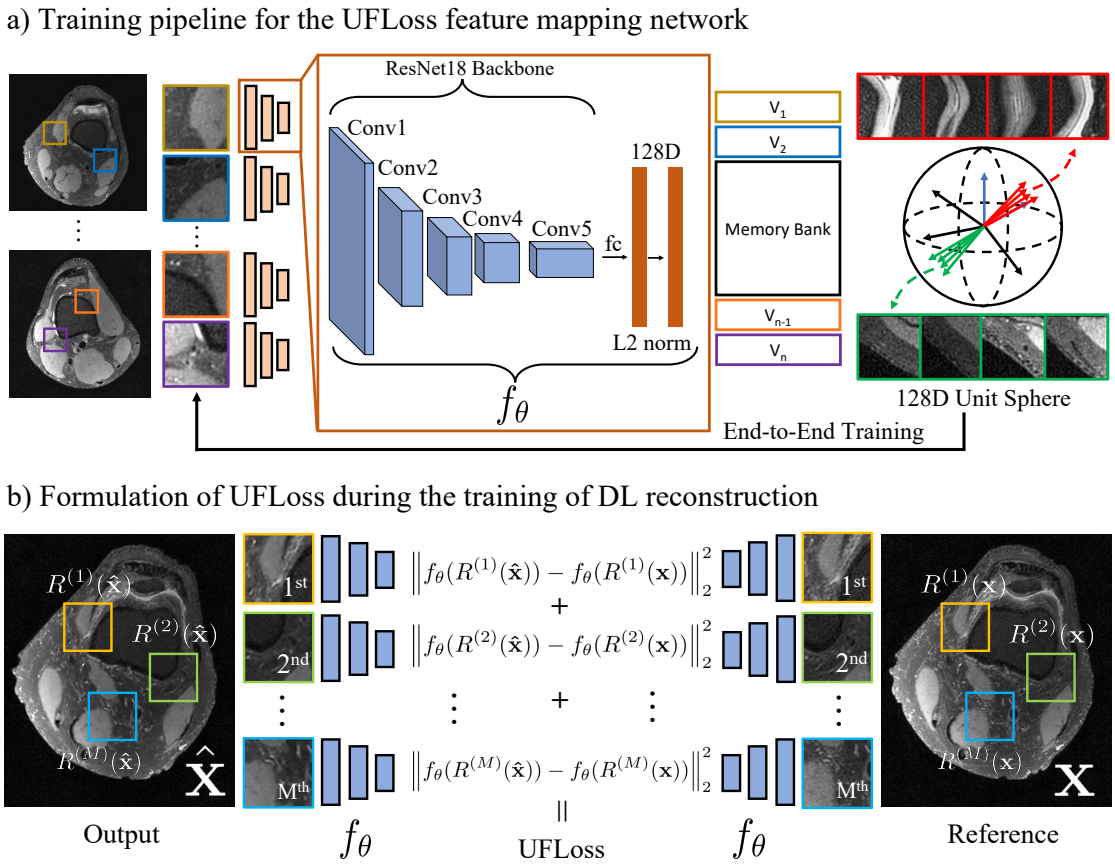


Figure 4.2: a) **Training pipeline for the UFloss feature mapping network.** Patches cropped from the fully sampled images are separately passed through a ResNet 18 [39] backbone followed by an ℓ_2 normalization layer to map the patches to features on a low-dimensional unit sphere (128-dimension unit-norm features in this work). A memory bank is used to store the features from all the training patches to save computation when computing the softmax loss function (Equation 4.9). Then, end-to-end training is performed such that each patch is maximally separated from other patches in the 128D unit-norm feature space. Similar patches will naturally cluster in the low-dimensional space. b) **Detailed formulation of the proposed UFLoss during the training of the DL-based reconstruction.** Operator R extracts a total of M patches from an image. These patches are extracted on a grid with a sliding window. Each patch from the reconstructed output and the fully-sampled reference will go through a pre-trained network f_θ and mapped to a low-dimensional feature space. The UFLoss corresponds to the sum of the ℓ_2 distance between the feature vectors from the output and the fully-sampled reference.

where τ is a temperature parameter that controls the extent of separation/concentration of the distribution in the feature space. The learning objective is set to maximize the joint probability $\Pi_{i=1}^{N} P_\theta(i|f_\theta(\mathbf{x}_i))$, which is equivalent to minimizing the negative log-likelihood over the training set:

$$J(\theta) = -\sum_{i=1}^{N} \log P(i|f_\theta(\mathbf{x_i})). \tag{4.9}$$

Note that in order to compute the probability $P(i|\mathbf{v})$ in Equation 4.8, features $\{\mathbf{v}_i\}$ from all the patches are required. Instead of exhaustively computing all the features every time, a memory bank $\mathbf{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_N\}$ is constructed to store all the feature vectors. During each training iteration, while the network parameters θ are optimized over the i^{th} patch, the i^{th} entry of the memory bank $\mathbf{v}_i$ is replaced by the output of the feature mapping network $\mathbf{f}_\theta(\mathbf{p}_i) \rightarrow \mathbf{v}_i$.

Once trained, the UFLoss network can be used as a perceptual loss term in other supervised reconstruction tasks, as described next.

4.4 Deep learning-based reconstruction with UFLoss

The UFLoss network is designed to maximally separate patches in the low-dimensional unit-sphere feature space. Perceptually similar patches are mapped to similar features.

Consider the under-sampled reconstruction using an unrolled network in Equation 4.5. Suppose we have the ground truth fully-sampled image $\mathbf{x}_i$, and the output of the unrolled network $\hat{\mathbf{x}}_i = G_w(\mathbf{y}_i, \mathbf{E}_i)$. Since the inputs of the UFLoss network are image patches (Figure 4.2b), we first extract M overlapping image patches from both $\mathbf{x}_i$ and $\hat{\mathbf{x}}_i$, obtaining two patch groups: $\{\mathbf{p}_i^1, \mathbf{p}_i^2, ..., \mathbf{p}_i^M\}$ and $\{\hat{\mathbf{p}}_i^1, \hat{\mathbf{p}}_i^2, ..., \hat{\mathbf{p}}_i^M\}$. The patches are extracted on a grid with N_s pixel strides horizontally and vertically.

During each training step, random shifts between 0 to N_s pixels are applied with equal shifts to both $\mathbf{x}_i$ and $\hat{\mathbf{x}}_i$. This choice has the effect of averaging out the blocking artifacts and achieves the same performance as extracting all the patches [64, 111].

Since we use inner products to measure the distance in the hyperspherical feature space, the UFLoss can be formulated as the average of the negative inner products over all the patches. On top of that, we add a constant 1 in front of our loss function:

$$L_{UFLoss}(\mathbf{x}_i, \hat{\mathbf{x}}_i) = \frac{1}{M} \sum_j 1 - \langle f_\theta(\mathbf{p}_i^j), f_\theta(\hat{\mathbf{p}}_i^j) \rangle, \tag{4.10}$$

where $\langle \cdot, \cdot \rangle$ is the inner product operation between two unit-norm vectors and f_θ is the pre-trained UFLoss mapping network. As both $f_\theta(\mathbf{p}_i^j)$ and $f_\theta(\hat{\mathbf{p}}_i^j)$ have unit norms, the above loss function can be also written as a mean-squared-error (MSE) in the feature space, or:

$$
\begin{aligned}
L_{UFLoss}(\mathbf{x}_i, \hat{\mathbf{x}}_i) &= \frac{1}{M}\sum_j 1 - \langle f_\theta(\mathbf{p}_i^j), f_\theta(\hat{\mathbf{p}}_i^j)\rangle \\
&= \frac{1}{2M}\sum_j \left\|f_\theta(\mathbf{p}_i^j)\right\|_2^2 - 2\langle f_\theta(\mathbf{p}_i^j), f_\theta(\hat{\mathbf{p}}_i^j)\rangle + \left\|f_\theta(\hat{\mathbf{p}}_i^j)\right\|_2^2 \\
&= \frac{1}{2M}\sum_j \left\|f_\theta(\mathbf{p}_i^j) - f_\theta(\hat{\mathbf{p}}_i^j)\right\|_2^2 .
\end{aligned} \tag{4.11}
$$

Following the per-pixel ℓ_2 loss and UFLoss mentioned above, the full objective loss function for the DL-based reconstruction can be written as:

$$
\begin{aligned}
L_{Recon} &= L_{MSE-all} + 2\mu L_{UFLoss-all} \\
&= \sum_i L_{MSE}(\mathbf{x}_i, \hat{\mathbf{x}}_i) + 2\mu \sum_i L_{UFLoss}(\mathbf{x}_i, \hat{\mathbf{x}}_i) \\
&= \sum_i \|G_w(\mathbf{y}_i, \mathbf{E}_i) - \mathbf{x}_i\|_2^2 + \mu \sum_i \frac{1}{M}\sum_j \left\|f_\theta(\mathbf{p}_i^j) - f_\theta(\hat{\mathbf{p}}_i^j)\right\|_2^2 ,
\end{aligned} \tag{4.12}
$$

where μ is the weighting factor on the contribution of the UFLoss. End-to-end training is then performed on this total loss to optimize the reconstruction network G_w.

4.5 Datasets and implementations

Imaging datasets

We trained and evaluated our proposed UFLoss on both 2D and 3D fully-sampled knee datasets with retrospective under-sampling. We used the fastMRI [141] high-resolution knee data set for our 2D experiments. A total of 5700 fully-sampled slices from 380 cases were split into 320 cases (6080 slices) for training, 40 cases (640 slices) for validation, and 20 cases (320 slices) for testing. Image normalization was performed such that the 95% percentile of the intensity values was scaled to 1 for each subject. The training dataset includes data from two different contrasts: proton-density with (PDFS) and without (PD) fat suppression. Relevant imaging parameters are described in the fastMRI [141] paper. For the unrolled reconstruction task, retrospective under-sampling was performed by applying a 1D five times accelerated random under-sampling mask (20% sampling rate) with an 8% fully sampled k-space center. Sensitivity maps were computed using ESPIRiT [116] using BART [115] with a 24×24 calibration region.

We conducted our 3D experiments on 20 fully sampled 3D knee scans (available at mridata.org) [98] with retrospective under-sampling. The k-space data was acquired on a 3T GE Discovery MR 750, with an 8-channel HD knee coil. Scan parameters include a matrix size of 320×320×256, and TE/TR of 25ms/1550ms. A total of 5120 slices from 16 cases were used for training, 640 slices from 2 cases were used for validation, and 640 slices from

the remaining 2 cases were used for testing. We normalized each 3D volume with respect to the 95% percentile of the intensity values for the entire volume. Each 3D volume was under-sampled with a different 8× Poisson-disk sampling mask (12.5% sampling rate) with a 24×24 calibration region. Sensitivity maps were computed using ESPIRiT [116] with a 24×24 calibration region using BART [115]. Note that we train both the UFLoss network and the DL-based reconstructions on the entire training set and use fully-sampled coil-combined images as ground truth.

Implementation of UFLoss feature mapping network

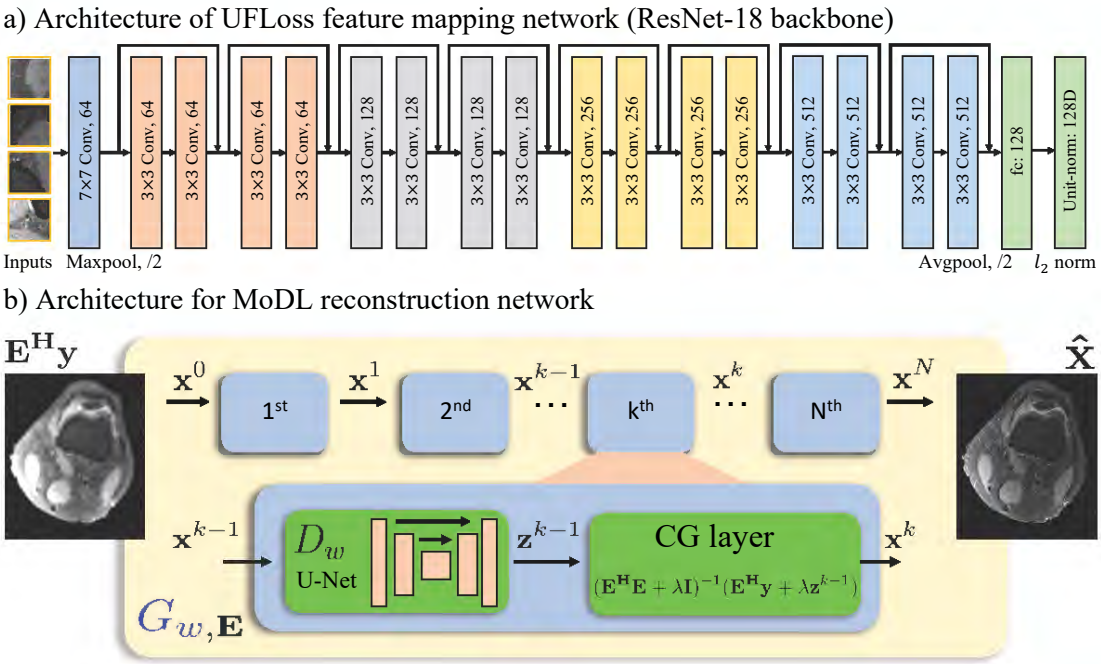


Figure 4.3: **Architectures for UFLoss feature mapping network and MoDL.** a) The UFLoss feature mapping network is based on a ResNet 18 network structure [39] and followed by an ℓ_2 normalization layer to map the input patches to the 128D unit-norm feature space. b) Architecture of the MoDL [2] reconstruction network. A data consistency Conjugate Gradient Descent (CG) module is inserted after a CNN-based denoiser D_w. D_w follows the structure of U-Net [94] with two input channels that represent the real and imaginary parts of the complex-valued image data.

In all our networks, the input coil-combined complex-valued MR images/patches $\mathbf{x} \in \mathbb{C}^{\mathbf{N}}$ are converted into a two-channel representation $\mathbf{x} \in \mathbb{R}^{\mathbf{2N}}$, where the real and imaginary

components are treated as two individual channels. As illustrated in Figure 4.3a), we implemented the UFLoss network using a ResNet 18 [39] backbone followed by a ℓ_2 normalization layer to map the input patches to 128 dimension unit-norm features. Based on the FOV and resolution difference, the input patch sizes of the 2D fastMRI knee dataset and 3D knee dataset were set to 60×60 and 40×40 pixels, respectively. The UFLoss networks for the 2D fastMRI and 3D knee datasets were trained separately due to the differences in image content. Eighty patches were extracted from each slice at random locations, resulting in 409,600 patches used to train the UFLoss network. Other hyperparameters include temperature τ of 1 (Equation 4.8), batch size 16, the number of epochs of 100, and the learning rate of 1e−4 with Adam [52] optimizer.

Implementation of DL-based reconstruction with UFLoss

For the unrolled reconstruction network architecture, we used the structure from the MoDL paper [2], where a CG block was inserted after a CNN-based denoiser, and unrolled with a fixed number of iterations. In this work, we used 5 unrolls and 6 CG steps. As shown in Figure 4.3b), a U-Net [94] architecture was adopted for the CNN-based denoiser D_w.

The training of MoDL was performed by minimizing the proposed loss function L_{Recon} (Equation 4.12) over the training set for 50 epochs, with an empirical weighting parameter $\mu = 1.5$, and Adam [52] optimizer with a learning rate of 1e−4.

To compute the UFLoss, patches are extracted on a grid across the image with 5-pixel strides in both vertical and horizontal directions. At each training step, both output and reference images are randomly shifted from 0 to 5 pixels in the vertical and horizontal directions to eliminate blocking artifacts. In this work, we chose the weighting parameter to balance the values of $L_{MSE-all}$ and $L_{UFLoss-all}$ so that they are on par after the training converges. During inference, a zero-filled reconstruction is passed through the MoDL reconstruction network. Note that training with UFLoss does not change the network architecture, so the inference time remains the same as MoDL with pure ℓ_2 loss.

All the proposed algorithms were implemented using Pytorch 1.2 [86], and were run on 12 GB Nvidia Titan Xp graphics processing units (GPUs).

4.6 Evaluation of the proposed UFLoss

UFLoss as valid loss function

To evaluate whether UFLoss is also a valid loss function for comparing two images at the intensity level, we study how the UFLoss changes with different sizes of perturbations in two representative types:

1. Additive white Gaussian noise.

   A perturbed image $\mathbf{x_p}$ is generated from the original image $\mathbf{x_o}$ by adding different levels of additive Gaussian noise $\mathbf{n}_\sigma$:

$$\mathbf{x_p} = (1-\beta)\mathbf{x_o} + \beta\mathbf{n}_\sigma, \tag{4.13}$$

where β is the noise level parameter in the range of $0-10\%$, and noise $\mathbf{n}_\sigma$ follows normal distribution: $\mathbf{n}_\sigma \sim \mathcal{N}(0,1)$. We study how $L_{UFLoss}(\mathbf{x_o}, \mathbf{x_p})$ changes as β increases.

2. Image blurring.

   A perturbed low-resolution image $\mathbf{x_p}$ is generated by cropping and zero-padding the k-space of the original image $\mathbf{x_o}$. The k-space cropping rate $\mathbf{R}$ ranges from 1-4. $\mathbf{R} = 4$ indicates that only 25% of k-space samples in both horizontal and vertical dimensions are kept. A higher $\mathbf{R}$ corresponds to more blurring and a coarser resolution. We study how $L_{UFloss}(\mathbf{x_o}, \mathbf{x_p})$ varies with different $\mathbf{R}$'s.

In addition, we evaluate whether, by minimizing the objective UFLoss between the original and perturbed images $L_{UFloss}(\mathbf{x_o}, \mathbf{x_p})$, we are able to guide the perturbed version towards the original version without falling into local minima. The starting perturbed image $\mathbf{x_{p-0}}$ is generated by image blurring where $\mathbf{R} = 4$. We update it per gradient descent with respect to $L_{UFloss}(\mathbf{x_o}, \mathbf{x_{p-k}})$ in an iterative fashion:

$$\mathbf{x_{p-k+1}} = \mathbf{x_{p-k}} - \alpha \frac{\partial L_{UFLoss}(\mathbf{x_o}, \mathbf{x_{p-k}})}{\partial \mathbf{x_{p-k}}}, \tag{4.14}$$

where $\mathbf{x_{p-k}}$ is the perturbed image after $\mathbf{k}$ steps of gradient descent.

Perceptual Similarity

In order to better interpret and understand the perceptual features learned for the UFLoss, we performed a patch retrieval experiment to evaluate and show patch pairs with high and low UFLoss feature similarities. First, we constructed a feature database (memory bank) by running all training patches through the pre-trained UFLoss network. Then, given an input patch from the testing set, we passed it through the network and queried its neighbors from the training patches based on their distances (inner products) in the feature space. We picked and visualized patches of the highest feature inner products with the input patch and also counter-examples with relatively low inner products.

To further evaluate the UFLoss sensitivity and perceptual similarity for different anatomies and contrasts, we constructed correlation maps by computing the feature correlation (inner product) between a source patch and all patches in different images and visualized them as heatmaps. This experiment helps us better understand how anatomy and structure similarities relate to UFLoss feature similarities.

Specifically, we first extracted a source patch from a source image. Then, we computed the feature correlations between the source patch and all patches on a grid from 1) the same source image; 2) the target image with the same contrast but from a different subject; and 3) the target image with different contrast and also from a different subject. Patches closer to the source patch in the feature space correspond to higher inner products. We evaluated

this experiment on both PDFS and PD scans. For comparisons, we also conducted the same experiments for the SSIM feature, where we computed the SSIM score between the source patch and all patches from different images.

Unrolled Reconstructions with UFLoss

To quantitatively evaluate our proposed UFLoss on under-sampled MRI reconstruction, we implemented both PICS [64] and MoDL [2]. In the unrolled reconstruction experiments, MoDL with our proposed UFLoss was compared with PICS and with MoDL using only per-pixel ℓ_2 loss. The PICS method was implemented using the BART Toolbox [115] with wavelets as the sparse transform. In order to further demonstrate the performance of our UFLoss, MoDL with ℓ_2 + perceptual VGG loss [49] was also included in our comparisons. To compute the perceptual VGG loss, both the real and imaginary parts are scaled from 0 to 255 and duplicated 3 times to serve as the inputs of the pre-trained VGG network. The VGG network is pre-trained on ImageNet classification. VGG loss corresponds to the ℓ_2 distance between the relu_22 features from the output and the ground truth image.

For all the experiments, reconstruction performance was evaluated using different quantitative metrics, which reflect different aspects of image quality. The normalized root mean squared error (NRMSE) was used to measure the overall pixel-wise errors. SSIM [130] was used to assess the local image similarity with respect to the fully sampled reference. At the same time, we also computed our proposed UFLoss between the reconstructed images and the fully sampled references.

4.7 Results

UFLoss as a valid loss function

4.4 indicates that our proposed UFLoss could be used as a valid loss function by itself. As shown in 4.4a), UFLoss between the perturbed and original clean images increases in a convex way with respect to more Gaussian noise and increases in a near-convex way with respect to more blurring. Even though the UFLoss feature mapping network is not specifically trained for any such perturbations, it learns low-level perceptual similarities between images, where a larger intensity perturbation corresponds to a larger UFLoss. On the other hand, 4.4b) indicates that by minimizing the UFLoss between the perturbed and target images, we are able to successfully restore the blurred image towards the clean one without falling into any local minimum. Intermediate deblurred image samples are shown in the along with the UFLoss evolution curve.

Perceptual Similarity

Figure 4.5a) shows the feature similarity results using the UFLoss feature. The feature space inner products between the input patch and the retrieved patches are shown as different colors

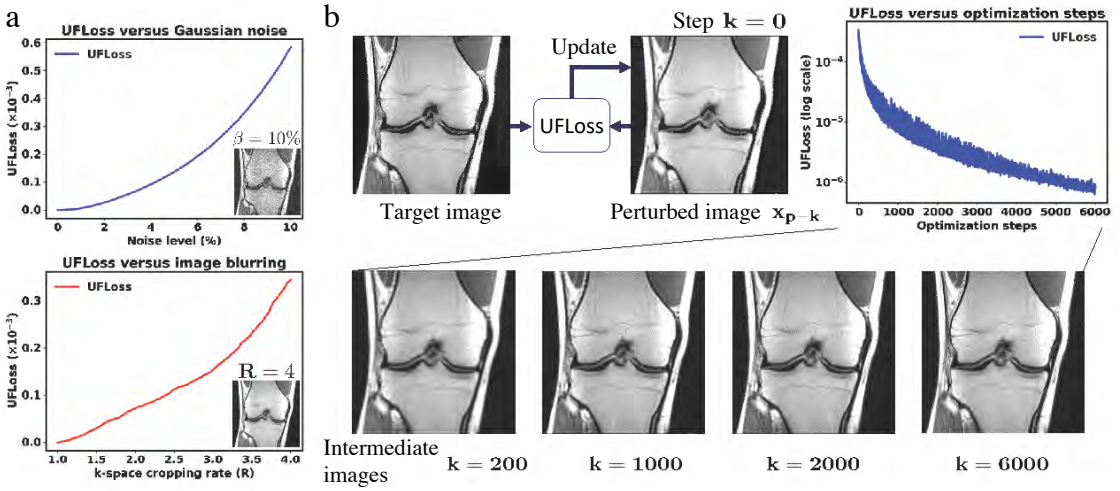


Figure 4.4: **UFLoss can be used as a valid loss function.** a) Evaluation of UFLoss with different levels of perturbations. **Upper**: additional Gaussian noise, **Lower**: image blurring through k-space cropping. UFLoss evolution curves indicate that UFLoss increases in a convex way with respect to more Gaussian noise and increases in a near-convex way with respect to more blurring. b) Evaluation of UFLoss in guiding a blurred image $\mathbf{x_{p-0}}$ to the target high resolution image. Gradient descent is performed on $\mathbf{x_{p-k}}$ to reduce the UFLoss with respect to the target image in an iterative way. Intermediate images show that UFLoss is able to gradually guide the blurred image to the target without falling into any local minimum.

of the borders. As seen in the figure, patches with similar perceptual structures (*e.g.*, edges, bone structures) are mapped closer to each other in the feature space.

Figure 4.5b) (PDFS) and Figure 4.6 (PD) show the feature correlation maps (UFLoss and SSIM) between different patches. Two source patches, indicated with green and blue edges, were chosen from each source image in the left column. The heatmaps under to each image, with corresponding green and blue edges, show the corresponding maps for each source patch from the source image. For the UFLoss results, we only show the positive inner products for visualization purposes, while in principle, the inner products range from -1 to 1. As shown in the UFLoss feature correlation maps, patches containing meniscus from both the same contrast and different contrast show high correlations with the input patch of the meniscus (blue border) while, on the other hand, patches from other anatomy show low correlation with it. These UFLoss feature correlation maps indicate that our unsupervised feature mapping is able to capture the perceptual structure similarities across different subjects and across different contrasts. In contrast, SSIM feature correlation maps do not successfully capture perceptual similarities across anatomies and contrasts (*e.g.*, meniscus).

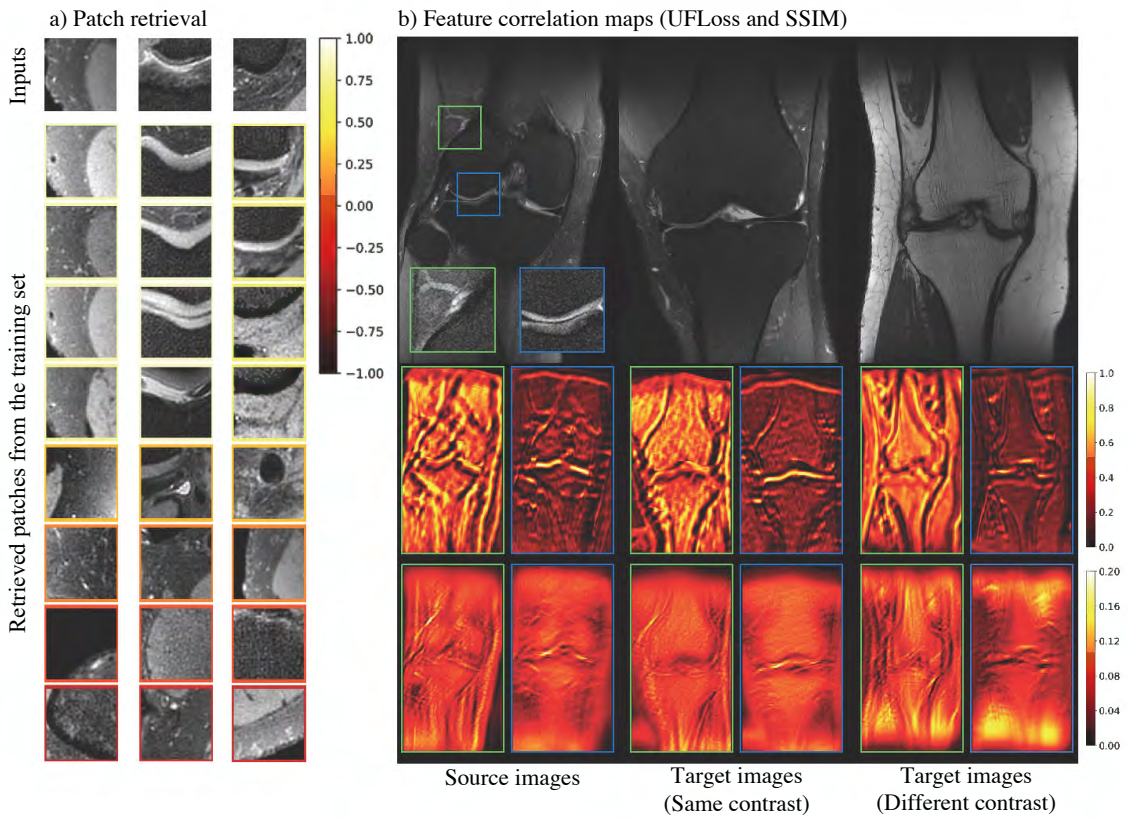


Figure 4.5: **UFLoss is able to capture perceptual similarities across anatomies and contrasts.** a) Feature clustering results using UFLoss feature mapping where, given an input patch, neighbor patches from the training set can be queried based on their feature space distance. The top four patches are the closest neighbors with the input patch and have the highest inner products. At the same time, we also show four counterexamples with relatively low inner products with the input patch. The feature space inner products between the input patch and the retrieved patches are shown as different colors of the borders. The color bar on the right indicates that a brighter border corresponds to a higher correlation while a darker border corresponds to a lower correlation. b) Feature correlations between different patches. The heat maps under a certain image show the feature correlations (feature space inner products for UFLoss) between all the patches from the image and the reference patches from the source image (first column). The heat maps with green/blue borders correspond to different source patches whose borders have the same colors. The correlation results for PDFS contrast using UFLoss and SSIM features are shown in the top and bottom rows, respectively.

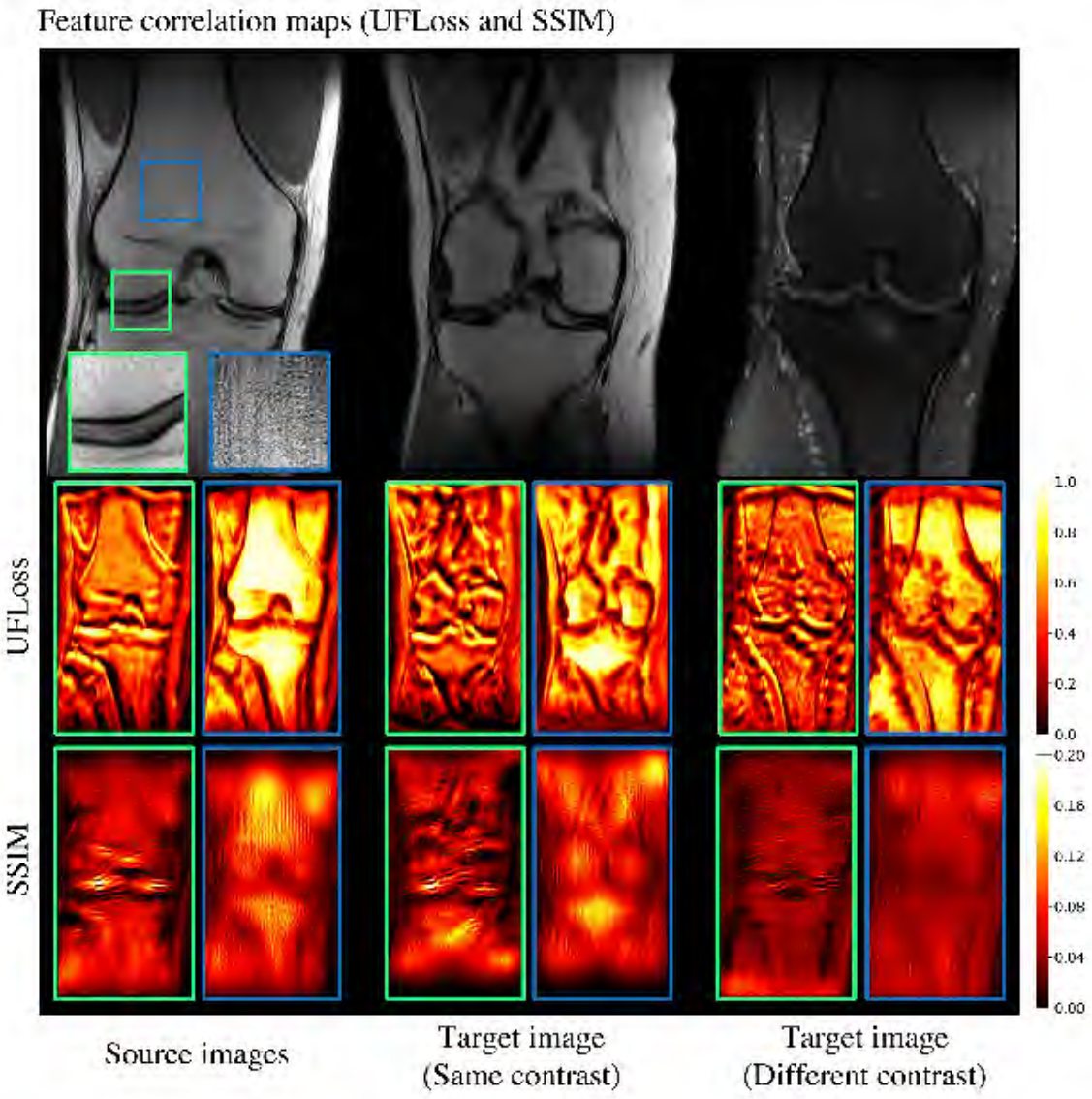


Figure 4.6: **UFLoss is able to capture perceptual similarities across anatomies and contrasts.** The heat maps under a certain image show the feature correlations between all the patches from the image and the source patches from the source image (first column). The heat maps with green/blue borders correspond to different source patches whose borders have the same colors. The correlation results for PD contrasts using UFLoss and SSIM features are shown in the top and bottom rows, respectively.

More specifically, as shown in supporting Figure 4.7, patch with the highest UFLoss

feature correlation (top) shows very similar anatomical textures of the meniscus compared to the source patch. At the same time, because SSIM focuses more on the local signal statistics instead of high-level perceptual similarity, the patch with the highest SSIM (bottom) has totally different textures from a different anatomical region.

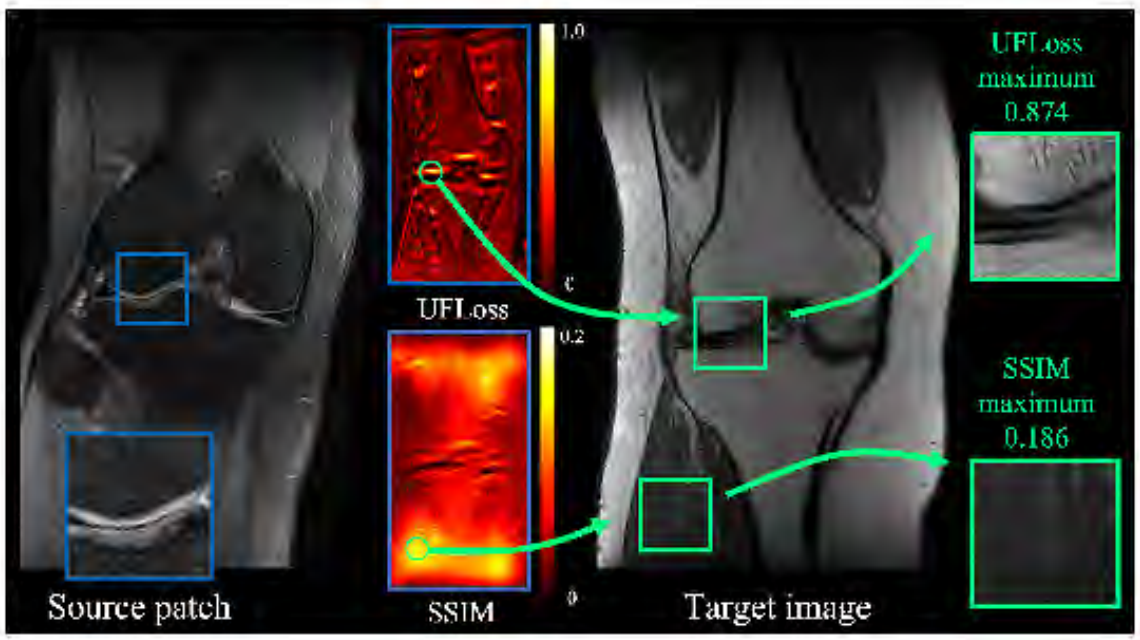
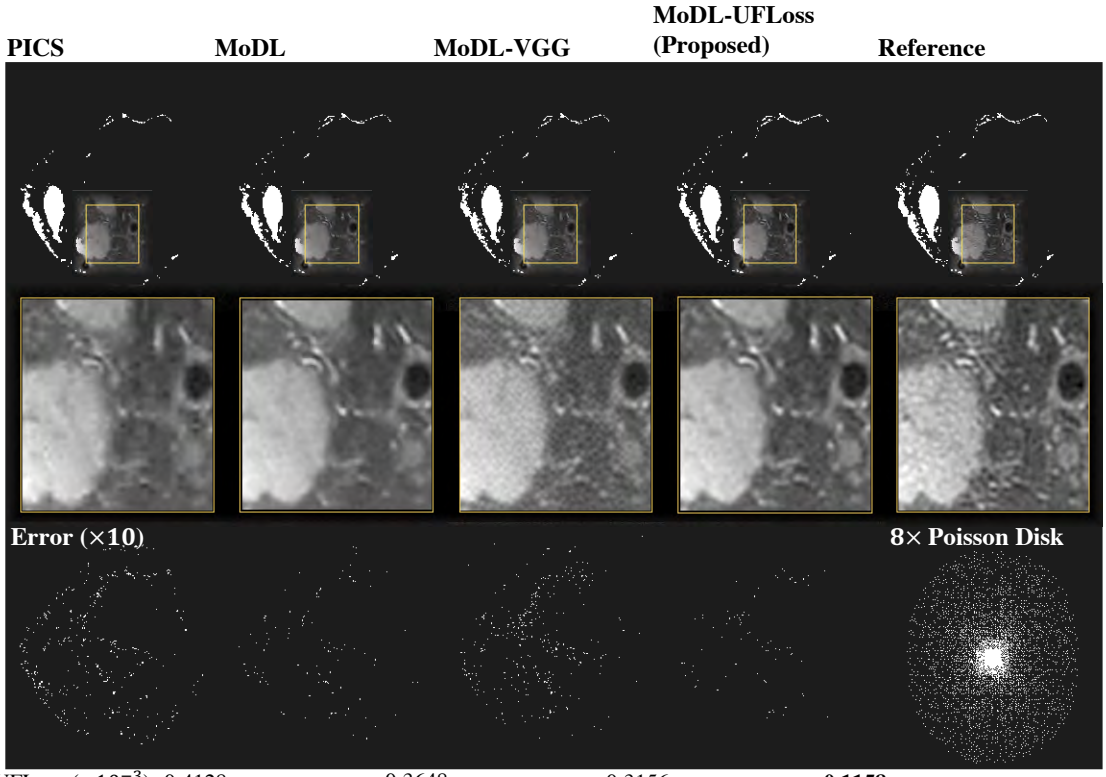


Figure 4.7: **UFLoss retrieves patches with closer structural similarity compared to SSIM across different contrasts.** The heat maps alongside the PD image show the feature correlation values between all the patches from the PD image and the source patch from the PDFS image (first column). The correlation results using UFLoss and SSIM features are shown on the right. Patches with the highest UFLoss and SSIM feature correlations in the PD image are visualized as zoomed-in patches with light blue borders. Feature correlation value are shown under each patch.

Unrolled reconstructions with UFLoss

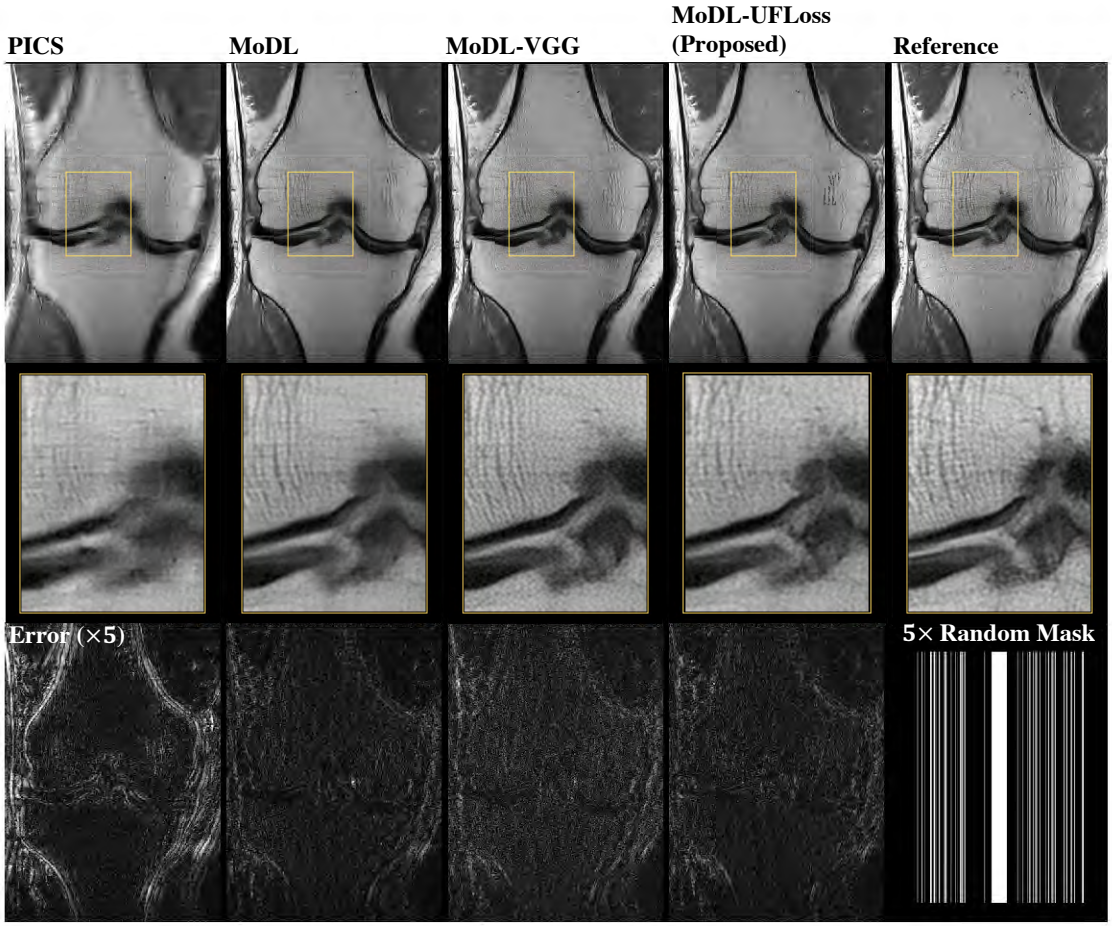
Figure 4.8 shows reconstruction comparisons between different methods (PICS, MoDL, MoDL with VGG, MoDL with UFLoss) for a representative 3D knee scan with under-sampling rate of $R = 8$. Quantitative metrics (NRMSE, SSIM) are shown under the images. As indicated in the zoomed images and error maps, MoDL with UFLoss shows finer structural details, sharper edges, and higher perceptual agreement with the fully-sampled reference images compared to the other reconstruction methods. Without our UFLoss, pure ℓ_2 loss at



	PICS	MoDL	MoDL-VGG	MoDL-UFLoss (Proposed)
UFLoss ($\times 10^{-3}$)	0.4128	0.3648	0.3156	**0.1158**
NRMSE	0.1247	**0.1120**	0.1433	0.1199
SSIM	0.8797	0.8865	0.8441	**0.9003**

Figure 4.8: **Representative 3D knee reconstruction results from different methods.** A fully-sampled scan is retrospectively under-sampled with a Poisson under-sampling mask by a factor of 8. From left to right are reconstructions by: PICS, MoDL with ℓ_2 loss, MoDL with ℓ_2+perceptual VGG loss, and MoDL with ℓ_2+our proposed UFLoss. NRMSE, SSIM, and UFLoss for each method are computed with respect to the fully sampled reference and shown under the image for reference. As shown in the zoomed images and error maps, our proposed MoDL with UFLoss showed sharper edges and more detailed structures with high perceptual similarity compared to the reference image.

this under-sampling rate leads to blurring and perceptual quality degradation. MoDL with the VGG perceptual loss [49] shows higher perceptual quality compared with MoDL, but generates unintended checkerboard structured artifacts, which is consistent with findings in [108, 80]. In terms of the training time and GPU memory cost for 3D reconstruction experiments, under the same setup, MoDL with UFLoss takes 92 minutes for a single epoch using 8.1 GB GPU memory, while MoDL with ℓ_2 loss takes 58 minutes using 5.5 GB and MoDL



	PICS	MoDL	MoDL-VGG	MoDL-UFLoss (Proposed)
UFLoss ($\times 10^{-3}$)	1.0242	0.2238	0.1683	**0.0501**
NRMSE	0.1621	**0.1142**	0.1246	0.1192
SSIM	0.8272	0.8998	0.8940	**0.9106**

Figure 4.9: **Representative examples of 2D PD knee reconstruction results using different methods.** A fully-sampled slice is retrospectively randomly under-sampled by a factor of 5. From left to right are reconstructions by PICS, MoDL with ℓ_2 loss, MoDL with perceptual VGG loss, and MoDL with our proposed UFLoss. NRMSE, SSIM, and UFLoss for each method are shown below the figure for references. As shown in the zoom-in views and error maps, our proposed MoDL with UFLoss can provide more realistic and natural-looking textures, while MoDL with ℓ_2 loss alone tends to blur out some high-frequency textures.

with perceptual VGG loss takes 61 minutes using 5.7 GB. In inference time, it takes around 25 ms and 0.9 GB for all methods.

Figure 4.9 shows the comparison of different reconstruction methods for a representative

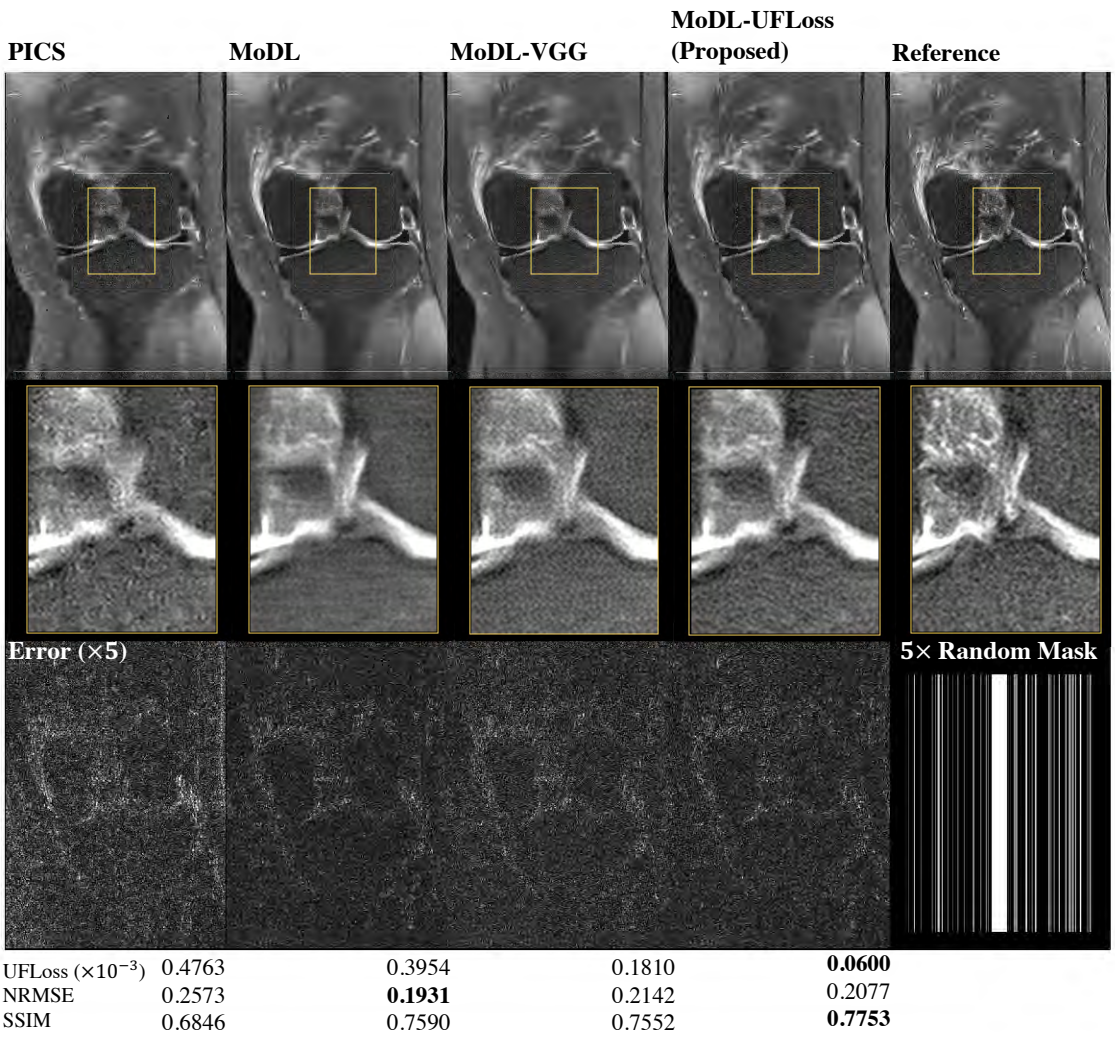


Figure 4.10: **Representative examples of 2D PDFS knee reconstruction results using different methods at under-sampling rate R=5.** nRMSE, SSIM, and UFLoss for each method are shown in the figure. Quantitative metrics indicate that MoDL with UFLoss has the highest SSIM and the lowest UFLoss, as well as the highest perceptual quality of the reconstructed image. Meanwhile, as shown in the zoom-in images and error maps, our proposed MoDL with UFLoss reconstruction looks more natural with a more faithful contrast than other methods.

2D PD slice from the fastMRI dataset [141]. The retrospective 2D under-sampling rate is 5, where around 20% of the k-space data is sampled. At this acceleration rate, PICS failed to effectively recover the fine bone structures, and MoDL with ℓ_2 loss alone also suffers blur-

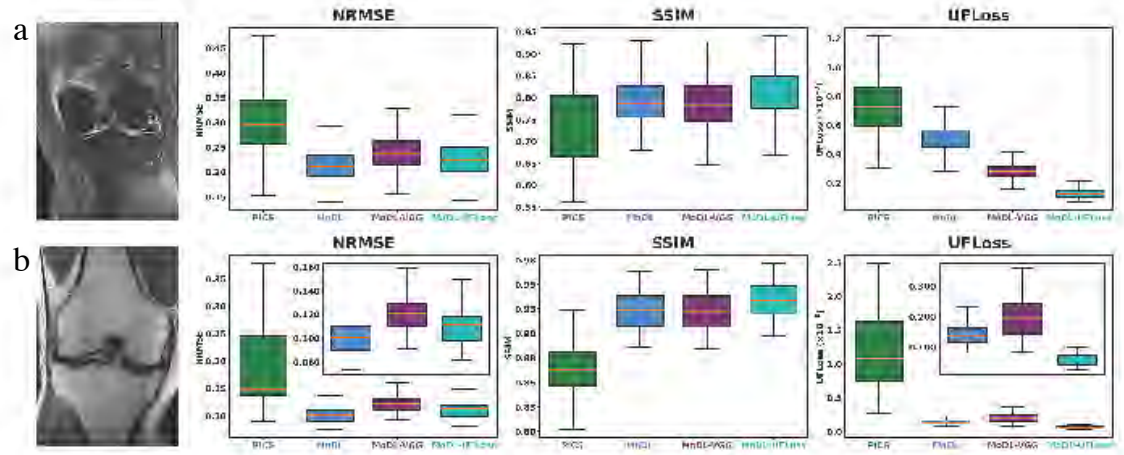


Figure 4.11: **MoDL with UFLoss shows competitive results in the metric comparisons for both a) PD and b) PDFS experiments.** Two representative fully-sampled scans (10 PD and 10 PDFS) with 15 slices each are randomly under-sampled by a factor of 5 and reconstructed using PICS, MoDL, MoDL with perceptual VGG loss, and MoDL with UFLoss. NRMSE, SSIM, and UFLoss are calculated with respect to fully sampled reference images and shown in the plot. We use zoomed-in plots to show more clear comparisons for some sub-plots. For both contrasts, MoDL with UFLoss outperforms both PICS and MoDL with ℓ_2 loss in terms of SSIM and UFLoss and can achieve comparable performance in terms of NRMSE.

ring artifacts. In contrast, MoDL with UFLoss demonstrates more realistic reconstruction performance with more detailed texture everywhere, including the bone.

Figure 4.10 shows the reconstruction comparisons for a representative 2D PDFS slice from the fastMRI dataset [141]. Quantitative comparisons are shown at the bottom of the figure. Due to the suppression of the fat signal, the SNR of the data is relatively low, where high-frequency features can be mixed up with the noise. The zoomed-in views and the corresponding error maps indicate that PICS results in a high level of artifacts. Meanwhile, MoDL with ℓ_2 loss alone misses fine detailed structures. Similar to the analysis above, MoDL with the VGG feature loss is capable of recovering subtle structures but generates unintended structured artifacts. In contrast, MoDL with UFLoss can effectively recover the detailed texture and have the most realistic reconstructions. In terms of the training time and GPU memory cost for 2D fastMRI experiments, under the same setup, MoDL with UFLoss takes 143 minutes for a single epoch using 11.9 GB GPU memory, while MoDL with ℓ_2 loss takes 104 minutes using 7.3 GB and MoDL with perceptual VGG loss takes 108 minutes using 7.5 GB. In inference time, it takes around 40 ms and 1.4 GB for all methods.

So far, for all of our experiments, we used a fixed UFLoss weighting factor ($\mu = 1.5$) for Equation 4.12. Supporting Figure 4.12 shows two representative reconstruction results with

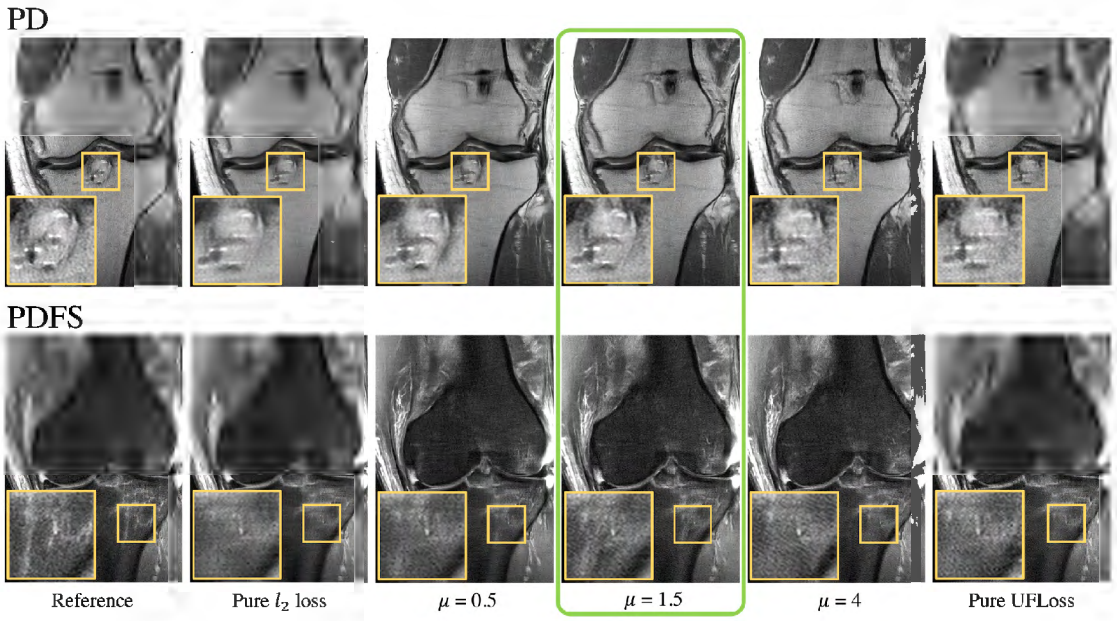


Figure 4.12: **Representative examples of 2D PD and 2D PDFS knee reconstruction with different UFLoss weighting factors during the training.** Fully-sampled slices are retrospectively randomly under-sampled by a factor of 5, and reconstructed using MoDL with different weights of UFLoss. Pure ℓ_2 loss, combined ℓ_2 and UFLoss with μ=0.5,1.5,4, and pure UFLoss are included for evaluations. Zoomed-in details are shown along with each image.

different UFLoss weighting factors during the training. We can clearly see that neither pure ℓ_2 loss nor pure UFLoss achieves the best image quality. By combining these two terms, our model is able to take advantage of both the per-pixel intensity information and patch-level perceptual similarities.

Figure 4.11 shows the quantitative metric (NRMSE, SSIM, UFLoss) comparisons for the 2D unrolled reconstruction experiments. For both a) PD and b) PDFS experiments, ten representative testing scans with 15 slices each are used to calculate the quantitative metrics. As indicated in the figure, for both contrasts, MoDL with UFLoss outperforms both PICS and MoDL with ℓ_2 loss in terms of SSIM and UFLoss and can achieve comparable performance in terms of NRMSE.

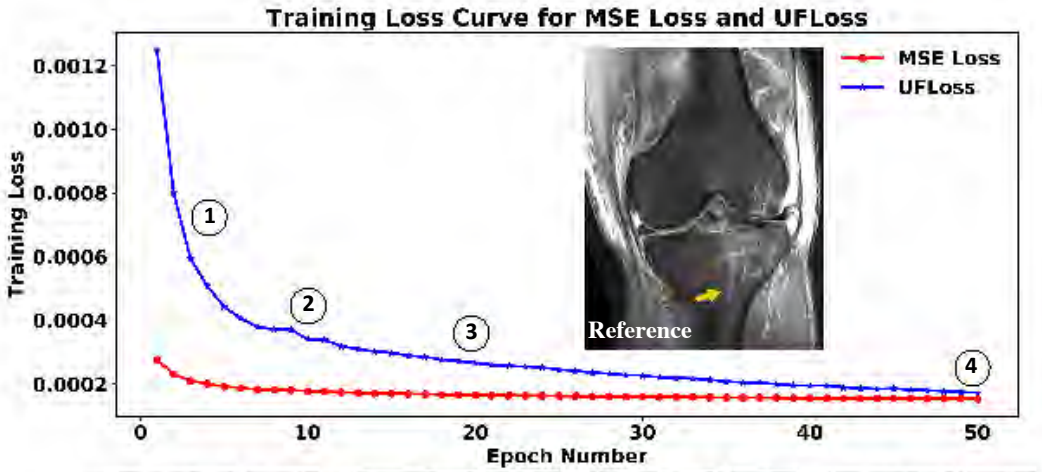
4.8 Discussion

In this work, we presented a novel patch-based perceptual loss function, which we call Unsupervised Feature Loss or UFLoss. UFLoss corresponds to the ℓ_2 distance in a low dimensional feature space. Feature vectors are mapped from image patches through a pre-trained mapping network. The mapping network aims to maximally separate all the patches in the feature space, where similar patches become closer to each other, capturing high-level perceptual similarities. As indicated in Figure 4.5, unlike ℓ_2 distance, which focuses on the pixel-wise values, our proposed UFLoss agrees better with human visual judgment, where similar-looking patches have lower UFLoss in the feature space. By incorporating UFLoss into the training of DL-based reconstructions, we are able to recover finer textures, smaller features, and sharper edges with higher overall image quality compared to conventional per-pixel losses. By leveraging a memory bank to store all the features, the training of our mapping network becomes feasible for a large dataset: The UFLoss network training required less than 500 MB GPU memory and was easily trained within two hours. In terms of computation costs, our UFLoss imposes about 50% increase in training time and memory requirements during training. However, in inference time, the UFLoss has no penalty at all on the reconstruction time as well as the memory requirements.

As we mentioned before, another important class of feature losses for DL-based reconstruction is adversarial loss or GAN loss [33]. Adversarial losses have shown great success in capturing perceptual properties of ground-truth images and could be used to improve the reconstruction quality. However, due to the min-max loss function, the convergence of GANs is generally underdetermined, and it is difficult to determine the stop criterion for GANs' training [54]. In contrast, the convergence and stop criterion of training with UFLoss is clear and straightforward, simply when the loss function (pixel loss + UFLoss) converges. Another important distinction is that GAN loss is an instance-to-set loss, which means that so long as the reconstruction is similar to any of those ground-truth training images, the loss would be small, which is undesirable for reconstruction [20, 72, 28, 96]. In comparison, UFLoss is an instance-wise discriminative loss, comparing the reconstruction to the specific ground truth image in the feature space, which provides clear guidance and is more constraining during the training.

In this study, UFLoss can be viewed as a separate module and be easily incorporated into other learning frameworks. The performance of UFLoss was demonstrated for accelerating 2D and 3D knee imaging by comparing the reconstruction results with respect to fully sampled references. The in-vivo results show that the addition of UFLoss during the network's training allows realistic texture recovery and improves overall image quality compared to a reconstruction network trained without UFLoss. Our UFLoss network trained on specific anatomy and contrast may yield suboptimal results when applied to a different contrast/anatomy. Therefore, in the ideal case, one may want to use different networks for different types of images. Fortunately, the UFLoss can be trained on the same ground truth images that are used to train reconstruction networks, therefore it does not require additional data sets to do so. Finally, the training of a UFLoss network takes less than two hours to

train, so the overhead is negligible.



UFLoss ($\times 10^{-3}$)	0.1788	0.1269	0.1005	0.0885
NRMSE	0.2160	0.2069	0.2012	0.1989
Epochs	4	10	20	50

Figure 4.13: **Training loss curves for the l_2 MSE loss and our proposed UFLoss.** A 2D fully-sampled slice is randomly under-sampled by a factor of 5 and reconstructed at different training epochs. NRMSE and UFLoss are shown as quantitative metrics under each reconstructed image. Yellow arrows point at the same representative textures at different reconstructions. UFLoss continues improving the reconstructed image quality after l_2 MSE loss converged.

Another interesting finding of the UFLoss comes from how the training losses evolve, as shown in Figure 4.13. The total loss consists of two different components, the per-pixel ℓ_2 MSELoss and our proposed UFLoss, which are shown in the top sub-figure as red and blue curves, respectively. The bottom sub-figure shows the testing reconstruction results at different epochs. As indicated from the curve, the MSELoss remains almost constant after ten epochs, while our proposed UFLoss still decreases continuously. Inspecting the reconstructed images at different training epochs, we can see that the image quality continues

to improve with the further reduction of the UFLoss. At the same time, the quantitative metrics indicate that those reconstructed images have very similar NRMSE compared with the fully-sampled reference but a much more significant difference in their UFLoss values. A low UFLoss value corresponds to better image quality. These results indicate that using the ℓ_2 MSE loss alone is not optimal. Therefore, the UFLoss can be potentially used as a better perceptual comparison criterion and help further improve the reconstruction quality.

Limitations of this study include: 1)The training of DL-based reconstructions with UFLoss is time-consuming (around 1.5×) and memory-inefficient (around 1.5×) due to the extraction and feed-forwarding of a large number of patches within a single step. This can be potentially improved by using fully-convolutional image-scale networks, GPU parallel computing, and efficient memory-time trade-off [127]. 2)In this work, we haven't thoroughly investigated the sensitivity of different hyperparameters (*e.g.*, patch size, temperature parameter, UFLoss network depth) to the training and final reconstructions. Supporting Figure S3 demonstrates how UFLoss weighting parameter contributes to the reconstruction results. A more thorough parameter search and analysis will be explored in the future. 3) Even though empirical evidence for both 2D and 3D knee results has demonstrated that UFLoss can effectively encourage finer texture and sharper edges, we have not investigated the theoretical performance guarantee of UFLoss on enhancing the texture sharpness and image quality in this paper; however, our observation is supported by other perceptual loss methods in the literature[145, 49].

4.9 Conclusion

In summary, a novel patch-based feature loss, Unsupervised Feature Loss or UFLoss, is proposed, and it can be easily incorporated into the training of any existing DL-based reconstruction frameworks without any modification to the model architecture. UFLoss is based on an unsupervised pre-trained feature mapping network without any external supervision. With the addition of our proposed UFLoss, we are able to reconstruct high fidelity images with sharper edges, more faithful contrasts, and better image quality overall.

Chapter 5

Memory-efficient learning for high-dimensional MRI reconstruction

5.1 Introduction

As we introduced in Chapter 2, DL-based unrolled reconstructions have emerged as a widely-used type of DL-based reconstruction method. These techniques, as demonstrated in studies such as [26, 99, 2, 37, 110, 57], have demonstrated remarkable success in under-sampled MRI reconstruction.

These methods are often formulated by unrolling the iterations of an image reconstruction optimization[37, 2, 110]. It has been shown that increasing the number of unrolls improves upon finer spatial and temporal textures in the reconstruction[37, 2, 95].

Similar to compressed sensing and other low-dimensional representations, DL-based unrolled reconstructions can take advantage of additional structure in very high-dimensional data (*e.g.*, 3D, 2D+time, 3D+time) to further improve image quality. However, these large-scale DL-based unrolled reconstructions are currently limited by GPU memory required for gradient-based optimization using backpropagation. Therefore, most DL-based unrolled reconstructions focus on 2D applications or are limited to a small number of unrolls.

In this Chapter, we employ our recently proposed memory-efficient learning (MEL) framework[51, 143] to reduce the memory needed for backpropagation, which enables the training of DL-based unrolled reconstructions for 1) larger-scale 3D MRI; and 2) 2D+time cardiac cine MRI with a large number of unrolls (Figure 5.1). We evaluate the spatiotemporal complexity of our proposed method on the Model-based Deep Learning (MoDL) architecture [2] and train these high-dimensional DL-based unrolled reconstructions on a single 12GB GPU. Our training uses far less memory while only marginally increasing the computation time. To demonstrate the advantages of high-dimensional reconstructions to image quality, we performed experiments on both retrospectively and prospectively under-sampled data for 3D MRI and cardiac cine MRI. Our in-vivo experiments indicate that by exploiting high-dimensional data redundancy, we can achieve better quantitative metrics and improved

image quality with sharper edges for both 3D MRI and cardiac cine MRI.

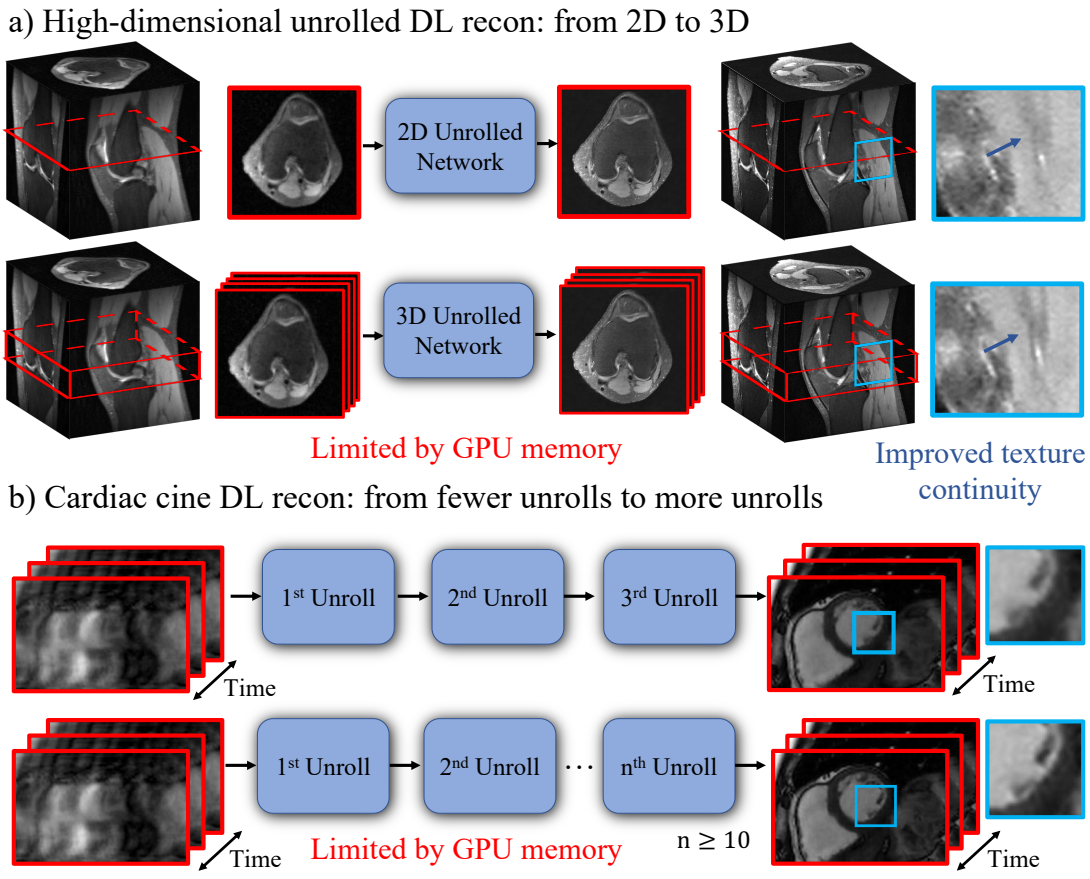


Figure 5.1: **GPU memory limitations for high-dimensional DL-based unrolled reconstructions.** a) Compared to a 2D unrolled network, the 3D unrolled network uses a 3D slab during training to leverage more 3D structural redundancy, but is limited by GPU memory. b) 2D+time Cardiac cine DL-based unrolled reconstructions are often performed with a small number of unrolls due to memory limitations.

5.2 Memory-efficient learning (MEL) framework

As shown in Figure 5.2 a), DL-based unrolled reconstructions are often formulated by unrolling the iterations of an image reconstruction optimization [37, 2]. Each unroll consists of two submodules: CNN-based regularization layer and data consistency (DC) layer. In conventional backpropagation, the gradient must be computed for the entire computational graph, and intermediate variables from all N unrolls need to be stored at a significant mem-

ory cost. By leveraging MEL, we can process the full graph as a series of smaller sequential graphs. As shown in Figure 5.2 b), first, we forward propagate the network to get the output $\mathbf{x}^{(N)}$ without computing the gradients. Then, we rely on the invertibility of each layer (required) to recompute each smaller auto-differentiation (AD) graph from the network's output in reverse order. MEL only requires a single layer to be stored in memory at a time, which reduces the required memory by a factor of N. Notably, the required additional computation to invert each layer only marginally increases the backpropagation runtime.

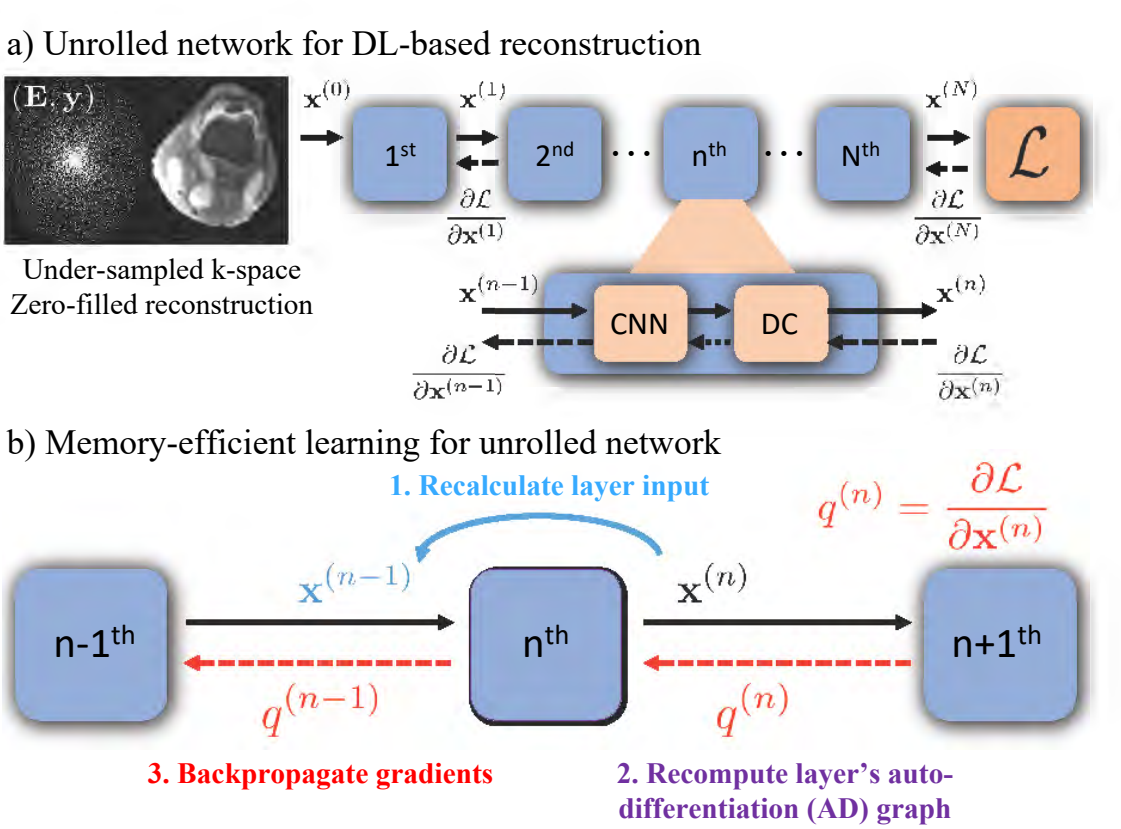


Figure 5.2: **Gradient backpropagation of conventional training and MEL.** a) In conventional DL-based unrolled reconstruction training, gradients of all layers are evaluated as a single computational graph, requiring significant GPU memory. b) In MEL, we sequentially evaluate each layer by: i) Recalculate the layer's input $\mathbf{x}^{(n-1)}$, from the known output $\mathbf{x}^{(n)}$. ii) Reform the AD graph for that layer. iii) Backpropagate gradients $q^{(n-1)}$ through the layer's AD graph.

Memory-efficient learning for MoDL

Here, we use a widely used DL-based unrolled reconstruction framework: MoDL [2]. We formulate the reconstruction of $\hat{\mathbf{x}}$ as an optimization problem and solve it as below:

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \|\mathbf{E}\mathbf{x} - \mathbf{y}\|_2^2 + \mu\|\mathbf{x} - R_w(\mathbf{x})\|_2^2, \tag{5.1}$$

where $\mathbf{E}$ is the system encoding matrix, $\mathbf{y}$ denotes the k-space measurements and R_w is a learned CNN-based denoiser. For multi-channel MRI reconstruction, $\mathbf{E}$ can be formulated as $\mathbf{E} = \mathbf{PFS}$, where $\mathbf{S}$ represent the multi-channel sensitivity maps, $\mathbf{F}$ denotes Fourier Transform and $\mathbf{P}$ is the undersampling mask used for selecting the acquired data. MoDL solves the minimization problem by an alternating procedure:

$$\mathbf{z}_n = R_w(\mathbf{x}_n) \tag{5.2}$$

$$\begin{aligned} \mathbf{x}_{n+1} &= \arg\min_{\mathbf{x}} \|\mathbf{E}\mathbf{x} - \mathbf{y}\|_2^2 + \mu\|\mathbf{x} - \mathbf{z}_n\|_2^2, \\ &= (\mathbf{E}^{\mathbf{H}}\mathbf{E} + \mu\mathbf{I})^{-1}(\mathbf{E}^{\mathbf{H}}\mathbf{y} + \mu\mathbf{z}_n) \end{aligned} \tag{5.3}$$

which represents the CNN-based regularization layer and DC layer respectively. In this formulation, the DC layer is solved using Conjugate Gradient (CG) [101], which is unrolled for a finite number of iterations. For all the experiments, we used an invertible residual convolutional neural network (RCNN) introduced in [32, 84, 39], whose architecture is composed of a 5-layer CNN with 64 channels per layer.

The residual CNN is inverted using the fixed-point algorithm as described in [51], while the DC layer is inverted through:

$$\mathbf{z}_n = \frac{1}{\mu}((\mathbf{E}^{\mathbf{H}}\mathbf{E} + \mu\mathbf{I})\mathbf{x}_{n+1} - \mathbf{E}^{\mathbf{H}}\mathbf{y}). \tag{5.4}$$

Training and evaluation of memory-efficient learning

With IRB approval and informed consent/assent, we trained and evaluated MEL on the both retrospective and prospective 3D knee and 2D+time cardiac cine MRI. We conducted 3D MoDL experiments with and without MEL on 20 fully-sampled 3D knee datasets (320 slices each) from mridata.org [98]. 16 cases were used for training, 2 cases were used for validation, and the other 2 for testing. Around 5000 3D slabs with size 21×256×320 were used for training the reconstruction networks. All data were acquired on a 3T GE Discovery MR 750 with an 8-channel HD knee coil. An 8x Poisson Disk sampling pattern was used to retrospectively undersample the fully sampled k-space. Scan parameters included a matrix size of 320×256×320 and TE/TR of 25ms/1550ms. In order to further demonstrate the feasibility of our 3D reconstruction with MEL on realistic prospectively under-sampled scans, we reconstructed 8× prospectively under-sampled 3D FSE knee scans (available at

mridata.org) with the model trained on retrospectively under-sampled knee data. Scanning parameters includes: Volume size: 320×288×236, TR/TE = 1400/20.46ms, Flip Angle: 90°, FOV: 160 mm×160 mm× 141.6 mm.

For the cardiac cine MRI, fully-sampled bSSFP cardiac cine datasets were acquired from 15 volunteers at different cardiac views and slice locations on 1.5T and 3.0T GE scanners using a 32-channel cardiac coil. All data were coil compressed [146] to 8 virtual coils. Twelve of the datasets (around 190 slices) were used for training, 2 for validation, and one for testing. k-Space data were retrospectively under-sampled using a variable-density k-t sampling pattern to simulate 14-fold acceleration with 25% partial echo. We also conducted experiments on a prospectively under-sampled scan (R=12) which was acquired from a pediatric patient within a single breath-hold on a 1.5T scanner.

We compared the spatiotemporal complexity (GPU memory, training time) with and without MEL. In order to show the benefits of high-dimensional DL recons, we compared the reconstruction results of PICS, 2D and 3D MoDL with MEL for 3D MRI, and 2D+time MoDL with 4 unrolls and 10 unrolls for cardiac cine MRI. For both 2D MoDL and 3D MoDL with MEL, we used 5 unrolls, 10 CG steps and Residual CNN as the regularization layer. A baseline PICS reconstruction was performed using BART [115]. Sensitivity maps were computed using BART [115] and SigPy [81]. Common image quality metrics such as Peak Signal to Noise Ratio (pSNR), Structural Similarity (SSIM) [44] and Fréchet Inception Distance (FID) [40] were reported. FID is a widely used measure of perceptual similarity between two sets of images. All the experiments were implemented in Pytorch [86] and used Nvidia Titan XP (12GB) and Titan V CEO (32GB) GPUs. Networks were trained end-to-end using a per-pixel l_1 loss and optimized using Adam [52] with a learning rate of 1×10^{-4}.

5.3 Results: spatiotemporal complexity

We first evaluate the spatiotemporal complexity of MoDL with and without MEL (Figure 5.3). Without MEL, for a 12GB GPU memory limit, the maximum slab size decreases rapidly as the number of unrolls increases, which limits the performance of a 3D reconstruction. In contrast, using MEL, the maximum slab size is roughly constant. Figure 5.3 b) and c) show the comparisons from two different perspectives: 1)GPU memory usage; 2)Training time per epoch. Results indicate that for both 3D and 2D+time MoDL, MEL uses significantly less GPU memory than conventional backpropagation while marginally increasing training time. Notably, both MoDL with and without MEL have the same inference time.

5.4 Results: reconstruction comparisons with MEL

Figure 5.4 shows a comparison of different methods for 3D reconstruction. Instead of learning from only 2D axial view slices (Figure 5.1 a), 3D MoDL with MEL captures the image features

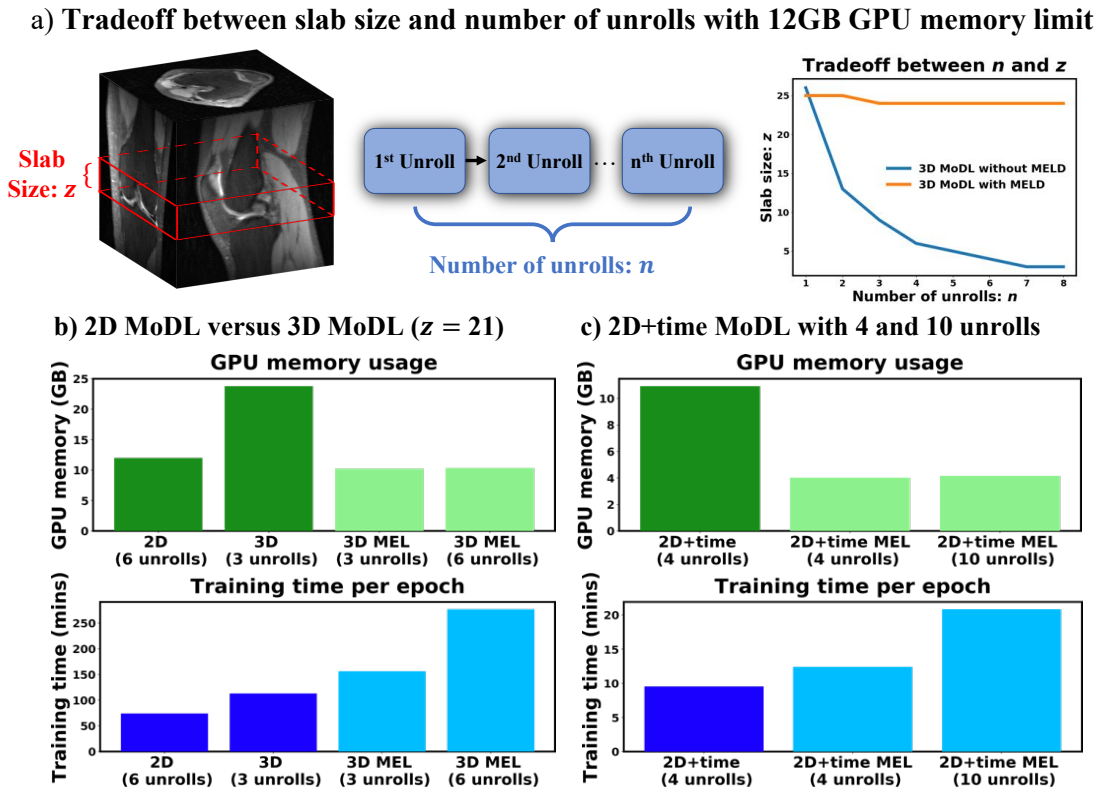


Figure 5.3: **Spatio-temporal complexity of MoDL with and without MEL.** a) Tradeoff between 3D slab size z and a number of unrolls n with a 12GB GPU memory limitation. b) and c) show the memory and time comparisons for MoDL with and without MEL.

from all three dimensions. Zoomed-in details indicate that 3D MoDL with MEL is able to provide more faithful contrast with more continuous and realistic textures as well as higher pSNR over other methods.

Figure 5.5 demonstrates that MEL enables the training of 2D+time MoDL with a large number of unrolls (10 unrolls), which outperforms MoDL with 4 unrolls with respect to image quality and y-t motion profile. With MEL, MoDL with 10 unrolls resolves the papillary muscles (yellow arrows) better than MoDL with 4 unrolls. Also, the y-t profile of MoDL with 10 unrolls depicts motion in a more natural way while MoDL with 4 unrolls suffers from blurring. Meanwhile, using 10 unrolls over 4 unrolls yields an improvement of 0.6dB in validation pSNR.

Table 5.1 shows the quantitative metric comparisons (pSNR, SSIM and FID) between different methods on both 3D MRI and cardiac cine MRI reconstructions. Here, we also included feed-forward U-Net [19] as a baseline. The results indicate that both 3D MoDL with MEL and 2D+time MoDL with MEL outperforms other methods with respect to pSNR,

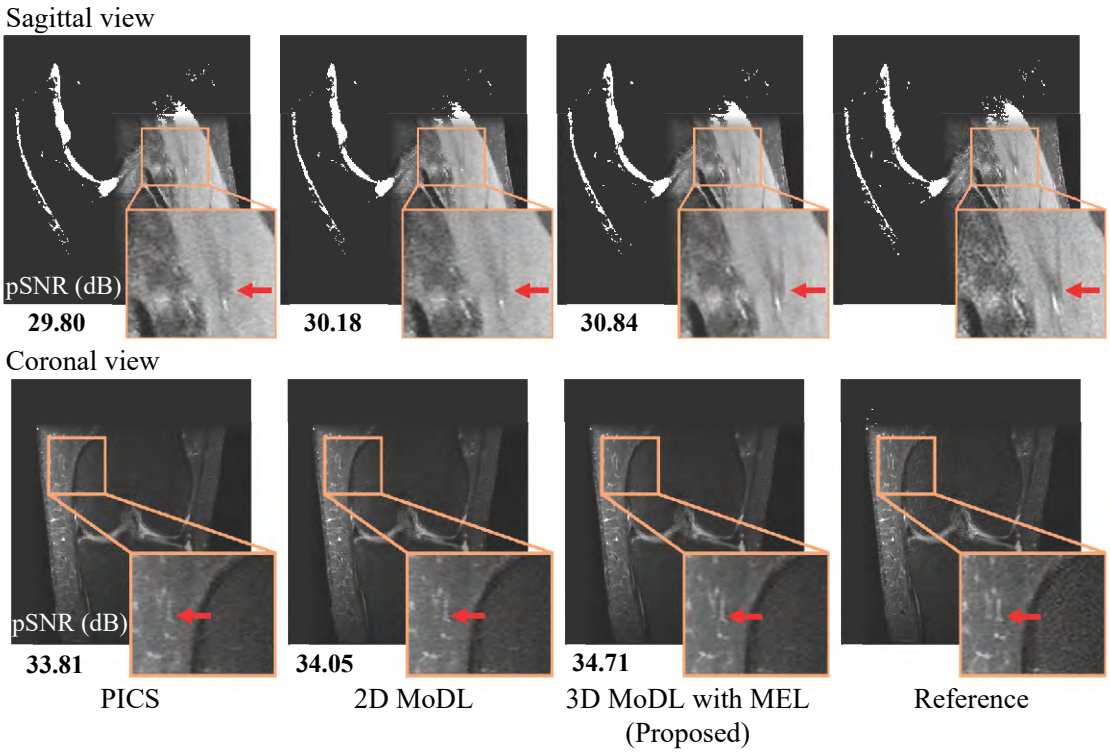


Figure 5.4: **A representative comparison of different methods on 3D knee reconstruction.** From the left to the right, we compare PICS, 2D MoDL, and 3D MoDL with MEL. The sagittal view and The coronal view are visualized, while pSNRs are shown under each reconstructed image. 3D MoDL with MEL is able to provide more faithful contrast with more continuous and realistic textures as well as higher pSNR over other methods.

SSIM, and FID.

Figure 5.6 a) show the reconstruction results on two representatives prospectively under-sampled 3D FSE knee scans. Note that in this scenario, there is no fully-sampled groud truth. Despite there exists some differences between the training and testing (*e.g.*, matrix size, scanning parameters), 3D MoDL with MEL is still able to resolve more detailed texture and sharper edges over traditional PICS and learning-based 2D MoDL. Figure 5.6 b) and Vedio results show the reconstruction on a representative prospective under-sampled cardiac cine scan. We can clearly see that enabled by MEL, 2D+time MoDL with 10 unrolls can better depicts the finer details as well as a more natural motion profile.

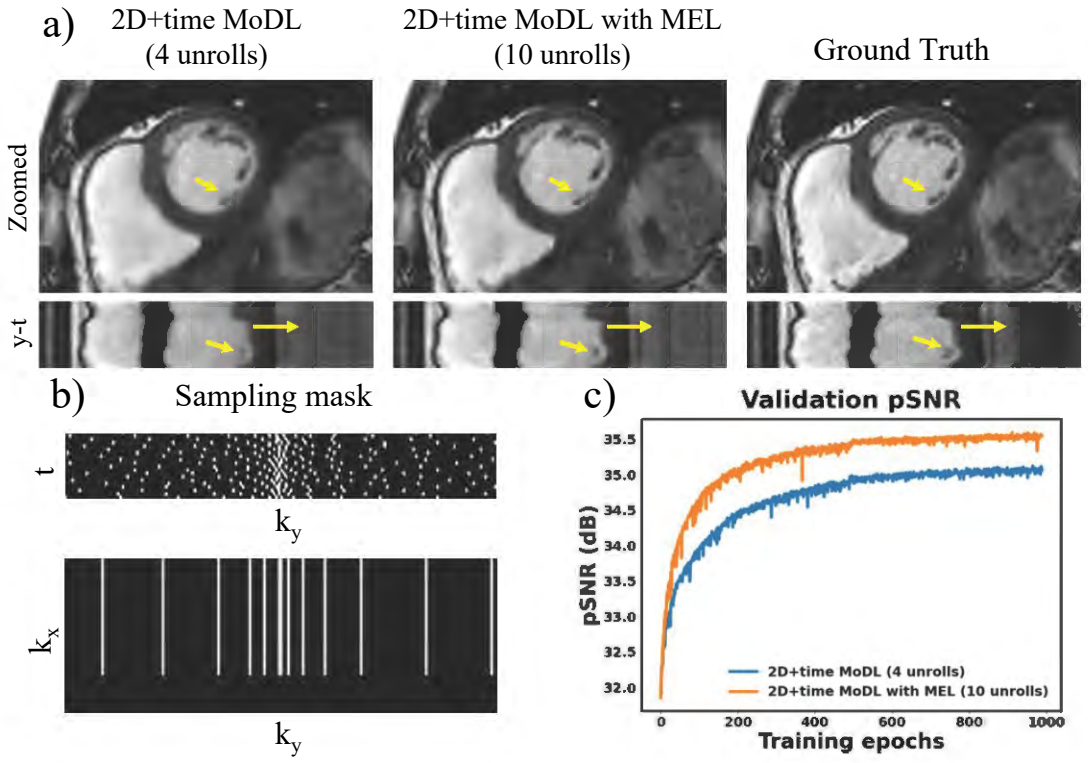


Figure 5.5: **Results on 2D+time cardiac cine reconstruction.** a) Short-axis view cardiac cine reconstruction of a healthy volunteer on a 1.5T scanner. k-Space data was retrospectively under-sampled to simulate 14-fold acceleration with 25% partial echo (shown in b) and reconstructed by: 2D+time MoDL with 4 unrolls, 2D+time MoDL with MEL and 10 unrolls. c) Validation pSNR of MoDL with 4 unrolls and MoDL with 10 unrolls.

metric	method	3D MRI	2D cardiac cine MRI
pSNR (dB)	PICS	31.01±1.97	24.69±2.74
	2D MoDL	31.44±2.07	-
	3D U-Net	29.55±1.86	-
	3D MoDL with MEL	**32.11±2.05**	-
	2D+time MoDL: 4 unrolls	-	26.87±2.98
	2D+time MoDL with MEL: 10 unrolls	-	**27.42±3.21**
SSIM	PICS	0.816±0.046	0.824±0.071
	2D MoDL	0.821±0.044	-
	3D U-Net	0.781±0.039	-
	3D MoDL with MEL	**0.830±0.038**	-
	2D+time MoDL: 4 unrolls	-	0.870±0.042
	2D+time MoDL with MEL: 10 unrolls	-	**0.888±0.042**
FID	PICS	46.71	39.40
	2D MoDL	43.58	-
	3D U-Net	60.10	-
	3D MoDL with MEL	**41.48**	-
	2D+time MoDL: 4 unrolls	-	36.93
	2D+time MoDL with MEL: 10 unrolls	-	**31.64**

Table 5.1: **Quantitative metrics comparisons.** Quantitative metrics (pSNR, SSIM and FID) of different methods on 3D MRI and cardiac cine MRI reconstructions (mean ± standard deviation of pSNR and SSIM).

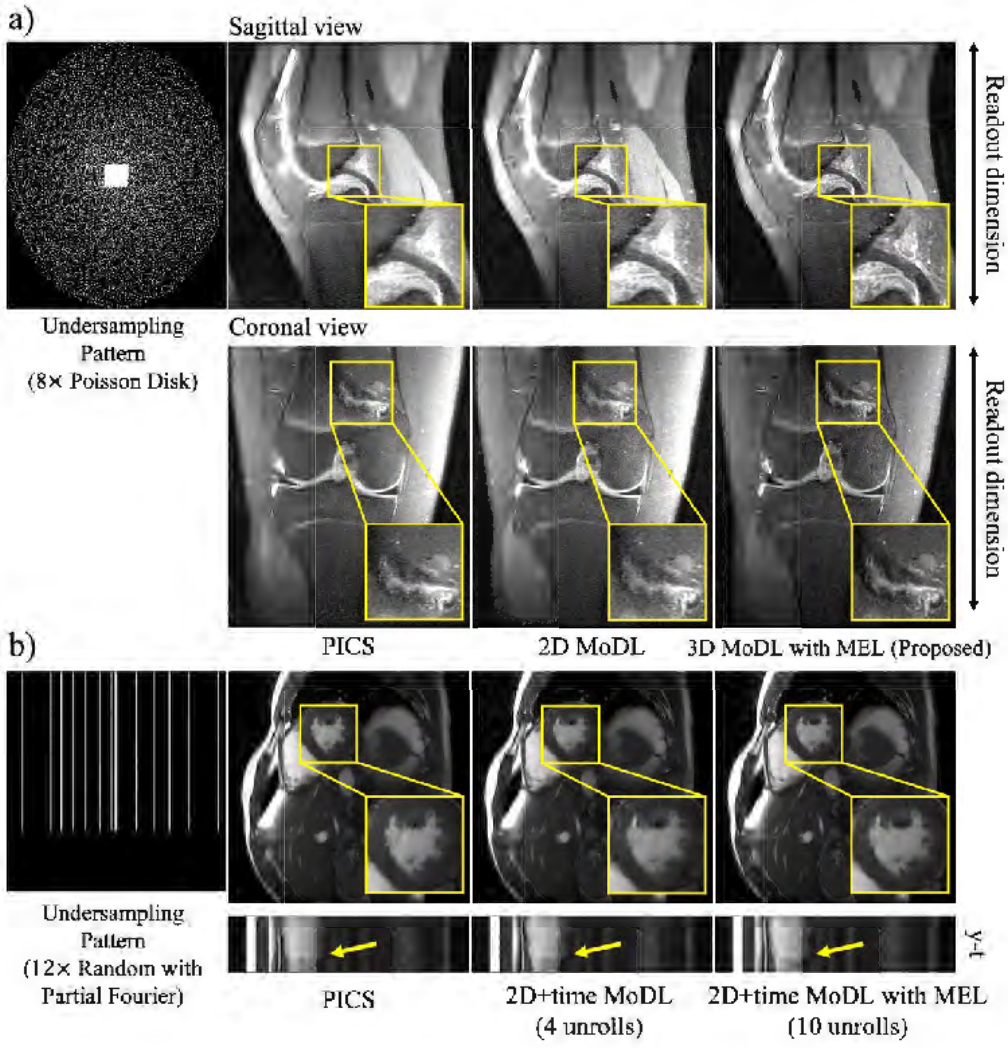


Figure 5.6: **Results for prospectively under-sampled reconstruction.** a) Representative reconstruction results on a prospectively under-sampled 3D FSE knee scan using different methods (PICS, 2D MoDL and 3D MoDL with MEL). b) Representative reconstruction results on a prospectively under-sampled cardiac cine dataset. y-t motion profiles are shown along with the reconstructed images.

5.5 Conclusions

In this work, we show that MEL enables learning for high-dimensional MR reconstructions on a single 12GB GPU, which is not possible with standard backpropagation methods. We demonstrate MEL on two representative large-scale MR reconstruction problems: 3D volumetric MRI, 2D cardiac cine MRI with a relatively large number of unrolls. By leveraging the high-dimensional image redundancy and a large number of unrolls, we were able to get improved quantitative metrics and reconstruct finer details, sharper edges, and more continuous textures with higher overall image quality for both 3D and 2D cardiac cine MRI. Furthermore, 3D MoDL reconstruction results from prospectively undersampled k-space show that the proposed method is robust to the scanning parameters and could be potentially deployed in clinical systems. Overall, MEL brings a practical tool for training large-scale high-dimensional MRI reconstructions with much less GPU memory and is able to achieve improved reconstructed image quality.

Chapter 6

Rigorous uncertainty estimation for MRI reconstruction

6.1 Introduction

In the preceding chapters, we discussed the impressive capabilities of DL-based reconstruction techniques, which have demonstrated remarkable potential in significantly reducing scan time while preserving high image quality [2, 37, 124, 127]. These methods have outperformed traditional optimization-based approaches, offering new possibilities in the field of medical imaging.

Despite their merits, DL-based reconstructions have notable limitations, as they carry a significant risk of hallucination, manifesting as fabricated structures within the reconstructed images [72]. Moreover, these methods may unintentionally remove genuine structures, resulting in potential inaccuracies that could hinder clinical adoption. These issues highlight the importance of carefully examining and addressing the limitations of these methods, ensuring their reliability and validity in medical applications.

In this regard, as one possible solution, precise uncertainty estimation for various regions of the image can substantially enhance diagnostic confidence, further bolstering the adoption of these methods in clinical settings. Over the past decade, numerous approaches have been proposed to provide uncertainty maps alongside the reconstruction results [28, 46, 88, 73].To name a few, [28] construct the DL models using Variational Autoencoder (VAE) [53] to develop a probabilistic reconstruction scheme. By utilizing Monte Carlo sampling to generate reconstructions, the VAE models inherently produce pixel variance maps that serve as uncertainty maps. [88] adapts deep learning models to output a value distribution for each pixel rather than a singular value, enabling the variance of the distribution to represent the uncertainty associated with that pixel.

While existing methods have shown promising results in estimating uncertainty maps, most of them are either based on running multiple reconstructions to sample their distribution or require modification of the network architecture. Meanwhile, none of them provides

confidence levels to ensure the quantitative accuracy of the uncertainty estimation.

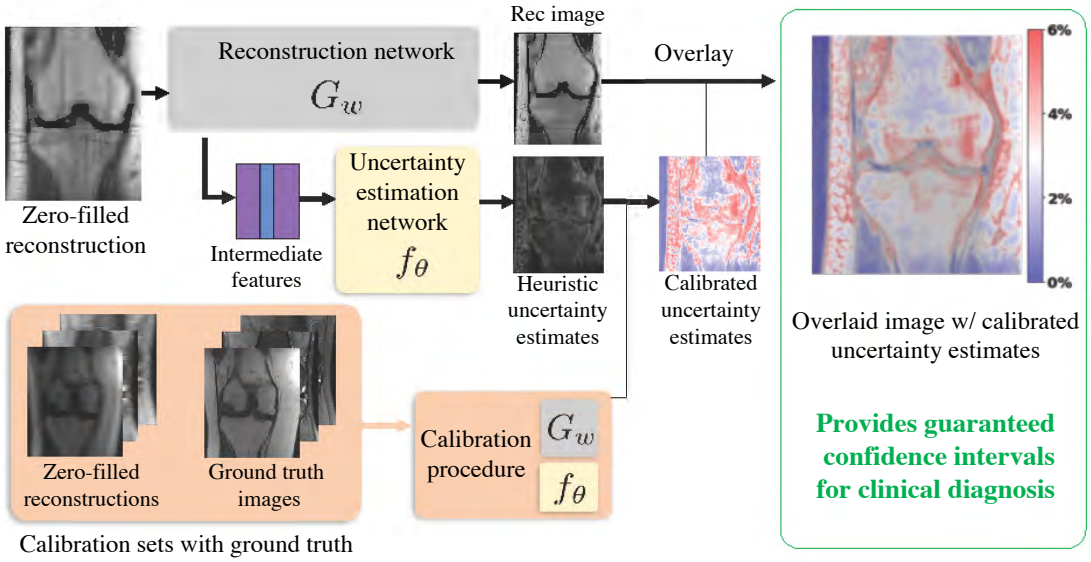


Figure 6.1: **Overview of the proposed model-specific rigorous uncertainty estimation framework for general DL-based reconstruction models.** After training f_θ, our networks output the heuristic uncertainty estimates alongside the reconstructed image in one forward pass. By developing a new form of Risk-Controlling Prediction Set to calibrate the uncertainty estimates, our calibrated uncertainty estimates provide guaranteed confidence intervals that contain at least (1-γ) (*e.g.*, 95%) of the ground truth pixel values.

To address the aforementioned challenges, in this chapter, we introduce a straightforward and rigorous uncertainty estimation framework that can be seamlessly integrated into the existing reconstruction network without necessitating any modification or retraining (Figure 6.1). This streamlined approach not only maintains the integrity of existing network architectures but also enhances efficiency by delivering both the image and its uncertainty estimation simultaneously.

Our technique provides a rigorous finite-sample statistical guarantee. Our key contribution is the development of a new form of Risk-Controlling Prediction Set (RCPS) [6, 8] tailored to MRI reconstruction that outputs image-valued confidence intervals containing at least (1-γ) (*e.g.*, 95%) of the ground truth pixel values.

We showcase the effectiveness of our proposed framework by applying it to the fastMRI knee and brain datasets [141] within the context of the MoDL reconstruction framework [2].

In-vivo experimental results indicate a strong correspondence between our uncertainty estimation outcomes and the absolute residual error. Furthermore, our approach refines the heuristic uncertainty estimation, quantitatively guaranteeing the desired confidence levels.

This demonstrates the potential of our framework in enhancing the reliability and accuracy of deep learning-based reconstructions in medical imaging applications.

Our method trains an uncertainty estimation network, then calibrates that network to achieve a rigorous guarantee. We will now detail these two subroutines.

6.2 Training the uncertainty estimation network

Given a pre-trained reconstruction network (*e.g.*, MoDL [2]), the uncertainty estimation network f_θ takes the intermediate features and predicts the absolute residual error for that network (Figure 6.2a). The pre-trained network G_w takes the zero-filled reconstruction and maps it to $\hat{\mathbf{x}}_i$, an estimate of the ground truth image $\hat{\mathbf{x}}_i$. Our uncertainty estimation network f_θ is trained to output an estimate $\mathbf{err}_i$ of the magnitude of the residual error $|\hat{\mathbf{x}}_i - \hat{\mathbf{x}}_i|$.

In practice, the input provided to f_θ consists of multiple concatenated features extracted from each iteration of G_w. Once the training process is complete, our framework is capable of mapping new, unseen under-sampled inputs to both reconstructed images and corresponding uncertainty estimates in a single forward pass, ensuring efficiency and accuracy in clinical settings. It is important to note, however, that there is no inherent guarantee that $\mathbf{err}i$ effectively estimates the pixel-wise error. As a result, it is crucial to calibrate the uncertainty estimation network to ensure its accuracy and reliability.

6.3 Calibration of the heuristic uncertainty estimates

Once the uncertainty estimation network is trained, we aim to calibrate its output using RCPS (Figure 6.2b) to achieve a statistical guarantee. We first select a subset of the validation set to form the calibration set $(\mathbf{x}_i, \hat{\mathbf{x}}_i, \mathbf{err}_i), i = 1, 2, 3, 4, ...N$ (typically $N \gtrapprox 1000$). Then, we calibrate a global scalar $\hat{\alpha}$ from the calibration set to ensure that, on average, at least $(1 - \gamma)$ of all pixels from the reference are within its confidence intervals:

$$I_i^{(m,n)} = [\hat{\mathbf{x}}_i^{(m,n)} - \hat{\alpha} \cdot \mathbf{err}_i^{(m,n)}), \hat{\mathbf{x}}_i^{(m,n)} + \hat{\alpha} \cdot \mathbf{err}_i^{(m,n)})], \tag{6.1}$$

for all pixel locations (m, n) in an image of size $M \times N$. For example, choosing $\gamma = 0.05$ and $\delta = 0.1$ will result in 95% of the pixels being contained in their intervals with 90% probability. The detailed calibration procedure is described as follows. For a given image $\mathbf{x}_i$, we first define the loss:

$$L(\alpha)_i = \frac{|(m, n) : \mathbf{x}_i^{(m,n)} \notin I_i^{(m,n)}|}{MN} \tag{6.2}$$

as the fraction of pixels not included in their respective intervals. We compute the empirical risk over the calibration dataset and use the Upper Confidence Bound (UCB) [132, 42] procedure from [6, 8] with the Waudby-Smith and Ramdas (WSR) bound from [132] to choose the smallest α that gives a RCPS,

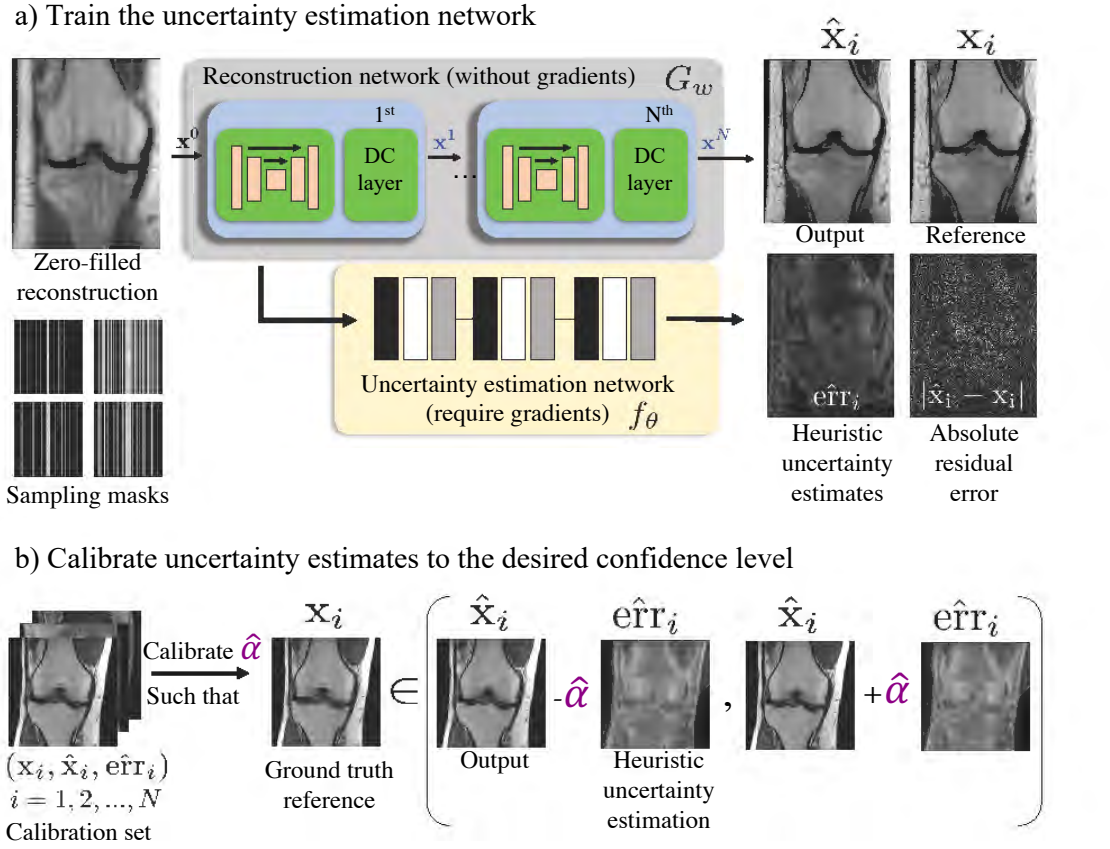


Figure 6.2: **Detailed subroutines for the proposed framework.** a) we first train an uncertainty estimation network f_θ to predict the pixel-wise residual of a pre-trained reconstruction model G_w, where we name the output as heuristic uncertainty estimates. b) After training, we calibrate the uncertainty estimates to form finite-sample confidence intervals, which ensures that on average, (1-γ) of pixels are covered within the confidence interval with high probability regardless of the distribution of the training data.

$$\mathbb{P}[\hat{R}^+(\alpha) \geq R(\alpha)] \geq (1-\delta), \tag{6.3}$$

where δ here is the desired violation rate (*e.g.*, $\delta = 0.1$). In short, the method involves computing the UCB $\hat{R}^+(\alpha)$ using a pointwise concentration inequality, then picking

$$\hat{\alpha} = \min\{\alpha : \hat{R}^+(\alpha') < \gamma, \forall \alpha' > \alpha\}. \tag{6.4}$$

Deploying this choice of $\hat{\alpha}$ guarantees risk-control; we defer the proof of this to [8].

6.4 Datasets and experimental setups

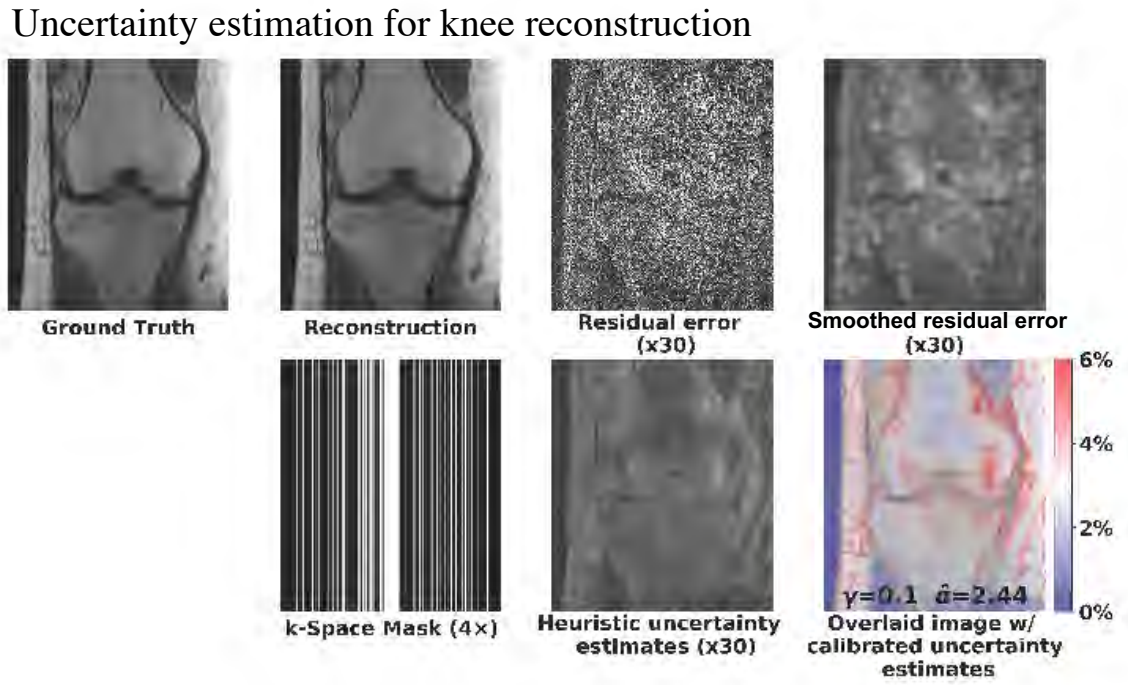


Figure 6.3: **Representative uncertainty estimation comparisons for knee reconstructions.** We compare our uncertainty estimate and the absolute residual error for the Proton density sequence. We also visualize the smoothed absolute residual error for comparison. We overlaid the MoDL reconstructed images and the calibrated uncertainty estimates for better visualization. Colorbar along with the overlaid image indicates the guaranteed confidence interval with respect to the maximum value of the image.

We assessed the performance of our proposed uncertainty estimation framework on both 2D knee and brain fastMRI [141] datasets to demonstrate its effectiveness and applicability across different anatomical structures.

Initially, we trained MoDL for both anatomies using 5120 distinct slices, ensuring the reconstruction network was well-suited to handle the variations present in knee and brain MRI data. Subsequently, we trained the uncertainty estimation network f_θ using the same training set, employing an acceleration factor of 4 to expedite the training process. After training the networks, we proceeded to calibrate the heuristic uncertainty estimates using a calibration set consisting of 1000 slices. Meanwhile, we established a validation set containing 2000 slices to facilitate a comprehensive comparison.

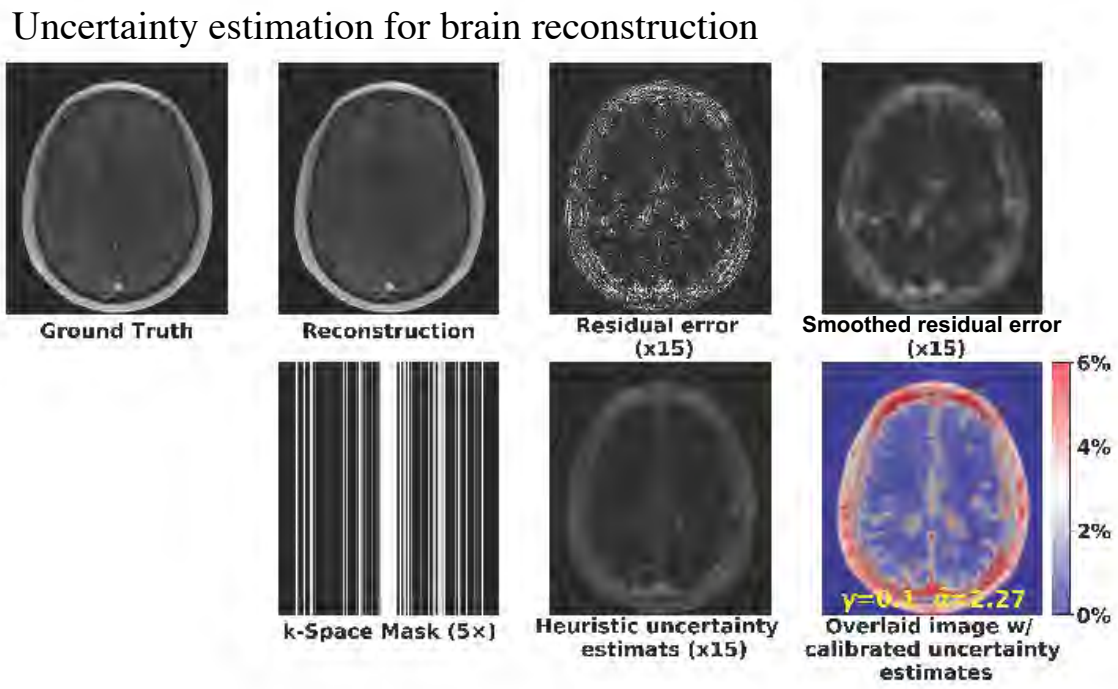


Figure 6.4: **Representative uncertainty estimation comparisons for brain reconstructions.** We compare our uncertainty estimate and the absolute residual error for the Proton density sequence. We also visualize the smoothed absolute residual error for comparison. We overlaid the MoDL reconstructed images and the calibrated uncertainty estimates for better visualization. Colorbar along with the overlaid image indicates the guaranteed confidence interval with respect to the maximum value of the image.

To evaluate the calibration procedure's efficacy, we randomly split the validation set 2000 times. For each iteration, we calibrated a $\hat{\alpha}_j, j = 1, 2, 3, ..., 2000$ and assessed the empirical risk $\hat{R}_j$ on the remaining validation set (evaluation set). By presenting a histogram of the empirical risks, we were able to evaluate the empirical violation rate $\hat{\delta}$, which served as a key metric for assessing the performance of our calibration process.

6.5 Results

Figure 6.3 and 6.4 present the uncertainty estimation outcomes for the knee and brain reconstructions, providing a comprehensive comparison between our uncertainty estimates and the absolute residual error for the Proton Density sequence. Our findings exhibit a robust correlation between the uncertainty estimates and the attenuated residual error. Additionally, we have incorporated a visual representation of the softened absolute residual error to

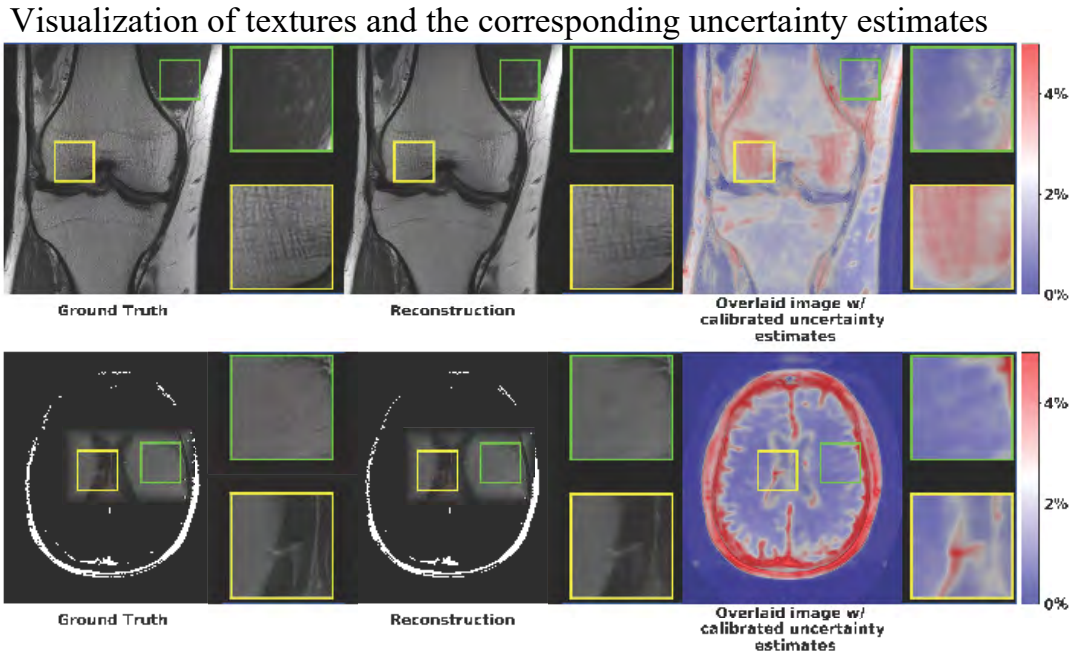


Figure 6.5: **Visualization of textures and the corresponding uncertainty estimates from two representative images.** As can be seen in the zoomed-in details, the reconstructions of the green-outlined patches are highly similar to the ground truth ones, while those of the yellow-outlined patches are of lower quality, since some of the high-frequency details are missing or blurred out. This is reflected by the overlaid calibrated uncertainty estimates, where the yellow-outlined patches have much higher uncertainty levels than the green ones.

facilitate an effective comparison. To improve the visualization experience, we have combined the MoDL-reconstructed images with their respective calibrated uncertainty estimates in a cohesive display. The accompanying color bar serves as a guide to indicate the guaranteed confidence intervals relative to the maximum value found within the image, further enhancing the overall understanding of the data presented.

Figure 6.5 provides a detailed visualization of the textures and their associated uncertainty estimates, effectively highlighting the intricate connections between them. By examining the zoomed-in sections of these images, it becomes evident that regions with higher uncertainty coincide with areas where the reconstructed images were unable to successfully capture the finer textures and subtle details.

Figure 6.6 shows the empirical risk distribution given different splits of calibration and evaluation sets. Histograms show that the empirical violation rate $\hat{\delta}$ hits nearly exactly $\hat{\delta}$ for both $\hat{\gamma}$, which demonstrates the tightness and validity of our calibration procedure.

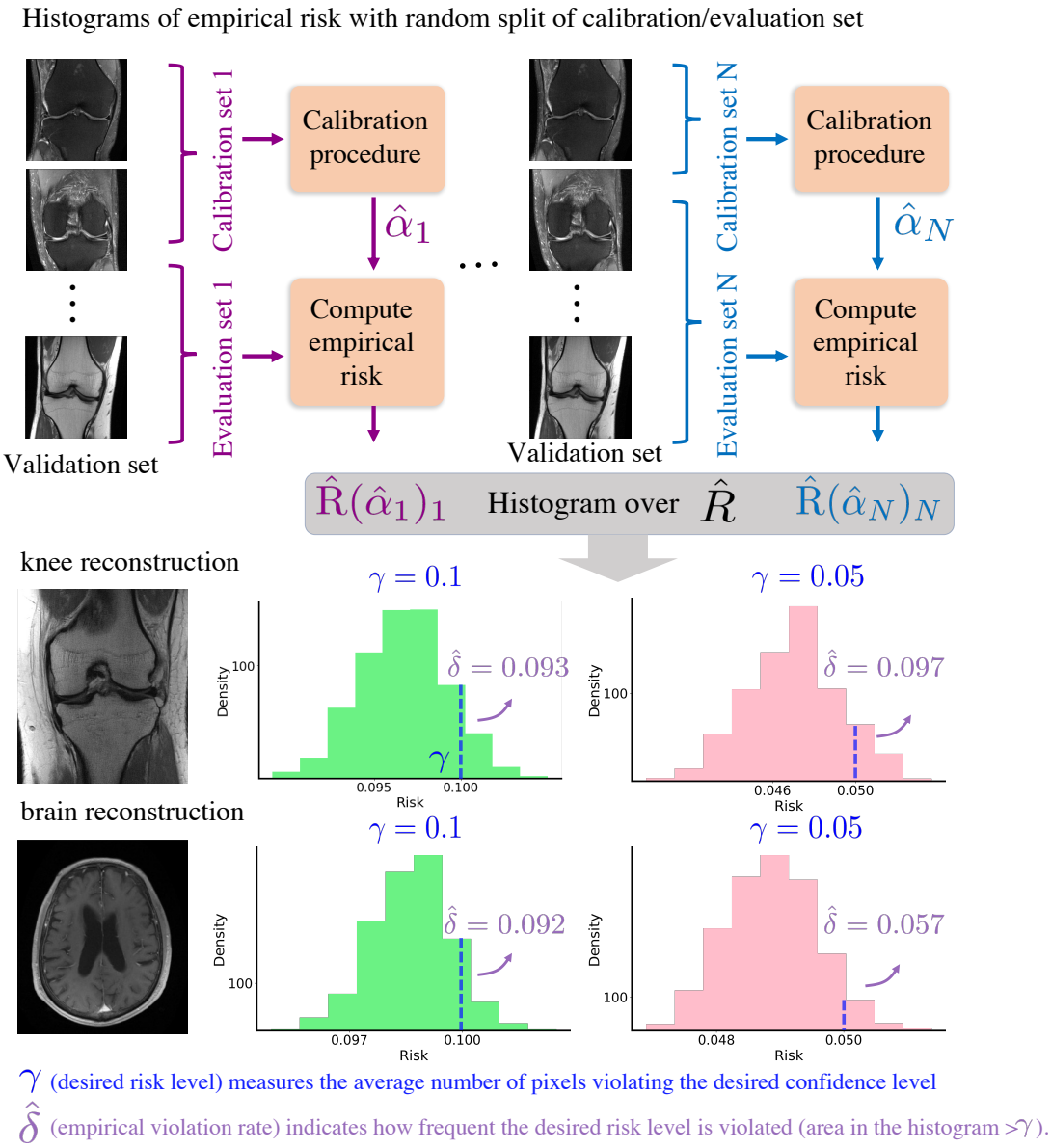


Figure 6.6: **Empirical risk distribution under 2000 random split of calibration/evaluation sets for brain and knee datasets.** Each split of the calibration set outputs an $\hat{\alpha}$ and the corresponding empirical risk $\hat{R}$, which roughly describes the number of pixels violating the desired risk/confidence level. Given a desired violation rate, the empirical violation rate $\hat{\delta}$ indicates how frequently the desired risk/confidence levels are violated. Comparisons of two desired risk/confidence levels $\gamma = 0.1, 0.005$ are presented.

6.6 Conclusions

In this chapter, we have introduced a comprehensive and robust uncertainty estimation framework, designed to offer precise uncertainty estimates supported by finite-sample guarantees. Remarkably, our framework functions as a versatile plug-and-play module, imposing no constraints on the reconstruction model. As a result, it holds the potential to substantially enhance the accuracy of diagnoses and clinical interpretations derived from deep learning-based reconstructions.

By providing a solid foundation for understanding the limitations and uncertainties inherent in reconstructed images, our innovative approach offers valuable insights that can be employed to improve clinical decision-making processes, ultimately contributing to better patient care and outcomes.

Chapter 7

Complex-valued Scattering Representations

7.1 Introduction

So far, this dissertation has introduced various DL-based image reconstruction techniques. In this particular chapter, I venture beyond the box and place emphasis on the distinctiveness of general complex-valued DL approaches for MR images, with a specific focus on addressing the challenge of limited training data.

Unlike deep learning applied to real-valued natural images, MR imaging inherently involves complex-valued images due to the underlying physics principles. The phase of these images carries crucial information for a broad range of applications, such as flow imaging and susceptibility-weighted imaging. However, most existing networks are real-valued, which results in the loss of important phase information. As a result, complex-valued deep learning has emerged as a potent approach for modeling complex-valued data, taking advantage of the distinctive algebraic operations and properties of complex-valued data to develop more precise and efficient models.

On the other hand, MRI faces challenges arising from the limited availability of fully-sampled ground truth and well-annotated data. To date, the largest raw MRI dataset - fastMRI [141] contains 1,594 volumes, which is way less than natural image datasets (*e.g.*, ImageNet [24], LAION-5B [100]).

The present chapter aims to concentrate on complex-valued deep learning, introducing Complex-valued Scattering Representations (CSR) as a universal representation for diverse input domains. This approach achieves exceptional performance in image classification tasks, particularly in the context of limited training data.

Prior to delving into the specifics, let us begin with a brief overview of complex-valued deep learning. Complex-valued deep learning is an emerging field that extends traditional deep learning models to handle complex-valued data, enabling more accurate and efficient systems for a wide range of scientific and engineering applications. Recent developments in

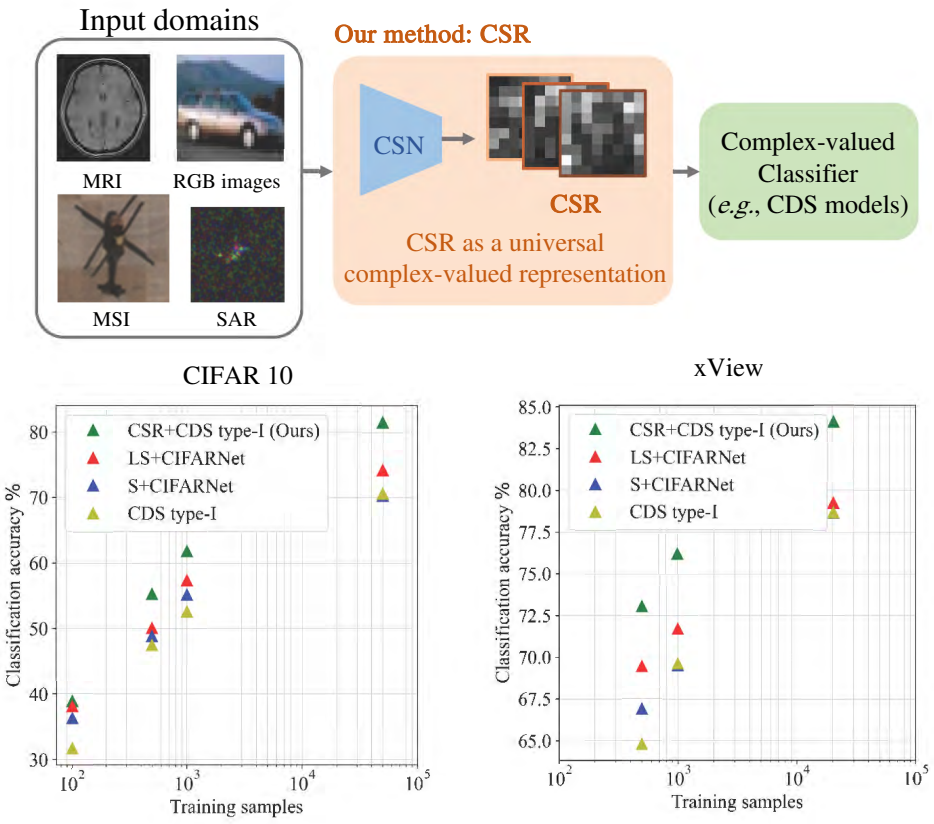


Figure 7.1: **Complex-valued Scattering Representations (CSR) serve as universal complex-valued representations for a wide range of input domains. TOP:** Given image from an input domain (*e.g., RGB image, MRI, MSI*), our Complex-valued Scattering Networks (CSN) output CSR, which is then fed into the complex-valued classifiers as universal complex-valued representations. **Bottom:** By incorporating CSR with Co-domain Symmetry (CDS) models, our approaches significantly outperform CDS and other real-valued counterparts with different training samples on CIFAR 10 and xView benchmarks.

manifold geometry [15, 105] and group theory have further advanced the field, leading to the creation of leaner and better classifiers with novel complex-valued layer functions and network architectures [119, 114, 105, 15].

While complex-valued deep learning was initially developed to better model naturally complex-valued data such as MRI, and Synthetic Aperture Radar (SAR), it has recently been shown to be effective for real-valued input data such as RGB [105] or multispectral images [106] through complex-valued representations, delivering leaner and better classifiers with novel complex-valued layer functions and network architectures.

Singhal *et al.* [105] introduced "sliding" and "LAB" encodings to convert RGB color space

to complex-valued representations. "Sliding" encoding maps adjacent color channels to the real and imaginary parts of a complex-valued channel to exploit the inter-channel correlations. "LAB" encoding converts RGB space to a real-valued luminance and a complex-valued chromaticity, which leads to color distortion robustness without color jitter augmentations.

Despite notable progress achieved with current encoding methods, the resulting complex-valued representations for real-valued inputs are rudimentary and limited to channel characteristics, lacking the ability to model any spatial and spatial-frequency properties of the input data.

A promising solution that can capture both spatial and spatial-frequency features at the same time is Wavelet transform. Wavelet transforms achieve this by using a set of filters that can decompose an image into different spatial-frequency bands at different scales, allowing for the extraction of both spatial and spatial-frequency features simultaneously. Building on this property, Bruna et al. [12] proposed real-valued Wavelet Scattering Networks (WSNs), which have achieved notable success in extracting non-learned features for image classification tasks, particularly when training with limited labeled data [85, 31].

Inspired by Wavelets and WSNs, we propose learnable Complex-valued Scattering Representations (CSR) as a universal complex-valued representation to model the spatial and spatial-frequency properties of the input data. We introduce the term Complex-valued Scattering Networks (CSNs) to refer to the networks that produce CSR as their output for convenience. As shown in Figure 7.1, we further integrate CSR with complex-valued deep learning models, such as complex-valued Co-domain Symmetry models (CDS) [105], for downstream image classifications.

As shown in Figure 7.2, we construct filters based on complex-valued Morlet wavelets [85, 12, 31]. Instead of limiting the filters to the half-frequency domain as in real-valued WSNs, CSN extends them to the entire spatial-frequency domain. Motivated by [31], we learn the geometry parameters of the complex-valued filters (*i.e.* orientations and aspect ratios) to better fit the dataset.

Additionally, in contrast to using a fixed absolute value as the non-linear activation function, which can discard important phase information, we introduce a learnable high-dimensional complex-valued ReLU function as the non-linear activation module. This enables our CSR to better adapt to the complexities of the input data by preserving the phase information.

We integrated CSR into complex-valued models (Linear layer (LL) and CDS) and achieved significant classification performance improvements compared to CDS and other real-valued WSN-based models, especially on tasks with limited labeled data. Our evaluation includes various benchmarks from different domains such as CIFAR 10/100 [55], xView MSI classification [106], and a newly introduced complex-valued MRI Patch classification dataset.

To summarize, we make the following contributions:

- We propose CSR, a novel and universal complex-valued representation for extracting both spatial and spatial-frequency features from diverse input image domains in complex-valued deep learning.

- We introduce a novel learnable high-dimensional Complex-valued ReLU function as the non-linear activation module for our CSR. This module enhances the network's ability to effectively adapt to the complexities of the input data.
- By integrating CSR with complex-valued models, our approach outperforms complex-valued models and real-valued WSNs in CIFAR10/100, xView MSI, together with a new evaluation benchmark of complex-valued MRI patch classification.

7.2 Related Work

Complex-valued networks

Complex-valued neural networks (CVNNs) are an extension of traditional real-valued neural networks designed to handle complex-valued data. Due to the importance of complex numbers in engineering and scientific disciplines [74], CVNNs have been an active topic since the early days of deep learning research.

The paper [77] analyzes CVNNs in the context of the XOR problem and finds that the real and imaginary components of the decision boundary of a CVNN are orthogonal. Further works demonstrate better optimization properties [76] and representational capacity [78]. We refer the reader to [7] for a deeper review of CVNNs.

A central question in this literature is how to adapt real-valued deep learning to complex numbers. Previous works [119, 147, 114] redefine basic building blocks for complex-valued networks, such as complex-valued convolution, batch normalization, and non-linear activations. However, those methods are not robust against complex-valued scaling. SurReal [15] addressed this issue by modeling the complex value space as a manifold to enable robustness to complex-valued scaling. Meanwhile, Singhal et al. [105] developed equivariant and invariant neural network layers for co-domain transformation that outperform other complex-valued networks in image classification tasks. Additionally, they proposed novel complex-valued encodings of real-valued color space, such as RGB, to effectively exploit inter-channel correlations for real-valued inputs.

While current complex-valued encoding approaches have made notable progress, they still lack the ability to effectively model both spatial and spatial-frequency properties of the input data. Our proposed CSR, on the other hand, successfully captures both types of features.

Scattering representations

Scattering representations, as proposed by Bruna and Mallat in [12], leverage pre-determined wavelet filters to create powerful hierarchical representations and extract features from both the spatial and spatial-frequency domains. From a mathematical perspective, these representations satisfy translation invariance up to a particular scale and are stable to deformations, making them a simple yet effective tool for signal analysis in various fields such as image

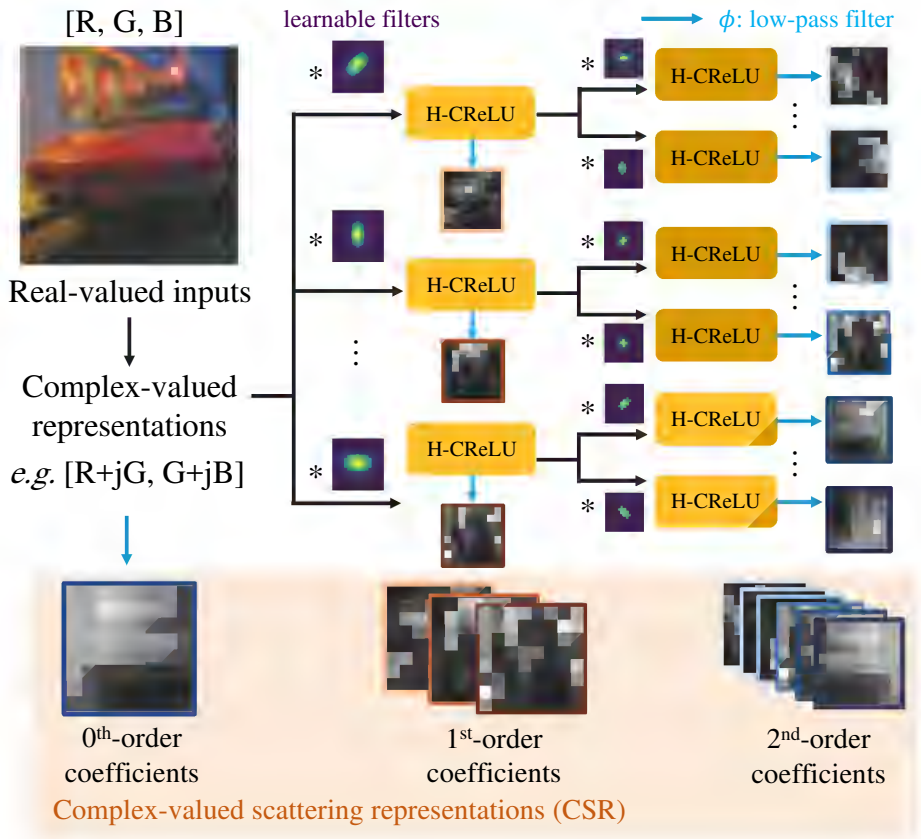


Figure 7.2: **Diagram of obtaining CSR from real-valued inputs.** Real-valued inputs are first converted to complex-valued representations using "Sliding" encodings [105]. We then convolve with learnable filters and apply our high-dimensional Complex ReLU (H-CReLU) module to extract the scattering coefficients up to 2^{nd} order. H-CReLU lifts a complex number to high-dimensional space, applies point-wise CReLU, and maps back to a complex number. Coefficients from different orders are then concatenated to form CSR.

and audio processing [12, 4, 41, 29]. Benefiting from the well-designed filters, scattering representation-based models have shown promising results in applications with limited labeled data.

Oyallon et al. [85] introduce hybrid networks, which demonstrate the effectiveness of scattering transforms as early layers of learned CNNs. Hybrid networks outperform other predefined representations and show comparable results with end-to-end learned CNNs. McEwen et al. [70] constructed scattering networks on the sphere, providing a powerful representational space for spherical data. Gauthier et al. [31] learns the geometric parameters of wavelet filters (*e.g.* orientation, aspect ratio), achieving new state-of-the-art results in a low-data regime.

Our CSR can be seen as an extension of real-valued scattering representations to the complex-valued domain, providing a universal and powerful representation for complex-valued deep learning.

7.3 Complex-valued Scattering Representations (CSR)

Figure 7.2 visualizes how we obtain CSR from real-valued inputs (RGB image as an example). For simplicity, we limit our focus to 2D CSNs and only consider their up to 2^{nd} order coefficients. It's worth noting that previous studies have demonstrated that higher-order coefficients yield negligible energy [12]. We start with a real-value image I of m channels, we turn it into a complex-valued image of $m-1$ channels through "sliding" color encoding [105]:

$$\begin{aligned} I(u) &= [I_1, I_2, ..., I_m] \\ &\rightarrow [I_1 + jI_2, I_2 + jI_3, ..., I_{m-1} + jI_m], \end{aligned} \tag{7.1}$$

where u is the spatial position index, $j = \sqrt{-1}$.

Our CSN starts with the complex-valued representation $I(u)$, a scaling integer $J \in \mathbb{N}$, and an integer $L \in \mathbb{N}$ representing the number of wavelet angular orientations. CSN computes the scattering coefficients $S^0 I$, $S^1 I$, and $S^2 I$ of orders 0, 1, and 2, respectively, which can be interpreted as the result of convolving $I(u)$ with 0, 1, and 2 wavelet filters. J represents the spatial scale of the scattering transform.

As shown in Figure 7.2, to compute the 0^{th} order coefficient, we use a low pass filter ϕ_J with a spatial window of scale 2^J (here is the Gaussian smoothing function). To obtain the coefficient, we convolve the input signal $I(u)$ with ϕ_J, and then downsample the result by a factor of 2^J. This operation can be expressed as $S^0 I(u) = I * \phi_J(2^J u)$. To recover the high-frequency information that S^0 discards, higher-order coefficients are introduced using wavelets.

A Morlet wavelet family is derived by scaling and rotating a complex-valued mother wavelet ψ. Specifically, we obtain a particular Morlet wavelet at scale $j \geq 0$, rotation θ, and aspect ratio γ by dilating the mother wavelet as follows:

$$\psi_{j,\theta,\gamma}(u) = \frac{1}{2^{2j}} \psi_\gamma(r^{-\theta} \frac{u}{2^j}), \tag{7.2}$$

where $r^{-\theta}$ represents the rotation by $-\theta$. For real-valued WSNs, it's important to note that the spatial-frequency domain exhibits conjugate symmetry. As a result, the rotation angle θ is constrained to range from $[0, \pi)$. In CSNs, we design θ to range from $[0, 2\pi)$.

To compute the 1^{st}-order scattering coefficients, we convolve the input signal with one of the complex-valued wavelets $\psi_{j_i,\theta_i,\gamma_i}(u)$ and downsample the response by the scale 2^{J-j_i}.

Next, we apply a pointwise activation function $f(\cdot)$ to the downsampled signal to add non-linearity. Finally, the smoothed signal is obtained by convolving it with the low-pass filter $\phi_J(2^J u)$.

For real-valued WSNs, $f(\cdot)$ is usually a complex modulus, which takes the absolute value $|\cdot|$ of a complex number. However, complex modulus discards its phase information which can be crucial for complex-valued applications where phase carries important information. Here, we propose a learnable activation function $f_w(\cdot)$, where w is the learnable parameters (§ 7.4). Mathematically, the 1^{st}-order coefficients can be expressed as:

$$S^1 I(u) = f_w(I * \psi_{j_i,\theta_i,\gamma_i}) * \phi_J(2^J u). \tag{7.3}$$

Similarly, as illustrated in Figure 7.2, we perform a second wavelet transform on each channel of the 1^{st}-order coefficients before applying the low-pass filter. This can be written as:

$$S^2 I(u) = f_w(f_w(I * \psi_{j_i,\theta_i,\gamma_i}) * \psi_{j_k,\theta_k,\gamma_k}) * \phi_J(2^J u), \tag{7.4}$$

where $\psi_{j_k,\theta_k,\gamma_k}$ is the second filter we apply. Due to the spatial-frequency supports of filters, only coefficients with $j_i < j_k$ have significant energy [12].

Motivated by [31], we let the network learn the orientation θ and aspect ratio γ of each wavelet to enable better adaptions to particular datasets. θ is initialized to be equally spaced on $[0, 2\pi]$, while γ is initialized as a constant $\frac{4}{L}$. We adapted Kymatio software package [5] to implement CSNs.

7.4 Learnable high-dimensional complex ReLU

As we pointed out, the complex modulus for SNs discards the important phase information from the signal. One alternative is to use complex-valued ReLU (CReLU) [1, 114]. However, CReLU destroys the phase information other than the first quadrant. Thus, instead of using a hand-crafted function, we proposed a learnable high-dimensional CReLU (H-CReLU) module (Orange block in Figure 7.2).

Motivated by other high-dimensional lifting methods [109, 97], H-CReLU operates on a complex number $z \in \mathbb{C}$ by first lifting it to a higher-dimensional space using linear mapping. Specifically, we use a trainable matrix $\mathrm{UP}_{N_h} \in \mathbb{C}^{N_h \times 1}$ to transform z into a N_h-dimensional representation, where N_h is set to 16 in our experiments. After lifting the input, we apply point-wise CReLU to the high-dimensional intermediate results. Finally, we map the high-dimensional intermediate results back to the original space using a trainable matrix $\mathrm{DOWN}_{N_h} \in \mathbb{C}^{1 \times N_h}$. The resulting activation function, $f_w(z)$, can be then written as:

$$f_w(z) = \mathrm{DOWN}_{N_h} \cdot \mathrm{CReLU}(\mathrm{UP}_{N_h} \cdot z), \tag{7.5}$$

where $\{\mathrm{UP}_{N_h}, \mathrm{DOWN}_{N_h}\}$ are the learnable matrices with $2N_h$ complex-valued learnable parameters. Ablation studies demonstrate the effectiveness of H-CReLU as $f_w(z)$.

7.5 CSR for downstream image classification

We integrate CSR with complex-valued models for downstream image classification tasks. We convert real-valued inputs, such as CIFAR 10/100 and xView, to complex-valued representations using "sliding" encoding 7.1 before computing CSR. On the other hand, we extract CSR directly from complex-valued data such as MRI patches.

In our experiments, we integrate CSR with two different types of recently proposed complex-valued networks: 1) Complex-valued linear layer; 2) Type-I CDS with CIFARNet architecture from [105]. We also include CDS-Large with Wide Residual Network (WRN) [140] architecture from [105] for comparisons.

7.6 Experiments

We compare the performance of our approach against real-valued scattering representations and previous complex-valued models (without scattering) on four diverse image datasets: CIFAR 10, CIFAR 100, xVIEW MSI [106, 58], and a newly introduced dataset for MRI patch classification. In addition, we assess the model's performance under limited-labeled training data.

CIFAR 10 and CIFAR 100 are well-established natural RGB image classification benchmarks that have been used for similar analysis [31, 12, 85]. xView MSI is a large-scale 8-band MSI dataset. Each channel within an 8-band image contains measurements obtained from a different electromagnetic spectrum. Following [106], xView consists of 60 total classes, from which we select 10 supercategories.

MRI classification dataset We created a new complex-valued dataset for MRI patch classification to showcase the effectiveness of CSR on naturally complex-valued data. To tackle the scarcity of labeled MRI data, we performed automatic labeling by slicing a volumetric MRI dataset into its cross-sections.

MRI data itself is complex-valued due to the physics and acquisition process involved in capturing it. The magnitude of an MR image pixel provides information about different tissues, but the phase component is also important for applications such as flow imaging and susceptibility imaging.

To create our MRI patch dataset, we used complex-valued multi-echo 3D MRI volume data from [102], which was originally intended for susceptibility mapping. The dataset includes 144 3D GRE scans from eight healthy subjects, all acquired using a single MR scanner. We take the first echo volumes and slice 2D images from three different orientations (*i.e.*, Sagittal, Axial, Coronal). Then, as shown in Figure 7.3, 2D patches (32×32) are extracted for each orientation. The objective is to train a classifier that can correctly identify the orientation of the complex-valued patch (3-classes classification task). Our training set consists of 26,640 patches extracted from 4 subjects, while the testing set consists of 17,520 patches from another 4 subjects.

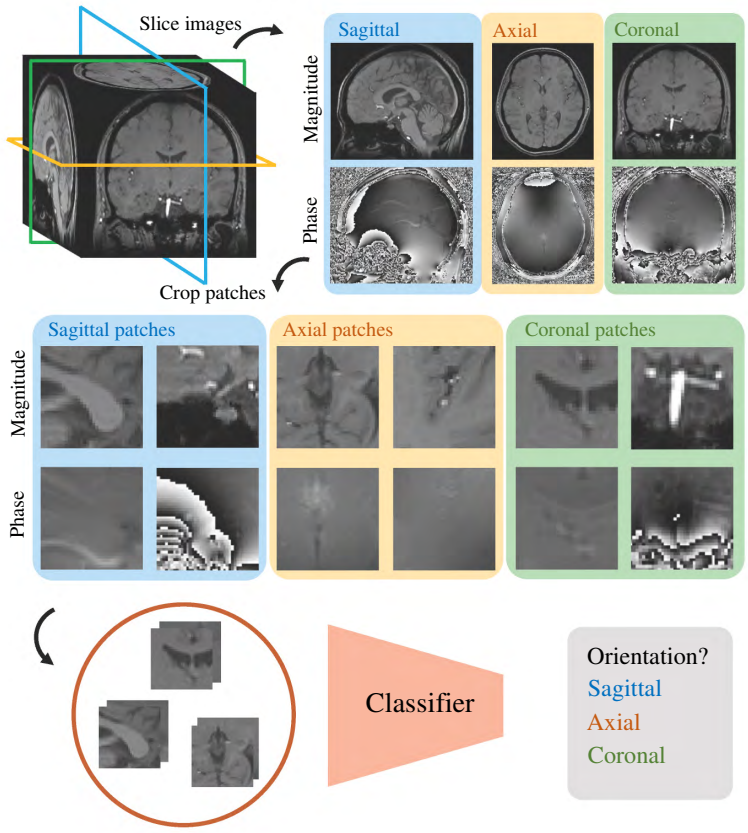


Figure 7.3: **Construction of complex-valued MRI patch classification dataset.** We start from complex-valued multi-echo 3D MRI volumes obtained from [102]. To create our dataset, we sliced 2D images from different anatomical orientations, including sagittal, axial, and coronal. We then cropped patches from these images to generate our dataset of complex-valued patches. The objective is to train a classifier that can correctly identify the anatomical orientation of the input complex-valued patch.

We evaluate CSR by utilizing them with two common complex-valued models. In the first model, we considered CSR as the input of a simple LL. This configuration of LL helps us understand the linear separability of CSR. In the second case, we integrate CSR with recently proposed complex-valued CDS networks [105], where we experimented on Type-I CDS. For both models (LL and CDS), we compare our CSNs with their real-valued counterparts, including conventional scattering (S) and recently proposed learnable scattering (LS) [31]. We design the networks to have a similar number of parameters for fair comparisons. For reference, we also compare our approach to CDS-Large (complex-valued) and WRN-16 (real-valued).

All of our models and experiments are implemented in PyTorch [87] and optimized using the AdamW optimizer [52, 63] with an initial learning rate of 3×10^{-3}, decayed by a factor

0.3 every 10k iterations. We use a batch size of 256 for 50k training iterations.

Benchmark comparisons

Method	CIFAR 10				CIFAR 100			
-	100 samples	500	1000	All	1000 samples	5000	10000	All
Scattering + Linear layers								
S [12]+LL	35.78±0.62	48.32±0.30	53.52±0.24	65.46	17.03±0.74	33.00±0.50	37.98±0.22	41.12
LS [31]+LL	37.87±0.55	52.88±0.26	56.94±0.20	69.68	18.96±0.71	33.95±0.63	39.83±0.18	43.65
CSR+LL †	**39.84** ±0.54	**56.23** ±0.32	**60.01**±0.16	**74.30**	**20.07** ±0.82	**34.54**±0.49	**41.18**±0.29	**47.81**
Scattering + CIFARNet								
S [12]+CIFARNet	36.23±0.70	48.88±0.54	55.17±0.18	70.23	17.29±0.93	30.44±0.39	36.23±0.28	40.76
LS [31]+CIFARNet	38.06±0.68	50.92±0.58	57.34±0.26	74.07	17.90±0.85	32.05±0.51	38.45±0.30	42.81
CSR+CDS type-I †	**38.87**±0.49	**55.26**±0.45	**61.78** ±0.14	**81.52**	**18.68**±0.77	**34.24**±0.40	**40.03**±0.37	**46.80**
CDS and large models (no scattering)								
CDS type-I [105]	31.67±0.50	47.53±0.21	52.57±0.31	70.55	15.52±1.01	29.77±0.36	33.98±0.23	37.14
CDS large [105]	**33.32**±0.98	**48.65**±0.27	**60.23**±0.13	93.27	**17.30**±0.65	33.73±0.72	48.19±0.33	71.03
WRN-16 [140]	32.55±1.13	44.19±0.83	59.57± 0.40	**96.34**	17.03±1.38	**36.99** ±1.04	**53.98** ±0.57	**76.35**

† ours ; S: Scattering [12]; LS: Learnable Scattering [31]; CSR: Complex-valued Scattering Representations (ours).
# Parameters for CIFAR 10 (CIFAR 100): 156k (1.6M) for S+LL; 156k (1.6M) for LS+LL; 207k (2.1M) for CSR+LL; 124k (136k) for S+CIFARNet; 124k (136k) for LS+CIFARNet; 122k (145k) for CSR+CDS type-I; 105k (128k) for CDS type-I; 1.7M (1.8M) for CDS large; 17.1M (22.4M) for WRN-16

Table 7.1: **Classification accuracy for CIFAR 10 and CIFAR 100 benchmarks (mean ± std.).** We report results from models trained with varying sample sizes to demonstrate the effectiveness of CSNs. To calculate the standard error in limited-data regimes, we trained our models using 10 different seeds. **Bold** highlights the best results in each category, while **Bold** represents the best results across all categories. CSNs outperform their real-valued counterparts and CDS (without CSR) in all training setups.

CIFAR 10 (and CIFAR 100) consists of 10 (100) classes containing 6,000 (600) images from each class. Each image has a size of 32×32. Both datasets are split into a training set of 50,000 images and a test set of 10,000 images. The training images are augmented with horizontal flipping, and random cropping. We use "sliding" color encoding [106] and set spatial scale $J = 2$. Additionally, we also evaluate the performance of CSRs in small data regimes with limited labeled data. We train our models on a small random subset of the training data but evaluate their performance on the entire testing set as done in previous works [31, 85]. To account for the randomness in data selection, we train the same model using 10 different seeds for the small-size experiments. The set of seeds used is consistent for models trained with the same sample size. We compute the mean and standard error across 10 different seeds. We evaluate training size of $\{100, 500, 1000, 50k\}$ for CIFAR 10, and $\{1000, 5000, 10000, 500k\}$ for CIFAR 100.

Method	xView			MRI patch classification	
-	500 samples	1000	All	100 samples	500
Scattering + Linear layers					
S [12] + LL	62.55±2.35	68.45±1.47	74.30	56.79±0.88	68.95±0.34
LS [31] + LL	67.69±2.01	71.14±1.88	75.78	67.03±0.64	85.40±0.42
CSR + LL †	**71.83**±2.70	**74.86**±1.17	**80.04**	**74.22**±0.57	**91.73**±0.33
Scattering + CIFARNet					
S [12] + CIFARNet	66.88±2.65	69.54±1.60	78.68	59.62±0.80	83.53±0.96
LS [31] + CIFARNet	69.49±2.05	71.72±1.68	79.25	71.86±0.98	94.74±0.60
CSR + CDS type-I †	**73.07** ±1.79	**76.18** ±1.21	**84.13**	**84.80** ±1.06	**99.18** ±0.15
CDS and large models					
CDS type-I [105]	64.80±2.45	69.65±1.33	78.69	54.49±0.34	69.77±0.38
CDS large [105]	**68.45**±2.32	**72.77**±0.98	81.80	**82.68**±0.43	**98.45**±0.17
WRN-16 [140]	61.13±3.74	70.46±1.52	**84.25**	39.25±1.43	55.66±2.35

† ours ; S: Scattering; LS: Learnable Scattering; CSN: Complex-valued Scattering Network (ours). # Parameters for xView (MRI Patch classification): 415k (31k) for S+LL; 415k (31k) for LS+LL; 726k (31k) for CSR+LL; 364k (76k) for S+CIFARNet; 364k (76k) for LS+CIFARNet; 357k (73k) for CSR+CDS type-I; 111k (102k) for CDS type-I; 1.8M (1.7M) for CDS large; 17.1M (17.1M) for WRN-16

Table 7.2: **Classification accuracy for xView and MRI patch classification dataset (mean ± std.).** XView models were trained with sample sizes of 500, 1000, and full size, while MRI patch classification models used 100 and 500 samples. For both datasets, our CSNs significantly outperform their real-valued counterparts. Table layouts and symbols are the same as Table 7.1.

Table 7.1 summarizes the results under different training setups. In the first LL category, our proposed CSR+LL outperforms the previous state-of-the-art LS method under all training setups. CSR+LL achieves a $> 4\%$ accuracy gain for the full-size ($50k$) training and set a new state-of-the-art for scattering methods in small data training regimes. LL results demonstrate the superior linear separability of our scattering representation and provide the most interpretable comparisons between different models.

For the second category CIFARNet (and CDS type-I) comparisons, Our proposed approach significantly outperforms its real-valued counterparts as well as the CDS model in all the comparisons. When comparing results on CIFAR 10, CSR+CDS type-I outperformed CSN+LL in the full-size dataset and the 1000 samples senorita, getting comparable results in $\{100, 500\}$ samples regimes. However, in CIFAR 100, CSR+LL consistently outperforms CSR+CDS type-I. This may be attributed to LL's larger network capacity (2.1M) under CIFAR-100 setups. Besides, we also present the results of CDS type-I [105], CDS large [105],

WRN-16 [140] as references and comparisons. While large networks tend to excel when trained on ample amounts of data, they often fall short when the available data is limited, such as in the case of CIFAR 10 (100, 500, 1000 examples) or CIFAR 100 (1000 examples). In such scenarios, scattering-based methods tend to yield superior results.

xView MSI dataset Multi-band MSI remote sensing images consist of multiple bands in addition to RGB color images. These images contain highly correlated channels that, when analyzed together, reveal structures with greater clarity compared to the limited information available in 3-band color images. xView MSI dataset contains a total of 86,980 images (size 32×32), with 20,431 images for training, 2,270 images for validation, and 63,279 images for testing. We use a spatial scale $J = 2$ and compare models trained with $\{500, 1000, \text{full size}\}$ samples.

Our experimental results (shown in Table 7.1) demonstrate that CSRs consistently outperform real-valued networks as well as CDS without CSR across all training settings by a substantial margin. Furthermore, we found that our CSR+CDS model achieves the same level of accuracy as WRN-16 on full-sized training data while using only 2% of the parameters. Meanwhile, in small data regimes, our method outperforms WRN-16 by a significant margin.

MRI patch classification Previous sections evaluated CSR on real-valued benchmarks. Here, we evaluate CSR on our complex-valued MRI patch classification dataset, where we directly extract CSR from naturally complex-valued MRI data. MRI patch classification is typically considered an easier task compared to natural image classification tasks like CIFAR [55] and ImageNet [24], primarily due to the lower complexity and diversity of the data. Thus, we create two small-data training regimes: (1) using 100 samples from a single subject, and (2) using 500 samples from 5 scans of a single subject. We use a spatial scale $J = 2$.

Table 7.2 shows that our Complex-Valued Neural Networks (CSR) significantly outperform their real-valued counterparts, CDS (without CSR) and achieve better results compared with large models (*i.e.,* WSN and CDS large). For LL experiments, CSR+LL improves classification accuracy by 7% and 6% for 100 and 500 samples training, respectively. In CIFARNet (CDS type-I) experiments, CSR+CDS produces a notable 13% accuracy gain (for 100 samples training) compared to LS [31]. It's noteworthy that, given the large network capacity and inadequate training data, WRN-16 exhibits poor performance on this task.

Understanding CSR

To gain a better understanding of CSR, we conduct an analysis of the learnable filters and H-CReLU modules. Figure 7.4 showcases the visualization of data-specific scattering filters of CSNs in Fourier space that were trained with linear classification layers. The filters displayed in the figure were trained on CIFAR 10 (full size) and MRI Patch (500 samples) datasets.

Unlike real-valued scattering methods that only cover half of the Fourier space [12, 31, 85] due to the conjugate symmetry, CSN initializes filters equally across the entire Fourier space. As illustrated in Figure 7.4, the filters optimized for both CIFAR 10 and MRI Patch

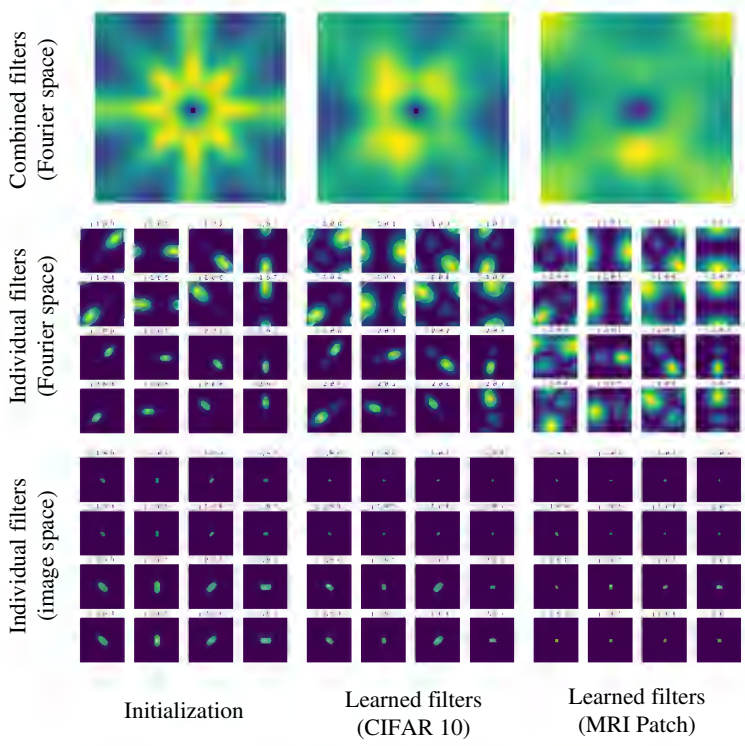


Figure 7.4: **Visualization of learned data-specific filters.** We visualize the learned filters of CSR trained with linear classification layers on CIFAR 10 and MRI Patch dataset. From top to bottom, we present combined filters in Fourier space, individual filters in Fourier space, and individual filters in image space. We initialize the filters equally spaced across the entire Fourier space. After learning, the scattering filters for both CIFAR 10 and MRI Patch datasets exhibit wider bandwidths in the Fourier domain compared to their initialization. Filters optimized for CIFAR 10 have higher spectral energy in the low-frequency regions, while filters optimized for the MRI Patch dataset focus more on high-frequency regions.

present wider bandwidths than the initial filters, resulting in better coverage of the Fourier domain. Furthermore, we observe that the learned filters for both datasets exhibit higher spectral energy in high-frequency regions compared to the initial filters, indicating the ability of our CSR in capturing and representing high-frequency features.

When comparing between the filters optimized for CIFAR 10 and MRI Patch, we notice that the filters designed for MRI Patch present an even higher concentration of high-frequency energy, whereas the filters for CIFAR 10 focus more on low-frequency regions. This observation implies that the classification of MRI patches heavily relies on high-frequency details, while CIFAR 10 classification is more sensitive to low-frequency features.

Figure 7.5 visualizes how our H-CReLU $f_w(\cdot)$ maps the complex numbers. Following [97], we generate an initial set of points on a spiral trajectory on the complex plane. Each point

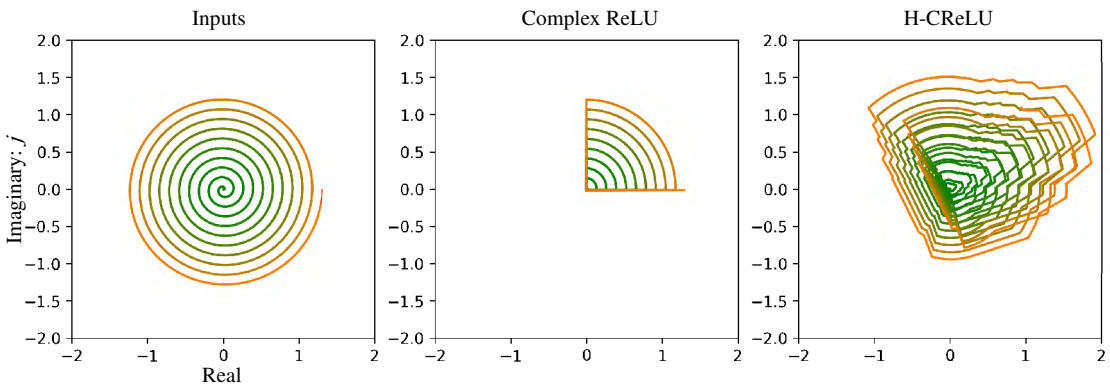


Figure 7.5: **Visualization of H-CReLU in mapping points on the complex plane.** We generate an initial set of points on a spiral trajectory on the complex plane, where each point corresponds to a unique complex number. We then visualize how Complex ReLU (CReLU) and our H-CReLU map (CIFAR 10 with linear layers) the input complex numbers to their outputs. The same color corresponds to the same points across figures. CReLU results in certain input points collapsing into each other, while H-CeLU successfully avoids information loss.

of coordinate (x, y) corresponds to a unique complex number $x + jy$. Next, we apply two activation functions, C-ReLU and H-CReLU, to the initial set of points. These functions transform the input points into output points, which we plot on the complex plane. Our H-CReLU is obtained from CSR with a linear layer classifier trained on the full-sized CIFAR 10 dataset. More results can be found in the supplementary. From the figure, We've noticed that the CReLU function can cause certain input points to collapse into each other, destroying the phase information other than the first quadrant. In comparison, H-CReLU avoids information loss, where the outputs have better coverage of the complex plane.

Phase information for MRI patch classification

In Section 7.6, we introduce a novel dataset for classifying complex-valued MRI patches. The goal is to train a classifier that can accurately identify the anatomical orientation of the complex-valued input patches. This section aims to analyze whether the inherent phase information can enhance patch classification performance.

Table 7.3 compares the performance of the CSR+LL and CSR+CDS models trained with magnitude-only MRI patches against the model trained using complex-valued input data.

Due to physics and acquisition factors, complex-valued MRI images can have random phase offsets. Therefore, two phase maps, ϕ_1 and $\phi_2 = \phi_1 + \alpha$ (where α is a constant phase ranging from $[0, 2\pi)$), provide the same phase information. To reduce the phase sensitivity, during training, we augment the phase of the input patches on the fly by multiplying each

Method	Phase	Aug.	MRI Patch	
			100 samples	500
CSR+ LL†	✗	-	75.97±0.63	89.33±0.40
	✓	✗	74.22±0.57	91.73±0.33
	✓	✓	**78.53**±0.35	**92.25** ±0.30
CSR+ CDS	✗	-	85.60±0.73	95.60±0.19
	✓	✗	84.80±1.06	99.18±0.15
	✓	✓	**88.95**±0.78	**99.32**±0.11

†: CDS type-I [105]; Aug.: random and constant phase augmentation

Table 7.3: **Classification accuracy of complex-valued MRI patch dataset with and without phase information.** To mitigate the sensitivity to phase, we also incorporate a model trained on complex-valued inputs with phase augmentation, resulting in the highest accuracy and demonstrating the importance of phase information in MRI patch classification.

patch with random constant phase $e^{j\theta}$, where θ is uniformly distributed between $[0, 2\pi)$. We include the results in Table 7.3.

The results indicate that when trained on only 100 samples, models trained on complex-valued inputs without phase augmentation (second row) show slightly lower accuracy than those trained on magnitude-only patches (first row) due to phase sensitivity. However, incorporating phase augmentation can considerably enhance the performance of models trained on complex-valued inputs (third row), resulting in the highest accuracy and highlighting the significance of phase information.

When moving to a larger training set of 500 samples, increased data diversity inherently mitigates the phase sensitivity. In this scenario, models trained on complex-valued inputs without phase augmentation already outperform those trained on magnitude-only images by a large margin, further highlighting the advantages of incorporating phase information in patch classification.

Robustness to geometric deformations

Scattering transforms [12, 31] have been shown to be stable to small deformations as a built-in feature. This section explores the stability of CSR to geometric deformations and compares it with S [12] and LS [31]. Following [31], we include rotation, scaling, and shearing as our deformations.

To study the robustness of deformations, we apply deformations of strength l to a given image I, resulting in a deformed image denoted as $\tilde{I}(l)$. When evaluating rotation and

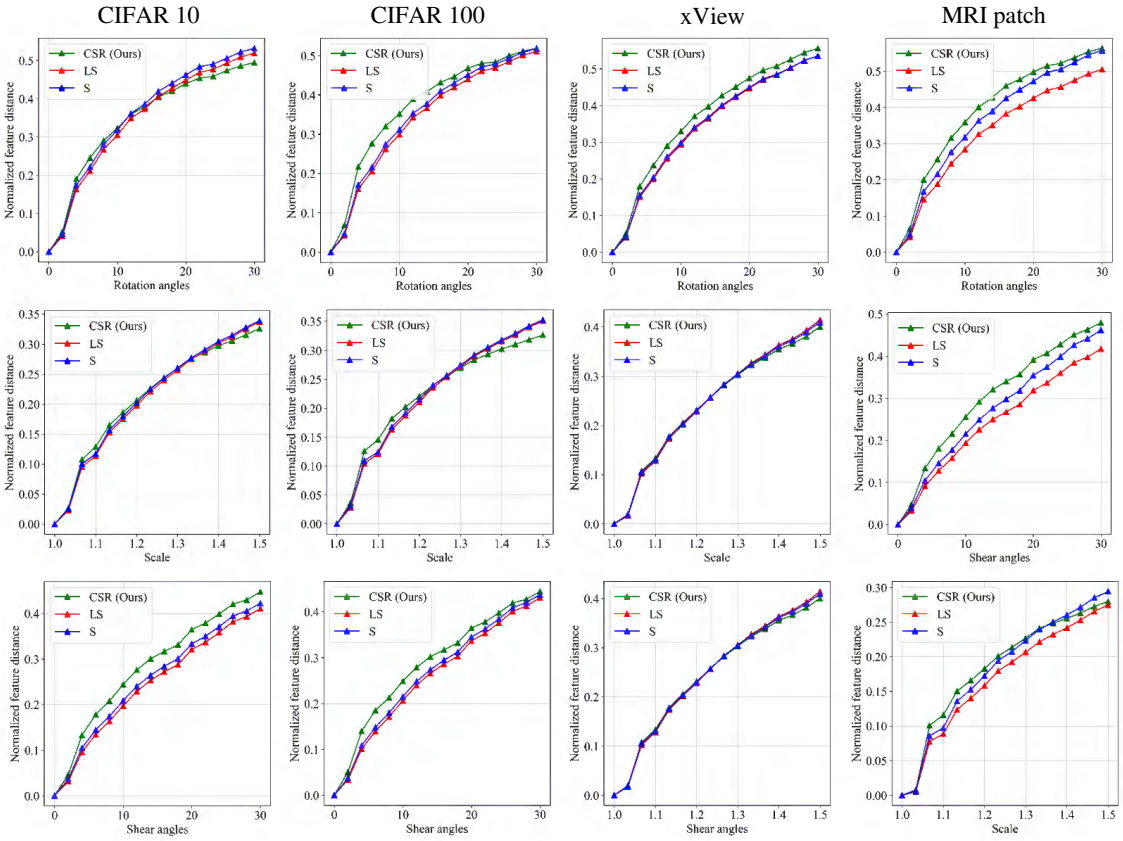


Figure 7.6: **Normalized distance comparisons of different deformations (*i.e.*,rotating, scaling, shearing).** We compare the deformation stability of CSR with LS [31] and S [12] evaluated on various datasets. From left to the right, we evaluate CIFAR 10, CIFAR 100, xView, and MRI patch classification. The plots illustrate the change in normalized distances with respect to deformation levels. CSR roughly matches the deformation stabilities of LS and S.

shearing, the deformation angle l ranges from $[0, 30°)$, while the scale parameter l ranges from $[1, 1.5]$ when assessing scaling. All the deformations are implemented using torchvision.

We then compute the normalized Euclidean distance $D(l)$ between CSRs of I and $\tilde{I}(l)$ as a function of l:

$$D(u) = \frac{\|\mathrm{CSR}(I) - \mathrm{CSR}(\tilde{I}(l))\|_2}{\|\mathrm{CSR}(I)\|_2}. \tag{7.6}$$

We compute the average $D(l)$ across the entire dataset and plot it against l. Figure 7.6 depicts the results obtained from all four datasets using CSR+LL: CIFAR 10, CIFAR 100, xView, and MRI patch. We include the results from S and LS for comparison.

Our observation suggests that CSR is generally on par with LS and S in terms of defor-

mation stability across datasets. More specifically, CSR exhibits slightly better deformation robustness (lower distance values) in CIFAR 10 rotation, CIFAR 10 scaling, CIFAR 100 scaling, xView scaling, and shearing, while it shows slightly worse stability in some other scenarios.

CSR for few-shot learning

Few-shot learning is a popular machine learning sub-field that aims to train models capable of recognizing and classifying new objects or categories with only a few examples or instances.

In this section, we evaluate the effectiveness of CSR for few-shot learning on the CIFAR 10 dataset and compare its performance with S [12], LS [31] and CDS.

We begin by training CSR and other models on images from 5 subclasses in the CIFAR 10 dataset, which include 25,000 training images from the following classes: airplane, automobile, bird, cat, and deer. Next, we fine-tune the models on few-shot images (5 and 10 samples from each class) from the remaining 5 classes: dog, frog, horse, ship, and truck. Finally, we evaluate the classification accuracy of the fine-tuned models on images of the second set of 5 classes (2,500 images).

Table 7.4 summarizes the results for both the 5 samples and 10 samples experiments. It can be observed that CSR outperforms its real-valued counterparts and CDS. Moreover, CSR+CDS outperforms CDS by 8.78% and 8.13% in the 5 and 10 samples few-shot learning experiments, respectively, which highlights the potential of CSR for few-shot learning.

Ablation studies

We evaluate the contributions of learnable filters and our proposed H-CReLU in CSR through ablation studies. We report the results on CIFAR 10 (full size) and MRI patch classification (100 samples) for CSR + LL and CSR + CDS type-I. More results can be found in the supplementary.

Our experimental setup for CSR (Table 7.5) includes the following configurations: 1) fixed filters and complex modulus as activation function (−,−); 2) learnable filters and complex modulus (✓,−); 3) fixed filters and H-CReLU (−,✓); and 4) learnable filters and H-CReLU (✓,✓).

Table 7.5 demonstrates that integrating learnable filters and H-CReLU into CSR results in enhanced performance with minimal parameter increase. Specifically, compared to the baseline model, the learnable filters module only adds 32 extra parameters, while the H-CReLU module introduces an additional 64 parameters. Furthermore, the combination of both modules can further enhance the performance, resulting in an accuracy gain of 8.28%/7.01% (LL/CDS) for CIFAR 10 and 16.06%/22.25% (LL/CDS) for MRI patch classification.

Tabel 7.6 further compare H-CReLU with other complex-valued activation functions: 1) Complex modulus (one used for real-valued SNs); 2) Complex ReLU (CReLU); 3) learnable Generalized Tangent ReLU proposed in [105]. To ensure fairness, we keep the learnable

Method	CIFAR 10	
-	5 samples	10 samples
Scattering + Linear layers		
S [12] + LL	52.70±3.01	64.56±1.27
LS [31] + LL	53.52±3.33	66.16±1.32
CSR + LL †	**55.62**±2.94	**68.74** ±1.48
Scattering + CIFARNet		
S [12] + CIFARNet	58.49±2.61	66.60±2.70
LS [31] + CIFARNet	58.95±3.30	65.45±1.90
CSR + CDS type-I †	**60.12** ±2.53	**68.04**±1.38
CDS type-I		
CDS type-I [105]	51.34±3.22	59.91±2.02

† ours ; S: Scattering; LS: Learnable Scattering; CSR: Complex-valued Scattering Representations (ours).

Table 7.4: **Few-shot classification accuracy on subset of CIFAR 10.** We pre-train the models on 25,000 images from 5 subclasses in the CIFAR 10 dataset. Next, we fine-tune the models on few-shot images from the remaining 5 classes and evaluate on the testing images (2,500) of the second set of 5 classes. In both training setups, our CSR outperforms CDS and other real-valued scattering counterparts.

filter module for all the experiments. Our findings suggest that CReLU is not as effective as modulus in producing higher accuracy due to the phase information loss. GTReLU slightly outperforms modulus in certain experiments. In comparison, H-CReLU yields the most significant improvement compared to other methods, demonstrating its superiority as a non-linear activation module for CSR.

Dimensionality of H-CReLU

Thus far, we have set $N_h = 16$ for H-CReLU in all our experiments. In this section, we explore the impact of N_h on the classification results, with a particular focus on the CIFAR 10 dataset. We conduct experiments with both non-learned and learned H-CReLU with $N_h = 2, 4, 8, 16, 32, 64$.

Table 7.7 presents the classification results of CSR+LL and CSR+CDS on the CIFAR 10 dataset. We note that, for learned H-CReLU, $N_h = 16$ yields the highest accuracy. Remarkably, the results of $N_h = 2$ are only 1.41% lower than $N_h = 16$ (CSR+LL), and still surpass other real-valued scattering counterparts and complex-valued networks discussed

Method	L. F.	H-C.	CIFAR 10	MRI Patch
CSR + LL	-	-	66.02	58.16±0.44
	✓	-	71.23 ↑5.21	68.85±0.68 ↑10.69
	-	✓	70.35 ↑4.33	70.07±0.46 ↑11.91
	✓	✓	**74.30** ↑**8.28**	**74.22**±0.57 ↑**16.06**
CSR + CDS†	-	-	74.51	62.55±0.74
	✓	-	77.60 ↑2.89	74.03±0.56 ↑11.48
	-	✓	79.02 ↑4.51	78.40±0.91 ↑15.85
	✓	✓	**81.52** ↑**7.01**	**84.80**±1.06 ↑**22.25**

†: CDS type-I [105]; L.F.: Learnable filters; H-C.: High-dimensional C-ReLU (H-CReLU)

Table 7.5: **Ablation studies of different CSR components.** We analyze the contributions of learnable filtering and H-CReLU for CSNs on CIFAR 10 and MRI patch benchmarks. **Bold** corresponds to the best results, ↑ shows the accuracy gain compared to the baseline model (fixed filters and complex modulus).

Method	Activation	CIFAR 10	MRI Patch
CSR +LL	Modulus [12]	71.23	68.85±0.68
	CReLU	70.88 ↓0.35	65.33±0.88 ↓3.52
	GTReLU [105]	71.04 ↓0.19	69.08±0.45 ↑0.24
	H-CReLU (Ours)	**74.30** ↑**3.07**	**74.22**±0.57 ↑**5.37**
CSR +CDS	Modulus [12]	77.60	74.03±0.56
	CReLU	75.08 ↓2.52	71.01±0.90 ↓3.02
	GTReLU [105]	78.24 ↑0.64	76.40±0.78 ↑2.37
	H-CReLU (Ours)	**81.52** ↑**3.92**	**84.80**±1.06 ↑**10.07**

Table 7.6: **Ablation studies of different non-linear activation functions with learnable filters.** We compare our H-CReLU with other complex-valued activation functions: complex modulus, CReLU, and learnable Generalized Tangent ReLU (GTReLU) from [105]. H-CReLU yields the best results. ↑ and ↓ indicate an increase and decrease in classification accuracy, respectively.

Method	H-CReLU	CIFAR 10					
		$N_h = 2$	4	8	16	32	64
CSR +LL	non-learned	66.62	68.25	70.54	72.39	**72.48**	70.77
	learned	72.89	73.28	73.79	**74.30**	73.75	73.40
CSR +CDS	non-learned	70.77	74.56	77.99	80.63	80.52	**80.80**
	learned	79.18	80.86	81.40	**81.52**	81.05	80.87

Table 7.7: **Ablation studies on different H-CReLU dimensionalities.** We compare the classification results of H-CReLU with varying dimensionalities (N_h), including non-learned and learned H-CReLU using CSR+LL and CSR+CDS on the CIFAR 10 dataset. **Bold** indicates best result in each row.

in § 7.6. This highlights the benefits of H-CReLU when N_h is low. We note that, when $N_h > 16$, the accuracy starts to saturate and decrease. We hypothesize that H-CReLU is more susceptible to overfitting when N_h is high.

On the other hand, for non-learned H-CReLU, we observe a significant gap in accuracy across different dimensionalities, where $N_h = 2, 4$ yields markedly poorer results. We note that the accuracy begins to saturate when $N_h \geq 16$. Specifically, for CSR+LL, $N_h = 32$ yields the highest results, while for CSR+CDS, $N_h = 64$ produces the highest accuracy.

7.7 Conclusion

In this work, we propose Complex-valued Scattering Representations (CSR) as a novel and universal complex-valued representation for a wide range of input domains, including RGB, MRI, and MSI, in the field of complex-valued deep learning. The incorporation of tunable data-specific wavelet filters and H-CReLU enables CSR to effectively capture both spatial and spatial-frequency properties of input data. By integrating CSR into complex-valued models for image classification, we have achieved significant performance gains compared to real-valued counterparts and complex-valued models without CSR, especially under limited labeled training data settings. Therefore, CSR can greatly enhance complex-valued networks on a broader range of applications. In this study, Compressive Sensing Reconstruction (CSR) is primarily applied to image classification tasks. Extending the application of CSR to image-to-image translation tasks, such as image segmentation and image reconstruction, presents promising opportunities for future research and advancements in the field.

Chapter 8

Summary and future work

In this dissertation, I showcase a collection of research projects dedicated to enhancing MRI image reconstruction's fidelity and efficiency by developing advanced deep learning techniques.

Our proposed approaches make significant strides in addressing the limitations of current deep learning methods by improving the fidelity and efficiency of image reconstruction and expanding the range of possible applications (*e.g.*, high-dimensional MR reconstruction). These advancements not only contribute to the current state-of-the-art but also create opportunities for further exploration and innovation in the field, opening up new directions for future research and clinical applications. In this concluding chapter, I summarize the contributions and outline some promising directions for future research.

8.1 Summary of contributions

The contributions of this dissertation are summarized in this section.

Direct contrast synthesis from MRF

I introduced a supervised learning-based method (N-DCSNet) that synthesizes multiple contrast-weighted images (T1w, T2w, and FLAIR) from a single, short MRF scan, significantly reducing examination time while preserving image quality. By training a network to generate contrast-weighted images directly from MRF data, our method eliminates the need for model-based simulations, thus avoiding reconstruction errors caused by simulation inaccuracies.

In-vivo experiments demonstrate that N-DCSNet produces high-fidelity contrast-weighted images with sharp contrasts and minimal artifacts (in-flow and spiral off-resonance artifacts), outperforming simulation-based contrast synthesis and PixelNet both visually and metrically. Furthermore, our proposed method inherently mitigates some off-resonance artifacts

within MRF data, resulting in high-quality contrast-weighted images with minimal residual artifacts.

Unsupervised Feature Loss for DL-based MRI reconstruction

I introduced the Unsupervised Feature Loss (UFLoss), a novel patch-based unsupervised learning-based feature loss, designed to address the limitations of existing hand-crafted loss functions, particularly their inability to capture high-level perceptual information and to preserve high-dimensional perceptual similarities.

Our UFLoss can be seamlessly integrated into the training of any existing deep learning-based reconstruction frameworks without necessitating modifications to the model architecture. UFLoss is founded on an unsupervised pre-trained feature mapping network that operates independently of external supervision.

By incorporating our proposed UFLoss, we successfully reconstruct high-fidelity images characterized by sharper edges, more accurate contrasts, and overall enhanced image quality.

Memory-efficient learning for high-dimensional MRI reconstruction

In order to address the memory challenges associated with unrolled reconstruction, I utilized the Memory-Efficient Learning (MEL) framework, which significantly reduces memory consumption during backpropagation in the training process. Our approach facilitates the training of unrolled deep learning-based reconstruction for 1) large-scale 3D MRI and 2) 2D+time cardiac cine MRI with an extensive number of unrolls, thereby enhancing the efficiency and capabilities of MRI reconstruction techniques. Our *in-vivo* experiments indicate that by exploiting high-dimensional data redundancy, we can achieve better quantitative metrics and improved image quality with sharper edges for both 3D MRI and cardiac cine MRI.

Rigorous uncertainty estimation for MRI reconstruction

To guarantee the reliability and validity of deep learning-based reconstructions, I introduce a straightforward and rigorous uncertainty estimation framework that can be effortlessly integrated into existing reconstruction networks without requiring any modification or retraining. Notably, our technique, unlike previous work, provides a rigorous finite-sample statistical guarantee on the predicted confidence intervals, ensuring that the confidence intervals contain at least $(1-\gamma)$ (*e.g.*, 95%) of the ground truth pixel values.

In-vivo experimental results reveal a strong correlation between our uncertainty estimation outcomes and the absolute residual error. Moreover, our approach refines heuristic uncertainty estimation, quantitatively guaranteeing the desired confidence levels. This highlights the potential of our framework in enhancing the reliability and accuracy of deep learning-based reconstructions for medical imaging applications.

Complex-valued Scattering Representations (CSR)

Deep learning for MRI applications consistently faces challenges arising from complex-valued data inputs and limited training data availability. To address these challenges, I introduced Complex-valued Scattering Representations (CSR) as a novel and universal complex-valued representation for a broad range of input domains.

CSR employs tunable data-specific wavelet filters and H-CReLU, showcasing significant performance improvements on image classification tasks compared to conventional CNNs, particularly in situations with limited labeled training data. Thus, CSR exhibits immense potential for representation learning in MRI applications.

8.2 Suggestions for future works

In this dissertation, we have showcased various advancements aimed at improving the fidelity and efficiency of MRI reconstruction. Despite our progress, numerous challenges and opportunities still lie ahead. The topics of this dissertation pave the way for exciting future research endeavors. Here are some promising ideas and projects for future research.

Direct contrast synthesis trained on large-scale diagnostic imaging datasets

In Chapter 3, we introduce a supervised learning approach for high-fidelity direct contrast synthesis from MRF. Nevertheless, in our work, we only trained our network on a relatively small dataset acquired from healthy volunteers.

This could lead to generalization challenges when handling images with pathologies. As such, a direct approach to address this issue would involve expanding the training dataset and incorporating more diagnostic images containing various pathologies. Additionally, comprehensive and in-depth clinical assessments would be highly beneficial for DCS and its integration into clinical practice.

Perceptual metric for MR image assessment

Quantitative metrics hold a crucial role in various aspects of image analysis, including image assessment, algorithm evaluation, and the adoption of new techniques in clinical settings.

Nonetheless, the most prevalent metrics for comparing two MR images remain pixel-wise (ℓ_1, ℓ_2 loss) or those based on local statistics (SSIM loss). While these metrics are straightforward, they have been demonstrated to be inadequate for capturing the high-dimensional perceptual similarities that are crucial for a more comprehensive comparison. Zhang et al. [145] first demonstrated the effectiveness of using learned features as a perceptual metric for natural images. However, applying this perceptual metric to MR images is not straightforward due to their complexity and high dimensionality. To the best of my knowledge,

there have been limited attempts at designing perceptual metrics for MR image assessment, especially for high-dimensional MRI.

In Chapter 4, I describe an unsupervised feature loss designed to capture high-dimensional perceptual similarities without requiring any supervision. Therefore, we hypothesize that UFLoss can be also used as an effective perceptual metric for MRI, especially for high-dimensional MRI. Thorough analysis, fair comparisons, and well-designed radiologist studies are required to validate the hypothesis. One possible user study is Two-alternative forced choice (2AFC).

Uncertainty estimation in feature domains (latent space)

Chapter 6 presents a novel uncertainty estimation framework for MRI reconstruction. Our proposed uncertainty map resides in the image domain and effectively provides pixel-level confidence intervals.

One limitation of the uncertainty map is that it does not provide high-level information, such as the accuracy of the recovered pathologies. This makes it difficult to assess the overall reconstruction quality in terms of clinically relevant features.

To address these challenges, we propose two directions: 1) Uncertainty estimation in the feature domain or latent space, which can be directly translated into uncertainty in downstream tasks, such as pathology classification or anatomical segmentation. 2) Designing pipelines to propagate pixel-valued uncertainty to downstream tasks. As a more concrete example, given a pre-trained pathology detection network, one promising avenue is to propagate the pixel-valued uncertainty to the uncertainty for pathology detection, thereby providing more informed and reliable assessments of the reconstructed images in terms of clinically relevant outcomes.

CSR for image-to-image translation tasks

In Chapter 7, I introduce CSR, which exhibits exceptional performance in image classification, particularly under limited data conditions. One straightforward and promising direction is to apply CSR to image-to-image translation tasks (*e.g.*, image segmentation, and image reconstruction).

A widely utilized image segmentation framework [17] involves using a pre-trained backbone to extract feature representations, which are then fed into an upsampling module. One potential direction is to replace the pre-trained backbone with CSR, while maintaining the upsampling module. Additionally, investigating the application of CSR in MR image reconstruction tasks, particularly high-dimensional MRI with limited data, presents a promising research avenue.

Bibliography

[1] Abien Fred Agarap. "Deep learning using rectified linear units (relu)". In: *arXiv preprint arXiv:1803.08375* (2018).

[2] Hemant K Aggarwal, Merry P Mani, and Mathews Jacob. "MoDL: Model-based deep learning architecture for inverse problems". In: *IEEE transactions on medical imaging* 38.2 (2018), pp. 394–405.

[3] Laith Alzubaidi et al. "Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions". In: *Journal of big Data* 8 (2021), pp. 1–74.

[4] Joakim Andén and Stéphane Mallat. "Deep scattering spectrum". In: *IEEE Transactions on Signal Processing* 62.16 (2014), pp. 4114–4128.

[5] Mathieu Andreux et al. "Kymatio: Scattering transforms in python". In: *The Journal of Machine Learning Research* 21.1 (2020), pp. 2256–2261.

[6] Anastasios N Angelopoulos et al. "Image-to-image regression with distribution-free uncertainty quantification and applications in imaging". In: *International Conference on Machine Learning*. PMLR. 2022, pp. 717–730.

[7] Joshua Bassey, Lijun Qian, and Xianfang Li. *A Survey of Complex-Valued Neural Networks*. 2021. DOI: 10.48550/ARXIV.2101.12249. URL: https://arxiv.org/abs/2101.12249.

[8] Stephen Bates et al. "Distribution-free, risk-controlling prediction sets". In: *Journal of the ACM (JACM)* 68.6 (2021), pp. 1–34.

[9] Amir Beck and Marc Teboulle. "A fast iterative shrinkage-thresholding algorithm for linear inverse problems". In: *SIAM journal on imaging sciences* 2.1 (2009), pp. 183–202.

[10] Ida Blystad et al. "Synthetic MRI of the brain in a clinical setting". In: *Acta radiologica* 53.10 (2012), pp. 1158–1163.

[11] Stephen Boyd et al. "Distributed optimization and statistical learning via the alternating direction method of multipliers". In: *Foundations and Trends® in Machine learning* 3.1 (2011), pp. 1–122.

[12] Joan Bruna and Stéphane Mallat. "Invariant scattering convolution networks". In: *IEEE transactions on pattern analysis and machine intelligence* 35.8 (2013), pp. 1872–1886.

[13] Guido Buonincontri and Stephen J Sawiak. "MR fingerprinting with simultaneous B1 estimation". In: *Magnetic resonance in medicine* 76.4 (2016), pp. 1127–1135.

[14] Tianle Cao et al. "Three-dimensional simultaneous brain mapping of T1, T2, and magnetic susceptibility with MR Multitasking". In: *Magnetic Resonance in Medicine* 87.3 (2022), pp. 1375–1389.

[15] Rudrasis Chakraborty, Yifei Xing, and X Yu Stella. "SurReal: Complex-valued learning as principled transformations on a scaling and rotation manifold". In: *IEEE Transactions on Neural Networks and Learning Systems* 33.3 (2020), pp. 940–951.

[16] Feiyu Chen et al. "Variable-density single-shot fast spin-echo MRI with deep learning reconstruction by using variational networks". In: *Radiology* 289.2 (2018), pp. 366–373.

[17] Liang-Chieh Chen et al. "Rethinking atrous convolution for semantic image segmentation". In: *arXiv preprint arXiv:1706.05587* (2017).

[18] Anthony G Christodoulou et al. "Magnetic resonance multitasking for motion-resolved quantitative cardiovascular imaging". In: *Nature biomedical engineering* 2.4 (2018), pp. 215–226.

[19] Özgün Çiçek et al. "3D U-Net: learning dense volumetric segmentation from sparse annotation". In: *International conference on medical image computing and computer-assisted intervention*. Springer. 2016, pp. 424–432.

[20] Joseph Paul Cohen, Margaux Luck, and Sina Honari. "Distribution matching losses can hallucinate features in medical image translation". In: *International conference on medical image computing and computer-assisted intervention*. Springer. 2018, pp. 529–536.

[21] Onat Dalmaz, Mahmut Yurt, and Tolga Çukur. "ResViT: residual vision transformers for multimodal medical image synthesis". In: *IEEE Transactions on Medical Imaging* 41.10 (2022), pp. 2598–2614.

[22] Salman UH Dar et al. "Image synthesis in multi-contrast MRI with conditional generative adversarial networks". In: *IEEE transactions on medical imaging* 38.10 (2019), pp. 2375–2388.

[23] Alfredo De Goyeneche et al. "ResoNet: Physics Informed Deep Learning based Off-Resonance Correction Trained on Synthetic Data". In: *Proceedings of the 30th Annual Meeting of ISMRM*. Vol. 555. 2022.

[24] Jia Deng et al. "Imagenet: A large-scale hierarchical image database". In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.

[25] Anagha Deshmane et al. "Partial volume mapping using magnetic resonance fingerprinting". In: *NMR in Biomedicine* 32.5 (2019), e4082.

[26] Steven Diamond et al. "Unrolled optimization with deep priors". In: *arXiv preprint arXiv:1705.08041* (2017).

[27] Alexey Dosovitskiy et al. "An image is worth 16x16 words: Transformers for image recognition at scale". In: *arXiv preprint arXiv:2010.11929* (2020).

[28] Vineet Edupuganti et al. "Uncertainty quantification in deep mri reconstruction". In: *IEEE Transactions on Medical Imaging* 40.1 (2020), pp. 239–250.

[29] Michael Eickenberg et al. "Solid harmonic wavelet scattering for predictions of molecule properties". In: *The Journal of chemical physics* 148.24 (2018), p. 241732.

[30] Sebastian Flassbeck et al. "Flow MR fingerprinting". In: *Magnetic Resonance in Medicine* 81.4 (2019), pp. 2536–2550.

[31] Shanel Gauthier et al. "Parametric scattering networks". In: *Proceedings of the IEEE and CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 5749–5758.

[32] Aidan N Gomez et al. "The reversible residual network: Backpropagation without storing activations". In: *arXiv preprint arXiv:1707.04585* (2017).

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

[34] Ian Goodfellow et al. "Generative adversarial networks". In: *Communications of the ACM* 63.11 (2020), pp. 139–144.

[35] T Granberg et al. "Clinical feasibility of synthetic MRI in multiple sclerosis: a diagnostic and volumetric validation study". In: *American Journal of Neuroradiology* 37.6 (2016), pp. 1023–1029.

[36] Mark A Griswold et al. "Generalized autocalibrating partially parallel acquisitions (GRAPPA)". In: *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* 47.6 (2002), pp. 1202–1210.

[37] Kerstin Hammernik et al. "Learning a variational network for reconstruction of accelerated MRI data". In: *Magnetic resonance in medicine* 79.6 (2018), pp. 3055–3071.

[38] Xiao Han. "MR-based synthetic CT generation using a deep convolutional neural network method". In: *Medical physics* 44.4 (2017), pp. 1408–1419.

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

[40] Martin Heusel et al. "Gans trained by a two time-scale update rule converge to a local nash equilibrium". In: *arXiv preprint arXiv:1706.08500* (2017).

[41] Matthew Hirn, Nicolas Poilvert, and Stéphane Mallat. "Quantum energy regression using scattering transforms". In: *arXiv preprint arXiv:1502.02077* (2015).

[42] Wassily Hoeffding. "Probability inequalities for sums of bounded random variables". In: *The collected works of Wassily Hoeffding* (1994), pp. 409–426.

[43] Jean JL Hsieh and Imants Svalbe. "Magnetic resonance fingerprinting: from evolution to clinical applications". In: *Journal of Medical Radiation Sciences* 67.4 (2020), pp. 333–344.

[44] Phillip Isola et al. "Image-to-image translation with conditional adversarial networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1125–1134.

[45] John I Jackson et al. "Selection of a convolution function for Fourier inversion using gridding (computerised tomography application)". In: *IEEE transactions on medical imaging* 10.3 (1991), pp. 473–478.

[46] Ajil Jalal et al. "Robust compressed sensing mri with deep generative priors". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 14938–14954.

[47] Yun Jiang et al. "MR fingerprinting using fast imaging with steady state precession (FISP) with spiral readout". In: *Magnetic resonance in medicine* 74.6 (2015), pp. 1621–1631.

[48] Joblib Development Team. "Joblib: running Python functions as pipeline jobs". In: (2020). URL: `https://joblib.readthedocs.io/`.

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

[50] Uday Kamath, John Liu, and James Whitaker. *Deep learning for NLP and speech recognition*. Vol. 84. Springer, 2019.

[51] Michael Kellman et al. "Memory-efficient learning for large-scale computational imaging". In: *IEEE Transactions on Computational Imaging* 6 (2020), pp. 1403–1414.

[52] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[53] Diederik P Kingma and Max Welling. "Auto-encoding variational bayes". In: *arXiv preprint arXiv:1312.6114* (2013).

[54] Naveen Kodali et al. "On convergence and stability of gans". In: *arXiv preprint arXiv:1705.07215* (2017).

[55] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. "Cifar-10 and cifar-100 datasets". In: *URl: https://www. cs. toronto. edu/kriz/cifar. html* 6.1 (2009), p. 1.

[56] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Communications of the ACM* 60.6 (2017), pp. 84–90.

[57] Thomas Küstner et al. "CINENet: deep learning-based 3D cardiac CINE MRI reconstruction with multi-coil complex-valued 4D spatio-temporal convolutions". In: *Scientific reports* 10.1 (2020), pp. 1–13.

[58] Darius Lam et al. "xview: Objects in context in overhead imagery". In: *arXiv preprint arXiv:1802.07856* (2018).

[59] Anders Boesen Lindbo Larsen et al. "Autoencoding beyond pixels using a learned similarity metric". In: *International conference on machine learning*. PMLR. 2016, pp. 1558–1566.

[60] Zhi-Pei Liang and Paul C Lauterbur. *Principles of magnetic resonance imaging*. SPIE Optical Engineering Press Bellingham, 2000.

[61] Fang Liu et al. "SANTIS: Sampling-augmented neural network with incoherent structure for MR image reconstruction". In: *Magnetic resonance in medicine* 82.5 (2019), pp. 1890–1904.

[62] Marc Moreno Lopez and Jugal Kalita. "Deep Learning applied to NLP". In: *arXiv preprint arXiv:1703.03091* (2017).

[63] Ilya Loshchilov and Frank Hutter. "Decoupled weight decay regularization". In: *arXiv preprint arXiv:1711.05101* (2017).

[64] Michael Lustig, David Donoho, and John M Pauly. "Sparse MRI: The application of compressed sensing for rapid MR imaging". In: *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* 58.6 (2007), pp. 1182–1195.

[65] Michael Lustig and John M Pauly. "SPIRiT: iterative self-consistent parallel imaging reconstruction from arbitrary k-space". In: *Magnetic resonance in medicine* 64.2 (2010), pp. 457–471.

[66] Dan Ma et al. "Fast 3D magnetic resonance fingerprinting for a whole-brain coverage". In: *Magnetic resonance in medicine* 79.4 (2018), pp. 2190–2197.

[67] Dan Ma et al. "Magnetic resonance fingerprinting". In: *Nature* 495.7440 (2013), pp. 187–192.

[68] Morteza Mardani et al. "Deep generative adversarial neural networks for compressive sensing MRI". In: *IEEE transactions on medical imaging* 38.1 (2018), pp. 167–179.

[69] Matteo Maspero et al. "Dose evaluation of fast synthetic-CT generation using a generative adversarial network for general pelvis MR-only radiotherapy". In: *Physics in Medicine & Biology* 63.18 (2018), p. 185001.

[70] Jason D McEwen, Christopher GR Wallis, and Augustine N Mavor-Parker. "Scattering networks on the sphere for scalable and rotationally equivariant spherical CNNs". In: *arXiv preprint arXiv:2102.02828* (2021).

[71] Mehdi Mirza and Simon Osindero. "Conditional generative adversarial nets". In: *arXiv preprint arXiv:1411.1784* (2014).

[72] Matthew J Muckley et al. "Results of the 2020 fastMRI challenge for machine learning MR image reconstruction". In: *IEEE transactions on medical imaging* 40.9 (2021), pp. 2306–2317.

[73] Dominik Narnhofer et al. "Bayesian uncertainty estimation of learned variational MRI reconstruction". In: *IEEE Transactions on Medical Imaging* 41.2 (2021), pp. 279–291.

[74] Tristan Needham. *Visual complex analysis*. Oxford University Press, 1998.

[75] Dwight G. Nishimura. *Principles of magnetic resonance imaging*. Stanford Univeristy, 1996.

[76] T Nitta. "On the critical points of the complex-valued neural network". In: *Proceedings of the 9th International Conference on Neural Information Processing, 2002. ICONIP'02.* Vol. 3. IEEE. 2002, pp. 1099–1103.

[77] Tohru Nitta. "The Computational Power of Complex-Valued Neuron". In: *Joint International Conference ICANN/ICONIP*. 2003. URL: `https://link.springer.com/content/pdf/10.1007/3-540-44989-2_118.pdf`.

[78] Tohru Nitta. "The computational power of complex-valued neuron". In: *Artificial Neural Networks and Neural Information Processing—ICANN/ICONIP 2003: Joint International Conference ICANN/ICONIP 2003 Istanbul, Turkey, June 26–29, 2003 Proceedings*. Springer. 2003, pp. 993–1000.

[79] John D O'Sullivan. "A fast sinc function gridding algorithm for Fourier inversion in computer tomography". In: *IEEE transactions on medical imaging* 4.4 (1985), pp. 200–207.

[80] Augustus Odena, Vincent Dumoulin, and Chris Olah. "Deconvolution and checkerboard artifacts". In: *Distill* 1.10 (2016), e3.

[81] Frank Ong and Michael Lustig. "SigPy: a python package for high performance iterative reconstruction". In: *Proc. ISMRM*. 2019.

[82] Frank Ong et al. "Extreme MRI: Large-scale volumetric dynamic imaging from continuous non-gated acquisitions". In: *Magnetic resonance in medicine* 84.4 (2020), pp. 1763–1780.

[83] Ricardo Otazo, Emmanuel Candes, and Daniel K Sodickson. "Low-rank plus sparse matrix decomposition for accelerated dynamic MRI with separation of background and dynamic components". In: *Magnetic resonance in medicine* 73.3 (2015), pp. 1125–1136.

[84] Yaniv Ovadia et al. "Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift". In: *arXiv preprint arXiv:1906.02530* (2019).

[85] Edouard Oyallon et al. "Scattering networks for hybrid representation learning". In: *IEEE transactions on pattern analysis and machine intelligence* 41.9 (2018), pp. 2208–2221.

[86] Adam Paszke et al. "Automatic differentiation in pytorch". In: (2017).

[87] Adam Paszke et al. "Pytorch: An imperative style, high-performance deep learning library". In: *Advances in neural information processing systems* 32 (2019).

[88] Kamlesh Pawar, Gary F Egan, and Zhaolin Chen. "Estimating Uncertainty in Deep Learning MRI Reconstruction using a Pixel Classification Image Reconstruction Framework". In: *International Society for Magnetic Resonance in Medicine & Society for MR Radiographers & Technologists Annual Meeting & Exhibition 2021*. 2021.

[89] Carlo Pierpaoli. *Quantitative brain MRI*. 2010.

[90] Klaas P Pruessmann et al. "SENSE: sensitivity encoding for fast MRI". In: *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* 42.5 (1999), pp. 952–962.

[91] Liangqiong Qu et al. "Synthesized 7T MRI from 3T MRI via deep learning in spatial and wavelet domains". In: *Medical image analysis* 62 (2020), p. 101663.

[92] Tran Minh Quan, Thanh Nguyen-Duc, and Won-Ki Jeong. "Compressed sensing MRI reconstruction using a generative adversarial network with a cyclic loss". In: *IEEE transactions on medical imaging* 37.6 (2018), pp. 1488–1497.

[93] Saiprasad Ravishankar and Yoram Bresler. "MR image reconstruction from highly undersampled k-space data by dictionary learning". In: *IEEE transactions on medical imaging* 30.5 (2010), pp. 1028–1041.

[94] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation". In: *International Conference on Medical image computing and computer-assisted intervention*. Springer. 2015, pp. 234–241.

[95] Christopher M Sandino et al. "Accelerating cardiac cine MRI using a deep learning-based ESPIRiT reconstruction". In: *Magnetic Resonance in Medicine* 85.1 (2021), pp. 152–167.

[96] Christopher M Sandino et al. "Compressed sensing: From research to clinical practice with deep neural networks: Shortening scan times for magnetic resonance imaging". In: *IEEE signal processing magazine* 37.1 (2020), pp. 117–127.

[97] Mark Sandler et al. "Mobilenetv2: Inverted residuals and linear bottlenecks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 4510–4520.

[98] Anne Marie Sawyer et al. "Creation of fully sampled MR data repository for compressed sensing of the knee". In: ().

[99] Jo Schlemper et al. "A deep cascade of convolutional neural networks for dynamic MR image reconstruction". In: *IEEE transactions on Medical Imaging* 37.2 (2017), pp. 491–503.

[100] Christoph Schuhmann et al. "Laion-5b: An open large-scale dataset for training next generation image-text models". In: *arXiv preprint arXiv:2210.08402* (2022).

[101] Jonathan Richard Shewchuk et al. *An introduction to the conjugate gradient method without the agonizing pain*. 1994.

[102] Yuting Shi et al. "Towards in vivo ground truth susceptibility for single-orientation deep learning QSM: A multi-orientation gradient-echo MRI dataset". In: *NeuroImage* 261 (2022), p. 119522.

[103] Efrat Shimron et al. "Implicit data crimes: Machine learning bias arising from misuse of public data". In: *Proceedings of the National Academy of Sciences* 119.13 (2022), e2117203119.

[104] Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556* (2014).

[105] Utkarsh Singhal, Yifei Xing, and Stella X Yu. "Co-domain symmetry for complex-valued deep learning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 681–690.

[106] Utkarsh Singhal et al. "Multi-Spectral Image Classification with Ultra-Lean Complex-Valued Models". In: *arXiv preprint arXiv:2211.11797* (2022).

[107] Daniel K Sodickson and Warren J Manning. "Simultaneous acquisition of spatial harmonics (SMASH): fast imaging with radiofrequency coil arrays". In: *Magnetic resonance in medicine* 38.4 (1997), pp. 591–603.

[108] Yusuke Sugawara, Sayaka Shiota, and Hitoshi Kiya. "Super-resolution using convolutional neural networks without any checkerboard artifacts". In: *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE. 2018, pp. 66–70.

[109] Johan AK Suykens. "Nonlinear modelling and support vector machines". In: *IMTC 2001. proceedings of the 18th IEEE instrumentation and measurement technology conference. Rediscovering measurement in the age of informatics (Cat. No. 01CH 37188)*. Vol. 1. IEEE. 2001, pp. 287–294.

[110] Jonathan I Tamir, Stella X Yu, and Michael Lustig. "Unsupervised Deep Basis Pursuit: Learning inverse problems without ground-truth data". In: *arXiv:1910.13110* (2019).

[111] Jonathan I Tamir et al. "T2 shuffling: Sharp, multicontrast, volumetric fast spin-echo imaging". In: *Magnetic resonance in medicine* 77.1 (2017), pp. 180–195.

[112] Jonathan I Tamir et al. "Targeted rapid knee MRI exam using T2 shuffling". In: *Journal of Magnetic Resonance Imaging* 49.7 (2019), e195–e204.

[113] Lawrence N Tanenbaum et al. "Synthetic MRI for clinical neuroimaging: results of the Magnetic Resonance Image Compilation (MAGiC) prospective, multicenter, multireader trial". In: *American Journal of Neuroradiology* 38.6 (2017), pp. 1103–1110.

[114] Chiheb Trabelsi et al. "Deep complex networks". In: *arXiv preprint arXiv:1705.09792* (2017).

[115] Martin Uecker et al. "Berkeley advanced reconstruction toolbox". In: *Proc. Intl. Soc. Mag. Reson. Med.* Vol. 23. 2486. 2015.

[116] Martin Uecker et al. "ESPIRiT—an eigenvalue approach to autocalibrating parallel MRI: where SENSE meets GRAPPA". In: *Magnetic resonance in medicine* 71.3 (2014), pp. 990–1001.

[117] Martin Uecker et al. "Image reconstruction by regularized nonlinear inversion—joint estimation of coil sensitivities and image content". In: *Magnetic Resonance in Medicine* 60.3 (2008), pp. 674–682.

[118] MI Vargas, J Boto, and BM Delatre. "Synthetic MR imaging sequence in daily clinical practice". In: *AJNR: American Journal of Neuroradiology* 37.10 (2016), E68.

[119] Patrick Virtue, X Yu Stella, and Michael Lustig. "Better than real: Complex-valued neural nets for MRI fingerprinting". In: *2017 IEEE international conference on image processing (ICIP)*. IEEE. 2017, pp. 3953–3957.

[120] Patrick Virtue et al. "Direct Contrast Synthesis for Magnetic Resonance Fingerprinting". In: *Proc. Intl. Soc. Mag. Reson. Med.* 2018.

[121] Athanasios Voulodimos et al. "Deep learning for computer vision: A brief review". In: *Computational intelligence and neuroscience* 2018 (2018).

[122] Fuyixue Wang et al. "Echo planar time-resolved imaging (EPTI)". In: *Magnetic resonance in medicine* 81.6 (2019), pp. 3599–3615.

[123] Guanhua Wang et al. "Synthesize high-quality multi-contrast magnetic resonance imaging from multi-echo acquisition using multi-task deep generative model". In: *IEEE transactions on medical imaging* 39.10 (2020), pp. 3089–3099.

[124] Ke Wang et al. "High fidelity direct-contrast synthesis from magnetic resonance fingerprinting in diagnostic imaging". In: *Proceedings of the 28th Annual Meeting of ISMRM*. Vol. 867. 2020.

[125] Ke Wang et al. "High-fidelity Direct Contrast Synthesis from Magnetic Resonance Fingerprinting". In: *arXiv preprint arXiv:2212.10817* (2022).

[126] Ke Wang et al. "High-Fidelity Reconstruction with Instance-wise Discriminative Feature Matching Loss". In: *Proc. Intl. Soc. Mag. Reson. Med* (2020).

[127] Ke Wang et al. "Memory-efficient Learning for High-Dimensional MRI Reconstruction". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VI 24*. Springer. 2021, pp. 461–470.

[128] Ke Wang et al. "OUTCOMES: Rapid Under-sampling Optimization achieves up to 50% improvements in reconstruction accuracy for multi-contrast MRI sequences". In: *arXiv preprint arXiv:2103.04566* (2021).

[129] Ke Wang et al. "Rigorous Uncertainty Estimation for MRI Reconstruction". In: *Proceedings of the Proceedings of the 30th Annual Meeting of ISMRM*. Vol. 749. 2022.

[130] Zhou Wang et al. "Image quality assessment: from error visibility to structural similar ity". In: *IEEE transactions on image processing* 13.4 (2004), pp. 600–612.

[131] JBM Warntjes et al. "Rapid magnetic resonance quantification on the brain: optimization for clinical usage". In: *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* 60.2 (2008), pp. 320–329.

[132] Ian Waudby-Smith and Aaditya Ramdas. "Estimating means of bounded random variables by betting". In: *arXiv preprint arXiv:2010.09686* (2020).

[133] Matthias Weigel. "Extended phase graphs: dephasing, RF pulses, and echoes-pure and simple". In: *Journal of Magnetic Resonance Imaging* 41.2 (2015), pp. 266–295.

[134] Jelmer M Wolterink et al. "Deep MR to CT synthesis using unpaired data". In: *International workshop on simulation and synthesis in medical imaging*. Springer. 2017, pp. 14–23.

[135] Zhirong Wu et al. "Unsupervised feature learning via non-parametric instance discrimination". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 3733–3742.

[136] Xin Yi, Ekta Walia, and Paul Babyn. "Generative adversarial network in medical imaging: A review". In: *Medical image analysis* (2019), p. 101552.

[137] Leslie Ying and Jinhua Sheng. "Joint image reconstruction and sensitivity estimation in SENSE (JSENSE)". In: *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* 57.6 (2007), pp. 1196–1202.

[138] Biting Yu et al. "Ea-GANs: edge-aware generative adversarial networks for cross-modality MR image synthesis". In: *IEEE transactions on medical imaging* 38.7 (2019), pp. 1750–1762.

[139] Yong Yu et al. "A review of recurrent neural networks: LSTM cells and network architectures". In: *Neural computation* 31.7 (2019), pp. 1235–1270.

[140] Sergey Zagoruyko and Nikos Komodakis. "Wide residual networks". In: *arXiv preprint arXiv:1605.07146* (2016).

[141] Jure Zbontar et al. "fastMRI: An open dataset and benchmarks for accelerated MRI". In: *arXiv preprint arXiv:1811.08839* (2018).

[142] David Y Zeng et al. "Deep residual network for off-resonance artifact correction with application to pediatric body MRA with 3D cones". In: *Magnetic resonance in medicine* 82.4 (2019), pp. 1398–1411.

[143] Kevin Zhang et al. "Memory-efficient learning for unrolled 3D MRI reconstructions". In: *ISMRM Workshop on Data Sampling and Image Reconstruction*. 2020.

[144] Richard Zhang, Phillip Isola, and Alexei A Efros. "Colorful image colorization". In: *European conference on computer vision*. Springer. 2016, pp. 649–666.

[145] Richard Zhang et al. "The unreasonable effectiveness of deep features as a perceptual metric". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 586–595.

[146] Tao Zhang et al. "Coil compression for accelerated imaging with Cartesian sampling". In: *Magnetic resonance in medicine* 69.2 (2013), pp. 571–582.

[147] Zhimian Zhang et al. "Complex-valued convolutional neural network and its application in polarimetric SAR image classification". In: *IEEE Transactions on Geoscience and Remote Sensing* 55.12 (2017), pp. 7177–7188.

[148] Jun-Yan Zhu et al. "Unpaired Image-To-Image Translation Using Cycle-Consistent Adversarial Networks". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. Oct. 2017.