



THE UNIVERSITY *of* EDINBURGH

This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e. g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

- This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.
- A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.
- This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.
- The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.
- When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.

Compressive MRI with Deep Convolutional and Attentive Models

Jingshuai Liu



Doctor of Philosophy

THE UNIVERSITY OF EDINBURGH

2023

Abstract

Since its advent in the last century, Magnetic Resonance Imaging (MRI) has demonstrated a significant impact on modern medicine and spectroscopy and witnessed widespread use in medical imaging and clinical practice, owing to the flexibility and excellent ability in viewing anatomical structures. Although it provides a non-invasive and ionizing radiation-free tool to create images of the anatomy of the human body being inspected, the long data acquisition process hinders its growth and development in time-critical applications. To shorten the scanning time and reduce the discomfort of patients, the sampling process can be accelerated by leaving out an amount of sampling steps and performing image reconstruction from a subset of measurements. However, the images created with under-sampled signals can suffer from strong aliasing artifacts which unfavorably affect the quality of diagnosis and treatment. Compressed sensing (CS) methods were introduced to alleviate the aliasing artifacts by reconstructing an image from the observed measurements via model-based optimization algorithms. Despite achieved success, the sparsity prior assumed by CS methods can be difficult to hold in real-world practice and challenging to capture complex anatomical structures. The iterative optimization algorithms are often computationally expensive and time-consuming, against the speed demand of modern MRI. Those factors limit the quality of reconstructed images and put restrictions on the achievable acceleration rates.

This thesis mainly focuses on developing deep learning-based methods, specifically using modern over-parametrized models, for MRI reconstruction, by leveraging the powerful learning ability and representation capacity of such models. Firstly, we introduce an attentive selection generative adversarial network to achieve fine-grained reconstruction by performing large-field contextual information integration and attention selection mechanism. To incorporate domain-specific knowledge into the reconstruction procedure, an optimization-inspired deep cascaded framework is proposed with a novel deep data consistency block to leverage

domain-specific knowledge and an adaptive spatial attention selection module to capture the correlations among high-resolution features, aiming to enhance the quality of recovered images. To efficiently utilize the contextual information hidden in the spatial dimensions, a novel region-guided channel-wise attention network is introduced to incorporate the spatial semantics into a channel-based attention mechanism, demonstrating a light-weight and flexible design to attain improved reconstruction performance.

Secondly, a coil-agnostic reconstruction framework is introduced to solve the unknown forward process problem in parallel MRI reconstruction. To avoid the estimation of sensitivity maps, a novel data aggregation consistency block is proposed to approximately perform the data consistency enforcement without resorting to coil sensitivity information. A locality-aware spatial attention module is devised and embedded into the reconstruction pipeline to enhance the model performance by capturing spatial contextual information via data-adaptive kernel prediction. It is demonstrated by experiments that the proposed coil-agnostic method is robust and resilient to different machine configurations and outperforms other sensitivity estimation-based methods.

Finally, the research work focusing on dynamic MRI reconstruction is presented. We introduce an optimization-inspired deep cascaded framework to recover a sequence of MRI images, which utilizes a novel mask-guided motion feature incorporation method to explicitly extract and incorporate the motion information into the reconstruction iterations, showing to better preserve the dynamic content. A spatio-temporal Fourier neural block is proposed and embedded into the network to improve the model performance by efficiently retrieving useful information in both spatial and temporal domains. It is demonstrated that the devised framework surpasses other competing methods and can generalize well on other reconstruction models and unseen data, validating its transferability and generalization capacity.

Lay Summary

Magnetic resonance imaging has revolutionized the field of medical imaging and radiology since its advent in the last century, by providing a radiation-free and non-invasive diagnostic tool to image the structures of human body and demonstrating its applications to a wide range of patients. However, MRI utilizes a strong magnetic field to deflect protons contained in molecules of water in biologic tissues and collects the signals emitted from the nuclei of the target to create images when the radio wave pulse is removed, which suffers from a long acquisition procedure and hinders its development in time-critical applications. It is common practice to accelerate the MRI data acquisition process by sampling a subset of signal measurements and imaging the anatomy with the collected data, which produces aliasing artifacts in the generated images due to data under-sampling and requires reconstruction methods to preserve the image quality.

To achieve MRI reconstruction with under-sampled measurement observations, we propose to leverage the learning and representation capacity of deep neural network models for accelerated MRI reconstruction and demonstrate the effectiveness of the proposed method in attaining high-quality images recovered from highly under-sampled signals. For parallel MR imaging, a coil-agnostic reconstruction framework is proposed to address the forward pass problem of multi-coil MRI reconstruction, which is shown to be robust and invulnerable to machine coil configurations while producing images with rich structures and textural details in accelerated acquisition diets. Dynamic MRI reconstruction imposes the demand for high resolution in both spatial and temporal domains, which amounts to recovering a sequence of MR images featuring the dynamic content of the underlying anatomy. To effectively exploit the redundancies and correlations among MRI signal sequences to improve the reconstruction performance, we propose a deep cascaded network framework to explicitly leverage the dynamic information contained in under-sampled measurement frames taken at consecutive time steps.

By experiments, it is shown that the devised model is able to preserve the dynamic content at aggressive acquisition acceleration rates. The generalization capacity and transferability of the proposed framework is substantiated, which potentially facilitates the application and deployment of our method in real-world scenarios, as a means to provide high-quality image creation of the human body being inspected under accelerated data acquisition configurations.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Dr Mehrdad Yaghoobi, for his continuous support and patience during my Ph.D. study. Additionally, I would like to express gratitude to Dr Chen Qin for her kind help and advice. Their immense knowledge and rich experience have encouraged me over the course of my academic research. Finally, my appreciation goes out to my family and friends for their tremendous understanding, encouragement, and support all through my studies.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

Jingshuai Liu

Contents

Abstract	ii
Lay Summary	iv
Acknowledgements	vi
Declaration	vii
Figures and Tables	xii
Acronyms and Abbreviations	xxi
Nomenclature	xxiv
1 Introduction	1
1.1 Research Problem Overview	1
1.1.1 Magnetic Resonance Imaging Introduction	1
1.1.2 Data Acquisition Acceleration and MRI Reconstruction	2
1.2 Main Contributions	3
1.2.1 Attentive Neural Network Models for MRI Reconstruction	4
1.2.2 Coil-Agnostic Parallel MRI Reconstruction Networks	5
1.2.3 Dynamic MRI Reconstruction Models	6
1.3 Thesis Overview	7
1.4 Publications	9
2 Background and Literature Review	11
2.1 Introduction	11
2.2 Deep Neural Network Models	12

CONTENTS	ix
2.2.1 Convolutional Neural Networks	12
2.2.2 Propagation-Based Neural Networks	15
2.2.3 Generative Adversarial Networks	17
2.2.4 Attention-Based Models	20
2.3 Compressive Sensing Methods for MRI Reconstruction	23
2.4 Deep Learning-Based Methods for MRI Reconstruction	29
2.4.1 End-to-End DL Models	29
2.4.2 Model-Based DL Frameworks	30
2.4.3 Plug-and-Play Methods	38
2.4.4 k-Space Prior Based DL Models	38
2.4.5 DL Methods for Multi-Coil MRI Reconstruction	40
2.4.6 DL Methods for Dynamic MRI Reconstruction	46
2.5 Conclusion	51
3 High-Fidelity MRI Reconstruction Using Deep Neural Networks	52
3.1 Introduction	52
3.2 Methodology	55
3.2.1 Problem Formulation	55
3.2.2 Attentive Selection Generative Adversarial Network (ASGAN)	56
3.2.3 Adaptive Spatial Attention Selection Network (ASASNet)	61
3.2.4 Region-Guided Channel-Wise Attention Network (RG-CANet)	72
3.3 Experiments	78
3.3.1 Accelerated MRI Reconstruction via Deep Cascaded Neural Networks	78
3.3.2 Region-Guided Attention Networks for MRI Reconstruction	92
3.4 Discussions and Conclusion	96
4 Coil-Agnostic Attention-Based Networks for Accelerated Parallel MR Imaging	98
4.1 Introduction	98
4.2 Methodology	100
4.2.1 Problem Formulation	100

4.2.2	Unrolled Optimization-Inspired Pipeline for Multi-Coil MRI	101
4.2.3	Data Aggregation Consistency Block (DACB)	105
4.2.4	Locality-Aware Spatial Attention Module (LoA-SAM)	114
4.2.5	Deep Cascaded Reconstruction Framework	118
4.2.6	Objective Function	121
4.3	Experiments	122
4.3.1	Multi-Coil MRI Reconstruction Comparisons	122
4.3.2	Ablation Studies on Model Components	125
4.3.3	Ablation Studies on Data Consistency Methods	126
4.3.4	Ablation Analysis on Relative Feature Aggregation	130
4.3.5	Robustness to Coil Configurations	131
4.4	Conclusion	132
5	Mask-Guided Motion Feature Incorporated Network for Dynamic MRI Recon-	
	struction	134
5.1	Introduction	134
5.2	Methodology	137
5.2.1	Problem Formulation and Unrolled Reconstruction Pipeline	137
5.2.2	(1+2)D Feature Pyramid Reconstruction Network	140
5.2.3	Spatio-Temporal Fourier Neural Block (ST-FNB)	141
5.2.4	Mask-Guided Motion Feature Incorporation (Mask-MFI)	147
5.3	Experiments	151
5.3.1	Implementations	151
5.3.2	Comparison Results	153
5.3.3	Ablation on Model Components	156
5.3.4	Ablation on Mask-MFI	158
5.3.5	Transferability of Mask-MFI	160
5.3.6	Model Generalization Validation	163
5.4	Conclusion	165

CONTENTS	xi
6 Conclusion	166
6.1 Summary	166
6.2 Future Works	168
6.2.1 MRI Reconstruction Using Diffusion Models	168
6.2.2 Model Learning with Limited Datasets	171
6.2.3 Reconstruction with Cartesian and Non-Cartesian Sampling Strategies .	172
Bibliography	173

Figures and Tables

Figures

2.1	Illustration of convolution variants with 3×3 kernels. (a) standard convolution, (b) strided convolution with a stride of 2, and (c) dilated convolution with a dilation rate of 2. Images taken from https://github.com/vdumoulin/conv_arithmetic	13
2.2	Illustration of recurrent neural networks (RNN) models.	15
2.3	Illustration of transformer.	20
2.4	Illustration of scaled dot-product attention and multi-head attention.	21
2.5	Illustration of shifted window partition.	23
2.6	Illustration of variational network (VN) architecture: (a) image processing pipeline, and (b) VN structure.	32
2.7	Illustration of MoDL architecture.	33
2.8	Illustration of VS-Net architecture. Each update step consists of three components: the denoising block (DB), the data consistency block (DCB), and the weighted average block (WAB).	36
2.9	Illustration of end-to-end variational network (E2E-VarNet) which consists of three constituent elements: the refinement CNN model $\mathcal{D}_{\theta^k}$, the data consistency block (DC), and the sensitivity map estimation module (SME).	41

3.1	Overview of the ASGAN architecture: a) the reconstruction is performed in two stages, where a U-net is adopted to initially recover the under-sampled image and produce the coarse reconstruction $\mathbf{x}'$ from the zero-filled $\mathbf{z}$. The final reconstruction $\mathbf{x}^{recons}$ is generated using an attention selection mechanism with the image and attention maps, termed $\mathbf{x}^i$ and $\mathbf{M}^i$, constructed from the collection of features extracted at multiple decoding levels, and b) deep de-aliasing large-field contextual feature integration (LCFI++).	56
3.2	Illustration of network modules: a) convolutional block attention module (CBAM), b) self-channel interaction (SCI), c) residual-in-residual dense block (RRDB).	57
3.3	Illustration of data consistency (DC) approaches: a) conventional data consistency block embedded in a reconstruction pipeline, b) generalized data consistency expression using the hard DC operator Ω , c) deep data consistency block (DCB).	63
3.4	Illustration of basic modules: a) densely connected block, b) deep data consistency block (DCB), and c) adaptive spatial attention selection module (ASASM). $C N^2$ indicates the channel size and spatial resolution of output features, where the spatial size is assumed to be $N \times N$, for simplicity.	64
3.5	Overview of the proposed adaptive spatial attention selection network (ASASNet) architecture. The input zero-filled is processed by a deep cascade of neural network models after being mapped to feature space by a head convolutional layer. Each reconstruction iteration has a U-shaped subnetwork featuring an adaptive spatial attention selection module (ASASM) and multi-level dense connections, and a deep data consistency block (DCB) as a prominent attribute. The final reconstruction is computed with the collection of all previous predictions by a tail convolutional block. The feature shape is indicated by $C N^2$ corresponding to the channel and spatial sizes with the input assumed to have a spatial size of 256×256 .	65
3.6	Illustration of the region-guided channel-wise attention module (RG-CAM).	74

3.7	Illustration of (a) region-guided channel-wise attention network (RG-CANet) and (b) U-shaped subnetwork. The input to the reconstruction frame is the zero-filled and the output is the image reconstructed with all previous predictions collected by subnetwork dense connections.	75
3.8	Comparison results for the proposed adaptive spatial attention selection network (ASASNet) using the knee MRI dataset under $8\times$ and $4\times$ acceleration.	81
3.9	Residual error maps ($2\times$ amplified) of knee MRI reconstruction under $8\times$ and $4\times$ acceleration for the proposed adaptive spatial attention selection network (ASASNet).	81
3.10	Comparison results for the proposed adaptive spatial attention selection network (ASASNet) using the brain MRI dataset under $8\times$ and $4\times$ acceleration.	83
3.11	Residual error maps ($2\times$ amplified) of brain MRI reconstruction under $8\times$ and $4\times$ acceleration for the proposed adaptive spatial attention selection network (ASASNet).	84
3.12	Illustration of network cascade categories. a) Sequentially connected, b) densely connected, and c) multi-level densely connected. For simplicity, the skip-connections within each subnetwork are not displayed in c).	87
3.13	Residual maps ($2\times$ amplified) of ablation cases for the proposed adaptive spatial attention selection network (ASASNet) under $8\times$ acceleration.	88
3.14	Ablation results for the proposed adaptive spatial attention selection network (ASASNet) with and without the presence of ASASM at $8\times$ acceleration.	89
3.15	Reconstructions using the proposed adaptive spatial attention selection network (ASASNet) with random Gaussian noise added to the measurement in inference.	90
3.16	Reconstruction comparisons for the region-guided channel-wise attention network (RG-CANet) using the knee MRI dataset under $8\times$ and $4\times$ acceleration.	92
3.17	Residual maps ($2\times$ amplified) of knee MRI reconstruction at $8\times$ and $4\times$ acceleration factor for the region-guided channel-wise attention network (RG-CANet).	93
3.18	Ablation residual maps ($2\times$ amplified) of $8\times$ accelerated reconstruction for the region-guided channel-wise attention network (RG-CANet).	95

3.19	Visualization of the spatial guiding masks from (left) penultimate decoding level of the 4-th subnetwork and (right) last decoding level of the 5-th subnetwork, by shading in different colors the non-zero values of the masks amalgamated into a single map.	95
4.1	Illustration of data consistency methods for multi-coil MRI: a) conventional data consistency (DC) method using the pre-computed sensitivity maps, b) coil-agnostic data consistency method based on the data aggregation consistency (DAC) operation, and c) data aggregation consistency block (DACB) generalizing the DAC method to feature level, which is embedded into each subnetwork of the reconstruction framework, as illustrated in Fig. 4.4, to correct the intermediate prediction. 108	
4.2	Illustration of basic modules: a) densely connected layers, b) locality-aware spatial attention module (LoA-SAM), and c) relative feature aggregation block (RFA). C and N^2 indicate the channel and spatial sizes of output features, noting that the spatial size is assumed to be $N \times N$, for simplicity.	115
4.3	Overview of the model architecture. The Root-Sum-of-Squares (RSS) computed with the measurements is first mapped via a head convolutional layer and passed to a network cascade consisting of U-shaped subnetworks with multi-level dense connections and the proposed data aggregation consistency blocks (DACB). For simplicity, only two subnetworks with three transitional levels are displayed. The feature size is indicated by $C N^2$ corresponding to the channel and spatial sizes. For convenience, the input is assumed to have a spatial size of 256×256	118
4.4	Illustration of the unrolled reconstruction network cascade built on a model-based algorithm. Deep neural network modules f_{θ_k} and data aggregation consistency blocks (DACB), illustrated in Fig. 4.1, are interleaved to form a coil-agnostic reconstruction framework, where the inter-connections between subnetworks are omitted, for simplicity. At each iteration, the intermediate prediction coming from the refinement network module f_{θ_k} is passed to the DACB where it is corrected with the multi-coil measurements $\{\mathbf{y}_i\}$ by applying the data aggregation consistency (DAC) method, also demonstrated in Fig. 4.1.	120

4.5	Comparison results of $8\times$ accelerated MRI reconstruction.	122
4.6	Comparison results of $4\times$ accelerated MRI reconstruction.	123
4.7	Ablation results on data consistency methods and data type at $8\times$ acceleration. . .	127
4.8	Artifacts caused by inaccurate sensitivity map estimation. First) RSS of fully sampled signal, second) positive component of the imaginary part of a sensitivity map, third) positive component of the imaginary part of the weighted average, and last) reconstruction produced by (E)-CSM.	129
4.9	Comparisons of $8\times$ accelerated MRI reconstruction with coil deployments randomly permuted.	132
5.1	Illustration of dynamic MRI reconstruction model. The structure of subnetworks and the mask-guided motion feature incorporation (Mask-MFI) are displayed in the bounding box. The motion features extracted from the under-sampled measurement sequences are incorporated into the reconstruction framework under the guidance of the motion masks generated with the measurement observations to explicitly leverage the dynamic information contained in MRI sequences and improve the reconstruction performance in terms of the recovery of the time-varying content.	139
5.2	Illustration of spatio-temporal Fourier neural block (ST-FNB). The input feature volume is split into two groups which respectively undergo the top spatial neural route and the bottom temporal neural route, where the information exchange among the global region in both spatial and time domains is achieved. The output feature volume is produced via a point-wise convolutional layer with the combination of the outputs coming from the two routes.	144
5.3	Reconstruction comparisons evaluated on OCMR dataset. GT refers to the ground truth image and ZF denotes the zero-filled.	154
5.4	Comparison of temporal profiles on OCMR dataset. GT refers to the ground truth image.	155
5.5	Ablation studies on model components evaluated on OCMR dataset under $16\times$ acceleration.	156

5.6	Ablation on mask-guided motion feature incorporation (Mask-MFI) using OCMR dataset under $16\times$ acceleration.	159
5.7	Illustration of transferring Mask-MFI to other models. The motion features and masks are generated with the under-sampled MRI sequences and incorporated into each reconstruction iteration by coupling with the features extracted by a convolutional layer at each reconstruction stage.	161
5.8	Transferability experiment on mask-guided motion feature incorporation (Mask-MFI) using OCMR dataset under $16\times$ acceleration.	162
5.9	Reconstruction comparisons evaluated on Harvard dataverse with $16\times$ acceleration and fine-tuning.	164

Tables

3.1	Quantitative evaluation on knee MRI reconstruction under $8\times$ and $4\times$ acceleration for the proposed adaptive spatial attention selection network (ASASNet). $4\times$ accelerated reconstruction has the same latency (s) and model size as $8\times$	82
3.2	Quantitative evaluation on brain MRI reconstruction under $8\times$ and $4\times$ acceleration for the proposed adaptive spatial attention selection network (ASASNet). $4\times$ accelerated reconstruction has the same latency (s) and model size as $8\times$	85
3.3	Ablation studies on the model components and structure of the proposed adaptive spatial attention selection network (ASASNet) using knee dataset.	88
3.4	Ablation study on data consistency blocks of the proposed adaptive spatial attention selection network (ASASNet) using knee dataset.	90
3.5	Ablation study on objective function using knee dataset for the proposed adaptive spatial attention selection network (ASASNet).	90

3.6	Quantitative evaluation on accelerated MRI reconstruction for the region-guided channel-wise attention network (RG-CANet). Separable convolutions are used for the variant RG-CANet-S in place of conventional convolutions in each subnetwork of RG-CANet, leading to a smaller model size to demonstrate that the improved performance is attributable to the devised region-guided attention mechanism and the network structure, and not simply owing to the parameter size. $4\times$ accelerated reconstruction has the same run-time and parameter size as $8\times$	94
3.7	Ablation studies on model components of the region-guided channel-wise attention network (RG-CANet) at $8\times$ acceleration factor.	94
4.1	Quantitative results of multi-coil knee MRI reconstruction. $4\times$ accelerated reconstructions have the same execution time and parameter size as $8\times$	124
4.2	Ablation on model components using $8\times$ accelerated multi-coil knee MRI reconstruction.	126
4.3	Ablations on data consistency methods and data type at $8\times$ acceleration. The proposed method involves no sensitivity estimation and is therefore free of ME-Arti.	127
4.4	Ablation analysis on residual feature connection in DACB under $8\times$ acceleration using multi-coil knee MRI reconstruction.	130
4.5	Ablation study on relative feature aggregation using $8\times$ accelerated multi-coil knee MRI reconstruction.	130
4.6	Evaluation results of $8\times$ accelerated multi-coil knee MRI reconstruction using random coil permutation.	131
5.1	Quantitative comparison results evaluated on OCMR dataset. Acceleration factors are respectively $8\times/16\times/24\times$. Running time (s) and parameter size (MB) are same for different acceleration rates.	153
5.2	Ablation studies on model components evaluated on OCMR dataset under $16\times$ acceleration. To provide a fair comparison, we adjust the width of a few convolutional layers in the competing model candidates to balance the parameter size, though it can introduce slightly more parameters.	157

5.3	Ablation on mask-guided motion feature incorporation (Mask-MFI) via OCMR data under $16\times$ acceleration.	158
5.4	Comparison between the models with and without the presence of Mask-MFI evaluated on OCMR dataset under $16\times$ acceleration.	161
5.5	Evaluating models trained on OCMR dataset using Harvard dataverse for testing and fine-tuning (FT) at $16\times$ acceleration factor (w/o FT/w FT).	164

Acronyms and Abbreviations

ASASM	Adaptive Spatial Attention Selection Module
ASASNet	Adaptive Spatial Attention Selection Network
ASGAN	Attentive Selection Generative Adversarial Network
CNN	Convolutional Neural Network
CRNN	Convolutional Recurrent Neural Network
CS	Compressed Sensing
DAC	Data Aggregation Consistency
DACB	Data Aggregation Consistency Block
DC	Data Consistency
DCB	Deep Data Consistency Block
(1+2)D-FPRN	(1+2)D Feature Pyramid Reconstruction Network
GAN	Generative Adversarial Network
GT	Ground Truth
LCFI++	Deep De-aliasing Large-Field Contextual Feature Integration
LoA-SAM	Locality-Aware Spatial Attention Module
Mask-MFI	Mask-Guided Motion Feature Incorporation
MFMM	Motion Feature Modulation and Masking
MRI	Magnetic Resonance Imaging
RBF	Gaussian Radial Basis Function
RG-CANet	Region-Guided Channel-Wise Attention Network
RFA	Relative Feature Aggregation

RG-CAM	Region-Guided Channel-Wise Attention Module
RNN	Recurrent Neural Network
RSS	Root Sum of Squares
ST-FNB	Spatio-Temporal Fourier Neural Block
ZF	Zero-Filled

Nomenclature

$[,]$	Feature channel concatenation
$\odot$	Element-wise multiplication
$\oplus$	Element-wise addition
$\otimes$	Matrix multiplication over the last two dimensions
$\circ$	Composition of operators
$*$	Convolution operator
$\langle, \rangle$	Inner-product
$conv$	Convolutional layer
$\mathbf{x}$	MR image
$\mathbf{y}$	MRI measurement
$\mathbf{z}$	Zero-filled
$\mathbf{A}$	Forward data acquisition process
$\mathbf{D}$	Encoding matrix
$\mathbf{F}$	Fourier transform matrix
$\mathbf{S}_i$	i -th coil sensitivity map
f_θ	Subnetwork of reconstruction framework
$\mathcal{D}_\theta$	Reconstruction network model
$\mathcal{D}_I$	Image domain neural network
$\mathcal{D}_F$	k-space neural network
$\mathcal{D}_S$	Sensitivity estimation neural network

ps	Pixel shuffle operation
q	Spatial position
δ	Moving footprint
Δ	Kernel grid
$\omega(\delta, q)$	Preference of position $q + \delta$
Ω	Hard data consistency
Ψ	Coil projection
Φ	Sensitivity-weighted average
σ	Sigmoid function
$\tanh$	Hyperbolic tangent function
$shrink_1$	Soft-threshold operator
$shrink_{vec,2}$	Vector shrinkage operator
div	Divergence operator
∇	Gradient operator
TV	Total variation

Chapter 1

Introduction

1.1 Research Problem Overview

1.1.1 Magnetic Resonance Imaging Introduction

Magnetic resonance imaging (MRI) employs the nuclear magnetic resonance (NMR) phenomenon to produce high-contrast images of body parts. The observation of NMR absorption in a molecular beam incentivized researches to utilize the phenomenon to provide accurate imaging in spectroscopy applications, using the gradient field to encode the spatial information contained in the radio waves emitted from the nuclei of the target, which provides a non-invasive and radiation-free tool for inspection of the underlying anatomy of human body.

During the scanning procedure of MRI, a strong magnetic field is exploited to align free protons in hydrogen contained in molecules of water presented in biologic tissues, and a radio wave pulse passing through the patient deflects the protons which return to the aligned state when the radio wave is removed and emit radio pulses whose frequency, time, and strength produce distinct signals used to create images.

The advantages of MRI include the capacity to image soft tissues at very high resolution and the lack of ionizing radiation, therefore providing a safe and versatile diagnostic tool, compared to computed tomograph (CT). The disadvantages include the expensive cost of devices, limited ability to image bones, long scanning procedure time which potentially causes discomfort to patients, and the tendency for movement artifacts.

1.1.2 Data Acquisition Acceleration and MRI Reconstruction

To produce images of the anatomy of the body, magnetic resonance imaging samples the Fourier transform to collect signals representing different types of tissues, which is often a long acquisition process and makes MRI a slow imaging modality against the speed requirement of modern medical imaging, possibly resulting in increased examine expenses and patient discomfort. Many types of strategies were proposed to reduce the scanning time by leaving out a number of sampling steps during data acquisition and predicting the missing components from the collected measurements to restore the target image.

However, under-sampled data in k -space can suffer from strong aliasing artifacts when being transformed into image domain, due to a violation of the Nyquist-Shannon sampling theorem, which can have deleterious effects on clinical diagnosis and treatment plan. To reduce the aliasing errors and generate clean images accurately representing the signals being measured, many reconstruction algorithms were put forward to solve the ill-posed inverse problems and have witnessed significant advances.

Among those methods, compressed sensing (CS) demonstrated the excellent performance in reconstructing MR images from the under-sampled k -space data. Although useful in practice, CS methods constrain by sparsity-based regularization the search space which is explored to find a solution to the under-determined problem, and achieve image reconstruction via iterative optimization algorithms which can be computationally expensive and time-consuming. The pre-defined and handcrafted sparsity prior leveraged by CS methods to model the intrinsic image features can be difficult to portray complicated anatomical structures. Those factors obstruct the development of CS approaches in modern MRI.

Parallel imaging (PI) exploits the correlations among different coil sensors to unfold an image of the underlying anatomy into multiple frames which are simultaneously measured at an acceleration rate to shorten the sampling process. It is widely used to resist aliasing artifacts in MR images by utilizing the spatial redundancies to compensate for the loss of information. Nevertheless, PI methods have to afford increased equipment expenses and can fail to remove strong aliasing errors at high acceleration factors.

In accelerated MRI scanning, it is highly demanded to perform high-quality reconstruction from a subset of measurements in hopes of creating images that are suitable to detect, analyze, and monitor disease progression in clinical practice. The reconstruction performance, which can be negatively impacted when the prior knowledge of the target signals is not appropriately modelled, is tightly associated with the image resolution and quality that can be attained, and consequently restricts the achievable acceleration factors. How to produce accurate and faithful reconstructions and improve the quality of recovered images still remains to be an open problem for fast MRI in modern medical imaging applications.

1.2 Main Contributions

To speed up the MRI data acquisition and reduce the scanning time, it is common practice to accelerate the sampling process by only taking a subset of k -space measurement values following a sampling pattern and creating images from the collected data. However, it suffers from aliasing artifacts in image domain when signals are taken in k -space with a sub-Nyquist rate, which inevitably leads to a decline in image quality and affects the clinical judgement and treatment plan. Several methods were suggested to recover the under-sampled signals, among which compressive sensing approaches gained popularity with the ability to reconstruct clean images from the observations corrupted by aliasing artifacts. However, the sparsity prior of images is normally assumed by compressed sensing methods to guarantee the success of reconstruction, which can be difficult to hold in practice and the adopted sparse transforms show the limited representation capacity to capture complicated anatomical structures and

textural details, making it challenging to produce high-quality reconstruction results. Recently, deep neural network models have demonstrated widespread use in vision tasks by utilizing the powerful representation and the excellent learning ability of deep architectures, also showing great potential in MRI reconstruction. Despite achieved success, it still remains challenging to provide high-fidelity reconstruction under aggressive acceleration diets to meet the demand of modern MR imaging.

1.2.1 Attentive Neural Network Models for MRI Reconstruction

To obtain high-quality reconstruction, we propose an attentive selection generative adversarial network, termed ASGAN, which exploits a devised large-field contextual information integration module and attention selection mechanism to generate images with realistic local details from under-sampled k -space measurements. Although effective in retrieving fine-grained results, the lack of the incorporation of the k -space data constraint leads to inconsistent image details and unwanted artifacts in reconstructions produced by ASGAN, which as a direct inversion model learns an end-to-end mapping from the sparsely sampled measurement to a clean image. To address this problem, an optimization-inspired framework is introduced by unrolling an iterative algorithm for regularized optimization with a novel deep data consistency block, dubbed DCB, to leverage domain-specific knowledge into the reconstruction pipeline aiming to encourage more faithful results. Additionally, we design an adaptive spatial attention selection module, termed ASASM, to capture the correlations among the spatial regions of high-resolution features, which is embedded in the unrolled network framework to efficiently retrieve contextual information without causing huge computational complexity, by performing data-dependent kernel prediction as a way to increase the diversity of attention patterns and provide improved reconstruction quality.

As an alternative to attention mechanism applied to learn the correlations among the spatial pixels, channel-wise attention methods were suggested to model the interdependencies between the feature channels to improve the reconstruction performance while circumventing the enormous computational costs, exhibiting an efficient application of attention-based

models. Nevertheless, channel-based attention mechanism suffers from the limited capacity to explore the rich and useful information among the spatial areas. To leverage the contextual hidden cues in the spatial dimensions, a novel region-guided channel-wise attention network is proposed, which is established by unfolding an optimization algorithm and endowed with the ability to incorporate the spatial semantics representing the structural features of the underlying anatomy being imaged into a channel-based attention mechanism, demonstrating a light-weight and efficient design to attain improvements in MRI reconstruction. Those works were introduced in the published papers [1, 2, 3, 4].

1.2.2 Coil-Agnostic Parallel MRI Reconstruction Networks

As one type of accelerated imaging methods, parallel imaging (PI) utilizes multiple coils to measure the signals emitted from tissues being inspected and leverages the spatial redundancies among different coil frames to resist aliasing artifacts when creating images via reconstruction algorithms. However, the forward data encoding process, which is normally needed to model the correlations of coil frames and expressed as a set of sensitivity maps, is unknown and can be estimated from the sampled measurements using model-based algorithms. Although widely used in practice, model-based sensitivity estimation methods are intrinsically time-consuming due to the calculations of the gradients and can offer inaccurate estimation results when sufficient k -space lines are unavailable at high under-sampling rates, therefore severely affecting the quality of restored images. To alleviate the inaccurate estimation of the forward process, deep neural networks, which have demonstrated their excellent performance in solving ill-posed inverse problems, were coupled in several works into a reconstruction framework to jointly predict the unknown sensitivity maps from the multi-coil measurements along with the reconstruction procedure. However, such methods are potentially susceptible to coil configurations varying across different scanning cases and can unfavorably affect the reconstruction performance. Moreover, the sensitivity estimation relying on neural network models introduce more parameters and inevitably increase computational overheads.

We propose a coil-agnostic reconstruction framework to solve the unknown forward process problem in multi-coil MRI using a novel data aggregation consistency block, dubbed DACB, which is devised to approximate the data consistency enforcement without the presence of coil sensitivities. A locality-aware spatial attention module is incorporated into the reconstruction model to capture spatial contextual information and endow the network with the data-adaptive properties, showing to enhance the reconstruction performance. We demonstrate through experiments that the devised coil-agnostic method is robust and invulnerable to machine configuration discrepancies, and surpasses other sensitivity estimation-based models. The associated publication is the paper [5].

1.2.3 Dynamic MRI Reconstruction Models

Dynamic MRI reconstruction is formulated as the recovery of a sequence of images embodying the dynamic properties of the organs being scanned from a collection of measurement frames sampled at consecutive time steps, which demands both spatial and temporal high resolutions and imposes a more challenging task. To improve the accuracy of reconstruction of the dynamic sequences, many methods proposed to exploit the redundant and complementary information contained in the neighboring frames, e.g. transmit the information across consecutive frames by utilizing the recurrent nature of RNN models, as a means to better preserve the time-varying content. However, the use of temporal redundancies can lead reconstruction models to ignore the variations of the individual temporal frames and has an adverse effect on the recovery of the dynamic features.

We propose an optimization-inspired deep cascaded framework to reconstruct dynamic MRI sequences from the k -space temporal views measured at consecutive time steps under an accelerated data acquisition diet. A novel mask-guided motion feature incorporation method, dubbed Mask-MFI, is devised to explicitly extract the dynamic cues from the under-sampled signals, which are incorporated into the reconstruction iterations to improve the model ability to preserve the motion of the organs in created images, potentially introducing a new direction to leverage the dynamic content hidden in the sparsely sampled data for MRI reconstruction. A

spatio-temporal Fourier neural block (ST-FNB) is devised to efficiently retrieve useful information from the global area in both spatial and time domains by extending the methodology of the Fourier neural operation to dynamic scenes, which by experiments demonstrates improved model performance. It is shown that the proposed method can outperform state-of-the-art models in terms of the reconstruction quality and the recovery of the dynamic content. The submitted paper [6] concludes those works.

1.3 Thesis Overview

The research of this thesis focuses on MRI reconstruction problems using deep neural network models, which is structured as follows:

In Chapter 2, we present a brief description of neural network models and introduce several typical types of network architectures. A review of compressive sensing methods for MRI reconstruction is presented, which is followed by an introduction of deep learning-based approaches to solve the ill-posed inverse problems, including the advances in the field of parallel and dynamic MRI reconstruction.

In Chapter 3, we introduce the methods for MRI reconstruction by exploiting the learning capacity and rich representations of deep neural networks to restore MR images from sparsely sampled signals measured under accelerated data acquisition scenarios. An attention selection generative adversarial network is devised to improve the quality of reconstructed images and achieve fine-grained results with sharp and realistic details. To resolve the lack of domain-specific knowledge which limits the model performance, we introduce an optimization-inspired deep cascaded network which utilizes a novel deep data consistency block to enforce the data consistency and an adaptive spatial attention selection module to retrieve the contextual information, showing to encourage more faithful reconstructions. To increase the capacity of channel-based attention mechanism which normally acts as an efficient alternative to attention methods applied to the spatial dimensions, we introduce a region-guided channel-wise attention network which has the ability to incorporate the implicitly learned spatial semantic information into

attention over channels via a light-weight and flexible network design, allowing the application of attention mechanism to high-resolution feature representations with the object of improving the reconstruction performance, validated qualitatively and quantitatively, without incurring heavy computational costs.

In Chapter 4, we introduce the work concentrating on multi-coil MRI reconstruction. A novel coil-agnostic reconstruction framework is presented to solve the unknown forward data acquisition process problem. We introduce a data aggregation consistency block to approximately perform the k -space data consistency enforcement without the presence of the estimation of coil sensitivities. A locality-aware spatial attention module is incorporated into the reconstruction network to capture the contextual information to improve the model performance while avoiding tremendous computational costs. We demonstrate through experiments the invulnerability and robustness of the proposed coil-agnostic method to machine configuration discrepancies.

In Chapter 5, the research focuses on dynamic MRI reconstruction. An optimization-inspired deep cascaded framework for the reconstruction of dynamic MRI sequences is presented, with a novel spatio-temporal Fourier neural block designed to efficiently retrieve useful information in spatial and temporal domains. A mask-guided motion feature incorporation method is introduced to extract the motion information from the under-sampled measurements and explicitly incorporate the dynamic cues into the reconstruction procedure, showing to remarkably improve the model capacity to recover the time-varying content contained in MRI sequences.

In Chapter 6, we conclude the works presented in this thesis and discuss potential directions for future works.

1.4 Publications

The research publications of this thesis are listed as follows:

Published Journal Paper:

1. **Jingshuai Liu**, Chen Qin, and Mehrdad Yaghoobi. “High-Fidelity MRI Reconstruction Using Adaptive Spatial Attention Selection and Deep Data Consistency Prior”. IEEE Transactions on Computational Imaging, 2023, vol. 9, pp. 298-313 [1].

Published Conference Papers:

1. **Jingshuai Liu**, Chen Qin, and Mehrdad Yaghoobi. “Coil-Agnostic Attention-Based Network for Parallel MRI Reconstruction”. Proceedings of the Asian Conference on Computer Vision (ACCV), 2022, pp. 2866-2883 [5].
2. **Jingshuai Liu**, Chen Qin, and Mehrdad Yaghoobi. “High-Fidelity MRI Reconstruction with the Densely Connected Network Cascade and Feature Residual Data Consistency Priors”. Machine Learning for Medical Image Reconstruction (MLMIR) 2022, Lecture Notes in Computer Science, vol. 13587, Springer [2].
3. **Jingshuai Liu**, Chen Qin, and Mehrdad Yaghoobi. “Region-Guided Channel-Wise Attention Network for Accelerated MRI Reconstruction”. Machine Learning in Medical Imaging (MLMI) 2022, Lecture Notes in Computer Science, vol. 13583, Springer [3].
4. **Jingshuai Liu** and Mehrdad Yaghoobi. “Fine-Grained MRI Reconstruction Using Attentive Selection Generative Adversarial Networks”. ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, Canada, 2021, pp. 1155-1159 [4].

Submitted Journal Paper:

1. **Jingshuai Liu**, Chen Qin, and Mehrdad Yaghoobi. “Mask-Guided Deep Network Cascade for Dynamic Cardiac MRI Reconstruction with the Fourier Neural Attention”. Submitted, 2023 [6].

Background and Literature Review

2.1 Introduction

Accelerating the MRI data acquisition process and shortening the scanning time are achieved by image reconstruction from signals taken under a sub-Nyquist rate, which can cause aliasing artifacts in image domain owing to the under-sampling in k -space if conventional reconstruction techniques are incorporated. To remove the unexpected artifacts and provide clean images for clinical diagnosis, it is common practice to recover the missing information via reconstruction algorithms which often require the modelling of the prior knowledge of target data to ensure the feasibility and quality of reconstruction results. Model-based methods such as compressed sensing are widely used in accelerated MRI to predict the missing frequency components by formulating the under-determined problem as a regularized optimization and search for a solution via iterative minimization algorithms, which are often of large amount of computational costs.

With significant advances in vision tasks, deep neural network models have shown the great potential to reconstruct MR images from a subset of the k -space measurements collected in a sub-Nyquist sampling scheme. By generalizing the neural network across a dataset of the target images, the trained model has the capacity to produce clean images from the corrupted inputs with the learned data prior encoded in the network parameters. Due to the excellent performance in achieving high-quality reconstruction, neural network models have gained popularity and achieved great success in the development of accelerated MRI.

In this chapter, we introduce the basic deep neural network models to provide a fundamental understanding of such acceleration techniques. Subsequently, a brief description of reconstruction methods based on compressed sensing theory is presented. In the ensuing section, deep learning-based approaches are introduced, including network models designed for multi-coil acquisition process and reconstruction approaches to recovering dynamic MRI sequences.

2.2 Deep Neural Network Models

Deep neural networks leverage the learning capacity of trainable weights and the powerful representation of deep architectures to achieve appealing performance in solving complex problems which are difficult to address using traditional methods. Several types of networks are introduced in this section to provide a fundamental understanding of typical neural network models.

2.2.1 Convolutional Neural Networks

Among the explorations of artificial neural network (ANN) models, convolutional neural networks (CNN) have gained great popularity due to their feasibility, flexibility, and outstanding performance in practice. The convolutional operation between the kernel ω and the input signal x is illustrated in Fig. 2.1 (a) and defined as follows,

$$(x * \omega)(p) = \sum_{\delta \in \Delta} \omega(\delta) x(p + \delta), \quad (2.1)$$

where p denotes the spatial position, δ indicates the moving footprint as the operation is applied in each local region, and Δ denotes the kernel grid such as $\{-1, 0, 1\}^2$ for 3×3 2-dimensional (2-D) kernels. Based on the operation defined in (2.1), convolutional layers are constructed as the constituent component to build up complicated CNN models. We denote by “features” the signal being processed by neural networks or the output from a network layer, which comprise

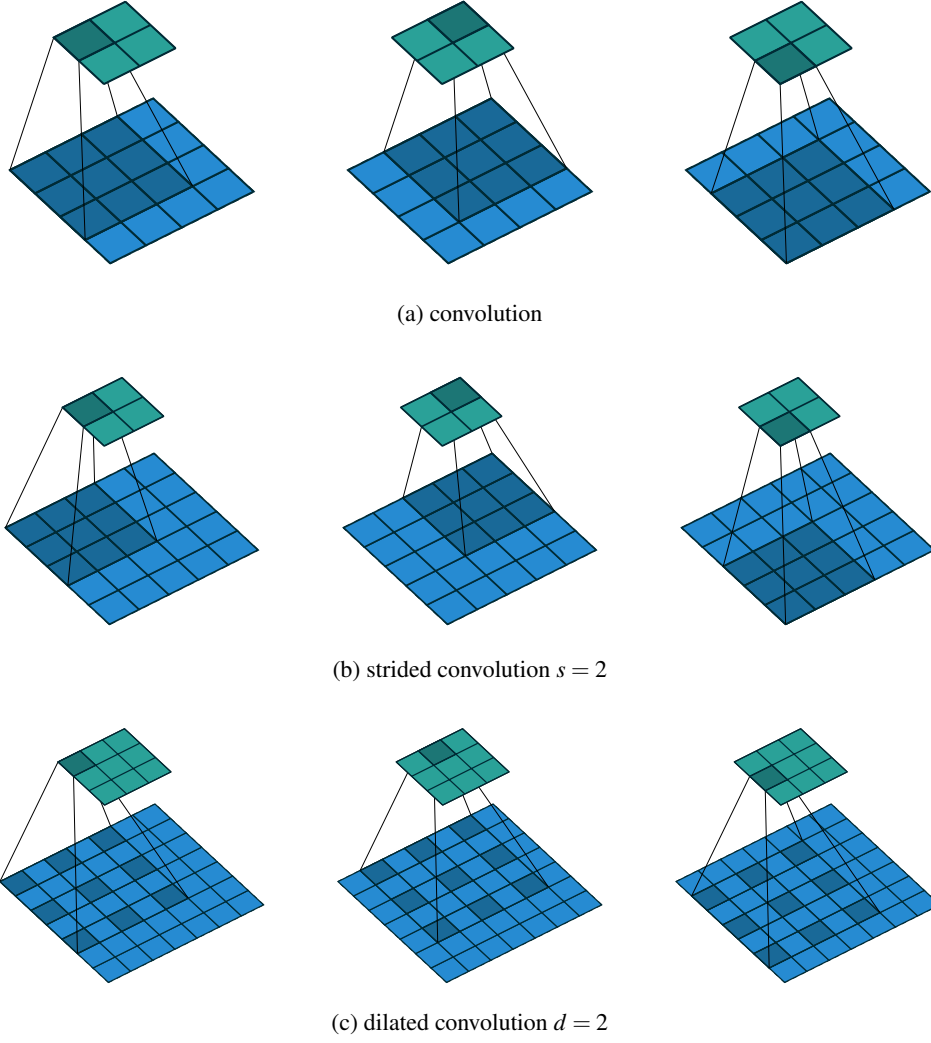


Figure 2.1: Illustration of convolution variants with 3×3 kernels. (a) standard convolution, (b) strided convolution with a stride of 2, and (c) dilated convolution with a dilation rate of 2. Images taken from https://github.com/vdumoulin/conv_arithmetic.

multiple feature maps located at different channels. To simplify the description, we assume 2-D data being processed and omit the batch dimension, leading to the signal size $C \times H \times W$ corresponding to $[channels, height, width]$. A convolutional layer is given by:

$$conv(x, W)_j = b_j + \sum_{i=1}^{C_{in}} x_i * \omega_i^j, \forall j \in \{0, 1, \dots, C_{out}\} \quad (2.2)$$

where j indicates the j -th channel of the output and i the i -th channel of the input. The bias b_j can be emitted for simplicity. Due to the dense connections between the input and output features, the layer has the kernel volume $W = \{w_i^j\}$ of size $C_{out} \times C_{in} \times k_y \times k_x$, e.g. $k_y = k_x = 3$ for 3×3 kernels. The locally executed convolutional operation using the sliding window endows the structures built upon convolutional layers with several inductive biases including: 1) weight sharing, 2) locality, and 3) shift-invariance, which enable effective and efficient model designs in application. The notation in (2.2) can simplify to $conv(x)$ with the kernels of definition elided when ensuring that there is no confusion. A variant of the layer defined above is the strided convolution [7] with selecting the sliding step s of the local window larger than 1, as displayed in Fig. 2.1 (b). It strikes a trade-off between computational efficiency and information retrieval by shifting the window with a step of s and uniformly “skipping” over positions. A grouped convolution was first introduced in AlexNet [8] to reduce the complexity of CNN and facilitate the training of large models. It divides the input features into g groups which separately convolve with their individual kernel volumes, aiming to reduce the parameters to be $C_{out} \times \frac{C_{in}}{g} \times k_y \times k_x$. The receptive field of each location to retrieve information is limited by the local window, which can be expanded by “inserting” holes between the kernel elements, leading to the dilated convolution illustrated in Fig. 2.1 (c) and defined below,

$$(x * \omega)(p) = \sum_{\delta \in \Delta} \omega(\delta) x(p + d\delta), \quad (2.3)$$

with the dilation rate d to determine how much the kernel is inflated. The size of the output can be computed with the hyperparameters $[padding, dilate, stride_size] = [p, d, s]$ by

$$\begin{aligned} H_{out} &= \frac{H_{in} + 2p - d(k_y - 1) - 1}{s} + 1 \\ W_{out} &= \frac{W_{in} + 2p - d(k_x - 1) - 1}{s} + 1, \end{aligned} \quad (2.4)$$

where p denotes half the amount of paddings applied to the input and a more detailed description of the configurations of applying a convolution over an input signal is available at <https://pytorch.org/docs/stable/generated/torch.nn.Conv2d.html>.

CNN-based models have been witnessing a monumental growth in enhancing the capacity of neural networks and achieving promising results in a wide range of applications in object classification, detection, and other vision tasks. Among those models, AlexNet [8] is one of the most typical deep convolutional networks applied to high-resolution (hig-res) images and achieving considerably better image classification results than previous methods on a highly challenging dataset. It also confirms the importance of network depth as removing a single layer leads to degraded performance in experiments. A deep convolutional network, dubbed VGG [9], was developed to explore the effect of the network depth on the accuracy in large-scale image recognition, which shows a substantial improvement with the use of deep visual representations and excellent generalization to a wide range of datasets and tasks. The architecture of U-shaped networks (U-net) was established in [10] which adopts the shortcuts in a symmetric structure to enable direct feature transmission. It shows to achieve record-breaking results in image segmentation when trained with very few samples.

2.2.2 Propagation-Based Neural Networks

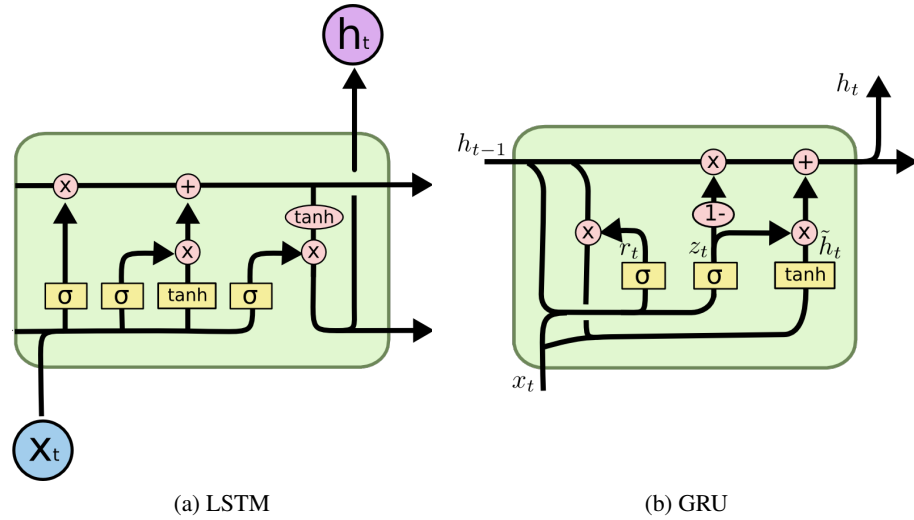


Figure 2.2: Illustration of recurrent neural networks (RNN) models.

Recurrent neural networks (RNN) are proposed to model the long-range dependencies contained in sequences of time series signals, e.g. in language translation, by reasoning about the previous events and passing the current output to a successor. In the past few years, RNN models have achieved great success in a variety of domains, e.g. language translation and speech recognition. However, RNN can be less capable of learning to connect the relevant information to the place where it is needed as the interval between them grows up to be very large. As a popular variant of RNN, long short-term memory (LSTM) [11] networks utilize a long-term unit suited to deal with dependencies with long-range intervals, as displayed in Fig. 2.2 (a). The update rule involved in each repeating cell of LSTM is shown below,

$$\begin{aligned}
i_t &= \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \\
f_t &= \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f) \\
c_t &= f_t \odot c_{t-1} + i_t \odot \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \\
o_t &= \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \\
h_t &= o_t \odot \tanh(c_t),
\end{aligned} \tag{2.5}$$

where x_t denotes the input at the t -th step, h_t is the hidden state and often adopted as the output, W refers to the weight matrix used here in the form of matrix-vector multiplication, $\odot$ refers to the element-wise multiplication, $\tanh$ refers to the hyperbolic tangent function, and σ denotes the Sigmoid function which scales the outputs in range $[0, 1]$ with a value of one representing “totally preserve the component” and a value of zero “completely forget it”, thus determining how much information is allowed to pass through. The gated recurrent unit (GRU) [12, 13] presented in Fig. 2.2 (b) offers a variant of LSTM as a streamlined version of the memory cell, which shows a comparable performance with the advantage of being faster to execute. The key equations are given below,

$$\begin{aligned}
z_t &= \sigma(W_{xz}x_t + W_{hz}h_{t-1}) \\
r_t &= \sigma(W_{xr}x_t + W_{hr}h_{t-1}) \\
\tilde{h}_t &= \tanh(W_{x\tilde{h}}x_t + W_{r\tilde{h}}(r_t \odot h_{t-1})) \\
h_t &= (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t.
\end{aligned} \tag{2.6}$$

To bridge the gap between time series signal modelling using RNNs and image processing, convolutional layers can be embedded into RNN-based models as introduced in ConvLSTM [14] and ConvGRU [15]. Taking ConvLSTM for instance, the iterative algorithm is modified as follows,

$$\begin{aligned}
i_t &= \sigma(W_{xi} * x_t + W_{hi} * h_{t-1} + W_{ci} \odot c_{t-1} + b_i) \\
f_t &= \sigma(W_{xf} * x_t + W_{hf} * h_{t-1} + W_{cf} \odot c_{t-1} + b_f) \\
c_t &= f_t \odot c_{t-1} + i_t \odot \tanh(W_{xc} * x_t + W_{hc} * h_{t-1} + b_c) \\
o_t &= \sigma(W_{xo} * x_t + W_{ho} * h_{t-1} + W_{co} \odot c_t + b_o) \\
h_t &= o_t \odot \tanh(c_t),
\end{aligned} \tag{2.7}$$

where W denotes the convolutional kernel and the operator $*$ can be generalized to a convolutional layer.

2.2.3 Generative Adversarial Networks

Generative adversarial networks [16] (GAN) have attracted great attention due to their significant capacity of producing photo-realistic images. The general design of a GAN-based model consists of a discriminator used to distinguish real data from the synthesized fake counterparts and a generator learning to produce realistic samples which are able to fool the discriminator. The objective of the vanilla GAN is expressible as a min-max game between the generator G_θ and discriminator D_ψ with trainable parameters denoted by θ and ψ :

$$\min_{\theta} \max_{\psi} E_{x \sim p_d} [\log(D_\psi(x))] + E_{z \sim p(z)} [\log(1 - D_\psi(G_\theta(z)))], \tag{2.8}$$

where $E_{x \sim p(x)}$ calculates the expectation over the variable x following the distribution $p(x)$ and $p(z)$ is normally a standard Gaussian from which a sample z is taken as input to the generator for image generation.

The distance between the generative distribution and the real data distribution is quantified by the Jensen-Shannon divergence (JSD) when the discriminator is trained to reach the optimal value. Despite successful cases in realistic image generation, the learning process is known to be unstable and slow, and prone to get stuck in a small area with very limited variety.

The Wasserstein GAN [17] (WGAN) aims to minimize the Wasserstein distance between the real data distribution p_d and the synthesized one p_θ , which is defined in the form of the earth mover's distance:

$$W(p_d, p_\theta) = \inf_{\gamma(x,y)} E_{(x,y) \sim \gamma} [\|x - y\|], \quad (2.9)$$

where γ is the joint distribution whose marginal densities are respectively p_d and p_θ . The distance can be regarded as the cost paid to transform one distribution to the other in an optimal transport scheme. It can alleviate the gradient vanishing problem caused by the support of high dimensional sample distributions confined to low dimensional manifolds, showing an intersection as a set of measure zero. The Kantorovich-Rubinstein duality allows us to compute the distance via the supremum of the expectation in (2.10) over all 1-Lipschitz continuous functions:

$$W(p_d, p_\theta) = \sup_{\|f\|_L \leq 1} E_{x \sim p_d} [f(x)] - E_{y \sim p_\theta} [f(y)], \quad (2.10)$$

where $\|\cdot\|_L$ denotes the Lipschitz constant. The objective of WGAN is therefore to minimize the Wasserstein distance estimated with a Lipschitz function f_ψ parametrized by ψ as below,

$$\min_{\theta} \max_{\psi} E_{x \sim p_d} [f_\psi(x)] - E_{z \sim p(z)} [f_\psi(G_\theta(z))], \quad (2.11)$$

where f_ψ serves as the discriminator, also named the critic, with the weights clamped to ensure a Lipschitz constant.

One of the challenges in training WGAN is to enforce the Lipschitz constraint. The model is sensitive to the clipping parameter: a large value makes it difficult for the weights to reach their optimality and a small one can easily diminish the gradients with an increase in the number of layers. Gradient penalty was introduced in WGAN-GP [17] to substitute the weight clipping, which penalizes the network when the gradient norm drifts away from the desired

value, obtaining the following objective function for the critic,

$$\min_{\psi} E_{z \sim p(z)} [f_{\psi}(G_{\theta}(z))] - E_{x \sim p_d} [f_{\psi}(x)] + \lambda E_{\hat{x}} [(\|\nabla_{\hat{x}} f_{\psi}(\hat{x})\|_2 - 1)^2], \quad (2.12)$$

where $\hat{x} = tG_{\theta}(z) + (1-t)x$ with t uniformly taken between 0 and 1. Despite the computational complexity added by the calculation of the gradient penalty, WGAN-GP shows to yield higher image quality and provide more stable training, compared with WGAN. Spectral normalization (SN) [18] is proposed to control the Lipschitz constant of a network by constraining the spectral norm of each layer to satisfy the Lipschitz constraint, as opposed to implicitly enforcing the Lipschitz continuity. This normalization technique proves to stabilize the training of GAN and is able to produce images of higher quality, compared to other stabilizing and normalization approaches. The findings in [18] show that the improved performance in terms of image quality can be achieved with the SN method applied to both the generator and discriminator.

To overcome the gradient vanishing problem, the least squares GAN [19] (LSGAN) adopts the least squares loss as the objective function:

$$\begin{aligned} & \min_{\psi} E_{x \sim p_d} [(D_{\psi}(x) - b)^2] + E_{z \sim p(z)} [(D_{\psi}(G_{\theta}(z)) - a)^2] \\ & \min_{\theta} E_{z \sim p(z)} [(D_{\psi}(G_{\theta}(z)) - c)^2], \end{aligned} \quad (2.13)$$

where a and b are the target values for the discriminator to label the fake and real samples and c is the value that the generator wants the discriminator to believe for the synthesized data, e.g. setting $a = 0$ and $b = c = 1$. It shows that minimizing the objective in (2.13) is equivalent to minimizing the Pearson χ^2 divergence between the two data distributions. LSGAN shows two advantages over regular GANs. First, the images generated by LSGAN are of better quality. Second, it exhibits a more stable and faster convergence during the learning process.

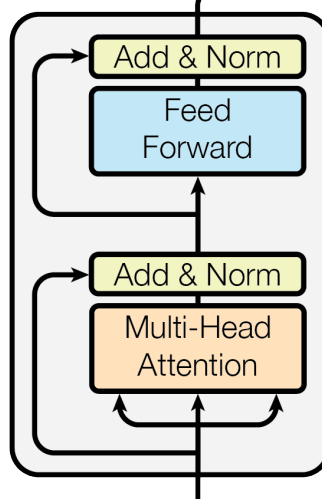
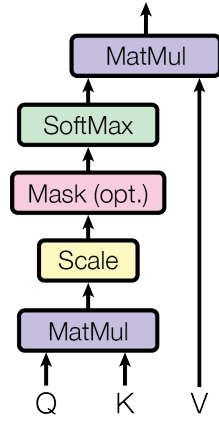


Figure 2.3: Illustration of transformer.

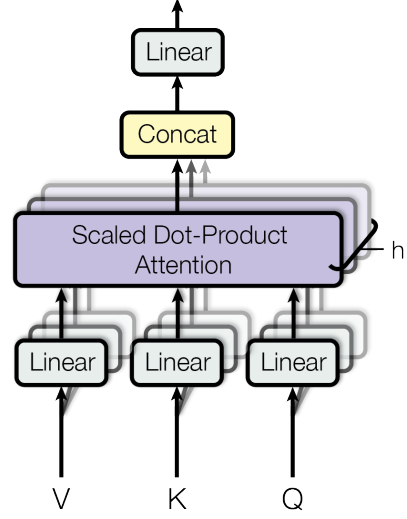
2.2.4 Attention-Based Models

As one of the most important and typical attention-based models, transformer was introduced in [20] as a novel method to model long-range dependencies potentially contained in an input data sequence by learning to pay more attention to relevant features while allowing it to attend to information at different positions, e.g. detection transformers (DETR) [21] for end-to-end object detection and Taming transformers [22] for image synthesis, demonstrating a distinct methodology to perform information exchange from recurrent neural networks. As displayed in Fig. 2.3, the basic transformer block consists of two constituent components, a multi-head attention mechanism and a feed forward network, designed to complement each other. The feed forward network is applied to the output of the multi-head attention and often composed of two linear mappings with a non-linear activation, e.g. ReLU, in between. A residual shortcut is employed around each of the two components, followed by normalization. The multi-head attention, as demonstrated in Fig. 2.4, projects the input queries, keys, and values, packed into matrices Q, K , and $V \in R^{N \times d}$ where d is the dimension of elements, also termed tokens, and N denotes the number of tokens, with different and learnable transforms $W_i^Q, W_i^K \in R^{d \times d_k}$ and $W_i^V \in R^{d \times d_v}$, and independently exhibits scaled dot-product attention h times, hence the subscript i in the projection matrices.

Scaled Dot-Product Attention



Multi-Head Attention

**Figure 2.4:** Illustration of scaled dot-product attention and multi-head attention.

The proposed scaled dot-product attention is used to measure the similarity between two tokens and the token-wise information is aggregated by the combination of the learned sequence representations weighted according to the computed attention scores. As depicted in Fig. 2.4, the attention function is computed with the packed, projected query, key, and value tokens and a softmax applied to obtain the weights on the values:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \quad (2.14)$$

The results from each scaled dot-product attention are concatenated and projected to d dimensions with a transform matrix $W^O \in \mathbb{R}^{(h \cdot d_v) \times d}$, yielding the final output:

$$\begin{aligned} \text{head}_i &= \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \\ \text{MultiHead}(Q, K, V) &= [\text{head}_1, \dots, \text{head}_h]W^O, \end{aligned} \quad (2.15)$$

where $\parallel$ denotes concatenation along the token dimension. The multi-head architecture allows attention-based models to jointly perform exploration in different representation spaces.

Although showing competitive performance with CNN-based models, transformer models have quadratic complexity with respect to the input sequence length, which limits their applications to vision tasks where high-resolution images are often involved and “tokenized” into a long sequence of image patches before being passed to attention modules. A vision transformer was proposed in [23] as an attempt to bridge the gap between natural language processing and image tasks, which attains good results in image classification with a pure transformer applied to sequences of partitioned image patches. However, the applications of such models are limited to high-level tasks where low-resolution image features are involved. A shifted window-based attention model, dubbed Swin Transformer, is introduced in [24] to serve as a general backbone for vision tasks, which adopts a hierarchical structure to capture attentions at different scales and shows computational complexity growing linearly with image size. It confines attention computation within a pre-defined non-overlapping local window and enables cross-window information exchange via a circular window shift scheme, as demonstrated in Fig. 2.5, alternating among stacked transformer layers, which noticeably brings application flexibility and computational efficiency. Swin transformers surpass other attention-based counterparts in vision tasks by a large margin and achieves a better trade-off between execution speed and accuracy [25, 26, 27]. However, the drawback of the shifted window partitioning method is that a cascade of Swin transformer layers are needed to expand the receptive field, which limits its modelling capacity and implies potential for further improvement. To efficiently communicate with other sub-windows, Twins transformer [28] employs an auxiliary token in each local window, which serves as the representative in the computation of global attention. It significantly reduces the complexity by a factor proportional to the partition size while compromising fine-grained token-wise attributes. Despite showing a reduction in computational sources required to train a model, such methods can still be expensive and cumbersome when processing high-resolution images or video sequences.

Concurrently, several extensive works have been introduced to reduce the computational costs by finding an approximation of self-attention [29, 30, 31, 32], but which typically achieve efficiency at the expense of model accuracy. Long-short transformer [33] was proposed to capture distant correlations with dynamic projection and model local context via a short-

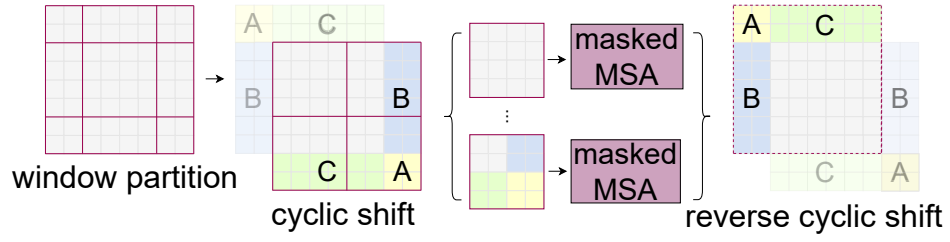


Figure 2.5: Illustration of shifted window partition.

term attention. The two types of attention are combined to perform an efficient attention mechanism with linear complexity for long input sequences. It avoids the prohibitive quadratic time and memory complexities with the modelling of long-range dependencies applied in low dimensions, which inevitably shows a reduction of expressiveness. The computational complexity and memory consumption still pose massive challenges in the respect of application of transformer-based models. It remains an open problem that how to efficiently model long-range dependencies in an effort to retrieve and exploit useful information.

2.3 Compressive Sensing Methods for MRI Reconstruction

The acquisition of MRI data is inherently slow, which restricts its growth in high-resolution imaging and time-critical applications. A common practice to accelerate the acquisition process and reduce the scanning time is to under-sample k -space data. However, the frequency components beyond half the Nyquist-Shannon sampling rate will become indistinguishable from the lower frequencies, which leads to aliasing artifacts in image domain. Therefore, exploring algorithms to recover the missing signals has become an important research direction in modern MR imaging. Several traditional reconstruction methods are reviewed here, and deep learning-based methods are introduced in the next chapters.

One of the crucial breakthroughs in the field of accelerating MRI is the introduction of the compressed sensing (CS) theory [34] applied to restore the under-sampled MR images. In recent years, reconstruction methods based on compressed sensing have made great progress to provide clean images containing distinct informative features for clinical diagnosis, showing

their maturity in this direction [35, 36]. The theme of those approaches is to exploit the data redundancy to reduce the sampling rate required to produce high-quality images. CS methods leverage the sparsity properties of signals to recover data from a subset of measurements where the Shannon sampling criterion is not satisfied. MRI reconstruction can be modelled as an optimization problem:

$$\min_{\mathbf{x}} R(\mathbf{x}) \text{ s.t. } \|\mathbf{Ax} - \mathbf{y}\| < \varepsilon, \quad (2.16)$$

where $\mathbf{A}$ denotes the encoding matrix typically consisting of the transformation to k -space and under-sampling, $\mathbf{y}$ is the observed measurement, $\mathbf{x}$ is the reconstruction, and ε refers to the tolerance relaxing the optimization constraint. The term $R(\mathbf{x})$ denotes the regularization used to presumably capture the signal prior and the constraint condition is adopted to limit the search space to make more faithful results and guarantee the uniqueness of solutions. CS methods normally adopt the sparsity prior as regularization and solve the problem using iterative algorithms [37]. We introduce some classic reconstruction approaches based on compressive sensing in the following.

The assumption that the signals being recovered are compact in image domain inspired the reconstruction model shown below, using the L_1 -norm to depict the image sparsity:

$$\min_{\mathbf{x}} \|\mathbf{x}\|_1 \text{ s.t. } \|\mathbf{Ax} - \mathbf{y}\| < \varepsilon. \quad (2.17)$$

Clean images are obtained by addressing the L_1 minimization problem via some convex relaxation approaches. The reconstruction performance can be improved by modelling the targets in some transformed domain where a compact representation is available to encode data redundancy:

$$\min_{\mathbf{x}} \|\Psi\mathbf{x}\|_1 \text{ s.t. } \|\mathbf{Ax} - \mathbf{y}\| < \varepsilon, \quad (2.18)$$

where Ψ denotes an analysis transformation matrix such that $\Psi\mathbf{x}$ has a sparse representation. It is clear that the special case of selecting Ψ to be the identity map simplifies to (2.17). In practice, the unconstrained form of the minimization problem is achieved by importing a regularization parameter λ to control the strength of the sparsity hypothesis:

$$\min_{\mathbf{x}} \frac{1}{2} \|\mathbf{Ax} - \mathbf{y}\|^2 + \lambda \|\Psi\mathbf{x}\|_1. \quad (2.19)$$

The ill-defined gradient at points with $\Psi\mathbf{x} = \mathbf{0}$ is a crucial issue in finding a solution to the minimization problem, which leads to the exploration of techniques as a new type of convex optimization algorithms. Several popular methods were developed to solve this problem, among which the alternating directional method of multiplier (ADMM) has been extensively applied to compressed sensing MRI reconstruction [38]. By defining an auxiliary variable, the problem in (2.19) is converted into the following form:

$$\begin{aligned} \min_{\mathbf{x}, \mathbf{u}} \quad & \frac{1}{2} \|\mathbf{Ax} - \mathbf{y}\|^2 + \lambda \|\mathbf{u}\|_1 \\ \text{s.t.} \quad & \mathbf{u} = \Psi\mathbf{x}, \end{aligned} \quad (2.20)$$

which gives the ADMM model by dividing (2.20) into several sub-problems as shown below,

$$\begin{aligned} \mathbf{x}^{k+1} &= \arg \min_{\mathbf{x}} \frac{1}{2} \|\mathbf{Ax} - \mathbf{y}\|^2 + \frac{\mu}{2} \|\boldsymbol{\xi}^k + \Psi\mathbf{x} - \mathbf{u}^k\|^2 \\ \mathbf{u}^{k+1} &= \arg \min_{\mathbf{u}} \lambda \|\mathbf{u}\|_1 + \frac{\mu}{2} \|\boldsymbol{\xi}^k + \Psi\mathbf{x}^{k+1} - \mathbf{u}\|^2 \\ \boldsymbol{\xi}^{k+1} &= \boldsymbol{\xi}^k + \mathbf{x}^{k+1} - \mathbf{u}^{k+1}. \end{aligned} \quad (2.21)$$

Each step presented in (2.21) has a closed-form solution, leading to the optimization algorithm summarized below,

$$\begin{aligned} \mathbf{x}^{k+1} &= (\mathbf{A}^H \mathbf{A} + \mu \mathbf{I})^{-1} (\mathbf{A}^H \mathbf{y} + \mu \mathbf{u}^k - \mu \boldsymbol{\xi}^k) \\ \mathbf{u}^{k+1} &= \text{shrink}_1(\boldsymbol{\xi}^k + \Psi\mathbf{x}^{k+1}, \frac{\lambda}{\eta}) \\ \boldsymbol{\xi}^{k+1} &= \boldsymbol{\xi}^k + \mathbf{x}^{k+1} - \mathbf{u}^{k+1}, \end{aligned} \quad (2.22)$$

where $shrink_1$ denotes the soft-threshold operator defined by:

$$shrink_1(x, \tau) = sign(x) \times \max\{|x| - \tau, 0\}. \quad (2.23)$$

In addition to the sparsity constraint, e.g. defined in (2.17) and (2.18), several works adopted the total variation (TV) as the penalty in reconstruction implementations using the difference operation to achieve sparse image representations [39, 40], which has several nice properties, e.g. it is convex and edge-preserving [41]. It is straightforward to revise the reconstruction model defined in (2.19), assuming the following form:

$$\min_{\mathbf{x}} \frac{1}{2} \|\mathbf{Ax} - \mathbf{y}\|^2 + \lambda TV(\mathbf{x}), \quad (2.24)$$

where $TV(\mathbf{x})$ is the total variation penalty defined by the L_1/L_p -mixed norm with $p \in \{1, 2\}$:

$$TV(\mathbf{x}) = \|\nabla \mathbf{x}\|_{1,p} = \sum_q \|\nabla \mathbf{x}(q)\|_p. \quad (2.25)$$

The gradient of the q -th pixel is denoted by $\nabla \mathbf{x}(q) \in R^d$ where d is the signal dimension. With a primal focus on the TV penalty, the constrained optimization is formulated by:

$$\begin{aligned} \min_{\mathbf{x}, \mathbf{u}(q)} \quad & \frac{1}{2} \|\mathbf{Ax} - \mathbf{y}\|^2 + \lambda \sum_q \|\mathbf{u}(q)\|_p \\ \text{s.t.} \quad & \mathbf{u}(q) = \nabla \mathbf{x}(q). \end{aligned} \quad (2.26)$$

The associated ADMM is provided below,

$$\begin{aligned} \mathbf{x}^{k+1} &= \arg \min_{\mathbf{x}} \frac{1}{2} \|\mathbf{Ax} - \mathbf{y}\|^2 + \frac{\mu}{2} \sum_q \|\boldsymbol{\xi}^k(q) + \nabla \mathbf{x}(q) - \mathbf{u}^k(q)\|^2 \\ \mathbf{u}^{k+1}(q) &= \arg \min_{\mathbf{u}} \lambda \|\mathbf{u}(q)\|_p + \frac{\mu}{2} \|\boldsymbol{\xi}^k(q) + \nabla \mathbf{x}^{k+1}(q) - \mathbf{u}(q)\|^2 \\ \boldsymbol{\xi}^{k+1}(q) &= \boldsymbol{\xi}^k(q) + \nabla \mathbf{x}^{k+1}(q) - \mathbf{u}^{k+1}(q). \end{aligned} \quad (2.27)$$

The solution to (2.27) is summarized as follows,

$$\begin{aligned}
\mathbf{x}^{k+1} &= (\mathbf{A}^H \mathbf{A} - \mu \mathbf{I})^{-1} (\mathbf{A}^H \mathbf{y} + \mu \operatorname{div}(\mathbf{u}^k - \boldsymbol{\xi}^k)) \\
\mathbf{u}^{k+1}(q) &= \begin{cases} \operatorname{shrink}_1(\boldsymbol{\xi}^k(q) + \nabla \mathbf{x}^{k+1}(q), \frac{\lambda}{\eta}), & \text{if } p = 1 \\ \operatorname{shrink}_{\text{vec},2}(\boldsymbol{\xi}^k(q) + \nabla \mathbf{x}^{k+1}(q), \frac{\lambda}{\eta}), & \text{if } p = 2 \end{cases} \\
\boldsymbol{\xi}^{k+1}(q) &= \boldsymbol{\xi}^k(q) + \nabla \mathbf{x}^{k+1}(q) - \mathbf{u}^{k+1}(q),
\end{aligned} \tag{2.28}$$

where div is the divergence operator and $\operatorname{shrink}_{\text{vec},2}$ denotes the vector shrinkage defined by:

$$\operatorname{shrink}_{\text{vec},2}(\mathbf{x}, \tau) = \frac{\mathbf{x}}{\|\mathbf{x}\|_2} \max\{\|\mathbf{x}\|_2 - \tau, 0\}. \tag{2.29}$$

Since the successful applications of total variation in MRI reconstruction, many studies were conducted to exploit the prior knowledge with the involvement of the combination of a fidelity term and a TV regularization to improve the accuracy of MRI reconstruction. The work in [42] introduced a composite of a Markov random field model and a TV-based regularization to solve the ill-posed problem using a computationally efficient iterative algorithm. On top of the split Bregman iteration algorithm, the approach proposed in [43] combines TV with two wavelet transformations as regularization to explore the sparsity prior of MR images in various transformed domains.

Although MRI imagery readily satisfies the condition that the signals being reconstructed should be sparse in some transformed space, which allows for successful applications of CS theory [44, 37], the limited expressiveness of handcrafted priors hinders the growth of CS in modern MRI. To address this issue, a dictionary learning-based method [45] was proposed in place of pre-determined sparsifying transforms, which learns the dictionary via the K-SVD algorithm [46] and bridges the gap between compressed sensing for MRI reconstruction and data-driven representation learning approaches. The work in [47] pioneered a distinct blind CS

approach which treats the underlying sparsifying transform as an unknown prior and proposes to simultaneously perform image reconstruction and transform learning with the highly under-sampled measurements. Compared with the dictionary learning model, it has shown to be a faster algorithm yielding promising reconstruction quality.

As another important medical imaging modality, dynamic MRI is used to monitor processes with temporal variations such as cardiac motion. Due to the limitation of scanning speed, a trade-off between the spatial resolution and temporal resolution needs to be obtained, which inevitably leads to highly aliased images when directly applying the inverse Fourier transform to the partially acquired k -space data instead of full sampling. Fortunately, the data redundancies not only are contained in separate spatial frames but also appear along the temporal dimension [48], which can be exploited to recover MR sequences with temporal variations. A model-based method for dynamic MRI, named k - t BLAST/SENSE [48], was introduced to achieve improved spatio-temporal resolution of MRI sequences. An extension of k - t BLAST/SENSE [48], dubbed k - t FOCUSS [49], was demonstrated to be a more general framework with residual encoding to estimate the residual signals of the initial prediction from very limited k - t samples, where a motion estimation and compensation scheme is utilized to significantly sparsify the residue.

Similar to TV-regularized models for MRI reconstruction by making use of CS theory, dynamic total variation [50, 51] was designed to exploit the redundancies in both spatial and temporal domains as an attempt to reduce the noise and artifacts in results. It solves the challenging problem via an accelerated re-weighted least squares algorithm [51, 52] motivated by the structure and properties of the Fourier transform in sparse MRI with the first reconstructed frame as the reference to guide the reconstruction of the successive frames, which shows to speed up the data acquisition procedure and enable an online reconstruction scheme with comparable performance to offline methods in terms of reconstruction accuracy. To improve the capacity of sparsity representation for dynamic MRI, the work in [53] focused on the learning of the sparsifying transform based on the blind compressive sensing ideology [54] where the sparse signal model is assumed to be unknown and jointly learned with the underlying

image being recovered. This method adopts wavelet and gradient sparsity as the regularization term and decomposes the original problem into several sub-problems having a closed-form solution that can be efficiently executed via an iterative variable splitting scheme, which yields improved reconstruction quality and evaluation results as compared to prior methods using fixed sparsifying transforms.

2.4 Deep Learning-Based Methods for MRI Reconstruction

With the advance of deep neural network models in a diversity of vision tasks such as image classification, segmentation, and synthesis, many works studied the potential of deep learning (DL) based approaches for MRI reconstruction [55]. Those methods can be categorized into three types comprising end-to-end DL models, model-based DL frameworks, and k -space based DL models, which are introduced in the following.

2.4.1 End-to-End DL Models

One typical example of end-to-end reconstruction models is to learn the direct mapping from the signals corrupted with aliasing artifacts caused by under-sampling to clean images by supervising a U-net with the ground truth (GT) data [56]. As an effort to produce more realistic details in reconstructions, DAGAN [57] incorporates an adversarial objective in the course of training, also using a VGG [9] network to retrieve more visually appealing results by matching the reconstructed images with the references in the learned deep representation spaces. The pixel-wise losses used to train a reconstruction model, e.g. the mean square error (MSE), mainly focus on low-frequency components and allow small differences lowering the image quality, since reconstruction assessment is conducted in pixel space, leading to excessively blurry and smooth results. To address this issue, the structural similarity index (SSIM) [58] was adopted in [59] to help in preserving structural details in reconstructed images. It also combines the residual-in-residual dense blocks (RRDB) [60] with a U-shaped network structure to benefit from both dense connections and residual shortcuts at multiple levels, showing to better recover

the structural content with finer details. A different approach was proposed in [61] which formulates the reconstruction problem as a deep density prior model, giving rise to:

$$\arg \max_{\mathbf{x}} \log p(\mathbf{x}|\mathbf{y}) = \arg \max_{\mathbf{x}} \{\log p(\mathbf{y}|\mathbf{x}) + \log p(\mathbf{x})\}, \quad (2.30)$$

where the Bayes' theorem is used and $\log p(\mathbf{y})$ is dropped since it is independent of $\mathbf{x}$ in the optimization. It can be rewritten as follows, assuming a Gaussian noise model,

$$\arg \min_{\mathbf{x}} \|\mathbf{Ax} - \mathbf{y}\|^2 - \lambda \log p(\mathbf{x}), \quad (2.31)$$

where $\arg \min$ is adopted to retrieve a solution to minimize the objective, therefore a negative sign is used for the data prior term $\lambda \log p(\mathbf{x})$. The first term in (2.31) corresponds to the reconstruction errors evaluated in the subspace spanned by the encoding matrix, and the second term requires to maximize the log-density of $\mathbf{x}$ representing the prior information about the estimate. A variational auto-encoder [62, 63] was used to learn the density function of an image as the probability prior, showing to yield more promising results and less sensitivity to under-sampling patterns.

2.4.2 Model-Based DL Frameworks

Although it achieved considerably improved image quality by utilizing the learning capacity of end-to-end trained neural networks, the reconstruction results can suffer from undesirable hallucinations, e.g. caused by the adversarial training, or become inconsistent with the observed measurements due to lacking of data consistency (DC) constraints. To overcome this problem, model-based iterative algorithms can be “unrolled” with neural network structures taking the place of some update steps [64, 65], with the intention of retaining the data consistency while leveraging deep architectures to retrieve higher performance in terms of both reconstruction accuracy and image quality.

A general model-based DL framework can be expressed as a cascaded network pipeline with trainable modules and data consistency operations interleaved with each other, where commonly adopted DC operations include replacing the pixel values supported by the sampling pattern with the measurements or mixing the predictions with the observed values via interpolation [64, 66]. It can be summarized below,

$$\begin{aligned}\mathbf{u}^{k+1} &= f_{\theta}(\mathbf{x}^k) \\ \mathbf{x}^{k+1} &= DC(\mathbf{u}^{k+1}; \mathbf{A}, \mathbf{y}),\end{aligned}\tag{2.32}$$

where f_{θ} is a network block parametrized by θ .

Deep neural networks such as CNNs are extensively used in accelerated MRI reconstruction, receiving a lot of attention in the literature due to their enormous potential and excellent performance in the field of vision tasks with images from both the daily life, e.g. image recognition [67, 68], and domain-specific data, e.g. medical image analysis and enhancement [69, 70, 71]. The use of neural networks as the forward model enables a systematically distinct approach from those that do not have a neural network for image recovery, by deriving an unfolded architecture based on optimization algorithms for inverse problems. In practice, CNN-based models with parameters trained on exemplars show to be well suited to capture the image features and account for the imperfections of the MRI data acquisition. We introduce several such reconstruction models below.

Generalizing the TV term used in (2.24) via the Field of Experts model [72], a variational minimization [73] for accelerated MRI reconstruction is derived:

$$\min_{\mathbf{x}} \frac{\lambda}{2} \|\mathbf{Ax} - \mathbf{y}\|^2 + R(\mathbf{x}),\tag{2.33}$$

where $R(\mathbf{x})$ is extended to N terms with filter kernels, given by:

$$R(\mathbf{x}) = \sum_{i=1}^N \langle \Phi_i(\mathbf{K}_i \mathbf{x}), \mathbf{1} \rangle,\tag{2.34}$$

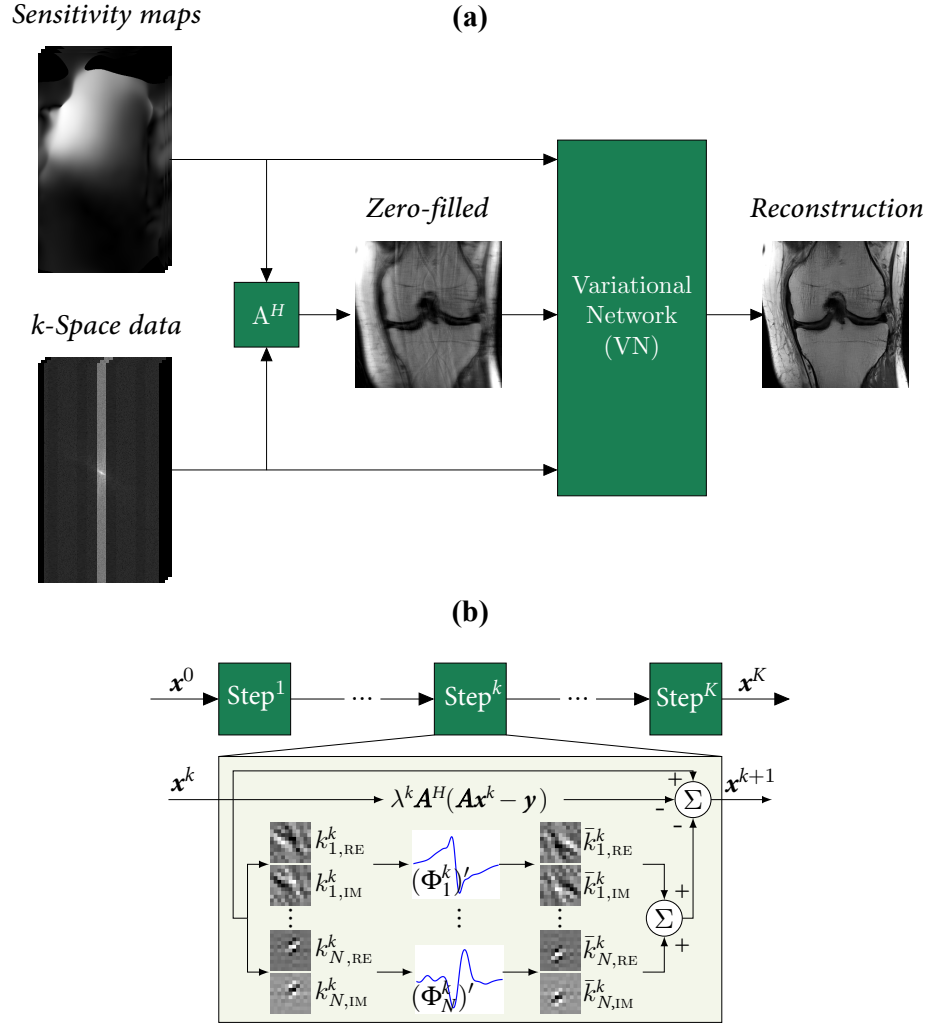


Figure 2.6: Illustration of variational network (VN) architecture: (a) image processing pipeline, and (b) VN structure.

where $\mathbf{K}_i$ models convolution with filter k_i , Φ_i is a non-linear function, $\mathbf{1}$ is a vector consisting of ones, and $\langle, \rangle$ denotes the inner-product. The convolution kernels and parametrization of the non-linear function are all learned from data. As the complex-valued signals are dealt with, separable convolution filters are adopted for the real and imaginary parts. Plugging the Field of Experts model defined in (2.34) into (2.33), the associated iterative algorithm is expressed as below,

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \sum_{i=1}^N (\mathbf{K}_i^k)^T (\Phi_i^k)' (\mathbf{K}_i^k \mathbf{x}^k) - \lambda^k \mathbf{A}^H (\mathbf{A} \mathbf{x}^k - \mathbf{y}), \quad (2.35)$$

where $(\Phi_i^k)'$ denotes the first derivative of Φ_i^k and is parametrized as a weighted combination of Gaussian radial basis functions (RBF), yielding

$$(\Phi_i^k)'(\mathbf{x}(q)) = \sum_{j=1}^{N_w} w_{ij}^k \exp\left(-\frac{(\mathbf{x}(q) - \mu_j)^2}{2\sigma^2}\right), \quad (2.36)$$

where μ_j is uniformly sampled from a predefined range $[-\mu_{\max}, \mu_{\max}]$ and $\sigma = \frac{2\mu_{\max}}{N_w - 1}$. The coil sensitivity maps are precomputed from the center of k -space data via the ESPIRiT algorithm [74] and subsumed within the encoding matrix $\mathbf{A}$ and its adjoint operator $\mathbf{A}^H$ to perform coil projection and sensitivity-weighted combination for a multi-coil MRI setting [73]. The variational network, dubbed VN, is obtained by unrolling the iterations in (2.35) where one network module accounts for one iteration of the reconstruction pipeline, as illustrated in Fig. 2.6. The model is trained with the ground truth data towards an optimal parameter set $\{\lambda^k, \{k_i^k\}_i, \{w_{ij}\}_{i,j}\}_k$ minimizing the reconstruction errors.

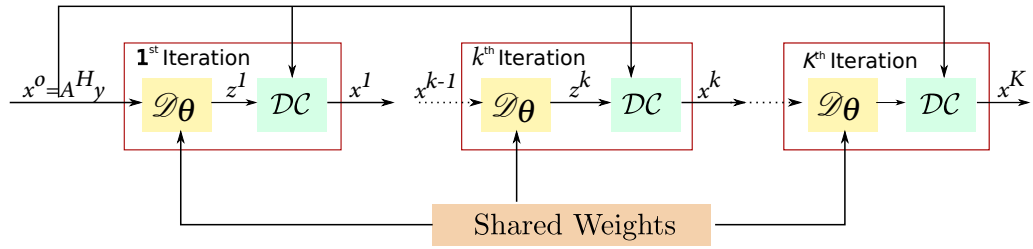


Figure 2.7: Illustration of MoDL architecture.

An CNN-based unrolled approach named MoDL [64] formulates the inverse problem using a CNN to generalize the domain-specific prior knowledge:

$$\mathbf{x}_{rec} = \arg \min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{y}\|^2 + \lambda \|\mathcal{N}_{\theta}(\mathbf{x})\|^2, \quad (2.37)$$

where $\mathcal{N}_{\theta}$ serves as a CNN-based estimator of noise and alias patterns contained in the reconstructed image, leading to a solution that is data consistent and minimally corrupted by noise and aliasing artifacts. The learned $\mathcal{N}_{\theta}(\mathbf{x})$ is expressed by:

$$\mathcal{N}_{\theta}(\mathbf{x}) = \mathbf{x} - \mathcal{D}_{\theta}(\mathbf{x}), \quad (2.38)$$

where $\mathcal{D}_\theta(\mathbf{x})$ gives the clean image with noise and alias patterns removed, relying on the trainable parameters θ . Let $\mathbf{x}^k$ be the reconstructed image at the k -th stage and $\mathbf{x} = \mathbf{x}^k + \Delta\mathbf{x}$, the estimate is updated by:

$$\mathbf{x}^{k+1} = \arg \min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{y}\|^2 + \lambda \|\mathbf{x} - \mathcal{D}_\theta(\mathbf{x}^k + \Delta\mathbf{x})\|^2. \quad (2.39)$$

The second term in (2.39) can be approximated as shown below when applying an alternation minimization solver to the regularized formulation,

$$\|\mathbf{x} - \mathcal{D}_\theta(\mathbf{x}^k + \Delta\mathbf{x})\|^2 \approx \|\mathbf{x} - \mathcal{D}_\theta(\mathbf{x}^k)\|^2 + \|\mathcal{J}(\mathbf{x}^k)\Delta\mathbf{x}\|^2, \quad (2.40)$$

where $\mathcal{J}(\mathbf{x}^k)$ denotes the Jacobian matrix at point $\mathbf{x}^k$ which is used to expand the term $\mathcal{D}_\theta(\mathbf{x}^k + \Delta\mathbf{x})$ via the Taylor series:

$$\mathcal{D}_\theta(\mathbf{x}^k + \Delta\mathbf{x}) \approx \mathcal{D}_\theta(\mathbf{x}^k) + \mathcal{J}(\mathbf{x}^k)\Delta\mathbf{x}. \quad (2.41)$$

The term $\|\mathcal{J}(\mathbf{x}^k)\Delta\mathbf{x}\|^2$ tends to diminish in the vicinity of $\mathbf{x}^k$, i.e. $\Delta\mathbf{x} \rightarrow 0$. The minimization in (2.39) can be written as follows with $\mathbf{z}^{k+1} = \mathcal{D}_\theta(\mathbf{x}^k)$

$$\mathbf{x}^{k+1} = \arg \min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{y}\|^2 + \lambda \|\mathbf{x} - \mathbf{z}^{k+1}\|^2. \quad (2.42)$$

The alternating minimization algorithm is then obtained:

$$\mathbf{z}^{k+1} = \mathcal{D}_\theta(\mathbf{x}^k) \quad (2.43a)$$

$$\mathbf{x}^{k+1} = (\mathbf{A}^H \mathbf{A} + \lambda \mathbf{I})^{-1} (\mathbf{A}^H \mathbf{y} + \lambda \mathbf{z}^{k+1}). \quad (2.43b)$$

The update in (2.43b) accounts for the DC operation and can be analytically calculated depending on the under-sampling pattern in k -space as follows,

$$(\mathbf{F}\mathbf{x}^{k+1})(q) = \begin{cases} (\frac{\mathbf{y} + \lambda \mathbf{F}\mathbf{z}^{k+1}}{1 + \lambda})(q) & \text{if sampling at the } q\text{-th position} \\ (\mathbf{F}\mathbf{z}^{k+1})(q) & \text{otherwise,} \end{cases} \quad (2.44)$$

where $\mathbf{F}$ denotes the Fourier transform matrix. Since the network only needs to learn the image redundancies, a simpler structure with fewer parameters suffices to have good results in practice. A strategy is to share the denoiser $\mathcal{D}_\theta$ at each iteration, which significantly reduces the amount of network parameters. The pipeline architecture is displayed in Fig. 2.7. The main benefits of the framework used in MoDL are the end-to-end training and the reduction in parameter size, which achieve improved reconstruction quality as compared to the methods without using end-to-end training or weight sharing [75, 76, 73, 77].

Similar to MoDL, the applicability of CNN models were evaluated in [77] for accelerated MRI reconstruction, where CNN blocks are concatenated to form a cascaded architecture with interleaved DC blocks. The output of the previous block is refined by the successive CNN with the network blocks iterating between intermediate reconstruction and data consistency steps. The resulting networks can be regarded an extension of dictionary learning (DL) methods [45] for compressive MRI where each CNN block expresses the learned representation. The residual connection [78] is used in CNN blocks, which performs identity mapping with several layers skipped over and summing the output prediction to the input to obtain performance gains from significantly increased network depth. By incorporating the fidelity prior knowledge into the learning process, it showcases that CNN-based models produce the results with more accurately recovered anatomical structures, while CS-based methods are incapable of ensuring such performance. To improve the reconstruction quality, a deep residual framework was used in [79] with data consistency applied to the initial predictions, ensuring that the reconstructed image conforms to the k -space observations. To solve the compressed sensing MR imaging problem, a model-based unrolled network was built up based on the approximate message passing (AMP) algorithm [80], where the CNN blocks are embedded into the optimization loop with the data consistency layer following each CNN and AMP blocks to obtain a faster convergence and higher reconstruction performance. A CNN-based cascaded framework termed BLIPS was proposed in [81] to refine the image details initially produced by a traditional dictionary

learning-based method as a way of combining the adaptive sparsity-based prior with the deep learned knowledge. Although involving more complicated learning steps, the improvements in reconstruction results imply the benefits of the learned features, suggesting the potential of investigating the complementary priors in conventional sparsity-based reconstruction.

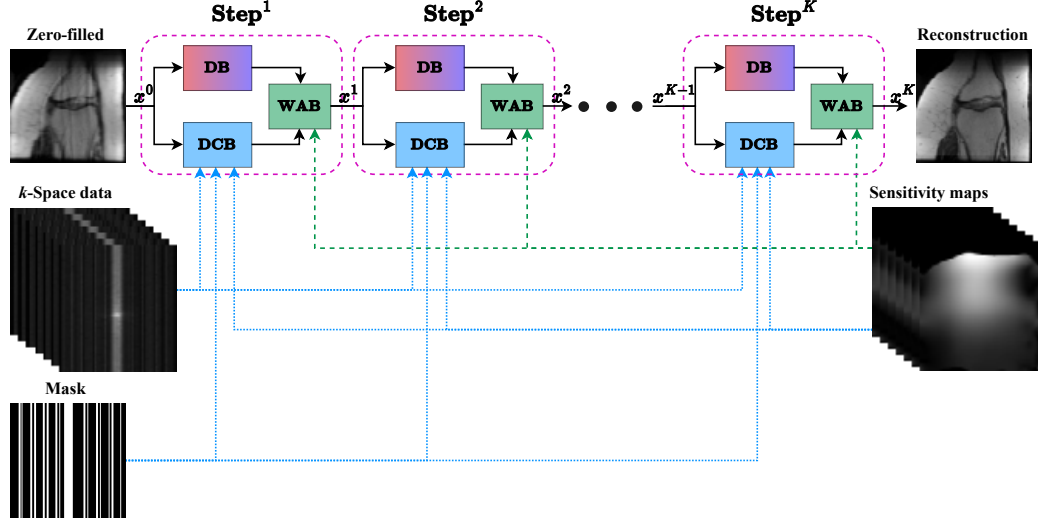


Figure 2.8: Illustration of VS-Net architecture. Each update step consists of three components: the denoising block (DB), the data consistency block (DCB), and the weighted average block (WAB).

The model named VS-Net [82] applies the DC operation to multi-coil MRI scenarios in a point-wise and analytical form, which yields a simpler implementation compared to works in [73, 64, 79] using gradient descent optimization and proximal approaches. Different from ADMM-Net [83] which is based on the augmented Lagrangian method (ALM), VS-Net adopts a variable splitting model using the penalty function method while embedding the iteration steps of an optimization process into a deep network framework. Following the convention of CS theory, the reconstruction of under-sampled images is estimated by solving the optimization problem below under a multi-coil MRI scenario:

$$\min_{\mathbf{x}} \sum_{i=1}^{n_c} (\lambda \|\mathbf{DFS}_i \mathbf{x} - \mathbf{y}_i\|^2) + R(\mathbf{x}), \quad (2.45)$$

where $\mathbf{y}_i$ denotes the i -th coil measurement, $\mathbf{S}_i$ is the sensitivity map, $\mathbf{D}$ denotes the under-sampling operation, and n_c is the number of coils. Two auxiliary variables $\mathbf{u}$ and $\mathbf{m}_i$ are introduced to efficiently optimize (2.45), converting the problem into the following form,

$$\begin{aligned} \min_{\mathbf{x}, \mathbf{u}, \mathbf{m}_i} \sum_{i=1}^{n_c} (\lambda \|\mathbf{D}\mathbf{F}\mathbf{m}_i - \mathbf{y}_i\|^2) + R(\mathbf{u}) \\ \text{s.t. } \mathbf{x} = \mathbf{u}, \mathbf{S}_i\mathbf{x} = \mathbf{m}_i. \end{aligned} \quad (2.46)$$

The first constraint $\mathbf{x} = \mathbf{u}$ is introduced to decompose $\mathbf{x}$ in regularization from the data fidelity term, enabling a denoising process in the formulation of the solution to (2.45). The second condition $\mathbf{S}_i\mathbf{x} = \mathbf{m}_i$ allows the decoupling of $\mathbf{S}_i\mathbf{x}$ from $\mathbf{D}\mathbf{F}\mathbf{S}_i\mathbf{x}$, which avoids dense matrix inversion involved in calculations. The equivalent form of (2.46) is obtained by adding the constraints back to the model with two Lagrange multipliers α and γ :

$$\min_{\mathbf{x}, \mathbf{u}, \mathbf{m}_i} \sum_{i=1}^{n_c} (\lambda \|\mathbf{D}\mathbf{F}\mathbf{m}_i - \mathbf{y}_i\|^2) + R(\mathbf{u}) + \alpha \|\mathbf{u} - \mathbf{x}\|^2 + \gamma \sum_{i=1}^{n_c} (\|\mathbf{m}_i - \mathbf{S}_i\mathbf{x}\|^2). \quad (2.47)$$

By alternating the optimization of $\mathbf{x}, \mathbf{u}, \mathbf{m}_i$, a solution to the multi-variable minimization is derived as follows,

$$\mathbf{u}^{k+1} = \text{denoiser}(\mathbf{x}^k) \quad \text{DB} \quad (2.48a)$$

$$\mathbf{m}_i^{k+1} = \mathbf{F}^H ((\lambda \mathbf{D}^H \mathbf{D} + \gamma \mathbf{I})^{-1} (\gamma \mathbf{F} \mathbf{S}_i \mathbf{x}^k + \lambda \mathbf{D}^H \mathbf{y}_i)) \quad \text{DCB} \quad (2.48b)$$

$$\mathbf{x}^{k+1} = (\alpha \mathbf{I} + \gamma \sum_{i=1}^{n_c} \mathbf{S}_i^H \mathbf{S}_i)^{-1} (\alpha \mathbf{u}^{k+1} + \gamma \sum_{i=1}^{n_c} \mathbf{S}_i^H \mathbf{m}_i^{k+1}) \quad \text{WAB.} \quad (2.48c)$$

The original problem is thus transformed into a denoising model as presented in (2.48a) and two subsequent equations. Due to the nature of diagonal matrix inversion, the computations involved in (2.48b) and (2.48c) can be efficiently executed in an element-wise manner. The consistency between the k -space data and the reconstruction in image domain is enforced by (2.48b), while the computations in (2.48c) can be regarded as a weighted average of the results of the first two steps. It is appropriate to establish the unrolled pipeline of VS-Net

with three constituent components: a denoising block (DB), a data consistency block (DCB), and a weighted average block (WAB), as displayed in Fig. 2.8, each of which addresses one corresponding equation in (2.48). In VS-Net, the regularization term is implicitly exploited in a CNN-based estimate process as compared to the VN [73] model that learns regularization with a set of filter kernels. It enables the use of more advanced and flexible CNN architectures to improve the model expressiveness for MRI reconstruction.

2.4.3 Plug-and-Play Methods

Plug-and-play (PNP) methods have the common ideology that a denoising forward pass is plugged as one step into an optimization-inspired algorithm for inverse problems [76]. Compared to traditional CS-based methods, PNP methods can capture rich image priors with application of a trained denoiser and show to be more suitable for compressive MRI recovery [84]. A plug-and-play framework was introduced in [85] to achieve an efficient and reliable treatment for MRI reconstruction by incorporating an automatic feedback mechanism into an iterative optimization algorithm. Concurrently, generative prior learned from a dataset of high-quality images was exploited to obtain accelerated MRI reconstruction, by searching the range of a pre-trained generative model, which takes the place of the learned modules plugged into a model-based optimizer, to provide a solution [86]. However, the aliasing artifacts have distinct patterns specific to the sampling schemes, which potentially imposes challenges on PNP methods. It can be an attractive avenue to explore the potential of such methods to produce high-quality MR images on par with end-to-end trained deep learning models.

2.4.4 k -Space Prior Based DL Models

An alternative direction of methods for MRI reconstruction is to solve the ill-posed inverse problems via k -space data interpolation. A representative method termed GRAPPA [87] was proposed to estimate the missing pixels directly in k -space by utilizing the fully sampled data in the auto-calibration signal (ACS) region. Let (x, y) denote the sampled k -space point and

$s(x, y, i)$ the corresponding value from the i -th coil. A neighborhood $N(x, y)$ is defined as a set of acquired points around (x, y) across all coil frames, whose values form a vector $\mathbf{s}_N(x, y)$. The value of the skipped point in position (x', y') of the i -th coil around (x, y) is computed by:

$$s(x', y', i) = \mathbf{G}(x' - x, y' - y, i) \mathbf{s}_N(x, y), \quad (2.49)$$

where $\mathbf{G}$ represents the linear convolution weights which are calibrated by solving a least squares problem utilizing the acquired ACS data.

To increase the learning capacity of interpolation methods, a CNN-based model named RAKI [88] was proposed whose premise is to replace the linear kernels used in GRAPPA with a series of CNN blocks consisting of a convolutional layer followed by a non-linear activation. It learns to interpolate skipped k -space points from adjacent pixels after being trained on subject-specific ACS data. The CNN architecture used in RAKI can be designed in an application-specific manner to increase the execution efficiency [89]. A hybrid linear and non-linear approach termed rRAKI for k -space deep learning was introduced in [90] to achieve scan-specific MRI reconstruction, where non-linear CNN blocks work with a linear convolution via a residual connection to perform k -space interpolation. The linear aspect presents a basic estimate of the missing values while a reduction in artifacts and noise coming from the linear part is provided by the non-linear module. It was shown that the explicit split between the linear and non-linear parts improves the quality of images with sharper details compared to conventional methods.

Although performing a data-free approach, RAKI and rRAKI require separate CNN training for the estimate of each slice and are less suitable for real-time applications. In contrast to RAKI, GrappaNet [91] utilizes a U-net to predict the local GRAPPA kernels from the ACS lines and jointly trains it with CNN models to refine the initial estimate from the k -space interpolation on a dataset, enabling a single-pass reconstruction forward. However, such methods often demand a fully sampled region larger than that typically acquired to estimate the kernels for interpolating missing samples, which limits the acceleration rate one can achieve. A reconstruction framework termed AUTOMAP [92] emerged as an automated transform

architecture unifying the recovery of corrupted images with the learning of the mapping from the data encoding space to the target domain. It is exhibited as a data-driven supervised task simultaneously learning to transform the measurements from k -space to image domain and remove the aliasing artifacts. Similarly, an interpolation network was introduced in [93] which adopts a U-net to predict the missing k -space pixels and maps the estimate to image space via the inverse Fourier transform.

2.4.5 DL Methods for Multi-Coil MRI Reconstruction

In recent years, parallel imaging (PI) was widely adopted to speed up MRI scanning, which acquires multiple views of the underlying anatomy simultaneously from different coil sensors. It exploits redundancies contained in multi-coil frames to reduce aliasing artifacts caused by under-sampling, complementary to compressed sensing methods which shorten the acquisition process by reconstructing images with fewer measurements. The combination of the two types of imaging tactics enables faster MRI scans than is available using either method alone. To provide high-quality images for accurate clinical diagnosis, the ill-posed inverse problems were cast as an optimization problem which inspired the use of neural networks in each iteration of model-based algorithms to estimate the clean image by leveraging the deep priors learned from training data. One limitation of such methods is the assumption that the forward process involving parallel image unfolding and under-sampling in k -space should be known. This assumption can be unrealistic for the multi-coil MRI reconstruction problem since the forward process is formulated with the presence of unknown coil sensitivity maps as shown in (2.45) and (2.48), which drives the need of explicitly modelling the sensitivity correlations as a necessary step prior to recovering signals.

2.4.5.1 Reconstruction Using Model-Based Sensitivity Map Estimation

Many methods were motivated to accommodate the need for the estimation of coil sensitivity maps in optimization-inspired reconstruction frameworks, among which ESPIRiT [74] was widely adopted to estimate the coil sensitivities via a model-based iterative algorithm from the ACS center corresponding to low frequencies encoding high-level image attributes such as object shapes and positions. The data consistency step defined in (2.43b) at each iteration of MoDL [64] uses the precomputed sensitivity maps for the forward process expressed in the form of the encoding matrix $\mathbf{A}$ and its adjoint operation. The data consistency block of VS-Net [82] requires to simulate the image projection onto coil sensors with the estimated sensitivity maps to exploit the physical constraint of parallel MR imaging, as presented in (2.48b). The sensitivity maps are also used to combine the coil-specific intermediate predictions in the weighted average block as shown in (2.48c). Despite achieved success in the literature, such estimation methods have a need for a fully sampled k -space region which is required to be sufficiently large for accurate estimation, hence limiting the maximum acceleration rate that can be attained [94, 95].

2.4.5.2 End-to-End Estimation Method for Multi-Coil MRI

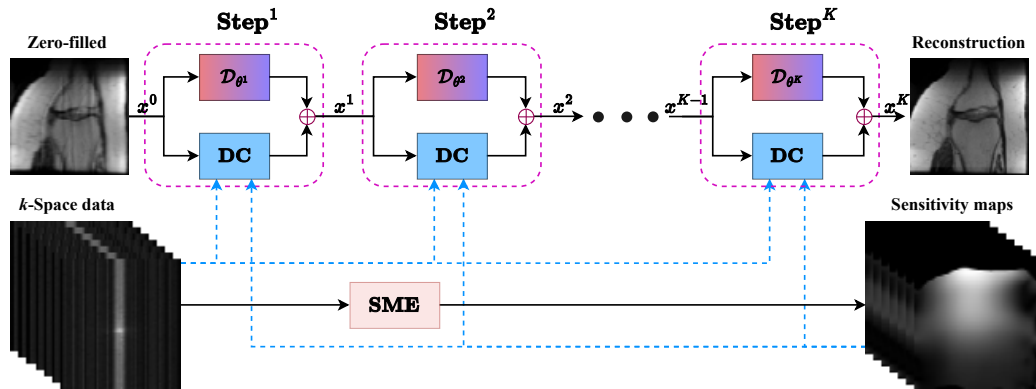


Figure 2.9: Illustration of end-to-end variational network (E2E-VarNet) which consists of three constituent elements: the refinement CNN model $\mathcal{D}_{\theta^k}$, the data consistency block (DC), and the sensitivity map estimation module (SME).

The work in [95] introduced a different approach which extends the variational network [73] for multi-coil MRI reconstruction to an end-to-end learning scheme, termed E2E-VarNet, to address the forward process issue posed by the unknown coil sensitivities, where the sensitivity maps are fully learned by neural networks in conjunction with image reconstruction. Motivated by the development of deep learning-based approaches to MRI reconstruction [96, 97, 73], it built up a deep network framework consisting of iterative update modules applied to the intermediate prediction from the previous reconstruction module, obtaining the iteration step shown below:

$$\begin{aligned} \mathbf{x}^{k+1} &= \mathbf{x}^k - \eta \mathbf{A}^H (\mathbf{A} \mathbf{x}^k - \mathbf{y}) + \mathcal{D}_{\theta^k}(\mathbf{x}^k) \\ &= \underbrace{\mathbf{x}^k - \eta \sum_{i=1}^{n_c} \mathbf{S}_i^H \mathbf{F}^H (\mathbf{D} \mathbf{F} \mathbf{S}_i \mathbf{x}^k - \mathbf{y}_i)}_{\text{data consistency}} + \overbrace{\mathcal{D}_{\theta^k}(\mathbf{x}^k)}^{\text{refinement}}, \end{aligned} \quad (2.50)$$

which is composed of two elements: the data consistency operation and the estimate refinement using a network parametrized by θ^k . The weight η can be a pre-determined hyperparameter or jointly learned from data. The update equation in (2.50) involves the use of the sensitivity maps that cannot be accurately estimated using traditional methods from very limited ACS lines, which is a necessary case to achieve high acceleration factors. To alleviate this issue, E2E-VarNet learns to predict the sensitivity maps using a parametric sensitivity map estimation module (SME) as part of the network. It employs a U-net G_ϕ in the SME module which takes as input the central region of k -space, dubbed $\mathbf{y}_i^{ACS} = \mathbf{M}^{ACS} \mathbf{y}_i$ with $\mathbf{M}^{ACS}$ zeroing out values except for the ACS lines, and maps it into the corresponding sensitivity map, yielding:

$$\mathbf{S}_i = \text{Norm}(G_\phi(\mathbf{F}^H \mathbf{y}_i^{ACS})), \quad (2.51)$$

where *Norm* normalizes the sensitivity maps, ensuring that the property presented in the following equation holds,

$$\sum_{i=1}^{n_c} \mathbf{S}_i^H \mathbf{S}_i = \mathbf{I}. \quad (2.52)$$

The framework uses the zero-filled as the initial estimate, given by (2.53), and is trained by minimizing the SSIM [58] loss between the final reconstruction and ground truth images,

$$\mathbf{x}^o = \sum_{i=1}^{n_c} \mathbf{S}_i^H \mathbf{F}^H \mathbf{y}_i. \quad (2.53)$$

The architecture of E2E-VarNet is presented in Fig. 2.9. Compared to model-based estimation methods whose reconstruction performance tends to drop with a smaller number of low-frequency lines due to a degradation in the quality of estimated sensitivity maps, it was demonstrated by experiments that E2E-VarNet is able to overcome this limitation and generate pleasing results with fewer ACS lines. Similar methods were proposed in [98, 99] which utilize a CNN-based network to estimate the sensitivity maps from the under-sampled k -space data measured by different coils and recover the image by recursively applying a reconstruction block. The work in [99] proposed to accumulate the outputs of all reconstruction blocks via skip-connections to produce the final result. Comparative experiments demonstrated that the framework combining coil sensitivity estimation and reconstruction refinement recovers MR images more accurately than traditional PI methods.

Although the sensitivity maps are suggested to be estimated in E2E-VarNet using a deep learning-based model, they are not jointly updated along with the reconstruction iterations. A network framework jointly performing sensitivity estimation and image reconstruction was proposed in [100] to regularize the inverse problem, e.g. the one formulated in (2.45), as follows:

$$\min_{\mathbf{x}, \{\mathbf{S}_i\}} \sum_{i=1}^{n_c} (\|\mathbf{D}\mathbf{F}\mathbf{S}_i\mathbf{x} - \mathbf{y}_i\|^2) + R(\mathbf{x}) + \sum_{i=1}^{n_c} R(\mathbf{S}_i). \quad (2.54)$$

The regularization terms with respect to the reconstructed image $\mathbf{x}$ and the sensitivity map $\mathbf{S}_i$ of the i -th coil are defined as follows,

$$\begin{aligned} R(\mathbf{x}) &= \lambda_I \|\mathbf{x} - \mathcal{D}_I(\mathbf{x})\|^2 + \lambda_F \|\mathbf{x} - \mathbf{F}^H \mathcal{D}_F(\mathbf{F}\mathbf{x})\|^2 \\ R(\mathbf{S}_i) &= \lambda_S \|\mathbf{S}_i - \mathcal{D}_S(\mathbf{F}^H \mathbf{y}_i)\|^2, \end{aligned} \quad (2.55)$$

where $R(\mathbf{x})$ learns to exploit the deep priors in both image domain and k -space via two neural networks $\mathcal{D}_I$ and $\mathcal{D}_F$, which has proven to increase the reconstruction performance [101], and the regularization of $\mathbf{S}_i$ is expressed as similar to that of $\mathbf{x}$ except that the deep prior model is structured based on a single forward call of $\mathcal{D}_S$ with the under-sampled data as input. The manipulations of coil frame pertaining to sensitivity maps have two basic operations: the projection of an MR image onto each coil and the sensitivity-weighted combination. We formulate the former one and the latter can be naturally derived via the adjoint operation. The coil projection can be regarded as the spatially-varying scaling of an image, which is given by:

$$\mathbf{S}_{\text{vect},i} \odot \mathbf{x} = \mathbf{S}_i \mathbf{x} = \mathbf{x}_{\text{mat}} \mathbf{S}_{\text{vect},i}, \quad (2.56)$$

where $\odot$ denotes the element-wise multiplication, and $\mathbf{S}_{\text{vect},i}$ and $\mathbf{x}_{\text{mat}}$ are respectively the vector and matrix forms of $\mathbf{S}_i$ and $\mathbf{x}$. It says that the sensing process of MRI at parallel coils can be written as the matrix-vector product of the corresponding sensitivity map and the underlying image as commutative, which is equivalent to the element-wise multiplication of the two vectorized variables. The minimization specified in (2.54) is divided into two sub-problems as shown below, alternating between optimizing the two variables,

$$\begin{aligned} \mathbf{x}^{k+1} = \arg \min_{\mathbf{x}} \sum_{i=1}^{n_c} (\|\mathbf{D}\mathbf{F}\mathbf{S}_i^k \mathbf{x} - \mathbf{y}_i\|^2) + \lambda_I^{k+1} \|\mathbf{x} - \mathcal{D}_I^{k+1}(\mathbf{x})\|^2 \\ + \lambda_F^{k+1} \|\mathbf{x} - \mathbf{F}^H \mathcal{D}_F^{k+1}(\mathbf{F}\mathbf{x})\|^2 \end{aligned} \quad (2.57a)$$

$$\mathbf{S}_{\text{vect},i}^{k+1} = \arg \min_{\mathbf{S}_{\text{vect},i}} \|\mathbf{D}\mathbf{F}\mathbf{x}_{\text{mat}}^{k+1} \mathbf{S}_{\text{vect},i} - \mathbf{y}_i\|^2 + \lambda_S^{k+1} \|\mathbf{S}_{\text{vect},i} - \mathcal{D}_S^{k+1}(\mathbf{F}^H \mathbf{y}_i)\|^2, \quad (2.57b)$$

which hold the matrix-vector multiplication in the encoding phase of the data fidelity term by resorting to the relationship presented in (2.56) while reversing the role of $\mathbf{x}$ and $\mathbf{S}_i$. The reconstruction regularization terms in (2.57a) respectively operate in the MR image domain and in the k -space domain, while $\mathcal{D}_S$ in (2.57b) regularizes each estimated sensitivity map. For simplicity, the transformations between the vector and matrix forms of the underlying variables are omitted when both of them are involved in the same equation. A gradient descent method can be used to solve the foregoing least squares problems, obtaining an iterative algorithm

shown below,

$$\begin{aligned}
\mathbf{x}^{k+1} &= \mathbf{x}^k - 2\mu_{k+1} \left[\sum_{i=1}^{n_c} ((\mathbf{S}_i^k)^H \mathbf{F}^H (\mathbf{D}\mathbf{F}\mathbf{S}_i^k \mathbf{x}^k - \mathbf{y}_i)) + \lambda_I^{k+1} (\mathbf{x}^k - \mathcal{D}_I^{k+1}(\mathbf{x}^k)) \right. \\
&\quad \left. + \lambda_F^{k+1} (\mathbf{x}^k - \mathbf{F}^H \mathcal{D}_F^{k+1}(\mathbf{F}\mathbf{x}^k)) \right] \\
\mathbf{S}_{\text{vect},i}^{k+1} &= \mathbf{S}_{\text{vect},i}^k - 2\nu_{k+1} [(\mathbf{x}_{\text{mat}}^{k+1})^H \mathbf{F}^H (\mathbf{D}\mathbf{F}\mathbf{x}_{\text{mat}}^{k+1} \mathbf{S}_{\text{vect},i}^k - \mathbf{y}_i) + \lambda_S^{k+1} (\mathbf{S}_{\text{vect},i}^k \\
&\quad - \mathcal{D}_S^{k+1}(\mathbf{F}^H \mathbf{y}_i))]. \quad (2.58)
\end{aligned}$$

The resulting reconstruction pipeline, termed Joint-ICNet, is established based on the iterations in (2.58) where the image reconstruction and the estimated sensitivity maps are jointly updated. Concurrently, an unrolled framework named Deep J-Sense [102] was introduced to refine the image in combination with the update of coil sensitivities, yielding the iterative algorithm as follows,

$$\begin{aligned}
\mathbf{S}_{\text{vect},i}^{k+1} &= \tilde{\mathbf{S}}_{\text{vect},i}^k - \mu (\tilde{\mathbf{x}}_{\text{mat}}^k)^H \mathbf{F}^H (\mathbf{D}\mathbf{F}\tilde{\mathbf{x}}_{\text{mat}}^k \tilde{\mathbf{S}}_{\text{vect},i}^k - \mathbf{y}_i) \\
\tilde{\mathbf{S}}_{\text{vect},i}^{k+1} &= \mathcal{D}_S(\mathbf{S}_{\text{vect},i}^{k+1}) \quad (2.59)
\end{aligned}$$

$$\begin{aligned}
\mathbf{x}^{k+1} &= \tilde{\mathbf{x}}^k - \nu \sum_{i=1}^{n_c} (\tilde{\mathbf{S}}_i^{k+1})^H \mathbf{F}^H (\mathbf{D}\mathbf{F}\tilde{\mathbf{S}}_i^{k+1} \tilde{\mathbf{x}}^k - \mathbf{y}_i) \\
\tilde{\mathbf{x}}^{k+1} &= \mathcal{D}_I(\mathbf{x}^{k+1}), \quad (2.60)
\end{aligned}$$

where the individual CNN-based models are shared by all iterations to reduce aliasing artifacts in the recovered image and correct the initially generated sensitivity maps with the interleaved model-based data consistency layers. Experimental results demonstrated improved image quality and comparative robustness to less ACS lines for sensitivity estimation, when implementing the refinement of sensitivity maps along with image reconstruction.

2.4.6 DL Methods for Dynamic MRI Reconstruction

As an imaging modality used to monitor the organ motions, dynamic MRI requires both high spatial and temporal resolutions, imposing different challenges associated with image reconstruction problems from static MRI. Deep learning-based models have witnessed immense progress in MRI reconstruction, which can be divided into two categories. One class of methods utilize the generalization of deep neural networks across exemplar data to learn the mapping from the under-sampled images with strong aliasing artifacts to high-quality clean reconstructions end-to-end. Due to the lack of domain-specific knowledge, a large amount of samples is demanded to achieve an optimal solution. In dynamic imaging scenarios, the data acquisition is particularly expensive, which conspires to exacerbate the situation and limit the model performance. The second type of methods establish a regularized optimization with some domain-specific prior to formulate the reconstruction problems and derive a deep cascade of neural network models by unfolding the iterative minimization procedure. They have shown potential to attain a high performance design, which is trained with fewer data, to recover MR images [103]. We introduce the optimization-based unrolling methods and other types of models for preserving dynamic MRI sequences in the following.

2.4.6.1 Unrolled Neural Networks for MRI Reconstruction

Among recent research trends, sparsity-based regularization is commonly adopted in the modelling of the ill-posed inverse problems. The associate iterative solvers are unrolled to provide an end-to-end reconstruction pipeline, where the underlying sparsity characteristic of the target images is expressed in the form of a parametric function [104, 105]. The sparsifying transformations learned from a dataset have demonstrated to provide more accurate and flexible sparse representations and therefore yield improved reconstruction performance, as compared to handcrafted transforms. Nevertheless, such sparsity-based priors are not specifically designed to mine the time-varying features contained in dynamic MRI sequences and can be less suited to recover the dynamic content. Different from 2-dimensional (2D) convolutions,

3-dimensional (3D) convolutional networks expand the receptive field in both spatial and temporal domains with 3D kernels, and are thus capable of simultaneously exploiting both spatial and temporal correlations. A deep cascade of 3D convolutional neural networks can be used to preserve dynamic MR images from the under-sampled signals in an accelerated acquisition procedure [106, 107, 108]. However, 3D convolutional layers introduce more parameters than 2D convolutions and immensely increase computational complexity [109, 110], which implies the need of efficient and effective networks particularly designed for the reconstruction of MRI sequences.

As a substitute for the sparsity prior of signals, which is normally expressed as the L_1 norm, total variation (TV) was proposed in compressed sensing MRI to retrieve the localized image sparsity enforcing piece-wise smoothed results [39, 40, 111, 112]. TV-based methods were extended to deal with the staircase effect of TV via high-order derivatives, obtaining the total generalized variation (TGV) prior which has proven to better recover edges and generate smoother images [113]. As a function of the spatio-temporal signals, TGV was used to model the sparsity of the moving content in dynamic MRI [114]. The method termed k - t TGV-TD [115] combines 3D total generalized variation (3D-TGV) with tensor decomposition (TD) to implement the sparsity regularization within and between individual temporal frames while making use of the redundancies contained in MRI sequences, which shows to diminish the staircase artifacts in reconstructed images by TV-based methods and yield improved reconstruction performance.

A class of unrolling approaches for dynamic MRI reconstruction were motivated by those works which couple the dependencies among consecutively sampled time frames into the regularized optimization problems. They divide the underlying image features in dynamic sequences into the time-coherent background and the time-varying dynamics. The static or slowly changing background is modelled via a norm-based minimization representing the underlying low-rank prior, which can be solved by low-rank decomposition of time sequences with a singular value threshold (SVT) [116, 117]. The dynamic content is normally expressed with the sparsity prior. Combining the sparsity regularization with the low-rank prior, LS-

Net [118] proposed a 3D CNN-based model to capture the sparsity property and a low-rank layer to achieve a reduction in the rank with a soft threshold optimized during the training phase, as opposed to an empirically selected parameter for SVT. Concurrently, a learned sparse transform was introduced in SLRNet [119] which adopts an ISTA-Net [104] with a soft threshold to address the low-rank constrained reconstruction problem. Due to the limited representation capacity of low rank, SVT-based methods are prone to neglect small components and lose subtle details, resulting in blurred reconstructions.

2.4.6.2 Recurrent Neural Networks for Dynamic MRI

Recurrent neural networks are specifically adopted for dynamic MRI reconstruction using the fantastic capacity of RNN models to transmit dependencies emanating from both the neighboring and distant references via a hidden state encoding the extracted representative information. Thanks to the update of the hidden state during iterations, RNN models can learn to keep useful information from the previous steps and “forget” irrelevant features, making it intrinsically suitable for the dynamic MRI process.

A bidirectional convolutional recurrent neural network, dubbed CRNN [120], was used in dynamic MRI reconstruction which demonstrated promising results in terms of modelling the organ motions from the under-sampled data via the forward and backward propagation of hidden information. Motivated by the breakthrough of dynamic MRI reconstruction using CRNN-based models, k - t -NEXT [121] was proposed to leverage spatio-temporal redundancies in image domain and enable the exploitation of complementary information by alternating the recurrence steps of CRNNs between x - f and image spaces. To fully make use of the complementary information contained in time sequences, CTFNet [122] employs two CRNN models which respectively perform recurrence over iterations and time dimension for images in x - t domain and over iterations for the representations mapped to x - f domain by the Fourier

transform applied to the temporal axis. The work introduced in [123] proposed a recurrent network framework with recurrence over time and iterations on top of an alternating procedure of a model-based optimization algorithm, which embeds ConvGRU [15] models in an encoding-decoding structure to enable the connections of data at different representation scales.

Although RNN-based reconstruction methods have attained encouraging performance, the hidden state transmitted to communicate information can struggle to encode complex motions and the propagation of the dynamic content can become less effective when the evolution of hidden state is dominated by a few inputs, which possibly leads to unbalanced historical accumulation [124]. The recurrent manner of RNN models can be less suited to process multiple frames in parallel, which often results in large memory caching [125, 124].

2.4.6.3 Motion-Guided Dynamic MRI Reconstruction Networks

In dynamic MRI sequences, the images sampled at consecutive time steps provide highly relevant information to depict consistent spatial deformation of the anatomy. Motivated by the ideology of motion flow estimation in computer vision tasks [126, 127], it was suggested to exploit the correlations among different temporal frames in the form of optical flow-based constraint to benefit MRI reconstruction [49, 128, 129, 130]. In a flow-based motion compensation diet where the misaligned neighboring frames are warped and incorporated to assist in the recovery of the current image, the model performance can be affected by the quality of motion estimation, e.g. it can be sensitive to image brightness. Traditional methods for motion flow prediction are often computationally expensive, which is prohibitive for real-time application of modern MRI.

To address those issues, several methods were proposed to query neural network models to predict the displacement fields delineating the potential organ movements between frames, which has shown to be more accurate and robust than conventional optical flow-based estimation approaches. GRAFT [131] proposed a neural network to learn the spatial deformation and

align a group of references towards the target frame to extend a model-based reconstruction procedure to dynamic scenes. Although it was demonstrated to improve the performance of conventional optimization methods, the quality of reconstructed images is still limited by the handcrafted priors of CS methods.

Recent works proposed to incorporate motion flow estimation to achieve motion-guided reconstruction compensation for dynamic MRI reconstruction while utilizing the learning capacity of deep network structures to improve the accuracy of predicting the underlying motion displacement. A motion compensation model was introduced in MODRN [123] to utilize the neighboring frames for the refinement of the initial prediction. It warps two adjacent frames by a differentiable sampling layer [132] with the estimated spatial deformation from a U-shape FlowNet [133] to reinforce motion consistency, and supervises the motion estimation network with a photometric loss [134]. To improve the performance of motion flow-based reconstruction models, an end-to-end training pipeline was used in [135], which jointly optimizes the reconstruction modules and the CNN-based optical flow estimation network, hence enabling the propagation of the gradients of motion-guided reconstruction to all model components. Compared to deformable convolutions [136], where convolutional layers are endowed with the capacity to retrieve useful information from the reference features aligned via offset estimation and spatial interpolation, motion-guided reconstruction models involve more parameters and have limited design flexibility. Moreover, the final result can be susceptible to inaccuracy flow-based alignment and devastatingly impacted by errors and aliasing artifacts in the initial reconstruction, causing degraded reconstruction performance [124, 137].

2.5 Conclusion

A primary introduction of different neural network architectures is presented in this chapter, followed by a review of compressed sensing methods using model-based optimization algorithms to solve the ill-posed inverse problems. An introduction of deep learning-based methods is given to review the advances of neural network models in the field of MRI reconstruction. In the following chapters, our proposed reconstruction approaches and their contributions are introduced.

High-Fidelity MRI Reconstruction Using Deep Neural Networks

3.1 Introduction

Magnetic resonance imaging provides a non-invasive and versatile diagnostic tool to image the anatomy of human body. Owing to the lack of ionizing radiation, as opposed to computed tomography (CT), it is applicable to a wide range of patients and has revolutionized medical imaging and radiology since its advent in the last century. MRI leverages a strong magnetic field and a radio wave pulse to create images by collecting and analyzing distinct signals emitted from the nuclei of the target, which suffers from a long data acquisition procedure and exhibits a low imaging modality, obstructing its application and development in time-critic practice. Parallel imaging is one of the widely used techniques to accelerate the acquisition process, which recovers a clean image from the under-sampled measurements [87] by utilizing the sensitivity correlations among different frames sensed by parallel coils. However, it can be burdened with complex and expensive equipment devices and the strong aliasing artifacts contained in images due to sub-Nyquist data sampling in k -space are difficult to remove using traditional reconstruction methods.

Compressed sensing (CS) methods were proposed to solve the ill-posed inverse problems by modelling the prior knowledge of data being recovered from a subset of k -space measurement values in the form of signal sparsity in image domain [36] or in some transformed space [138, 139, 140, 45]. Although providing an effectual treatment to reduce aliasing arti-

facts, CS achieves MRI reconstruction by solving nonlinear optimization via iterative model-based algorithms which are often computationally expensive and time-consuming, and relies on the pre-determined prior of target data which can be challenging to hold in real-world scenarios. Meanwhile, the handcrafted sparsity prior has limited capacity to represent complex structures and may restrict the achievable image quality. Those factors constrain the development of CS approaches in modern MRI.

In recent years, deep neural network models witnessed great breakthroughs in visual tasks and demonstrated outstanding ability to extract useful information from exemplar data [8, 9, 141, 60, 75]. Several works exploit the powerful representation of neural networks to model the image prior learned from training datasets to recover the under-sampled signals, yielding improved reconstruction performance as compared to traditional CS methods [142, 57, 79]. To obtain fine-grained MRI reconstruction from under-sampled k -space measurements in way to achieve accelerated acquisition process, an attentive selection generative adversarial network, termed ASGAN, is devised here to perform large-field contextual information integration and attention selection to encourage realistic local details in generated images, showing through experiments to attain high-quality reconstructions from highly corrupted images.

As a direct inversion model, ASGAN learns the mapping from the under-sampled measurement to a clean image in an end-to-end manner, which ignores the k -space data constraint and can produce inconsistent image features leading to degraded reconstruction performance. Contrasting with deep learning-based models which are directly trained on exemplar data to generate the target image, we introduce a novel network framework to pursue high-fidelity reconstruction, which is derived from a regularized optimization to incorporate domain-specific knowledge into the image recovery procedure to encourage more faithful results. To retrieve useful information from the spatial regions of high-resolution features, an adaptive spatial attention selection module (ASASM) is designed to efficiently capture contextual cues and perform data-adaptive kernel prediction to increase the diversity of attention patterns without introducing huge computational complexity. A deep data consistency block (DCB) is proposed to leverage the k -space condition into the reconstruction framework while avoiding

the bottleneck structure in conventional data consistency methods, which is demonstrated by experiments to impose a deleterious effect on the model performance. We show that the proposed deep cascaded network model, with the ASASM and DCB embedded in a multi-level densely connected architecture adopted to enable the transmission and reuse of features of different scales, outperforms other deep learning-based methods in terms of image quality and reconstruction accuracy and yields high-fidelity reconstruction under aggressive acceleration rates.

Contrasting with the attention methods applied to model the interdependencies among the spatial regions of the feature representations [143, 144, 145], channel-wise attention mechanism [66, 146], which controls the information propagation by recalibrating feature maps with the adaptively predicted kernels, was proposed as an attempt to circumvent the increased computational overheads. Although demonstrating an efficient implementation of attention mechanism, channel-wise attention methods ignore the useful information hidden in the spatial dimensions, which can limit the network capacity and lead to degraded reconstruction performance. To extract and incorporate the spatial cues potentially containing rich information of the underlying structures being imaged, a novel region-guided channel-wise attention network is devised based on the unrolling of an optimization algorithm, which has the ability to implicitly learn and leverage the spatial semantics to increase the diversities of channel-based attention mechanism. It is demonstrated through experiments that the suggested attention method achieves performance improvements in MRI reconstruction and surpasses other competing models qualitatively and quantitatively.

3.2 Methodology

3.2.1 Problem Formulation

The MRI data acquisition can be described as the process of taking samples in k -space, which is intrinsically slow and time-consuming. One acceleration strategy is to take a subset of measurements following a sampling pattern, which involves the data encoding process in k -space as shown below:

$$\mathbf{y} = \mathbf{D}\mathbf{F}\mathbf{x},$$

where $\mathbf{x}$ denotes the fully sampled image, also termed the reference, $\mathbf{y}$ denotes the measurement, $\mathbf{F}$ refers to the Fourier transform matrix, and $\mathbf{D}$ is the data encoding matrix depicting the under-sampling pattern. The under-sampled data contain strong aliasing artifacts when being transformed to image domain owing to the violation of the Nyquist-Shannon sampling theorem. To remove the aliasing artifacts and recover clean images, the ill-posed inverse problem is posed as a regularized optimization as shown below,

$$\min_{\mathbf{x}} \lambda \|\mathbf{D}\mathbf{F}\mathbf{x} - \mathbf{y}\| + R(\mathbf{x}), \quad (3.1)$$

where R denotes a regularization term and λ is a parameter controlling the strength of regularization. Compressed sensing methods normally express the regularization in the form of the sparsity prior of the target image and find the solution to the optimization problem via iterative algorithms, which are often computationally expensive due to the calculations of the gradients and unsuited to modern MRI.

3.2.2 Attentive Selection Generative Adversarial Network (ASGAN)

To provide an end-to-end approach to efficiently solve the ill-posed reconstruction problem in (3.1), a deep neural network, termed attentive selection generative adversarial network (ASGAN), is designed as a direct inversion model, which is trained to remove the aliasing artifacts in the under-sampled signals and recover the target images of high-quality. It extracts contextual cues to improve the image quality by modelling large-field spatial dependencies in feature spaces via a novel large-field contextual feature integration module, and performs a coarse to fine image recovery to achieve fine-grained reconstruction at low sampling rates.

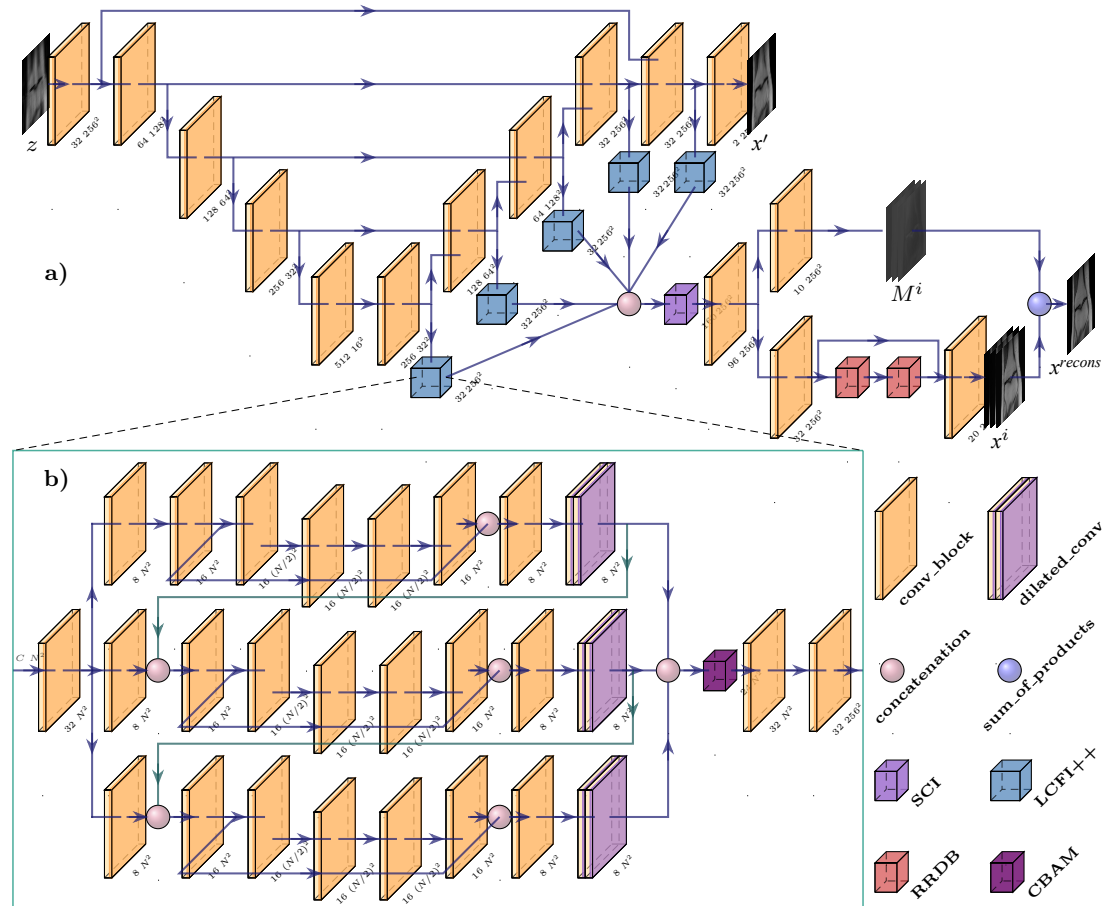


Figure 3.1: Overview of the ASGAN architecture: a) the reconstruction is performed in two stages, where a U-net is adopted to initially recover the under-sampled image and produce the coarse reconstruction $\mathbf{x}'$ from the zero-filled $\mathbf{z}$. The final reconstruction $\mathbf{x}^{recons}$ is generated using an attention selection mechanism with the image and attention maps, termed $\mathbf{x}^i$ and $\mathbf{M}^i$, constructed from the collection of features extracted at multiple decoding levels, and b) deep de-aliasing large-field contextual feature integration (LCFI++).

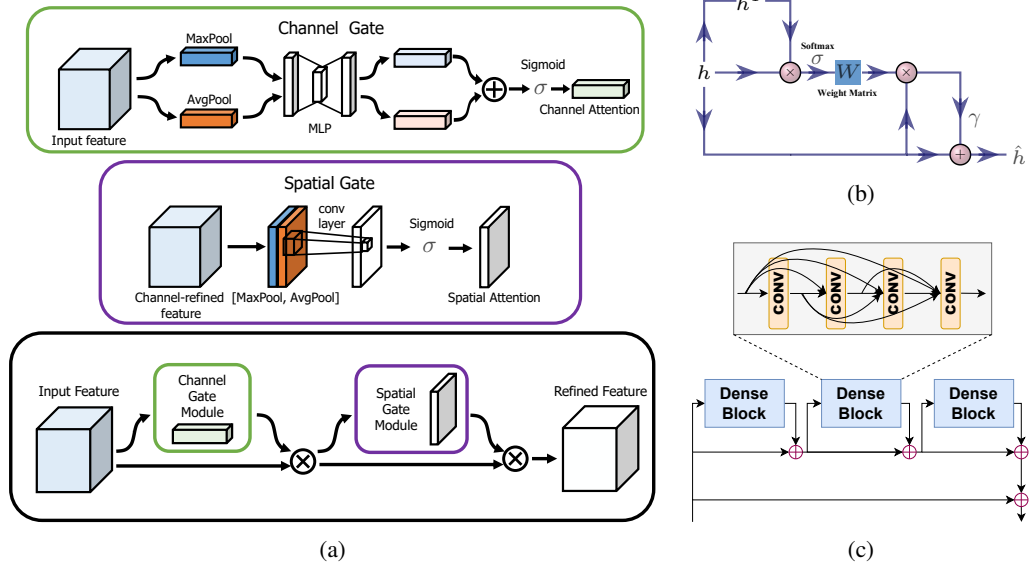


Figure 3.2: Illustration of network modules: a) convolutional block attention module (CBAM), b) self-channel interaction (SCI), c) residual-in-residual dense block (RRDB).

3.2.2.1 Multi-Level Feature Extraction

We propose to reconstruct a clean image from the highly aliased input in two stages, as demonstrated in Fig. 3.1. At the first stage, a U-shaped network (U-net) takes the zero-filled $\mathbf{z} = \mathbf{F}^H \mathbf{y}$ as input, which is expressed by two channels to accommodate complex-valued data, and produces the coarse reconstruction, termed $\mathbf{x}'$, by exploiting the excellent feature extraction ability of the encoding-decoding structure [10]. To preserve the information encoded in the deep feature representations and facilitate the final prediction, the extracted features at different decoding levels are collected and refined for the reconstruction at the second stage.

3.2.2.2 Deep De-aliasing Large-Field Contextual Feature Integration (LCFI++)

The features generated by the decoder of the U-net are refined as a way to provide more useful information. Motivated by the network design in [147] which proposed to capture long-range dependencies among feature maps to enhance the model capacity, a deep de-aliasing large-field contextual feature integration module, dubbed LCFI++, is devised to capture contextual information using a skip-connected parallel structure. As illustrated in Fig. 3.1 (b), the input

features are passed to three parallel branches for refinement where an encoding-decoding structure and a dilated convolutional layer are adopted to expand the receptive field and explore information across large regions. The outputs from each branch are concatenated and fused by a convolutional block attention module (CBAM) [148, 149], illustrated in Fig. 3.2 (a), which controls the contributions of feature maps via a cascade of channel and spatial gates.

3.2.2.3 Multi-Channel Attention Selection Reconstruction

A multi-channel attention selection mechanism is used here to produce the ultimate reconstruction, which potentially performs more sophisticated image synthesis by exploiting a larger generation space, as suggested in [150]. The predicted features spanning from a range of feature scales, as demonstrated in Fig. 3.1, are refined by the devised LCFI++, and aggregated using a self-channel interaction (SCI) block [150], depicted in Fig. 3.2 (b), which models the interdependencies among feature maps by suppressing the similarities and highlighting the distinct and complementary informative attributes.

As shown in Fig. 3.1, the output from the SCI block containing informative cues of various representation levels is used to produce a set of image and attention maps, dubbed $\mathbf{x}^i$ and $\mathbf{M}^i$, which formulates an approximate model ensemble learning method to promote diversity of modelling image features by exploring a larger generation space. Residual-in-residual dense blocks (RRDB) [60], illustrated in Fig. 3.2 (c), are employed in the generation of image maps to improve the model capacity. The final outcome $\mathbf{x}^{recons}$ is produced by combining the image and attention maps with the number N_a of image-attention pairs practically set to 10, yielding:

$$\mathbf{x}^{recons} = \sum_{i=1}^{N_a} \mathbf{M}^i \odot \mathbf{x}^i, \quad (3.2)$$

where $\odot$ denotes the element-wise multiplication.

3.2.2.4 Objective function

Let $\mathbf{x}^{GT}$ be the fully sampled reference and the reconstruction loss is computed as the combination of a pixel-wise metric and a structure-based index as shown below,

$$L_{rec} = (1 - \alpha)(L_1(\mathbf{x}^{recons}, \mathbf{x}^{GT}) + \beta \sum_{i=1}^{N_d} L_1(\mathbf{M}^i \odot \mathbf{x}^i, \mathbf{M}^i \odot \mathbf{x}^{GT})) + \alpha L^{SSIM}(\mathbf{x}^{recons}, \mathbf{x}^{GT}) + \gamma L_1(\mathbf{x}', \mathbf{x}^{GT}), \quad (3.3)$$

where L_1 denotes the L_1 -norm distance, L^{SSIM} is the structural similarity index metric termed SSIM, and the parameters are set to be $\alpha = 0.4$, $\beta = 0.05$, and $\gamma = 0.2$, where α is selected to balance the strengths of pixel-wise metric and structure-based evaluation, a small value is assigned to β to encourage the attention selection mechanism to produce consistent results, and γ is comparatively small since it only supervises the coarse reconstruction. To encourage the reconstructed image to agree with the measurement, we compute a consistency loss by applying the data encoding process to the reconstruction and comparing with the k -space observations:

$$L_{cyc} = \|\mathbf{y} - \mathbf{D}\mathbf{F}\mathbf{x}^{recons}\|_1. \quad (3.4)$$

Generative adversarial networks (GAN) [16] have achieved great breakthroughs in generating photo-realistic images, by exploiting a discriminator to distinguish real data from synthetic counterparts and a generator trained to fool the discriminator. We adopt a least squares GAN (LSGAN) in the training phase, which has shown to prevent the saturation problem as compared to vanilla GAN [19]. The adversarial loss is computed by

$$L_{adv}^D = \min_{\psi} E[(D_{\psi}(\mathbf{x}^{GT}) - b)^2] + E[(D_{\psi}(\mathbf{x}^{recons}) - a)^2] \\ L_{adv}^G = \min_{\theta} E[(D_{\psi}(\mathbf{x}^{recons}) - c)^2], \quad (3.5)$$

where θ and ψ refer to the learnable parameters of the reconstruction model and the discriminator, and the label values are selected to be $b = c = 1$ and $a = 0$. To enhance the visual quality of reconstructed images, a perceptual loss is utilized, which measures the differences between the generated image and the ground truth in feature spaces of a deep neural network, often a VGG pipeline [151, 57], providing more consistent representations with the human visual judgement:

$$L_{vgg} = \sum \|f_{vgg}^i(\mathbf{x}^{recons}) - f_{vgg}^i(\mathbf{x}^{GT})\|_2^2 + \|f_{gram}^i(\mathbf{x}^{recons}) - f_{gram}^i(\mathbf{x}^{GT})\|_2^2, \quad (3.6)$$

where f_{vgg}^i denotes the feature representation of the i -th layer in VGG and f_{gram}^i refers to the Gram matrix [151].

The perceptual loss defined in 3.6 is used to take the perceptual similarity in consideration and encourage more visually appealing results by exploiting the deep representations of a VGG encoder model which are learned from enormous examples. Although the training procedure of such a model is based on natural images, the VGG encoder learns a set of general representations which are consistent with the human visual system and have proven useful to enhance the quality of MRI reconstruction [57].

The total objective function is given by:

$$L = E_{\{\mathbf{x}^{recons}, \mathbf{x}^{GT}\}} [\lambda_{rec} L_{rec} + \lambda_{cyc} L_{cyc} + \lambda_{adv} L_{adv} + \lambda_{vgg} L_{vgg}]. \quad (3.7)$$

Considering that the reconstruction loss should dominate the training process to guarantee the success of reconstruction and the adversarial loss should be small to ensure training stability, the parameters are empirically selected to $\lambda_{rec} = 10$, $\lambda_{cyc} = 0.5$, $\lambda_{adv} = 0.01$, and $\lambda_{vgg} = 1$.

3.2.3 Adaptive Spatial Attention Selection Network (ASASNet)

Although effective in reconstructing high-quality MR images, deep neural networks, e.g. AS-GAN introduced above, which are directly trained on exemplar data to learn the mapping from under-sampled images to clean targets, can generate unwanted artifacts and lead to inferior reconstruction performance due to the lack of informative knowledge specific to MRI domain. We devise an optimization-inspired reconstruction framework which explicitly enforces the k -space consistency and therefore has the capacity to leverage domain-specific knowledge into the reconstruction pipeline to encourage more faithful results. The domain-specific knowledge is incorporated in the form of the unrolling of regularized optimization algorithms, enabling us to develop a deep cascaded adaptive spatial attention selection network, termed ASASNet, for image recovery. In the following, we give the details of the design of the unrolled reconstruction model which is characterized by three novel components: adaptive spatial attention selection module, deep data consistency block, and multi-level dense connections, which are subsequently introduced.

3.2.3.1 Unrolled Reconstruction Framework

With an auxiliary variable $\mathbf{u}$ imported into (3.1), the regularized optimization problem is converted into the following minimization,

$$\min_{\mathbf{x}, \mathbf{u}} \lambda \|\mathbf{D}\mathbf{F}\mathbf{x} - \mathbf{y}\|^2 + R(\mathbf{u}), s.t. \mathbf{x} = \mathbf{u}, \quad (3.8)$$

whose unconstrained form is given with a Lagrangian multiplier α by:

$$\min_{\mathbf{x}, \mathbf{u}} \lambda \|\mathbf{D}\mathbf{F}\mathbf{x} - \mathbf{y}\|^2 + R(\mathbf{u}) + \alpha \|\mathbf{x} - \mathbf{u}\|^2. \quad (3.9)$$

The problem defined in (3.9) is split into two sub-problems alternating the minimization with respect to $\mathbf{u}$ and $\mathbf{x}$ as follows,

$$\mathbf{u}^{k+1} \leftarrow \min_{\mathbf{u}} R(\mathbf{u}) + \alpha \|\mathbf{x}^k - \mathbf{u}\|^2 \quad (3.10a)$$

$$\mathbf{x}^{k+1} \leftarrow \min_{\mathbf{x}} \lambda \|\mathbf{D}\mathbf{F}\mathbf{x} - \mathbf{y}\|^2 + \alpha \|\mathbf{x} - \mathbf{u}^{k+1}\|^2. \quad (3.10b)$$

We exploit the representation capacity of deep neural networks to model the regularization in (3.10a), which learn the image prior from training datasets. The closed-form solution to (3.10b) can be obtained by taking the derivative with respect to $\mathbf{x}$, which realizes the data consistency (DC) constraint on the intermediate prediction. The iterative algorithm is provided as follows,

$$\mathbf{u}^{k+1} \leftarrow f_{\theta_k}(\mathbf{x}^k) \quad (3.11a)$$

$$\mathbf{x}^{k+1} \leftarrow \underbrace{\mathbf{u}^{k+1} - \eta \mathbf{F}^H \mathbf{D}^H (\mathbf{D}\mathbf{F}\mathbf{u}^{k+1} - \mathbf{y})}_{\text{DC operation}}, \quad (3.11b)$$

where f_{θ_k} refers to the k -th subnetwork of the reconstruction pipeline and $\eta = \frac{\alpha}{\lambda + \alpha}$. As shown in (3.11), the estimate of the target image is refined by a neural network model while the intermediate prediction is corrected with the k -space measurement via a DC operation.

3.2.3.2 Deep Data Consistency Block (DCB)

Incorporating the k -space condition into network pipelines can provide considerable improvements in MRI reconstruction [95, 82], by “correcting” the predictions with the observed measurements to ensure data consistency. As illustrated in Fig. 3.3 (a), a typical DC operation is expressed in the form of the update step in (3.11b) which can be rewritten with the feature

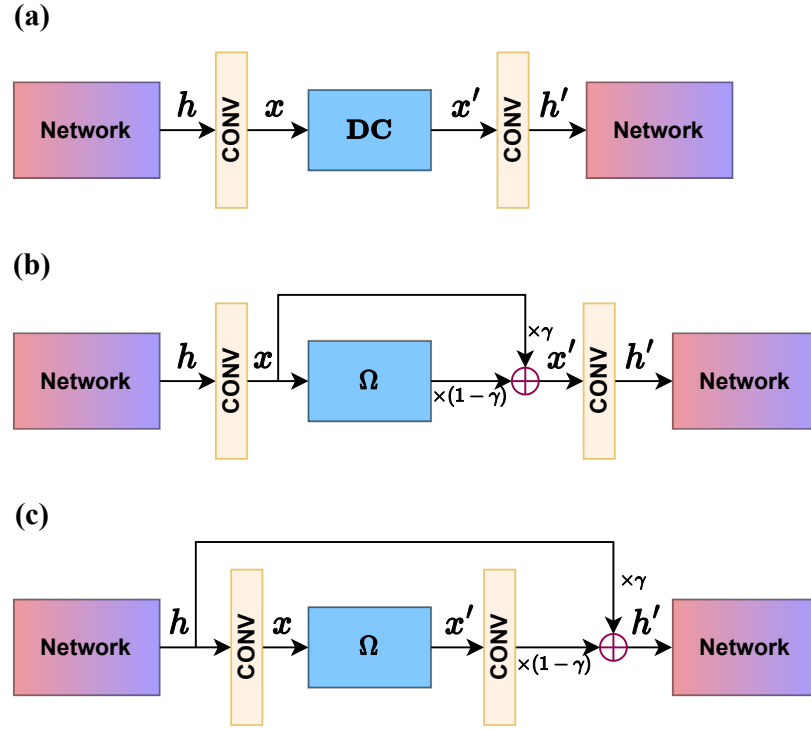


Figure 3.3: Illustration of data consistency (DC) approaches: a) conventional data consistency block embedded in a reconstruction pipeline, b) generalized data consistency expression using the hard DC operator Ω , c) deep data consistency block (DCB).

volume estimate, termed h , from the previous network model as follows,

$$\mathbf{x} = f_1(h) \quad (3.12a)$$

$$\mathbf{x}' = \mathbf{x} - \eta \mathbf{F}^H \mathbf{D}^H (\mathbf{D} \mathbf{F} \mathbf{x} - \mathbf{y}) : \text{DC operation} \quad (3.12b)$$

$$h' = f_2(\mathbf{x}'), \quad (3.12c)$$

where f_1 and f_2 are two parametric functions respectively connecting to the previous and successive subnetworks and can be treated as the learnable mappings between feature space and image domain, with the object of reducing and expanding feature channels to deal with complex-valued signals often represented using two feature channels. h' denotes the feature volume as input to the next reconstruction iteration. For simplicity, the single-coil configuration is adopted here for the expression of DC methods. Another type of DC methods update the predicted values at sampled k -space positions by interpolating with the measurements while

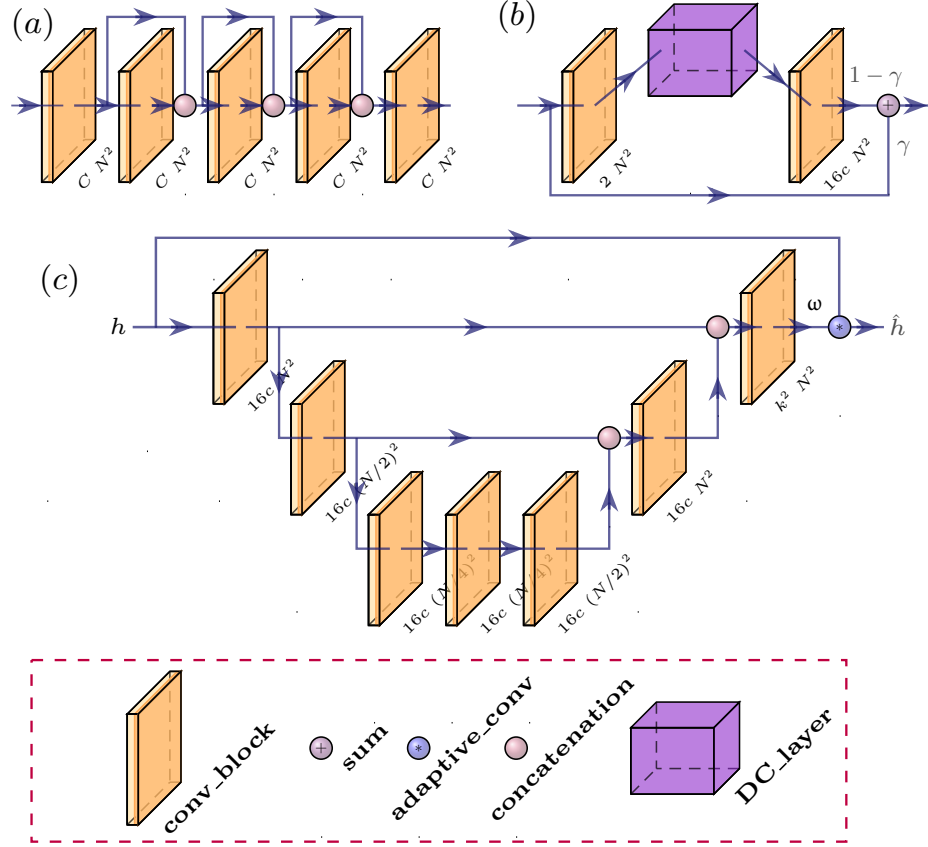


Figure 3.4: Illustration of basic modules: a) densely connected block, b) deep data consistency block (DCB), and c) adaptive spatial attention selection module (ASASM). $C N^2$ indicates the channel size and spatial resolution of output features, where the spatial size is assumed to be $N \times N$, for simplicity.

preserving the estimated values at missing positions [64, 66], yielding:

$$\mathbf{x}' = \mathbf{F}^H((\mathbf{I} - \mathbf{D})\mathbf{F}\mathbf{x} + \frac{\mathbf{D}\mathbf{F}\mathbf{x} + v\mathbf{y}}{1 + v}). \quad (3.13)$$

Mathematically, the DC operations in (3.12b) and (3.13) are equivalent to each other and can be expressed in a general form defined with the hard DC operator Ω shown below,

$$\Omega(\mathbf{x}; \mathbf{D}\mathbf{F}, \mathbf{y}) = \mathbf{F}^H((\mathbf{I} - \mathbf{D})\mathbf{F}\mathbf{x} + \mathbf{y}). \quad (3.14)$$

The general data consistency expression illustrated in Fig. 3.3 (b) is given by:

$$\mathbf{x}' = \gamma\mathbf{x} + (1 - \gamma)\Omega(\mathbf{x}; \mathbf{D}\mathbf{F}, \mathbf{y}), \quad (3.15)$$

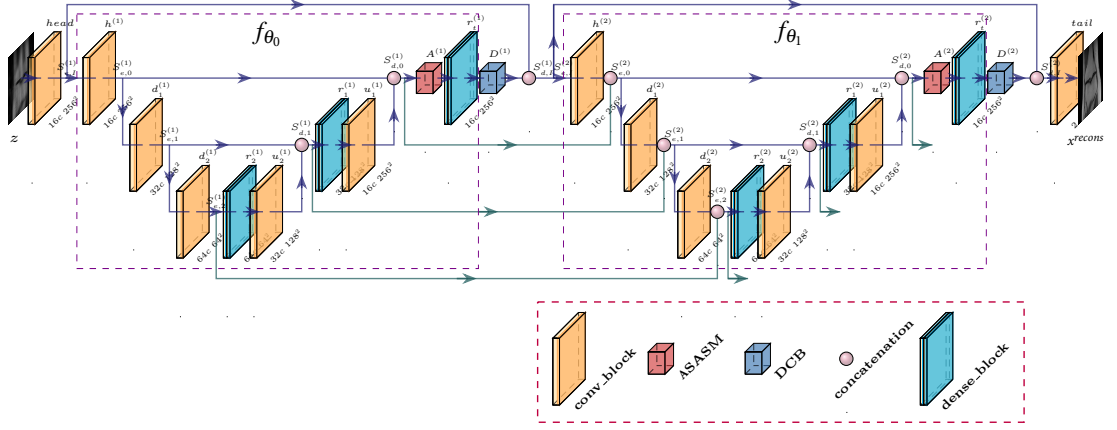


Figure 3.5: Overview of the proposed adaptive spatial attention selection network (ASASNet) architecture. The input zero-filled is processed by a deep cascade of neural network models after being mapped to feature space by a head convolutional layer. Each reconstruction iteration has a U-shaped subnetwork featuring an adaptive spatial attention selection module (ASASM) and multi-level dense connections, and a deep data consistency block (DCB) as a prominent attribute. The final reconstruction is computed with the collection of all previous predictions by a tail convolutional block. The feature shape is indicated by $C N^2$ corresponding to the channel and spatial sizes with the input assumed to have a spatial size of 256×256 .

where the parameters satisfy $\gamma = 1 - \eta = \frac{1}{1+\nu}$.

The foregoing k -space data enforcement methods inevitably require to transform the feature maps coming from a deep network model into image domain, which potentially disposes of useful information owing to the feature channel reduction which constructs a bottleneck structure when embedding data consistency layers into an unrolled reconstruction framework. Based on this observation, we devise a deep data consistency block (DCB), which is demonstrated in Fig. 3.3 (c) and Fig. 3.4 (b), to avoid the bottleneck design via a skip-connected structure utilizing the advantage of residual learning at feature level. The expression of the DCB is derived by extending the data consistency rule in (3.15) to feature space which potentially contains richer information, and the associated expression is given by:

$$h' = \gamma h + (1 - \gamma) f_2(\Omega(f_1(h); \mathbf{DF}, \mathbf{y})). \quad (3.16)$$

Considering that f_1 and f_2 are jointly optimized with the reconstruction model f_{θ_k} during training and can be conceptually subsumed into the adjacent network modules, leading to a totally non-parametric DC block, two convolutional layers are used for feature channel reduction and expansion.

3.2.3.3 Adaptive Spatial Attention Selection Module (ASASM)

To model the correlations among the deep feature representations, attentions modules, e.g. the transformer-based models originating from natural language processing, have been widely used for visual tasks, showing to deliver performance improvements. However, those models suffer from computational complexity increasing quadratically with the number of input tokens, e.g. the spatial pixels, which makes it challenging to apply attention mechanism to high-resolution features tightly associated with the quality of generated images. A convolutional block attention module (CBAM) [148] was proposed to reduce the computational costs by adopting a channel gate and a spatial gate to govern the contributions of feature channels and spatial positions. However, the CBAM models the importance of learned information via the statistics of extracted features, which can be difficult to capture the complicated relationships rich in feature space.

To avoid the enormous memory consumption and computational complexity of conventional attention mechanism, an adaptive spatial attention selection module termed ASASM is proposed to perform locality-aware spatial attention, which endows the model with the capacity to efficiently retrieve useful information contained in high-resolution feature representations. The architecture of ASASM is displayed in Fig. 3.4 (c), which we introduce in the following.

To capture contextual information among the extracted features, a neural network is proposed to predict spatially-varying attention patterns used to perform the attention selection in the form of the weighted combination of neighboring pixels. As illustrated in Fig. 3.4 (c), we utilize a shallow U-net as the kernel predictor which can effectively expand the receptive field via an encoding-decoding structure with limited parameter increase. The U-net for kernel

prediction has significantly fewer parameters as compared to the reconstruction model, which can provide a reasonable trade-off between model capacity and latency. Assuming that the size of the feature volume as input to ASASM, termed h , is $C \times H \times W$ where C denotes the number of channels and $H \times W$ refers to the spatial size, the kernel volume ω is computed with the input features in a data-adaptive way to exhibit the modelling of spatial interdependencies and has a size of $k^2 \times H \times W$. The spatial attention selection is performed by calculating the average of the spatial locations in the neighboring region with the endowed adaptive property and increased diversities in kernel prediction as an attempt to improve the deep representation of network models. The adaptive filtering operation is applied to produce the output features dubbed $\hat{h}$ with the captured spatial correlations expressed by the variance of attention patterns, yielding:

$$\hat{h}(q) = h(q) + \sum_{\delta \in \Delta} \omega(\delta, q) h(q + \delta), \quad (3.17)$$

where $h(q)$ and $\hat{h}(q)$ are two vectors formed with pixel values across the channel dimension at spatial position q , δ denotes the footprint moving in the kernel grid Δ , e.g. $\{-1, 0, 1\}^2$ for $k = 3$ as used in our experiments, representing the filtering sliding window, and $\{\omega(\delta, q)\}_{\delta \in \Delta}$ denotes the preferences of the neighboring pixels around location q .

Different from filter adaptive networks [152, 153] which predict a filter volume of size $(C \cdot k^2) \times H \times W$ and thus tremendously increase the computational burden, the devised ASASM computes independent attention patterns for each spatial position as a means to achieve spatial variability efficiently, as the kernels are shared for all feature channels whose information exchange has been realized by the kernel prediction network with enlarged receptive fields, leading to a great reduction in the computational overhead. In other word, the devised ASASM has the ability to perform data-adaptive filter prediction and encode the contextual information in attention patterns varying across feature regions, attaining both sample adaptive properties and kernel diversities without incurring huge computational complexity and costs.

3.2.3.4 Multi-Level Densely Connected Architecture

The deep representations learned by neural networks encode rich image attributes, which can be beneficial to improving the reconstruction performance when they are appropriately exploited for the estimate of the underlying anatomy not contained in the observed measurements, i.e. recovering the under-sampled images. As illustrated in Fig. 3.5, U-shaped networks are used as the reconstruction model, dubbed f_{θ_k} with k numbering the iteration, providing feature representations at various scales. Motivated by the improvements brought by densely connected layers [154] in vision tasks, we devise on top of the encoding-decoding design an architecture, which is constructed with multi-level dense connections as a means for “reusing” the features transmitted from the previous network modules to bring out the refinement of the feature learning with the current prediction.

The architecture of the proposed reconstruction pipeline is demonstrated in Fig. 3.5 where two subnetworks are presented for simplicity. The notations of the involved blocks are described in what follows. In the encoding part of the k -th subnetwork, the first convolutional layer is dubbed $h^{(k)}$ and the transitional convolutional layer used to reduce the spatial size is denoted by $d_i^{(k)}$ with $i \in \{1, \dots, L\}$ indicating the feature level from the first encoding space, i.e. $i = 1$, to the bottom level, i.e. $i = L$, where L transitional block pairs are adopted, i.e. $L = 2$ for the architecture in Fig. 3.5. In the decoding part, the densely connected block, as depicted in Fig. 3.4 (a), and the transitional convolutional layer are denoted by $r_i^{(k)}$ and $u_i^{(k)}$ with $i \in \{1, \dots, L\}$. The devised ASASM applied to the topmost feature space, i.e. image resolution, is denoted by $A^{(k)}$, which is followed a densely connected block $r_t^{(k)}$ and the DCB denoted by $D^{(k)}$. In addition, a head convolutional layer is introduced after the input layer and a tail one before the output layer.

To implement the dense shortcuts connecting predictions at multiple scales, a collection of items $\{S_I, S_0, S_1, \dots, S_L\}$ is defined to represent the feature volumes flowing across the cascaded reconstruction network via intra- and inter-connections, where S_I refers to the feature volume as input to each subnetwork and S_0 contains the features of image resolution, hence $L + 2$

feature levels in total. We utilize subscripts to specify the encoding and decoding processes. For instance, $S_{e,2}^{(1)}$ has the predictions coming from the second feature level until the encoder of the first subnetwork and $S_{d,0}^{(2)}$ delineates that it collects all features at image resolution generated within the U-shaped networks until the decoder of the second subnetwork.

The demonstration of the proposed multi-level dense connections starts from the zero-filled $\mathbf{z}$ which is the inverse Fourier transform of the zero-filled measurement given by:

$$\mathbf{z} = \mathbf{F}^H \mathbf{y},$$

where two channels are adopted to represent complex-valued signals. The computations involved in the encoding process of the first subnetwork are presented as follows, gradually extracting information from the previous output:

$$\begin{aligned} S_{e,I}^{(1)} &\leftarrow \text{head}(\mathbf{z}) \\ S_{e,0}^{(1)} &\leftarrow h^{(1)}(S_{e,I}^{(1)}) \\ S_{e,1}^{(1)} &\leftarrow d_1^{(1)}(S_{e,0}^{(1)}) \\ S_{e,2}^{(1)} &\leftarrow d_2^{(1)}(S_{e,1}^{(1)}). \end{aligned} \quad (3.18)$$

Starting from the features at the bottom, the decoding representations are calculated with the intra-connections to transmit information within the U-shaped model:

$$\begin{aligned} S_{d,1}^{(1)} &\leftarrow [u_2^{(1)} \circ r_2^{(1)}(S_{e,2}^{(1)}), S_{e,1}^{(1)}] \\ S_{d,0}^{(1)} &\leftarrow [u_1^{(1)} \circ r_1^{(1)}(S_{d,1}^{(1)}), S_{e,0}^{(1)}] \\ S_{d,I}^{(1)} &\leftarrow [D^{(1)} \circ r_I^{(1)} \circ A^{(1)}(S_{d,0}^{(1)}), S_{e,I}^{(1)}], \end{aligned} \quad (3.19)$$

where $\circ$ denotes the composite of network modules and $[,]$ refers to feature concatenation along the channel axis. As presented in (3.19), the intermediate prediction from f_{θ_0} expressed in the form of the learned feature representation is corrected by the DCB with the k -space measurements and combined with $S_{e,I}^{(1)}$ before being passed to the successive reconstruction

iteration. The encoding features of the second subnetwork are produced with all previous outputs, including encoding and decoding features, at the same levels, yielding:

$$\begin{aligned}
 S_{e,I}^{(2)} &\leftarrow S_{d,I}^{(1)} \\
 S_{e,0}^{(2)} &\leftarrow [h^{(2)}(S_{e,I}^{(2)}), S_{d,0}^{(1)}] \\
 S_{e,1}^{(2)} &\leftarrow [d_1^{(2)}(S_{e,0}^{(2)}), S_{d,1}^{(1)}] \\
 S_{e,2}^{(2)} &\leftarrow [d_2^{(2)}(S_{e,1}^{(2)}), S_{e,2}^{(1)}].
 \end{aligned} \tag{3.20}$$

In keeping with (3.19), the subsequent decoding process only entails intra-connections here as the recurrent use of predictions over all previous iterations evolves via the alternation between skip-connection and feature concatenation, giving rise to:

$$\begin{aligned}
 S_{d,1}^{(2)} &\leftarrow [u_2^{(2)} \circ r_2^{(2)}(S_{e,2}^{(2)}), S_{e,1}^{(2)}] \\
 S_{d,0}^{(2)} &\leftarrow [u_1^{(2)} \circ r_1^{(2)}(S_{d,1}^{(2)}), S_{e,0}^{(2)}] \\
 S_{d,I}^{(2)} &\leftarrow [D^{(2)} \circ r_I^{(2)} \circ A^{(2)}(S_{d,0}^{(2)}), S_{e,I}^{(2)}].
 \end{aligned} \tag{3.21}$$

The output from the last iteration, i.e. $S_{d,I}^{(2)}$ in Fig. 3.5, gathers all predictions coming from the iterative reconstruction steps together and is mapped into a two-channel image using the tail convolutional layer as follows,

$$\mathbf{x}^{recons} = \text{tail}(S_{d,I}^{(2)}), \tag{3.22}$$

where $\mathbf{x}^{recons}$ is the reconstruction result.

It is noteworthy that all previous predictions at the same feature level are combined without removing any element from the transmitted feature collection and concatenated with the current output to realize dense connections at multiple levels. The densely connected design is therefore implemented by alternating between skip-connection and feature concatenation.

3.2.3.5 Reconstruction Framework Design

The deep cascaded framework used in our experiments for MRI reconstruction is illustrated in Fig. 3.5, where densely connected layers are embedded in the decoding structure of each U-shaped subnetwork and the devised ASASM and DCB are adopted at the top decoding level. Information propagation within and between the reconstruction iterations is enabled via multi-level dense connections. Four subnetworks are used in the pipeline with the parameter c adjusting the network width set to be $c = 1$.

3.2.3.6 Objective Function

The fully sampled references dubbed $\mathbf{x}^{GT}$ are used in the training phase to optimize the reconstruction model towards generating faithful and high-quality reconstructions. The differences between the estimated image and the ground truth are quantified by the L_1 norm and structural similarity index metric (SSIM), leading to the reconstruction loss L_{rec} computed as follows,

$$L_{rec} = (1 - \alpha)L_1(\mathbf{x}^{recons}, \mathbf{x}^{GT}) + \alpha L^{SSIM}(\mathbf{x}^{recons}, \mathbf{x}^{GT}), \quad (3.23)$$

where α balancing the strength of each component is selected to be 0.4.

We leverage the data consistency loss to encourage the reconstruction to be consistent with the under-sampled k -space data, given that the forward data process should attain closely matching observations. The consistency loss is computed in k -space to measure the discrepancies between the reconstruction and observation in encoding space, as suggested in [155]. An alternative practice is to transform the signals to image domain [156] where the data consistency is enforced. We empirically found that they behave very similarly and make no fundamental difference. The data consistency loss L_{cyc} is given by:

$$L_{cyc} = \|\mathbf{y} - \mathbf{D}\mathbf{F}\mathbf{x}^{recons}\|_1, \quad (3.24)$$

where the Fourier transform and encoding operation are applied to each coil view when dealing with multi-coil configurations.

Generative adversarial networks [60] show great success in generating photo-realistic images. We encourage sharper details by training the model using the least squares GAN (LS-GAN). The adversarial loss is computed as below,

$$\begin{aligned} \min_{\psi} & E[(D_{\psi}(\mathbf{x}^{GT}) - 1)^2] + E[(D_{\psi}(\mathbf{x}^{recons}) - 0)^2] \\ \min_{\theta} & E[(D_{\psi}(\mathbf{x}^{recons}) - 1)^2], \end{aligned} \quad (3.25)$$

where θ denotes the collection of the parameters of all subnetworks and a patch-based discriminator D_{ψ} parametrized by ψ , used in CycleGAN[157], is adopted, which comprises a standard convolutional layer followed by three strided convolutional layers to decrease the spatial resolution. The total objective function is a weighted combination:

$$L = E_{\{\mathbf{x}^{recons}, \mathbf{x}^{GT}\}} [\lambda_{rec} L_{rec} + \lambda_{cyc} L_{cyc} + \lambda_{adv} L_{adv}], \quad (3.26)$$

where the training parameters are practically set to $\lambda_{rec} = 10$, $\lambda_{cyc} = 0.5$, and $\lambda_{adv} = 0.01$, using a large weight for the reconstruction loss to ensure the success of reconstruction and a small weight for the adversarial loss to prevent instability in training.

3.2.4 Region-Guided Channel-Wise Attention Network (RG-CANet)

Recently, deep learning-based methods have witnessed widespread use in medical imaging and achieved promising success in reconstructing MR images from sparsely sampled measurements with the aim of accelerating the data acquisition procedure by virtue of the powerful representations of deep structures. Although it is effective to achieve inspiring results, how to produce high-fidelity reconstructions from the observations measured at high acceleration rates still remains an open question. To improve the representation capacity of neural network models in an effort to more accurately recover the corrupted images, attention mechanism [143]

was adopted as a constituent component of network models to enhance the reconstruction quality, by leveraging the interdependencies among the spatial pixels of the learned feature representations [144, 145]. However, the computational complexity and memory consumption of attention mechanism increase quadratically with the amount of involved “tokens”, which is directly associated with the size of feature maps being processed. As a consequence, the significantly increased computational costs prohibit the implementation of attention modules to high-resolution features which are tightly linked with dense prediction tasks such as MRI reconstruction, limiting the achievable ability of deep learning-based models.

Alternatively, channel-wise attention approaches were proposed to exploit the relevant information among feature channels, which can be efficiently implemented, to achieve improved outcome [66, 146]. Such methods avoid the huge computational costs needed to capture the dependencies across feature regions, whereas the correlations between different spatial pixels, which are potentially characteristic of the structures of the anatomy being inspected, are not taken into consideration, restricting the effectiveness of attention mechanism applied over channels.

To address this issue, a novel channel-wise attention module is devised here, which incorporates the implicitly captured spatial semantics into a channel-based attention block as an attempt to increase the diversities of recalibration patterns used to scale feature maps to boost performance gains. By introducing region-based semantic information into attention mechanism over channels, we derive a region-guided channel-wise attention network, termed RG-CANet, to retrieve improved model performance in MRI reconstruction. The details of the proposed method are presented in the following.

3.2.4.1 Region-Guided Channel-Wise Attention Module (RG-CAM)

We propose to incorporate the implicitly captured spatial semantics into attention mechanism applied to feature channels, leading to a novel region-guided channel-wise attention module, dubbed RG-CAM. The architecture of the RG-CAM is presented in Fig. 3.6, where the input feature volume $h \in \mathbb{R}^{C \times H \times W}$ is passed to a backbone block to produce features termed $\hat{h}$, which are refined by the predicted attention gate dubbed K coming from an attention branch. A densely connected block is employed as the backbone due to its effective representation capacity [154]. The flexible design of the RG-CAM allows us to leverage other network modules, e.g. spatial attention blocks [143, 148, 146], in an effort to retrieve further improvements. The proposed RG-CAM entails the following operations.

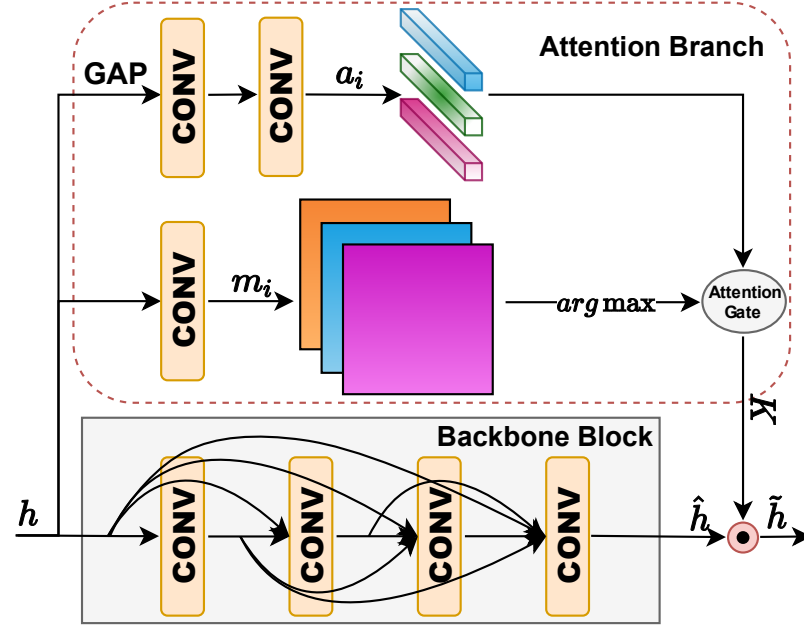


Figure 3.6: Illustration of the region-guided channel-wise attention module (RG-CAM).

D) Channel-Wise Attention Kernel Bank: To generate a set of attention kernels, termed a_i , a non-linear mapping consisting of two convolutions with a GELU activation function in between is applied to the input h after squeezing it in the spatial dimensions by the global average pooling (GAP) operation, as illustrated in Fig. 3.6. The resulting features, which are reshaped to be $M \times C \times 1 \times 1$ representing M predicted channel-wise attention kernels, termed

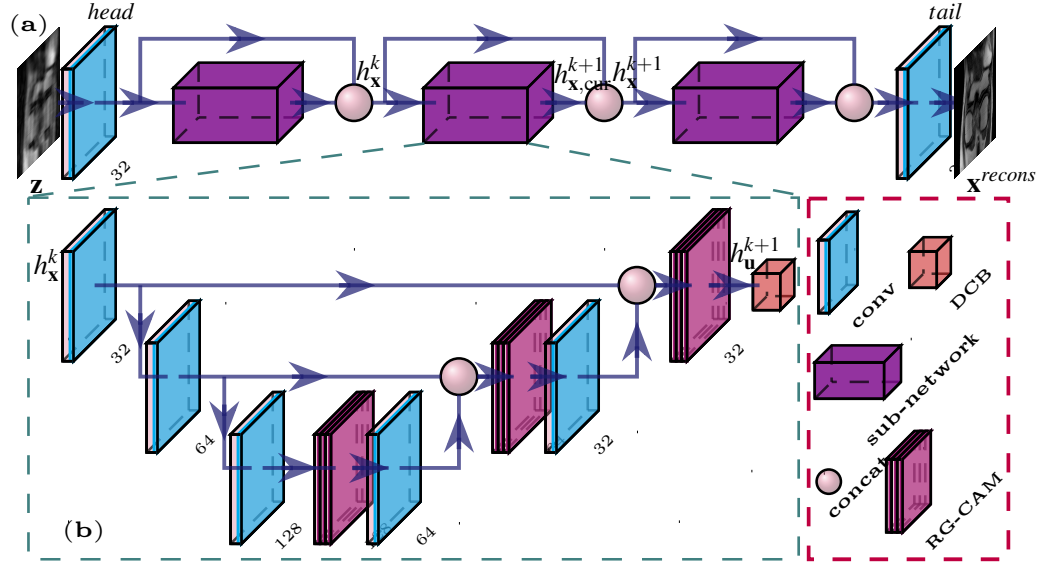


Figure 3.7: Illustration of (a) region-guided channel-wise attention network (RG-CANet) and (b) U-shaped subnetwork. The input to the reconstruction frame is the zero-filled and the output is the image reconstructed with all previous predictions collected by subnetwork dense connections.

a_i , having a size of $C \times 1 \times 1$, are mapped by a Sigmoid activation to $(0, 1)$, where 0 means no information allowed to pass through this channel and full attention is paid to channels with a value of 1. The kernel bank $\{a_i\}_{i=1}^M$ is subsequently utilized under the guidance of spatial guiding masks which are described below.

II) Spatial Guiding Mask: In parallel to the top route, as displayed in Fig. 3.6, where a bank of channel-wise attention kernels is produced, a convolutional layer is applied to generate a feature volume of size $M \times H \times W$ representing M spatial guiding masks, termed $m_i \in R^{H \times W}$. The pixel values at each spatial position q of the guiding masks are normalized as follows,

$$m_i(q) = \begin{cases} 1, & i = \arg \max_j m_j(q) \\ 0, & \text{otherwise.} \end{cases} \quad (3.27)$$

Due to the spatial invariance of convolution operations, pixels from the regions having similar semantics potentially lead to the same value in the normalized m_i and will share an attention pattern when conflating the guiding masks with the channel-based attention mechanism.

III) Region-Guided Attention Gate: The output from the backbone block is refined via a channel-wise attention method which is guided by the generated spatial masks, yielding a region-guided attention gate as presented in Fig. 3.6. The attention gate, dubbed $K \in R^{C \times H \times W}$, is constructed by stacking the attention kernels across the spatial area which are selected from the kernel bank based on the criterion shown below,

$$K(q) \leftarrow a_i, \text{ if } m_i(q) = 1. \quad (3.28)$$

Similar regions which are likely to produce the same value in the guiding masks will share an attention kernel, not only increasing the diversity of attention patterns but also forming the spatial semantics implicitly learned and encoded in the devised attention gate.

As demonstrated in Fig. 3.6, the output of the RG-CAM is computed by calibrating $\hat{h}$ using the generated attention gate:

$$\tilde{h} = K \odot \hat{h}. \quad (3.29)$$

By incorporating the region-guided attention method, the network model can be endowed with the ability to capture the spatial information when performing attention over channels. Note that the parameters and computational costs of the devised attention branch are negligible owing to the use of the GAP operation, which enables a light-weight, efficient, and flexible attention module simultaneously considering the spatial and channel-wise information without needing labeled data or extra supervision methods. The Softmax trick [158] is used to enable the gradient propagation with respect to the spatial guiding masks.

3.2.4.2 Densely Connected Reconstruction Framework

By virtue of the powerful representation of deep neural networks, a densely connected reconstruction framework is derived by unrolling the optimization algorithm expressed in (3.11), as illustrated in Fig. 3.7, which leverages the RG-CAM at decoding levels, where the number of the attention kernel and mask pairs is selected to 8, to refine the features transmitted from the encoder via skip-connections and adopts the deep data consistency block DCB to enforce the data consistency, as proposed in Sect. 3.2.3.2.

We adopt two convolutional layers, dubbed *head* and *tail*, to map the signals between image domain and feature space, aiming to expand and decrease the feature channels to accommodate complex-valued MRI data. The initial estimate $h_{\mathbf{x}}^0$ is computed with the zero-filled, obtaining a feature volume:

$$h_{\mathbf{x}}^0 = \text{head}(\mathbf{z}). \quad (3.30)$$

Owing to the DCB which is expressed as an extension of data consistency enforcement to feature space with the intention of retaining the rich information provided by deep representations, the reconstruction iteration formulated in (3.11) can be recast to meet the feature representations with the dense skip-connections employed in the network pipeline, giving rise to the following update steps:

$$h_{\mathbf{u}}^{k+1} \leftarrow f_{\theta_k}(h_{\mathbf{x}}^k) \quad (3.31a)$$

$$h_{\mathbf{x},\text{cur}}^{k+1} \leftarrow \text{DCB}(h_{\mathbf{u}}^{k+1}) \quad (3.31b)$$

$$h_{\mathbf{x}}^{k+1} \leftarrow [h_{\mathbf{x}}^k, h_{\mathbf{x},\text{cur}}^{k+1}], \quad (3.31c)$$

where $h_{\mathbf{u}}^{k+1}$ is the estimate from the k -th subnetwork, $h_{\mathbf{x},\text{cur}}^{k+1}$ is the current prediction computed by performing the data consistency enforcement, and the outcome $h_{\mathbf{x}}^{k+1}$ of the current iteration is given by concatenating the current prediction with the collection of all previous outputs. The final reconstruction $\mathbf{x}^{\text{recons}}$ of the cascaded framework is attained with the predictions from all

previous iterations:

$$\mathbf{x}^{recons} = tail(h_{\mathbf{x}}^N), \quad (3.32)$$

where N is the number of reconstruction stages and selected to be 5 in our experiments.

3.2.4.3 Objective Function

The objective function used to train the region-guided channel-wise attention framework is the combination of the reconstruction metric and the adversarial loss introduced in Sect. 3.2.3.6, which is given as follows,

$$L = E_{\{(\mathbf{x}^{recons}, \mathbf{x}^{GT})\}} [\lambda_{rec} L_{rec} + \lambda_{adv} L_{adv}]. \quad (3.33)$$

3.3 Experiments

3.3.1 Accelerated MRI Reconstruction via Deep Cascaded Neural Networks

The performance of the optimization-based framework introduced in Sect. 3.2.3 is evaluated in this section, where a single-coil knee dataset and a multi-coil brain database are used for both qualitative and quantitative assessments. ASGAN introduced in Sect. 3.2.2 as a direct inversion model is devised for single-coil MRI reconstruction and compared with deep cascaded networks derived by unfolding a model-based optimization algorithm to demonstrate the benefits of combining the forward encoding process with the learned deep prior. Comparison experiments against other state-of-the-art approaches, where the implementations are available or adequate information is provided for reproduction, and ablation studies are conducted to showcase the effectiveness of the devised method.

The acceleration factors are set to be $8\times$ and $4\times$ with the under-sampling process applied to k -space data along the x-axis, where 4% and 8% auto-calibration signal (ACS) lines in the center are preserved and periphery of k -space is sampled randomly, conforming to the conditions for compressed sensing and providing a fair and general comparison. Under-sampling is applied to the fully sampled k -space raw data loaded using the official code package [56] and the fast Fourier transform and its inverse are used for mappings between k -space and image domain. To approximate the forward encoding process in real-world applications, no pre-processing steps, e.g. using compressed images stored in a database and later loaded or non-negative magnitude images computed with parallel coil frames via the root sum of squares (RSS) method for synthesizing a new k -space [159], are employed in the simulated data acquisition procedure, in hopes of ensuring that the simulations comply with actual practice and has the potential to expedite the deployment of reconstruction models to real-world machines. The reconstruction models are trained for 35 epochs with a batch size of 5, using an Adam optimizer with $\beta_1=0.5$, $\beta_2=0.999$, and a learning rate of 2×10^{-4} .

Compressed sensing utilizes a sparsity-based signal prior to perform image reconstruction from the under-sampled measurements. Although effective in practice, the limited representation capacity of the assumed sparsity prior prevents CS-based algorithms from achieving high-quality reconstruction under aggressively accelerated acquisition scenarios where very little information is available in the observed measurements. We compared with deep learning-based reconstruction models and have not adopted compressed sensing methods in experiments, since such models have demonstrated superior performance over compressed sensing techniques both qualitatively and quantitatively [56, 64, 66, 82]. We used those models in our comparison experiments.

3.3.1.1 Single-Coil MRI Reconstruction

The FastMRI single-coil knee MRI dataset [56] is used in simulations, which contains rich textural and structural details. The dataset has about 1500 fully sampled MRI scanning cases collected on 3 and 1.5 Tesla magnets. Considering the computational resource consumed in experiments, 400 scanning cases are used for training and 164 different cases for test. The zero-filled, which is computed by the inverse Fourier transform with the zero-filled k -space data, is taken as input, where two channels are used to represent complex-valued signals.

The proposed deep cascaded network framework, termed ASASNet, is compared with the direct inversion model ASGAN and other deep learning-based approaches, including FastMRI U-net [56], MoDL [64], and MICCAN [66]. FastMRI U-net [56] adopts a U-net trained to produce a clean magnitude image from the absolute values of the inverse Fourier transform of the under-sampled data with unobserved positions filled with zeros. MoDL was derived from a model-based optimization algorithm by regularizing the under-determined problem via the learned representations of neural networks. MICCAN introduced a channel-wise attention module embedded in a neural network to enhance the reconstruction quality. Deep cascaded networks were utilized in MoDL and MICCAN with a closed-form solution to the data consistency constraint, which are used in experiments to testify the contribution of the devised deep data consistency block.

The reconstructed images at $8\times$ and $4\times$ acceleration factors are presented in Fig. 3.8, where it is shown that compared to other methods the devised framework is able to produce realistic and faithful reconstructions with sharper and more natural and complex textural details and better preserve the structural features, at both acceleration rates. The residual maps displayed in Fig. 3.9 show that fewer errors appear in the reconstruction results of the proposed method, particularly at a high under-sampling rate. Those observations demonstrate that the proposed model can yield high-fidelity reconstructions with highly under-sampled measurements.

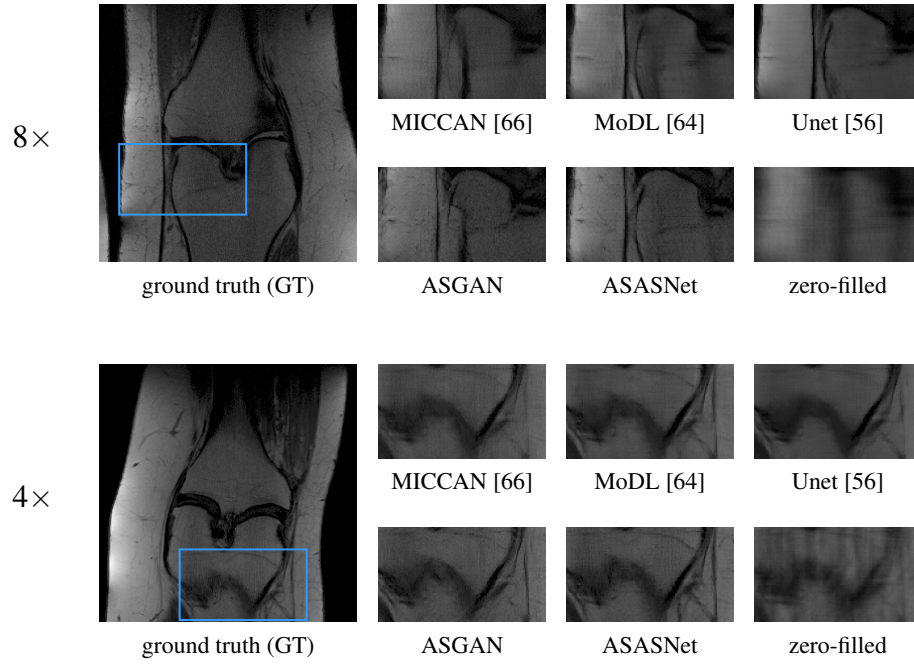


Figure 3.8: Comparison results for the proposed adaptive spatial attention selection network (ASASNet) using the knee MRI dataset under 8 $\times$ and 4 $\times$ acceleration.

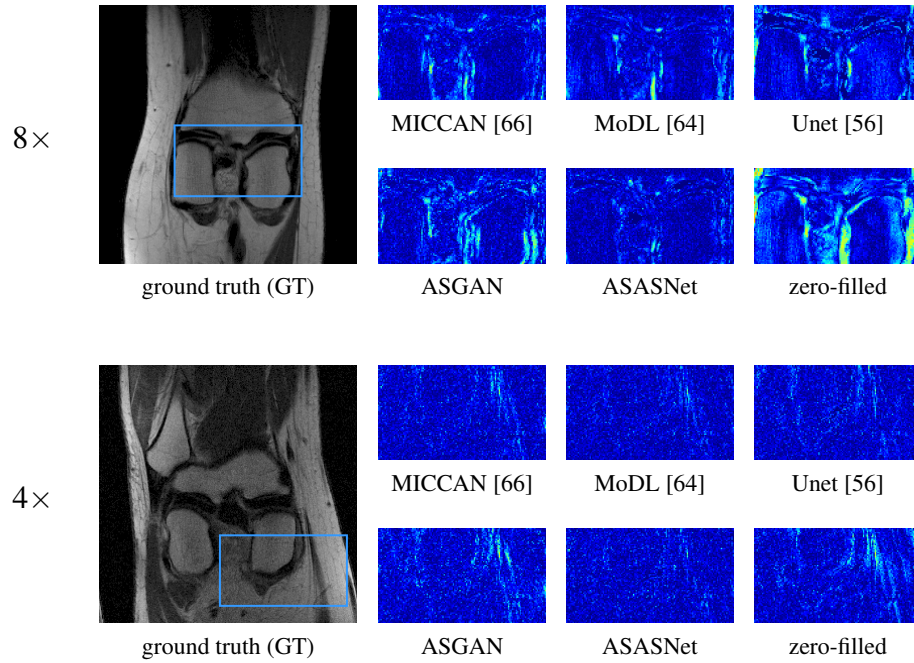


Figure 3.9: Residual error maps (2 $\times$ amplified) of knee MRI reconstruction under 8 $\times$ and 4 $\times$ acceleration for the proposed adaptive spatial attention selection network (ASASNet).

Although having the capacity to produce fine-grained details in recovered images, ASGAN can suffer from artifacts and more reconstruction errors. Contrasting with ASGAN falling in the category of neural network models which are developed as a means to directly generate a clean image from the under-sampled input and can consequently synthesize undesired artifacts and hallucinations, leading to less faithful results, e.g. the corrupted edge in the $8\times$ accelerated reconstruction shown in Fig. 3.8, owing to the lack of domain-specific knowledge, the devised deep cascaded framework not only relies on the learning of the recovery of high-quality images but also has the capacity to exploit domain-specific knowledge, expressed in the form of regularized optimization under the k -space constraint, to pursue more consistent reconstructions with the ground truths. In quantitative evaluation, ASGAN leads to inferior performance as compared to the optimization-inspired model, in agreement with the above observation.

The reconstruction performance of different models is quantified using PSNR and SSIM as evaluation metrics, in which higher values are preferred. Considering that visual quality can be an important aspect in evaluation of image recovery, particularly when dealing with texture-rich anatomies of human body, the Fréchet inception distance (FID) and the kernel inception distance (KID) [160], in which lower scores are better, are adopted to gauge the perceptual quality of reconstructed images by utilizing the deep representations of neural networks learned from a large-scale database.

Table 3.1: Quantitative evaluation on knee MRI reconstruction under $8\times$ and $4\times$ acceleration for the proposed adaptive spatial attention selection network (ASASNet). $4\times$ accelerated reconstruction has the same latency (s) and model size as $8\times$.

	method	PSNR $\uparrow$	SSIM $\uparrow$	FID $\downarrow$	KID $\downarrow$	Latency $\downarrow$	Size $\downarrow$
$8\times$	ASASNet	27.71	0.738	86.46	0.021	0.028	1183073
	ASGAN	25.45	0.638	104.34	0.036	0.056	4459479
	Unet [56]	25.82	0.703	160.35	0.121	0.013	2754384
	MoDL [64]	27.13	0.620	143.65	0.080	0.091	5849787
	MICCAN [66]	26.61	0.642	180.66	0.146	0.043	2649455
	zero-filled	20.54	0.388	423.32	0.533	-	-
$4\times$	ASASNet	31.62	0.841	61.43	0.005	-	-
	ASGAN	27.73	0.711	82.18	0.016	-	-
	Unet [56]	28.35	0.771	118.07	0.061	-	-
	MoDL [64]	30.34	0.745	98.86	0.042	-	-
	MICCAN [66]	30.11	0.711	99.44	0.040	-	-
	zero-filled	23.94	0.486	255.06	0.239	-	-

From the evaluation results presented in Table 3.1, we found that the proposed ASASNet consistently yields superior performance at two acceleration configurations, showing considerable improvements in all relevant metrics as compared to other approaches. From Table 3.1, it can be observed that MoDL has high PSNR scores while comparatively low SSIM values are obtained, which implies that a deep architecture which meets the k -space constraint by unrolling a model-based optimization procedure is able to recover consistent images with the observations, therefore achieving better results in terms of pixel-wise metrics. However, it fails to correlate well with the structural assessment which is indicative of the intricate properties of human visual perception [58], leading to degraded performance in terms of SSIM and visual metrics. We can also observe that the proposed framework attains superior performance with fewer parameters than other models, suggesting the efficient design of our method, and the competitive execution speed of our model shows the potential to enable real-time applications.

3.3.1.2 Multi-Coil MRI Reconstruction

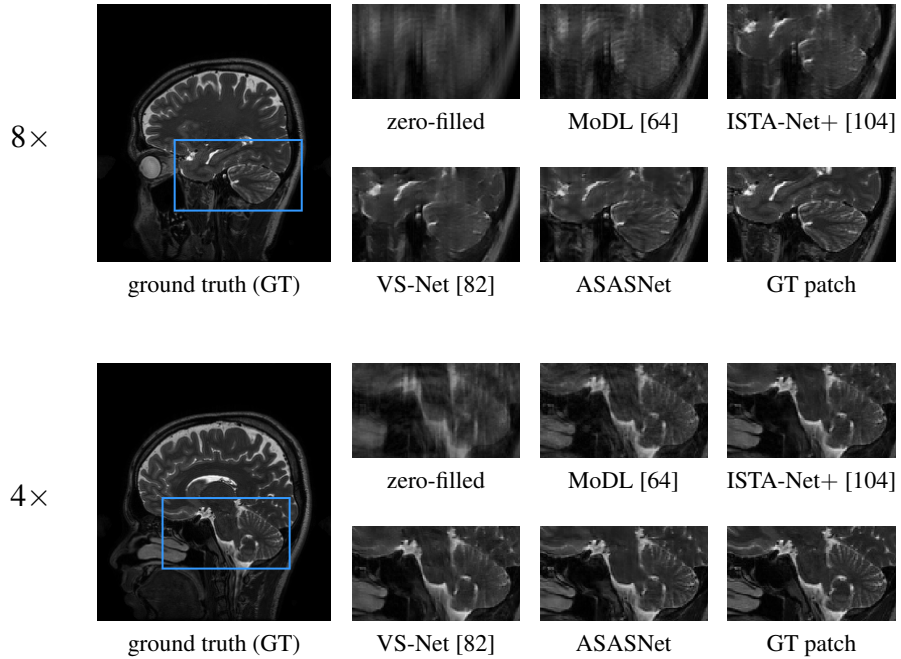


Figure 3.10: Comparison results for the proposed adaptive spatial attention selection network (ASASNet) using the brain MRI dataset under $8\times$ and $4\times$ acceleration.

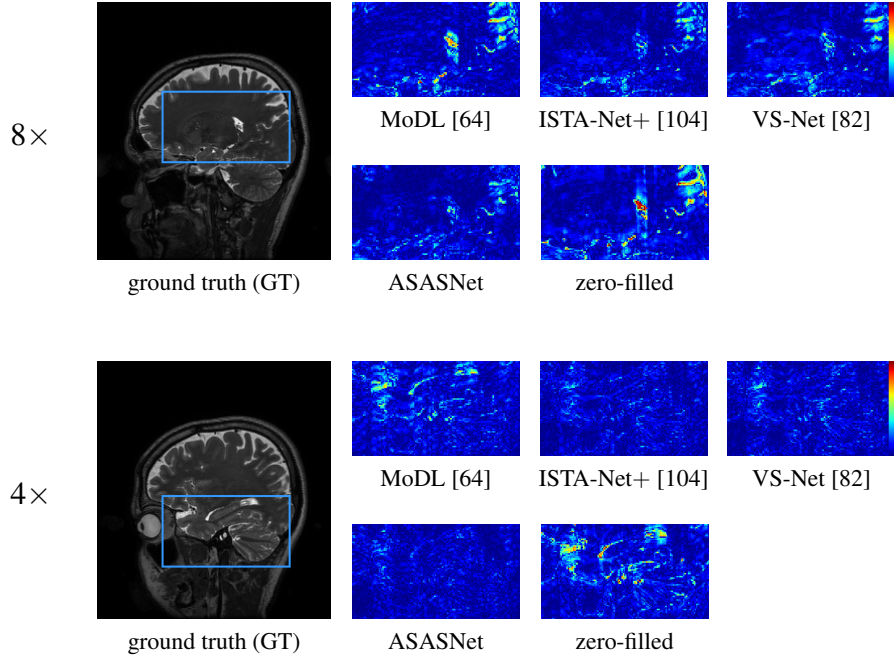


Figure 3.11: Residual error maps ($2\times$ amplified) of brain MRI reconstruction under $8\times$ and $4\times$ acceleration for the proposed adaptive spatial attention selection network (ASASNet).

The proposed reconstruction framework ASASNet is extensible to multi-coil MRI configurations, where the data encoding process occurs on the individual coil sensors measuring the underlying anatomy with distinct coil sensitivities. The forward data acquisition is delineated using the sensitivity maps dubbed $\mathbf{S}_i$ as the composite of coil projection and k -space sampling, which can be expressed by the following equation,

$$\mathbf{y}_i = \mathbf{D}\mathbf{F}\mathbf{S}_i\mathbf{x},$$

where $\mathbf{y}_i$ is the measurement from the i -th coil. To obtain a single image from a set of coil frames, the sensitivity-weighted average is used as input to the reconstruction model [64, 82], leading to the zero-filled computed with multi-coil data:

$$\mathbf{z} = \sum_{i=1}^{n_c} \mathbf{S}_i^H \mathbf{F}^H \mathbf{y}_i,$$

Table 3.2: Quantitative evaluation on brain MRI reconstruction under $8\times$ and $4\times$ acceleration for the proposed adaptive spatial attention selection network (ASASNet). $4\times$ accelerated reconstruction has the same latency (s) and model size as $8\times$.

	method	PSNR $\uparrow$	SSIM $\uparrow$	FID $\downarrow$	KID $\downarrow$	Latency $\downarrow$	Size $\downarrow$
$8\times$	ASASNet	27.65	0.853	89.80	0.057	0.031	1183073
	VS-Net [82]	27.56	0.833	147.65	0.129	0.035	1133232
	ISTA-Net+ [104]	27.48	0.831	171.58	0.170	0.189	385613
	MoDL [64]	25.28	0.711	238.32	0.283	0.094	5849787
	zero-filled	23.25	0.674	301.13	0.396	-	-
$4\times$	ASASNet	33.20	0.938	51.93	0.018	-	-
	VS-Net [82]	32.18	0.914	86.01	0.057	-	-
	ISTA-Net+ [104]	32.26	0.916	83.03	0.053	-	-
	MoDL [64]	29.89	0.801	150.11	0.152	-	-
	zero-filled	25.66	0.770	228.14	0.267	-	-

where n_c is the number of machine coils. Pursuant to the manipulations of coil sensitivities accounting for the accelerated data sampling process, the DCB is recast by applying the data consistency enforcement with the measurements observed at each coil to the projection of the intermediate estimate onto the associated coil sensitivity, giving rise to:

$$h' = \gamma h + (1 - \gamma) f_2 \left(\sum_{i=1}^{n_c} \mathbf{S}_i^H \Omega(\mathbf{S}_i f_1(h); \mathbf{DF}, \mathbf{y}_i) \right). \quad (3.34)$$

We evaluate the reconstruction models on a multi-coil brain MRI dataset [64], where the k -space data were acquired by multiple coil sensors using a 3D T2 CUBE sequence and the sensitivity maps were pre-computed. The dataset has 360 scanning samples for training and 164 samples available for test. Comparison experiments are performed under $8\times$ and $4\times$ acceleration against three deep model-based reconstruction frameworks, including VS-Net [82] which realizes the iteration of a variable splitting algorithm via a deep network architecture, ISTA-Net+ [104] which builds up a cascaded network by unfolding an iterative shrinkage-thresholding algorithm for resolving the under-determined inverse problems, and MoDL [64].

The reconstructions produced by different models are presented in Fig. 3.10, where we observed that the proposed framework has the ability to restore sharp and clear edges and textural details from corrupted observations, yielding more visually pleasing results as compared to other methods. In Fig. 3.11 displaying the residual maps of reconstructed images, it

is shown that the proposed model generates fewer reconstruction errors. From the evaluation results presented in Table 3.2, we found that our method delivers gains in PSNR and SSIM and obtains remarkable improvements in terms of FID and KID, showing to outperform competing approaches in quantitative evaluation and suggesting its contributions to the accuracy and perceptual believability of the recovery of accelerated MRI.

3.3.1.3 Ablation Studies on Model Components

We conduct the ablation studies to further substantiate the effectiveness of the devised model components and structure. To demonstrate the usefulness of the multi-level densely connected architecture, we compare with two candidates, termed (A) sequential and (B) denseNet. As illustrated in Fig. 3.12 (a), the subnetworks embedded in the reconstruction framework are sequentially cascaded, constructing the model (A) sequential. As a densely connected variant enabling dense connections of holistic networks, (B) denseNet recurrently uses the predictions from all previous steps by gathering their outputs together to produce the current estimate, as portrayed in Fig. 3.12 (b). To provide a fair comparison, we balance the model parameters by “copying” the feature volumes of each subnetwork for variant (A) and (B) at the positions where the inter-connections appear to form the densely connected design at multiple levels. We respectively remove the DCB and ASASM modules to verify their efficacy, which leads to two competing candidates (C) w/o DCB and (D) w/o ASASM. The shallow U-net used in ASASM for kernel prediction is preserved to construct a residual connection with the output channel number accordingly adjusted to meet the feature size, for a fair comparison. Similarly, the two convolutional layers in DCB are kept for (C). The performance of the foregoing variants is evaluated on the knee dataset, since it contains rich and complicated image features and is thus suitable to present fine-grained details and nuances in reconstructions.

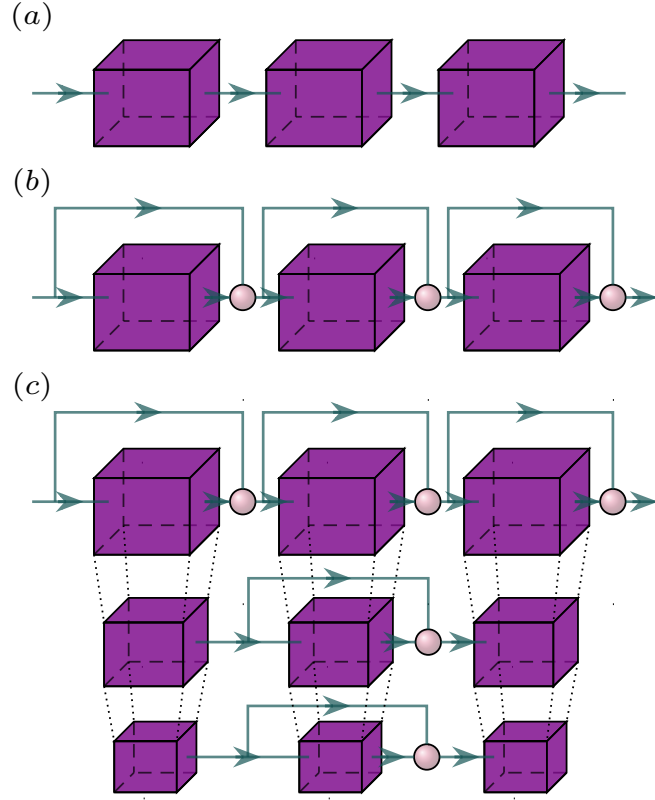


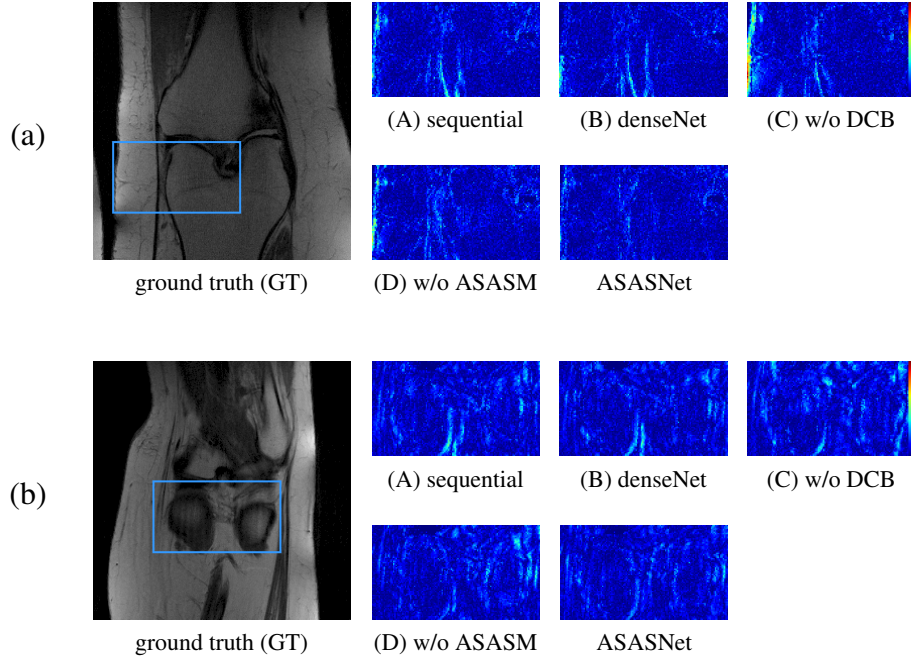
Figure 3.12: Illustration of network cascade categories. a) Sequentially connected, b) densely connected, and c) multi-level densely connected. For simplicity, the skip-connections within each subnetwork are not displayed in c).

The evaluation results are presented in Table 3.3, where it is shown that the devised multi-level dense connections deliver considerable more performance gains than candidate (B) only employing dense connections for the outputs of subnetworks, compared to model (A), substantiating the efficacy of information transmission and reuse at multiple feature levels. Comparing the proposed model with candidate (C) and (D), we found that incorporating the DCB and ASASM into the network framework can improve the reconstruction performance consistently in terms of all relevant metrics, confirming their effectiveness in improving the model capacity.

We display the residual error maps of images recovered by different candidates in Fig. 3.13 to visualize the performance gains owing to the devised model components. The results of the proposed method show fewer errors as compared to variant (A) and (B), which implies the improved model capacity using densely connected structures at feature levels. Comparing with

Table 3.3: Ablation studies on the model components and structure of the proposed adaptive spatial attention selection network (ASASNet) using knee dataset.

method	PSNR $\uparrow$	SSIM $\uparrow$	FID $\downarrow$	KID $\downarrow$
(A) sequential	27.25	0.725	91.33	0.024
(B) denseNet	27.49	0.731	89.91	0.023
(C) w/o DCB	26.13	0.705	95.50	0.029
(D) w/o ASASM	27.15	0.726	94.23	0.027
ASASNet	27.71	0.738	86.46	0.021

**Figure 3.13:** Residual maps ($2\times$ amplified) of ablation cases for the proposed adaptive spatial attention selection network (ASASNet) under $8\times$ acceleration.

candidate (C) and (D), it is demonstrated that removing the DCB produces much more errors in generated images and the proposed model with the presence of ASASM remarkably reduces reconstruction errors, which confirms the effectiveness of the devised DCB and ASASM in enhancing the reconstruction accuracy.

From Table 3.3 and Fig. 3.13, we observed that the DCB delivers prominent improvements and removing it leads to much more reconstruction errors presented in the residual maps, which suggests the advantage of the proposed method to enforce data consistency. Firstly, the k -space data constraint is explicitly satisfied by the DCB with the intermediate predictions corrected

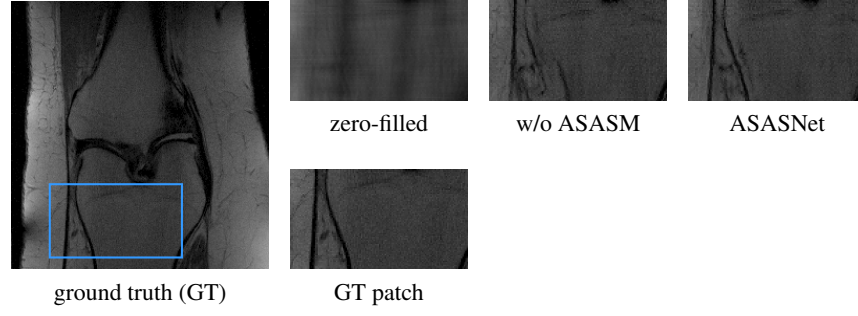


Figure 3.14: Ablation results for the proposed adaptive spatial attention selection network (ASASNet) with and without the presence of ASASM at $8\times$ acceleration.

with the measurements, which potentially produces more faithful reconstructions to the targets. Secondly, the DCB makes use of residual learning in feature space to avoid the bottleneck design existing in conventional data consistency methods and can further enhance the model performance, which we demonstrate in Sect. 3.3.1.4.

To visualize the contribution of ASASM, the images reconstructed by models with and without the presence of ASASM are presented in Fig. 3.14, where the candidate without using ASASM generates corrupted local structures, e.g. the dislocated edge, which can be restored when incorporating the ASASM module into the reconstruction pipeline. It indicates that the devised ASASM endows the model with the ability to capture contextual cues encoded in spatially-varying attention patterns and perform data-adaptive kernel prediction which shows to be beneficial to preserving structural details. Those observations showcase the usefulness of the proposed method and suggest a synergistic effect in obtaining performance advantages.

3.3.1.4 Ablation Study on Data Consistency Blocks

To validate the efficacy of the generalization of data consistency to feature space as used in the DCB, the residual feature skip-connection is removed from the DCB to implement data consistency enforcement in a conventional method, leading to the update shown below,

$$h' = f_2(\gamma f_1(h) + (1 - \gamma)\Omega(f_1(h); \mathbf{DF}, \mathbf{y})), \quad (3.35)$$

which we denote by w/o DC-shortcut. The comparison is presented in Table 3.4, which shows the inferior performance of the model using the DC method in (3.35) in all evaluation metrics, indicating the adverse impact of the bottleneck structure in conventional DC methods and confirming that the devised DCB with a feature-level residual connection can diminish this problem and push reconstruction performance via easier feature transmission.

Table 3.4: Ablation study on data consistency blocks of the proposed adaptive spatial attention selection network (ASASNet) using knee dataset.

method	DC	PSNR $\uparrow$	SSIM $\uparrow$	FID $\downarrow$	KID $\downarrow$
ASASNet	DCB	27.71	0.738	86.46	0.021
w/o DC-shortcut	conventional	27.21	0.730	93.51	0.026

Table 3.5: Ablation study on objective function using knee dataset for the proposed adaptive spatial attention selection network (ASASNet).

method	PSNR $\uparrow$	SSIM $\uparrow$	FID $\downarrow$	KID $\downarrow$
ASASNet	27.71	0.738	86.46	0.021
w/o L_{cyc}	27.57	0.734	88.06	0.023
w/o L_{adv}	27.88	0.750	106.96	0.049

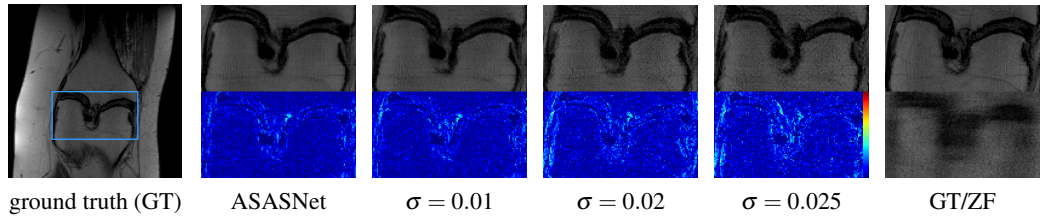


Figure 3.15: Reconstructions using the proposed adaptive spatial attention selection network (ASASNet) with random Gaussian noise added to the measurement in inference.

3.3.1.5 Ablation Studies on Objective Function

We present the ablation results on the objective function in this section. Since the reconstruction loss expressed in (3.23) measures the differences between the prediction and the fully sampled reference and ensures the success of the supervised training of the reconstruction framework, the model cannot produce meaningful results without it. We therefore respectively remove the data consistency loss and the adversarial loss from the training objective, resulting in the candidates dubbed w/o L_{cyc} and w/o L_{adv} , to manifest their roles in reconstruction tasks. From

the evaluation comparison displayed in Table 3.5, we discovered that adopting L_{cyc} leads to slightly improved performance in terms of all assessment metrics, showing the efficacy of the k -space data constraint embodied in the training objective. It can be found that the model trained without L_{adv} produces higher PSNR and SSIM values by a small margin, whereas the perceptual distances expressed in the form of FID and KID significantly increase, implying a reduction in the visual quality of generated images. To strike a better trade-off between reconstruction accuracy and image quality, the combination of the three training losses is adopted in our experiments.

3.3.1.6 Reconstruction Results with Random noise

By experiments, we demonstrated that the proposed method can achieve improved performance as compared to competitive models and generalize well on different anatomical structures under both single-coil and parallel imaging configurations. In real-world scenarios, the data observed at machine coils can be corrupted by unpredictable noise, which can possibly impact the quality of reconstructions. To demonstrate the model performance under harsh conditions, we display the results in Fig. 3.15 where random Gaussian noise of various strength levels are added to the k -space measurements in inference. Although effective in yielding high-quality predictions with well-preserved structures, the noise added to the input data cause more reconstruction errors as presented in the residual maps and the images recovered with strong random noise, e.g. $\sigma = 0.025$, show less faithful local textures, compared to the result without being corrupted by noise. Those observations have implications that noise occurring in k -space during the data encoding process impose a more challenging problem which requires enhancing the model robustness capacity to preserve the quality of reconstructed images. It can be an intriguing research direction to incorporate the de-noising task into the ill-posed problems, e.g. formed as a mixture of MR image de-noising and reconstruction, as an attempt to improve the generalization and robustness capacity of reconstruction models.

3.3.2 Region-Guided Attention Networks for MRI Reconstruction

3.3.2.1 Comparison Experiments

The comparison experiments are conducted to validate the performance of the region-guided channel-wise attention framework RG-CANet introduced in Sect. 3.2.4 using the FastMRI single-coil knee dataset [56], in comparison with ASGAN and other state-of-the-art methods including: MICCAN [66], MoDL [64], and FastMRI Unet [56]. The reconstruction results are presented in Fig. 3.16 where $8\times$ and $4\times$ acceleration diets are adopted. We can observe that the proposed model generates images with more faithful structural and textural details, as compared to other methods. The residual maps are displayed in Fig. 3.17 where it is shown that our model produces fewer spatial errors, particularly under a high acceleration factor, indicating that it can more accurately recover images from sparsely sampled measurements.

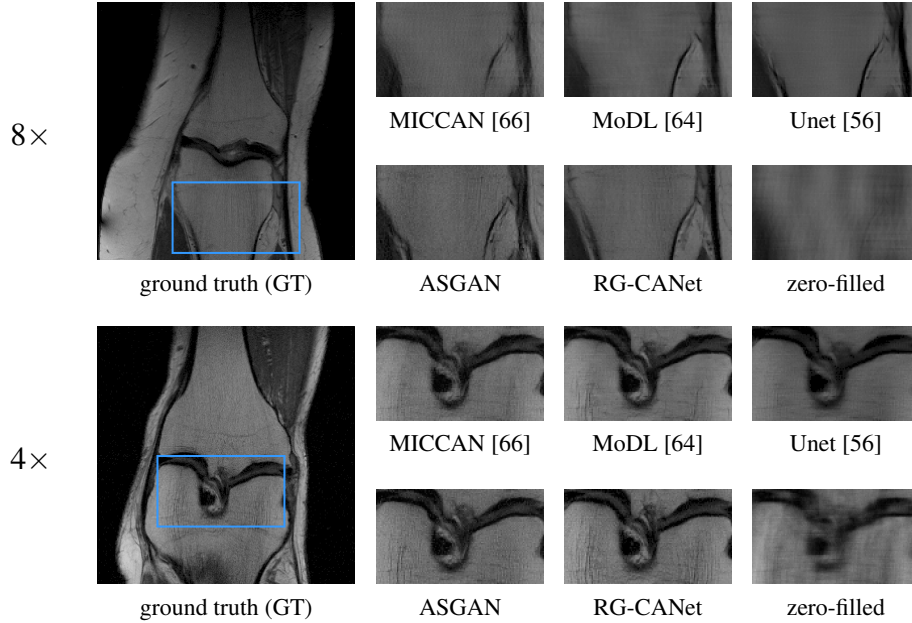


Figure 3.16: Reconstruction comparisons for the region-guided channel-wise attention network (RG-CANet) using the knee MRI dataset under $8\times$ and $4\times$ acceleration.

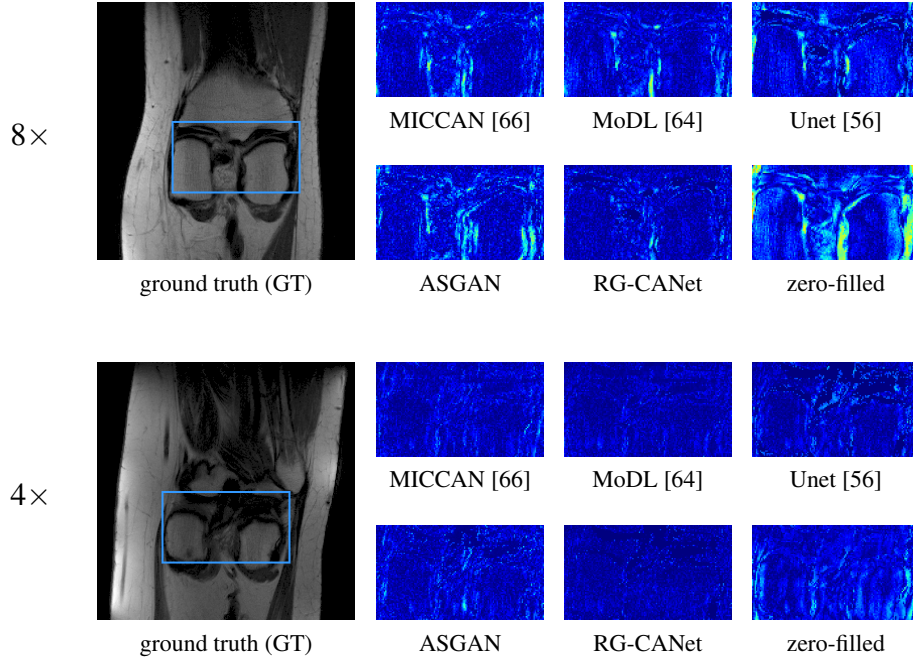


Figure 3.17: Residual maps ($2\times$ amplified) of knee MRI reconstruction at $8\times$ and $4\times$ acceleration factor for the region-guided channel-wise attention network (RG-CANet).

The quantitative evaluations displayed in Table 3.6 demonstrate that the devised reconstruction framework outperforms other approaches in terms of all relevant metrics and has competitive running speed owing to the encoder-decoder design. To achieve a better trade-off between accuracy and latency, we use the separable (depth-wise+point-wise) convolutional layer [161] to replace the conventional convolutions used in each iteration subnetwork, leading to the variant termed proposed-S. The superior results of variant proposed-S over other competing models show that the improved performance is attributable to the devised region-guided attention mechanism and the network structure, and not simply due to the parameter size.

3.3.2.2 Ablation Studies

To demonstrate the contributions of the devised region-based attention module and network structure, ablation studies are performed by respectively removing the model components: the DCB, dense skip-connections, and the RG-CAM, resulting in the variants dubbed w/o DCB, w/o dense, and w/o RG-CAM. For variant w/o dense, we “repeat” the feature channels to

Table 3.6: Quantitative evaluation on accelerated MRI reconstruction for the region-guided channel-wise attention network (RG-CANet). Separable convolutions are used for the variant RG-CANet-S in place of conventional convolutions in each subnetwork of RG-CANet, leading to a smaller model size to demonstrate that the improved performance is attributable to the devised region-guided attention mechanism and the network structure, and not simply owing to the parameter size. $4\times$ accelerated reconstruction has the same run-time and parameter size as $8\times$.

	method	PSNR $\uparrow$	SSIM $\uparrow$	FID $\downarrow$	KID $\downarrow$	Latency(s) $\downarrow$	Size(MB) $\downarrow$
$8\times$	RG-CANet	28.65	0.758	74.26	0.012	0.049	11673342
	RG-CANet-S	28.12	0.747	80.00	0.014	0.041	6007180
	ASGAN	25.45	0.638	104.34	0.036	0.056	4459479
	Unet [56]	25.82	0.703	160.35	0.121	0.013	2754384
	MoDL [64]	27.13	0.620	143.65	0.080	0.091	5849787
	MICCAN [66]	26.61	0.642	180.66	0.146	0.043	2649455
	zero-filled	20.54	0.388	423.32	0.533	-	-
$4\times$	RG-CANet	32.22	0.854	57.18	0.003	-	-
	RG-CANet-S	32.01	0.850	58.15	0.004	-	-
	ASGAN	27.73	0.711	82.18	0.016	-	-
	Unet [56]	28.35	0.771	118.07	0.061	-	-
	MoDL [64]	30.34	0.745	98.86	0.042	-	-
	MICCAN [66]	30.11	0.711	99.44	0.040	-	-
	zero-filled	23.94	0.486	255.06	0.239	-	-

maintain the parameter size for a fair comparison. The backbone block adopted in the RG-CAM has significantly more parameters than the attention branch and is therefore kept when removing the RG-CAMs from the model. For the same reason, the two convolutional layers in the DCB are also kept. The quantitative results are displayed in Table 3.7, showing that the proposed model components consistently enhance the reconstruction performance, validating their effectiveness.

Table 3.7: Ablation studies on model components of the region-guided channel-wise attention network (RG-CANet) at $8\times$ acceleration factor.

method	PSNR $\uparrow$	SSIM $\uparrow$	FID $\downarrow$	KID $\downarrow$
RG-CANet	28.65	0.758	74.26	0.012
w/o DCB	27.33	0.731	83.43	0.017
w/o dense	28.03	0.748	81.25	0.015
w/o RG-CAM	28.29	0.753	78.67	0.014
w/o RG	28.26	0.754	77.40	0.013

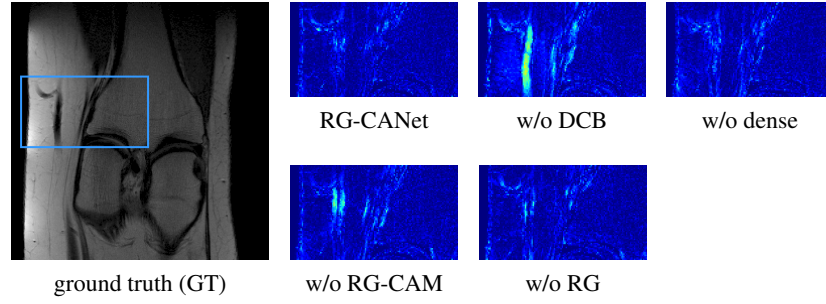


Figure 3.18: Ablation residual maps ($2\times$ amplified) of $8\times$ accelerated reconstruction for the region-guided channel-wise attention network (RG-CANet).

We utilize a variant w/o RG to evaluate the role of the region-based guidance in the RG-CAM by performing channel-wise attention using a single predicted kernel, similar to the approach used in [148, 146, 66]. The results shown in Table 3.7 demonstrate that the devised region-guided attention method improves the model performance, whereas the attention mechanism of variant w/o RG applied over channels without the use of the spatial guiding masks fails to obtain notable reconstruction improvements, comparing variant w/o RG-CAM with w/o RG. From the residual maps presented in Fig. 3.18, it can be observed that the devised region-guided attention model produces more accurate results with fewer reconstruction errors, as compared to other candidates. Those findings confirm the efficacy of the proposed RG-CAM and the resultant cascaded attention framework.

3.3.2.3 Region-Guided Attention Visualization

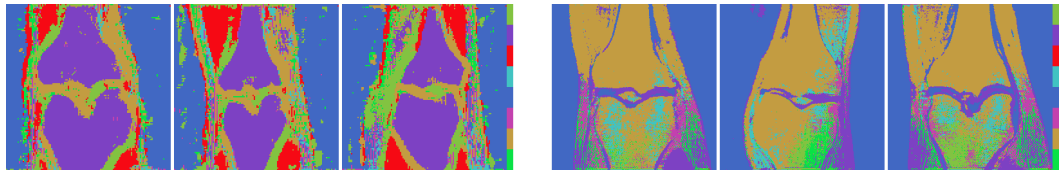


Figure 3.19: Visualization of the spatial guiding masks from (left) penultimate decoding level of the 4-th subnetwork and (right) last decoding level of the 5-th subnetwork, by shading in different colors the non-zero values of the masks amalgamated into a single map.

To demonstrate the spatial semantics learned by the proposed RG-CAM, we visualize the spatial guiding masks in Fig. 3.19 by amalgamating them into a single map with non-zero values shaded in different colors. From Fig. 3.19, it can be observed that the map of the underlying anatomy is “segmented” into several groups clearly indicating different structures, with an implication of the implicitly captured spatial information encoded in the guiding masks where pixels from similar regions are potentially assigned the same value and clustered to form semantic patterns. The region-based attention guidance is end-to-end learned along with the parameters of reconstruction models and applied to increase the ability of channel-wise attention mechanism without needing any annotations for supervision.

3.4 Discussions and Conclusion

In this chapter, we introduced deep learning-based models for accelerated MRI reconstruction. An attention selection generative adversarial network termed ASGAN was devised to produce images from under-sampled data with fine-grained details. However, the lack of domain-specific knowledge limits the model capacity, leading to degraded performance and artifacts in reconstructions. To address this issue, an optimization-inspired reconstruction framework was designed to pursue more faithful results, which is characterized by the proposed model components and structure: adaptive spatial attention selection module, deep data consistency block, and multi-level dense connections devised to encourage high-fidelity reconstructions and gain performance improvements. By experiments, the effectiveness of the proposed method was demonstrated, showing superior performance over other competing approaches.

The enormous computational costs and memory consumption of attention methods prohibit their application to high-resolution features which tend to portray local details and are closely linked with image reconstruction tasks, potentially limiting the representation of neural network models. To address this problem, a novel region-guided channel-wise attention network

was proposed to leverage a channel-wise attention mechanism under the guidance of the implicitly captured spatial cues via a light-weight and efficient design. By experiments, it was shown that the proposed model is able to yield superior performance as compared to other competing methods at different acceleration rates, validated both qualitatively and quantitatively.

It was demonstrated by experiments that adding noise to k -space measurement values can unfavorably affect the reconstruction quality and produce less faithful textures, albeit with the appearance and structural details retained, having an implication that it presents a more challenging problem as a combination of image reconstruction and de-noising. It is a promising research direction to couple image reconstruction with a de-noising task in hopes of improving the robustness and generalization capability of network models.

In experiments, Cartesian sampling patterns, which are often employed in real-world applications and used by comparison methods, were adopted. Non-Cartesian sampling methods such as radial sampling, which can be used as an alternative to Cartesian sampling strategies, are also common practice. Although demonstrating an effective sampling type, the forward data acquisition process relying on non-Cartesian masks can cause interpolation errors possibly propagated over the reconstruction pipeline and severely affecting the model performance [106], due to the use of the non-uniform fast Fourier transform and gridding [162, 163]. We therefore adopted Cartesian sampling methods in experiments and leave it for future work to explore the reconstruction models particularly suitable for other types of non-Cartesian sampling strategies.

Coil-Agnostic Attention-Based Networks for Accelerated Parallel MR Imaging

4.1 Introduction

As a non-invasive imaging modality, magnetic resonance imaging (MRI) is widely applied to visualize different types of anatomical structures and soft tissues. However, the acquisition speed of raw data is fundamentally restricted by both hardware and physiological conditions. It is a common practice to accelerate the process by under-sampling strategies in k -space and reconstruct images from a subset of measurements with the object of resisting aliasing artifacts in image domain due to the sampling procedure working below the Nyquist-Shannon rate. Compressed sensing (CS) methods model the reconstruction of high-quality MR images as a sparsity-regularized optimization problem which is normally solved via iterative model-based algorithms [138, 45, 139]. Such methods have several drawbacks including that the assumed sparsity prior can be difficult to meet in real-world scenarios and capture complex image features [57] and the iterative solvers are often computationally intensive, which hampers the development of CS-based approaches in modern MRI.

Parallel imaging (PI) exploits the connections among multiple coil sensors to recover the signals sampled in an accelerated acquisition configuration. A typical parallel imaging method is GRAPPA [87] which estimates the missing k -space pixels with the observed multi-coil lines in the auto-calibration signal (ACS) area. Despite useful in practice, those methods still struggle to prevent strong aliasing artifacts with high acceleration factors via traditional reconstruction approaches. In recent years, deep learning-based models have demonstrated great success and effectiveness in varieties of image tasks [8, 141, 164, 165, 166], which also showed great potential to improve the quality of MR images recovered from the under-sampled k -space observations by coupling deep neural networks into the reconstruction process, leveraging their excellent representation ability acquired to generalize across a dataset of samples [57, 167, 145, 168, 91].

In many PI methods for MRI reconstruction, the coil correlations are expressed in the form of sensitivity maps which are demanded in the forward process of image encoding and used to perform the k -space data constraint as a means to mine the information from the observations. It is a conventional tactic to estimate the unknown sensitivity maps from the acquired ACS lines using model-based algorithms, e.g. ESPIRiT [74], and incorporate them during the reconstruction process [64, 82]. A downside of this type of methods is that the model performance can be adversely affected when the sampled data are insufficient to guarantee the accuracy of estimation due to high acceleration rates, which potentially leads to unsatisfactory reconstruction results [94, 95].

A distinct class of approaches were proposed to mitigate the inaccurate estimation problem of traditional methods, which achieve improved reconstruction performance by jointly predicting an MR image as the target sensitivity-weighted combination of fully sampled coil frames and the associated sensitivity maps from the multi-coil measurements. Among them, Deep J-Sense [102] was introduced to jointly refine the image and the estimated coil sensitivities by unfolding an alternating optimization algorithm with neural networks, showing to yield promising results when fewer central lines are available to estimate the sensitivity maps. To increase the efficacy of sensitivity estimation and therefore enhance the reconstruction quality,

Joint-ICNet [100] updates the sensitivity maps at each step of an alternating framework, which is combined with the reconstruction procedure. Since neural networks rely on the multi-coil signals individually measured by coil sensors to predict the coil sensitivities, those methods can be sensitive to coil configuration discrepancies and influenced by estimation errors. Moreover, more parameters are required for the estimation of sensitivities, which inevitably increases computational costs.

We propose an optimization-inspired framework here to address the forward pass problem conventionally involved in multi-coil MRI reconstruction, which leverages a novel data aggregation consistency block termed DACB to enforce the k -space signal consistency in a completely coil-agnostic manner, hence invulnerable and robust to different coil configurations. A locality-aware spatial attention module is introduced in the reconstruction pipeline to capture spatial contexts with data-adaptive kernels and incorporated with the DACB in a multi-level densely connected architecture, which shows to improve the reconstruction performance. We conduct extensive experiments on texture-rich FastMRI knee dataset to verify the effectiveness of the proposed coil-agnostic method and demonstrate its superiority against other sensitivity estimation-based approaches.

4.2 Methodology

4.2.1 Problem Formulation

The forward process of parallel MRI can be expressed as the acquisition of a set of measurements simultaneously from multiple coils, which allows accelerated scanning by sampling a subset of k -space points and recover the signals by utilizing the redundant information contained in multiple coil views to alleviate aliasing artifacts. During this process, the sensitivity maps $\mathbf{S}_i$ are defined as the induced magnetic field at a specified point per unit supplied power, describing how the target image is projected onto the field of each coil, leading to different observations. Let $\mathbf{x}$ and $\{\mathbf{y}_i\}_i$ denote the desired image and its measurements with the subscript

i indexing coils, $\mathbf{F}$ denote the Fourier transform, and $\mathbf{D}$ refer to the under-sampling operation, i.e. the encoding matrix. The acquisition of multi-coil MRI is expressed as below,

$$\mathbf{y}_i = \mathbf{DFS}_i \mathbf{x}, \quad (4.1)$$

where the sampling pattern is shared by all coils and the sensitivity profile is made to vary across different coils. The inverse problem is to reconstruct the image from a collection of under-sampled measurements, which is normally cast as an optimization problem shown below, with a regularization term $R(\mathbf{x})$ to constrain the search space for a more accurate solution,

$$\min_{\mathbf{x}} \sum_{i=1}^{n_c} (\lambda \|\mathbf{DFS}_i \mathbf{x} - \mathbf{y}_i\|^2) + R(\mathbf{x}), \quad (4.2)$$

where the first part refers to the data fidelity term individually applied to multiple coil frames and n_c is the number of deployed coil sensors.

4.2.2 Unrolled Optimization-Inspired Pipeline for Multi-Coil MRI

To solve the minimization problem in (4.2), an auxiliary variable $\mathbf{u}$ is employed, which converts the original problem into the following form,

$$\min_{\mathbf{x}, \mathbf{u}} \sum_{i=1}^{n_c} (\lambda \|\mathbf{DFS}_i \mathbf{x} - \mathbf{y}_i\|^2) + R(\mathbf{u}) \text{ s.t. } \mathbf{x} = \mathbf{u}.$$

The unstrained expression is derived using a Lagrange multiplier α , yielding:

$$\min_{\mathbf{x}, \mathbf{u}} \sum_{i=1}^{n_c} (\lambda \|\mathbf{DFS}_i \mathbf{x} - \mathbf{y}_i\|^2) + R(\mathbf{u}) + \alpha \|\mathbf{u} - \mathbf{x}\|^2, \quad (4.3)$$

with λ and α controlling the regularization strength. We divide (4.3) into two sub-problems alternating the minimization of the two variables, obtaining:

$$\mathbf{u}^{k+1} \leftarrow \arg \min_{\mathbf{u}} R(\mathbf{u}) + \alpha \|\mathbf{u} - \mathbf{x}^k\|^2 \quad (4.4a)$$

$$\mathbf{x}^{k+1} \leftarrow \arg \min_{\mathbf{x}} \sum_{i=1}^{n_c} (\lambda \|\mathbf{DFS}_i \mathbf{x} - \mathbf{y}_i\|^2) + \alpha \|\mathbf{u}^{k+1} - \mathbf{x}\|^2. \quad (4.4b)$$

We take the advantage of the powerful representation of neural networks to learn the deep prior of MR images from data, and solve the sub-problem in (4.4a) via a forward pass of a neural network model. The solution of the second equation of (4.4) is encouraged to be consistent with the measurements observed at different coils and also preferred to be close to the updated variable $\mathbf{u}^{k+1}$. To solve (4.4b) in a closed form, the derivative of the minimization term with respect to $\mathbf{x}$ is computed and set to be zero, obtaining:

$$\sum_{i=1}^{n_c} \lambda \mathbf{S}_i^H \mathbf{F}^H \mathbf{D}^H (\mathbf{DFS}_i \mathbf{x} - \mathbf{y}_i) + \alpha (\mathbf{x} - \mathbf{u}^{k+1}) = \mathbf{0}, \quad (4.5)$$

which is converted into the following form:

$$\sum_{i=1}^{n_c} \mathbf{S}_i^H \underbrace{[\lambda \mathbf{F}^H \mathbf{D}^H \mathbf{DFS}_i \mathbf{x} + \alpha \mathbf{S}_i \mathbf{x} - \lambda \mathbf{F}^H \mathbf{y}_i - \alpha \mathbf{S}_i \mathbf{u}^{k+1}]}_{(a)} = \mathbf{0}, \quad (4.6)$$

by exploiting the nature of coil sensitivity and the properties of MRI encoding process expressed in (4.7),

$$\begin{aligned} \sum_{i=1}^{n_c} \mathbf{S}_i^H \mathbf{S}_i &= \mathbf{I} \\ \mathbf{D} \mathbf{y}_i &= \mathbf{y}_i \\ (\mathbf{I} - \mathbf{D}) \mathbf{y}_i &= \mathbf{0} \\ \mathbf{D}^H \mathbf{D} &= \mathbf{D}, \mathbf{F}^H \mathbf{F} = \mathbf{F} \mathbf{F}^H = \mathbf{I}, \end{aligned} \quad (4.7)$$

also noting $\mathbf{D}^H \mathbf{y}_i = \mathbf{D}^H \mathbf{D} \mathbf{y}_i = \mathbf{D} \mathbf{y}_i = \mathbf{y}_i$. We set (a) in (4.6) to be zero as a sufficient condition, which yields:

$$\lambda \mathbf{F}^H \mathbf{D} \mathbf{F} \mathbf{S}_i \mathbf{x} + \alpha \mathbf{S}_i \mathbf{x} = \lambda \mathbf{F}^H \mathbf{y}_i + \alpha \mathbf{S}_i \mathbf{u}^{k+1}. \quad (4.8)$$

We can observe that the left part of (4.8) consists of two elements all containing the term $\mathbf{S}_i \mathbf{x}$, one of which is transformed by $\mathbf{F}$ and mapped backward after being under-sampled via $\mathbf{D}$ and the other of which is simply weighted by a scalar. A similar structure can be seen from the right part except for $\mathbf{F}^H \mathbf{y}_i$ since $\mathbf{y}_i$ is already the under-sampled k -space signal, which allows us to simplify the equation by multiplying $\mathbf{F}$ to both sides and substituting $\mathbf{m}_i$ for $\mathbf{S}_i \mathbf{x}$, giving:

$$\begin{aligned} (\lambda \mathbf{D} + \alpha \mathbf{I}) \mathbf{F} \mathbf{m}_i &= \lambda \mathbf{y}_i + \alpha \mathbf{F} \mathbf{S}_i \mathbf{u}^{k+1} \\ \Rightarrow \mathbf{F} \mathbf{m}_i &= (\lambda \mathbf{D} + \alpha \mathbf{I})^{-1} (\lambda \mathbf{y}_i + \alpha \mathbf{F} \mathbf{S}_i \mathbf{u}^{k+1}). \end{aligned} \quad (4.9)$$

The burden in finding the solution to the iterative minimization is to compute the matrix inversion $(\lambda \mathbf{D} + \alpha \mathbf{I})^{-1}$. We leverage the property of diagonal matrix inversion to rewrite it in the following form:

$$\begin{aligned} (\lambda \mathbf{D} + \alpha \mathbf{I})^{-1} &= \frac{1}{\lambda + \alpha} \mathbf{D} + \frac{1}{\alpha} (\mathbf{I} - \mathbf{D}) \\ &= \frac{1}{\alpha} \mathbf{I} + \left(\frac{1}{\lambda + \alpha} - \frac{1}{\alpha} \right) \mathbf{D}. \end{aligned} \quad (4.10)$$

Combining (4.9) with (4.10) and utilizing the nice properties presented in (4.7), we derive the relationship shown below,

$$\begin{aligned} \mathbf{F} \mathbf{m}_i &= \left(\frac{1}{\lambda + \alpha} \mathbf{D} + \frac{1}{\alpha} (\mathbf{I} - \mathbf{D}) \right) (\lambda \mathbf{y}_i + \alpha \mathbf{F} \mathbf{S}_i \mathbf{u}^{k+1}) \\ &= \frac{\lambda}{\lambda + \alpha} \underbrace{\mathbf{D} \mathbf{y}_i}_{=\mathbf{y}_i} + \frac{\lambda}{\alpha} \underbrace{(\mathbf{I} - \mathbf{D}) \mathbf{y}_i}_{=\mathbf{0}} + \mathbf{F} \mathbf{S}_i \mathbf{u}^{k+1} + \left(\frac{\alpha}{\lambda + \alpha} - 1 \right) \mathbf{D} \mathbf{F} \mathbf{S}_i \mathbf{u}^{k+1} \\ &= \mathbf{F} \mathbf{S}_i \mathbf{u}^{k+1} - \frac{\lambda}{\lambda + \alpha} (\mathbf{D} \mathbf{F} \mathbf{S}_i \mathbf{u}^{k+1} - \mathbf{y}_i). \end{aligned} \quad (4.11)$$

Multiplying $\mathbf{F}^H$ to (4.11) and setting $\eta = \frac{\lambda}{\lambda + \alpha}$, we obtain:

$$\mathbf{m}_i = \mathbf{S}_i \mathbf{u}^{k+1} - \eta \mathbf{F}^H (\mathbf{D} \mathbf{F} \mathbf{S}_i \mathbf{u}^{k+1} - \mathbf{y}_i). \quad (4.12)$$

The closed-form solution to the iteration of $\mathbf{x}$ is now expressed in the form of its projected views onto each coil, which needs to combine all coil frames together to give the update of $\mathbf{x}$. The relationship between an image and its sensitivity-based representations is given by:

$$\mathbf{x} = \sum_{i=1}^{n_c} \mathbf{S}_i^H \mathbf{m}_i. \quad (4.13)$$

Incorporating (4.12) and (4.13) with the alternating optimization expressed in (4.4), the iterative algorithm of the inverse reconstruction problem is given as follows,

$$\mathbf{u}^{k+1} \leftarrow f_{\theta_k}(\mathbf{x}^k) \quad (4.14a)$$

$$\left. \begin{aligned} \mathbf{m}_i &= \overbrace{\mathbf{S}_i \mathbf{u}^{k+1}}^{\text{projection}} - \eta \mathbf{F}^H (\mathbf{D} \mathbf{F} \overbrace{\mathbf{S}_i \mathbf{u}^{k+1}}^{\text{projection}} - \mathbf{y}_i) \\ \mathbf{x}^{k+1} &= \underbrace{\sum_{i=1}^{n_c} \mathbf{S}_i^H \mathbf{m}_i}_{\text{combination}} \end{aligned} \right\} \text{data consistency}, \quad (4.14b)$$

where f_{θ_k} is a deep neural network used to refine the intermediate reconstruction from the previous iteration. The solution given in (4.14) comprises two iterative steps: the first step exploits a neural network to estimate the image from the previous reconstruction and the second one corrects the estimate by enforcing the data consistency (DC) with the collected measurements. Since the measurements are observed at multiple coils, the update step defined in (4.14b) involves the use of sensitivity maps which are required to be known in the forward and backward processes to perform coil projection and combination. However, the estimation of sensitivity maps can be time-consuming and sensitive to different scanning configurations. Inaccurate sensitivity estimation can have a negative impact on the final reconstruction, leading to degraded model performance. In the following section, we introduce a novel method for the data consistency iteration, which completely circumvents the estimation of sensitivity maps and induces a coil-agnostic reconstruction framework.

4.2.3 Data Aggregation Consistency Block (DACB)

Multi-coil MRI scanning models the forward process with the profiles of coil sensitivity, which are normally involved in model-based iterative algorithms to accommodate the k -space constraint, e.g. as presented in (4.14b). To avoid the problems associated with the estimation of sensitivity maps, we propose a data aggregation consistency block, dubbed DACB, to exhibit the data consistency operation without using the sensitivity maps. We present the details of DACB below.

4.2.3.1 Data Consistency in MRI Reconstruction

Data consistency plays an important role in deep learning-based MRI reconstruction methods built upon unrolling a model-based iterative algorithm, as an attempt to encourage the reconstructed image to be consistent with the observed k -space measurements, thus yielding more faithful results to the targets. A variety of DC operations are introduced in the literature, e.g. MoDL [64] and VS-Net [82], of which we derive a general expression to manifest the intrinsic mechanism of data consistency enforcement. Recalling the equation in (2.48b) of the DCB block in the iteration of VS-Net, we rewrite it with the sensitivity maps omitted to only focus on the update of the estimated data, giving rise to:

$$\mathbf{x} \leftarrow \mathbf{F}^H (\lambda \mathbf{D}^H \mathbf{D} + \alpha \mathbf{I})^{-1} (\lambda \mathbf{D}^H \mathbf{y} + \alpha \mathbf{F} \mathbf{x}). \quad (4.15)$$

Generally, the image being updated is denoted by $\mathbf{x}$ and the measurement imposing the k -space constraint is referred to as $\mathbf{y}$. The equivalence of it to (4.9) can be attained by utilizing the equations in (4.7) and applying $\mathbf{F}$ to both sides, which implies that it can be solved using a simpler solution following the expression in (4.12), obtaining:

$$\mathbf{x} \leftarrow \mathbf{x} - \eta \mathbf{F}^H (\mathbf{D} \mathbf{F} \mathbf{x} - \mathbf{y}), \quad (4.16)$$

which shows that the solution to the data fidelity based minimization is to correct the previous estimate using the difference from the observation when being mapped into k -space. We call the update term $\mathbf{F}^H(\mathbf{D}\mathbf{F}\mathbf{x} - \mathbf{y})$ the data consistency prior as it computes the errors between the reconstruction and the observed measurement and indicates the direction to a more faithful solution.

Similarly, the DC method adopted in MoDL takes the matrix inversion to solve the minimization as presented in (2.43b), giving:

$$\mathbf{x} \leftarrow (\mathbf{A}^H \mathbf{A} + \alpha \mathbf{I})^{-1} (\mathbf{A}^H \mathbf{y} + \alpha \mathbf{x}). \quad (4.17)$$

Taking $\mathbf{A} = \mathbf{D}\mathbf{F}$ into the matrix inversion, the relationship below can be induced:

$$\begin{aligned} (\mathbf{A}^H \mathbf{A} + \alpha \mathbf{I})^{-1} &= (\mathbf{F}^H \mathbf{D}^H \mathbf{D} \mathbf{F} + \alpha \mathbf{F}^H \mathbf{F})^{-1} \\ &= (\mathbf{F}^H (\mathbf{D}^H \mathbf{D} + \alpha \mathbf{I}) \mathbf{F})^{-1} \\ &= \mathbf{F}^H (\mathbf{D}^H \mathbf{D} + \alpha \mathbf{I})^{-1} \mathbf{F}. \end{aligned} \quad (4.18)$$

The equation in (4.17) can be expressed in the following form by substituting (4.18) into it:

$$\begin{aligned} \mathbf{x} &\leftarrow \mathbf{F}^H (\mathbf{D}^H \mathbf{D} + \alpha \mathbf{I})^{-1} \mathbf{F} (\mathbf{A}^H \mathbf{y} + \alpha \mathbf{x}) \\ &= \mathbf{F}^H (\mathbf{D}^H \mathbf{D} + \alpha \mathbf{I})^{-1} (\mathbf{F} \mathbf{F}^H \mathbf{D}^H \mathbf{y} + \alpha \mathbf{F} \mathbf{x}) \\ &= \mathbf{F}^H (\mathbf{D}^H \mathbf{D} + \alpha \mathbf{I})^{-1} (\mathbf{y} + \alpha \mathbf{F} \mathbf{x}), \end{aligned} \quad (4.19)$$

which can be easily shown to be equivalent to (4.15) when taking $\lambda = 1$ and noting $\mathbf{D}^H \mathbf{y} = \mathbf{y}$.

The DC operation used in MoDL can be analytically implemented as follows,

$$(\mathbf{F}\mathbf{x})(q) \leftarrow \begin{cases} (\frac{\mathbf{y} + \alpha \mathbf{F}\mathbf{x}}{1 + \alpha})(q) & \text{if sampling at the } q\text{-th position} \\ (\mathbf{F}\mathbf{x})(q) & \text{otherwise,} \end{cases} \quad (4.20)$$

which showcases the mechanism of data consistency methods: it keeps the estimated values at the k -space locations where the points are not sampled, and interpolates the update using the estimate and measurement otherwise. Due to the preservation of the predicted values in sampled positions, we call this type of DC methods the soft data consistency operation.

The data consistency constraint involved in the update steps introduced above in each iteration of model-based reconstruction algorithms can be implemented using the expression provided in (4.16), which can be rearranged as follows,

$$\begin{aligned}
 \mathbf{F}\mathbf{x} &\leftarrow \mathbf{F}\mathbf{x} - \eta(\mathbf{D}\mathbf{F}\mathbf{x} - \mathbf{y}) \\
 &= \mathbf{F}\mathbf{x} - \eta\mathbf{D}\mathbf{F}\mathbf{x} + \eta\mathbf{y} \\
 &= (\mathbf{I} - \eta\mathbf{D})\mathbf{F}\mathbf{x} + \eta\mathbf{y} \\
 &= ((1 - \eta)\mathbf{I} + \eta(\mathbf{I} - \mathbf{D}))\mathbf{F}\mathbf{x} + \eta\mathbf{y} \\
 &= (1 - \eta)\mathbf{F}\mathbf{x} + \eta \underbrace{((\mathbf{I} - \mathbf{D})\mathbf{F}\mathbf{x} + \mathbf{y})}_{(a)}. \tag{4.21}
 \end{aligned}$$

The term (a) in (4.21) can be analytically executed in k -space by:

$$(\mathbf{F}\mathbf{x})(q) \leftarrow \begin{cases} \mathbf{y}(q) & \text{if sampling at the } q\text{-th position} \\ (\mathbf{F}\mathbf{x})(q) & \text{otherwise,} \end{cases} \tag{4.22}$$

which shows to completely replace the predictions with the measurement values at sampled positions as compared to the soft DC method. We define the inverse transform of it as the hard data consistency operation as shown below:

$$\Omega(\mathbf{x}; \mathbf{A}, \mathbf{y}) = \mathbf{F}^H((\mathbf{I} - \mathbf{D})\mathbf{F}\mathbf{x} + \mathbf{y}). \tag{4.23}$$

In a multi-coil scenario, the expression of (4.23) is naturally rewritten as:

$$\Omega(\mathbf{S}_i\mathbf{x}; \mathbf{A}, \mathbf{y}_i) = \mathbf{F}^H((\mathbf{I} - \mathbf{D})\mathbf{F}\mathbf{S}_i\mathbf{x} + \mathbf{y}_i), \tag{4.24}$$

where the subscript i implies that the specified variables correspond to the i -th coil. The data consistency step in (4.16) can be computed using (4.23) by the following equation,

$$\mathbf{x} \leftarrow \gamma \mathbf{x} + (1 - \gamma) \Omega(\mathbf{x}; \mathbf{A}, \mathbf{y}), \quad (4.25)$$

where the weight is set to be $\gamma = 1 - \eta$.

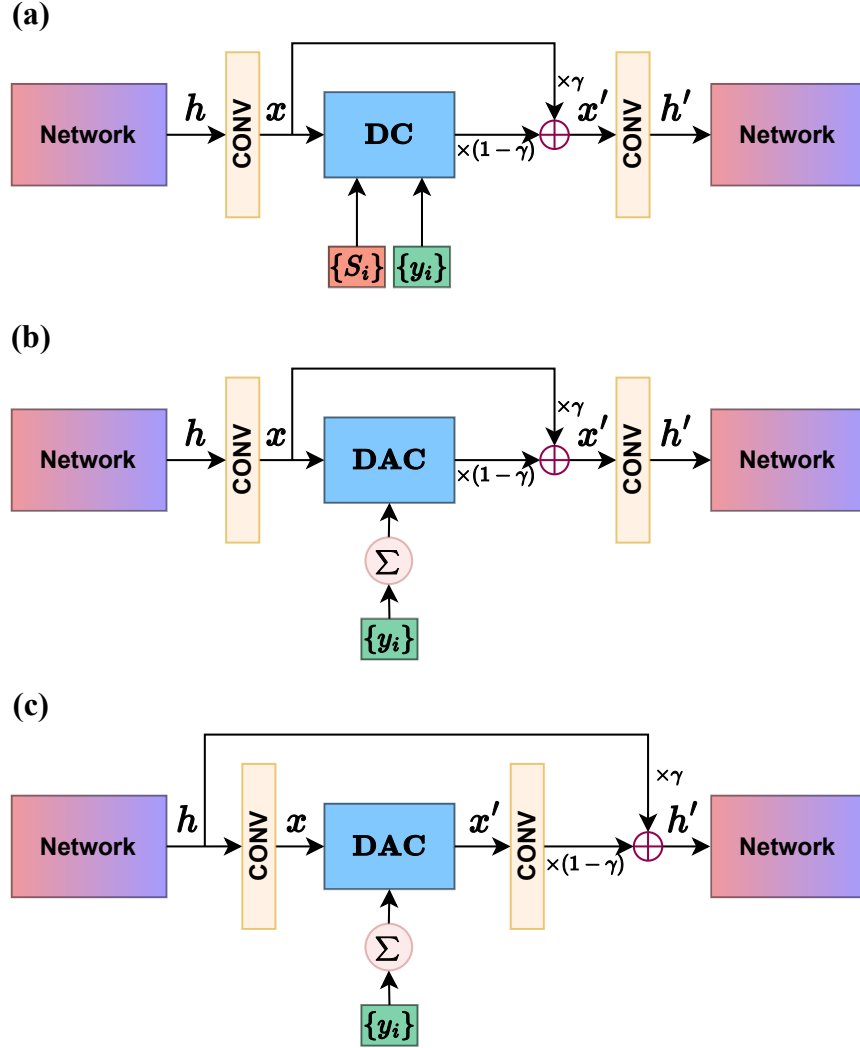


Figure 4.1: Illustration of data consistency methods for multi-coil MRI: a) conventional data consistency (DC) method using the pre-computed sensitivity maps, b) coil-agnostic data consistency method based on the data aggregation consistency (DAC) operation, and c) data aggregation consistency block (DACB) generalizing the DAC method to feature level, which is embedded into each subnetwork of the reconstruction framework, as illustrated in Fig. 4.4, to correct the intermediate prediction.

4.2.3.2 Multi-Coil Data Consistency Method

The expression of data consistency in multi-coil settings is derived in the following, which lays the foundation for the proposed coil-agnostic DC method. For multi-coil MRI reconstruction, the data fidelity is individually enforced with the measurements sensed by different coils, which needs the sensitivity maps to simulate the forward scanning process, and subsequently the sensitivity-weighted average is computed, as shown in (4.14b), to attain the updated image. The motivation of the adoption of the sensitivity-based combination is to reverse the sensitivity projection process to give the combined image, which can be obtained by the following equation,

$$\sum_{i=1}^{n_c} \mathbf{S}_i^H (\mathbf{S}_i \mathbf{x}) = \left(\sum_{i=1}^{n_c} \mathbf{S}_i^H \mathbf{S}_i \right) \mathbf{x} = \mathbf{x}, \quad (4.26)$$

where the normalization property of coil sensitivity, i.e. $\sum_{i=1}^{n_c} \mathbf{S}_i^H \mathbf{S}_i = \mathbf{I}$, is utilized. To provide a compact form of the multi-coil data consistency, we define the manipulations of signals using sensitivity maps as follows,

$$\begin{aligned} \Psi(\mathbf{x}; \mathbf{S}_i) &= \mathbf{S}_i \mathbf{x} \\ \Phi(\{\mathbf{x}_i\}; \{\mathbf{S}_i\}) &= \sum_{i=1}^{n_c} \mathbf{S}_i^H \mathbf{x}_i. \end{aligned} \quad (4.27)$$

The signal sensing process appearing at the i -th coil in the forward acquisition is delineated by Ψ , and the sensitivity-weighted average used to produce an image from multiple coil frames is modelled by Φ . Combining (4.24), (4.25), and (4.27) together, the data consistency enforcement for parallel MR imaging can be achieved using the following equation,

$$\mathbf{x} \leftarrow \Phi(\{\gamma \Psi(\mathbf{x}; \mathbf{S}_i) + (1 - \gamma) \Omega(\Psi(\mathbf{x}; \mathbf{S}_i); \mathbf{A}, \mathbf{y}_i)\}; \{\mathbf{S}_i\}). \quad (4.28)$$

Leveraging the linear property of the two sensitivity-based operations Ψ and Φ , the following relationships can be derived:

$$\begin{aligned}
 \Phi \circ \Psi(\mathbf{x}) &= \Phi(\{\Psi(\mathbf{x}; \mathbf{S}_i)\}; \{\mathbf{S}_i\}) \\
 &= \sum_{i=1}^{n_c} \mathbf{S}_i^H (\mathbf{S}_i \mathbf{x}) \\
 &= \left(\sum_{i=1}^{n_c} \mathbf{S}_i^H \mathbf{S}_i \right) \mathbf{x} \\
 &= \mathbf{I} \mathbf{x} = \mathbf{x},
 \end{aligned} \tag{4.29}$$

and let $\mathbf{x}_i = \mathbf{S}_i \mathbf{x}$, yielding:

$$\begin{aligned}
 \Psi \circ \Phi(\mathbf{x}_i) &= \Psi(\Phi(\{\mathbf{x}_j\}; \{\mathbf{S}_j\}); \mathbf{S}_i) \\
 &= \mathbf{S}_i \sum_{j=1}^{n_c} \mathbf{S}_j^H \mathbf{x}_j = \mathbf{S}_i \sum_{j=1}^{n_c} \mathbf{S}_j^H (\mathbf{S}_j \mathbf{x}) \\
 &= \mathbf{S}_i \left(\sum_{j=1}^{n_c} \mathbf{S}_j^H \mathbf{S}_j \right) \mathbf{x} \\
 &= \mathbf{S}_i (\mathbf{I} \mathbf{x}) = \mathbf{x}_i.
 \end{aligned} \tag{4.30}$$

The equations shown above demonstrate that the coil sensitivity operations Ψ and Φ describe the inverse process of each other, i.e. one reverses the effect of the other. In general, the combination of the projected views gives the wanted image and each coil view can be produced from the weighted average via the sensing process. As demonstrated in Fig. 4.1 (a), the update shown in (4.28) can be simplified as follows,

$$\mathbf{x} \leftarrow \gamma \mathbf{x} + (1 - \gamma) \underbrace{\Phi(\{\Omega(\Psi(\mathbf{x}; \mathbf{S}_i); \mathbf{A}, \mathbf{y}_i)\}; \{\mathbf{S}_i\})}_{(a)}, \tag{4.31}$$

which carries out the data consistency step via interpolation using the previous estimate and the sensitivity-weighted combination of coil projections of it with the k -space values at sampled positions replaced by the observed measurements, i.e. applying a hard DC to coil views coming from the simulated forward sensing process.

4.2.3.3 Coil-Agnostic Data Consistency Method

For multi-coil MRI reconstruction, the k -space constraint is normally executed with the use of sensitivity maps which are either estimated by model-based algorithms [64, 82, 73], or via deep neural networks with trainable parameters [95, 100]. The former approach introduces non-trivial computational costs and can have a severe impact on the final reconstructions when the accuracy of sensitivity estimation cannot be guaranteed with limited ACS lines. The latter method copes with this problem by embedding the sensitivity estimation into an end-to-end reconstruction framework, whereas it increases the model parameters and can be sensitive to scanning configurations, e.g. the number of coils and their deployment order, which limits the generalization ability of reconstruction systems.

The focal element of data consistency, as discussed above, is to “correct” the previously predicted values with the observed k -space data to ensure reconstruction fidelity, during which the forward process is required to be known in conventional methods. Instead of demanding the presence of sensitivity maps in data consistency operations, e.g. (4.14b) and (4.31), we propose a simple yet effective method to perform approximated data consistency without using sensitivity maps. Considering the embedding of data consistency layers in an unfolded neural network, the intermediate reconstruction $\mathbf{x}$ to be passed to a DC block, e.g. (4.31), is given by a parametric function f at the end of the previous reconstruction network, e.g. a convolutional layer of a CNN-based model, with a feature volume dubbed h , which normally reduces the feature channels to provide complex-valued signals and can be regarded as a mapping from feature space to image domain. Conventional DC methods need to employ the manipulations Ψ and Φ to deal with a collection of measurements and their sensitivity maps, which we propose to circumvent as a means to avoid the use of sensitivity maps, by learning a coil-invariant mapping end-to-end.

Expanding the term (a) in (4.31) obtains

$$\Phi(\{\Omega(\Psi(\mathbf{x}; \mathbf{S}_i); \mathbf{A}, \mathbf{y}_i)\}; \{\mathbf{S}_i\}) = \sum_{i=1}^{n_c} \mathbf{S}_i^H \mathbf{F}^H ((\mathbf{I} - \mathbf{D}) \mathbf{F} \mathbf{S}_i \mathbf{x} + \mathbf{y}_i), \quad (4.32)$$

where the sensitivity maps have to be separated from their adjoint operations $\mathbf{S}_i^H$ due to the spatial variance property. Considering a special case that the sensitivity maps are treated as the identity multiplied by a scalar, i.e. $\mathbf{S}_i = \alpha_i \mathbf{I}$ satisfying $\sum_{i=1}^{n_c} |\alpha_i|^2 = 1$, the equation in (4.32) can be converted into the following expression:

$$\begin{aligned} \Phi(\{\Omega(\Psi(\mathbf{x}; \mathbf{S}_i); \mathbf{A}, \mathbf{y}_i)\}; \{\mathbf{S}_i\}) &= \mathbf{F}^H((\mathbf{I} - \mathbf{D})\mathbf{F} \sum_{i=1}^{n_c} (\mathbf{S}_i^H \mathbf{S}_i \mathbf{x}) + \sum_{i=1}^{n_c} \mathbf{S}_i^H \mathbf{y}_i) \\ &= \mathbf{F}^H((\mathbf{I} - \mathbf{D})\mathbf{F} \mathbf{x} + \sum_{i=1}^{n_c} \mathbf{S}_i^H \mathbf{y}_i) \\ &= \Omega(\mathbf{x}; \mathbf{A}, \sum_{i=1}^{n_c} \mathbf{S}_i^H \mathbf{y}_i), \end{aligned} \quad (4.33)$$

which yields the resulting DC update:

$$\mathbf{x} \leftarrow \gamma \mathbf{x} + (1 - \gamma) \Omega(\mathbf{x}; \mathbf{A}, \sum_{i=1}^{n_c} \mathbf{S}_i^H \mathbf{y}_i), \quad (4.34)$$

where it only requires to utilize the sensitivity maps to aggregate the measurements prior to the implementation of the DC operation. Although the special case ensuring the validity of (4.34) is implausible in real applications, incorporating neural network models in the reconstruction pipeline allows to end-to-end learn a coil-invariant aggregation representation.

Inspired by the update rule in (4.31), we treat $\mathbf{x}$ generated with a volume of feature maps coming from the network module prior to the DC block as the complex-valued signal aggregation and execute the DC update using the summation of the measurements to avoid involving unknown coil sensitivities. As opposed to conventional deep multi-coil MRI reconstruction methods which respectively apply data consistency to each coil frame and combine their results together with sensitivity maps, we instead leverage the linearity of the Fourier transform, yielding that the aggregation of the transforms equals the transformed aggregation, and adopt the summation operator in place of coil combination. As illustrated in Fig. 4.1 (b), the updated estimate, dubbed $\mathbf{x}'$, is therefore computed by the equation shown below, termed the data

aggregation consistency (DAC) operation:

$$\begin{aligned}
 \mathbf{x}' &= \gamma \mathbf{x} + (1 - \gamma) \Omega(\mathbf{x}; \mathbf{DF}, \sum_{i=1}^{n_c} \mathbf{y}_i) \\
 &= \gamma f(h) + (1 - \gamma) \underbrace{\Omega(f(h); \mathbf{DF}, \sum_{i=1}^{n_c} \mathbf{y}_i)}_{\text{DAC}}, \tag{4.35}
 \end{aligned}$$

where γ is either learnable or a predefined hyperparameter, and f is jointly trained with the reconstruction model and can be subsumed into the previous network module, leaving the DC block non-parametric.

The data aggregation consistency method shown in (4.35) has two advantages: 1) it requires no sensitivity map to simulate the acquisition process and directly applies the DC operation to the intermediate prediction coming from the previous network model, which consequently avoids reconstruction errors caused by the inaccurate sensitivity estimation, 2) due to the permutation-invariance property of the linear aggregation operation, it is inherently robust to various coil configurations.

4.2.3.4 Feature-Level Generalization

The bottleneck design in data consistency blocks is used to reduce the feature channels and produce an image mapped from feature space, which can potentially have a deleterious effect on the model performance due to the potential loss of rich information appearing in feature representations. To diminish the impact of the bottleneck design, we suggest adopting a skip-connection for feature maps to preserve the useful information presented in feature space and benefit signal transmission. To take the advantage of residual learning, we apply f^{-1} to both sides of (4.35), obtaining:

$$\begin{aligned}
 f^{-1}(\mathbf{x}') &= \gamma f^{-1} \circ f(h) + (1 - \gamma) f^{-1}(\Omega(f(h); \mathbf{DF}, \sum_{i=1}^{n_c} \mathbf{y}_i)) \\
 \Rightarrow h' &= \gamma h + (1 - \gamma) f^{-1}(\Omega(f(h); \mathbf{DF}, \sum_{i=1}^{n_c} \mathbf{y}_i)), \tag{4.36}
 \end{aligned}$$

where $h' = f^{-1}(\mathbf{x}')$. Considering that the reconstruction framework is end-to-end trained, we use two parametric structures f_1 and f_2 to take the place of f and its inverse, which are jointly optimized in training phase without imposing any constraint. The data aggregation consistency block, presented in Fig. 4.1 (c), is eventually implemented as follows,

$$h' = \gamma h + (1 - \gamma) f_2(\Omega(f_1(h); \mathbf{DF}, \sum_{i=1}^{n_c} \mathbf{y}_i)), \quad (4.37)$$

where two convolutional layers are used in our experiments for f_1 and f_2 , which can suffice to model the coil-invariant mapping, thanks to the joint optimization with the reconstruction modules end-to-end.

4.2.4 Locality-Aware Spatial Attention Module (LoA-SAM)

As a focal constituent component of transformers [20], attention modules have gained increasing attention due to the excellent capacity to retrieve relevant and useful information from the learned feature representations by exploiting the correlations of the underlying “tokens”. Suffering from the quadratic complexity with the number of tokens involved in the computations of attention correlations, their implementations to high-resolution dense prediction tasks such as image reconstruction can be challenging, which impedes the development of attention-based methods in MRI reconstruction. To exploit the relationship among relevant features to improve the reconstruction quality without causing a substantial computational increase, we propose a locality-aware spatial attention module termed LoA-SAM which learns to capture the local contextual information in an adaptive and data-specific diet, showing to obtain performance gains. We introduce the design of LoA-SAM which is composed of two structures: a relative feature aggregation block and adaptive spatial attention selection.

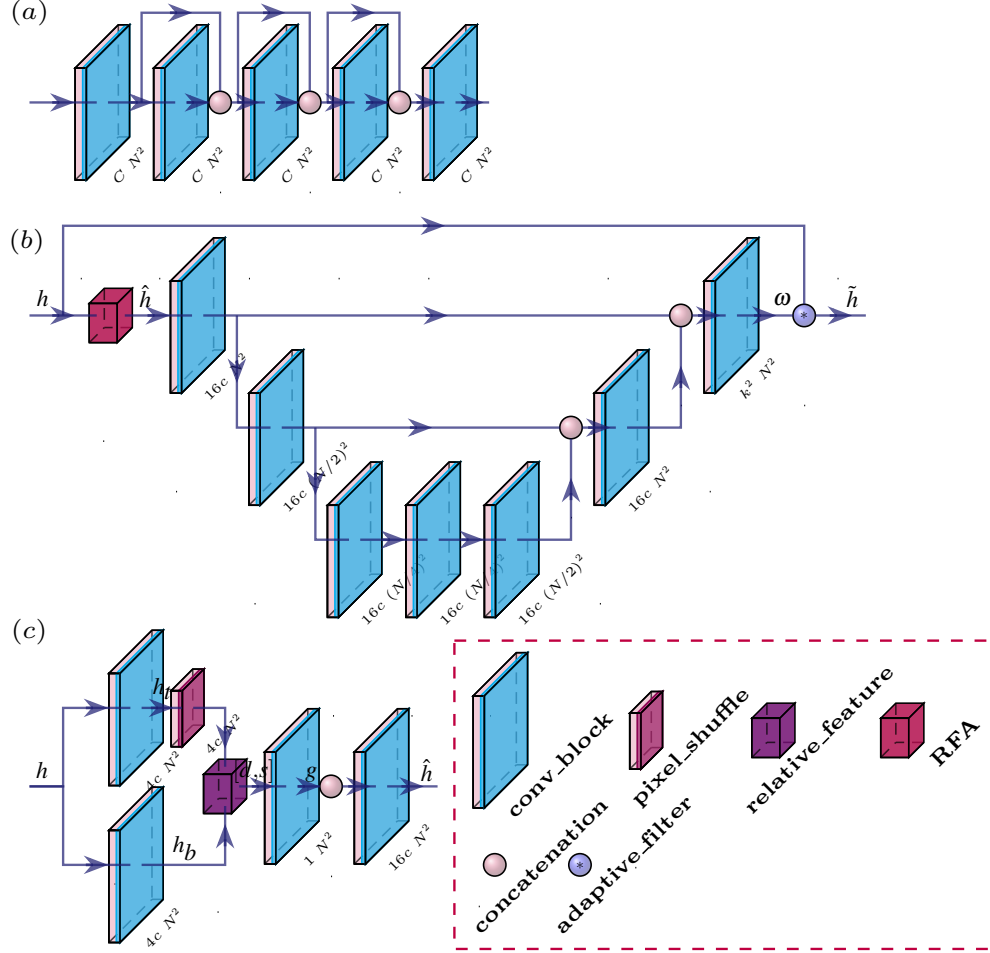


Figure 4.2: Illustration of basic modules: a) densely connected layers, b) locality-aware spatial attention module (LoA-SAM), and c) relative feature aggregation block (RFA). C and N^2 indicate the channel and spatial sizes of output features, noting that the spatial size is assumed to be $N \times N$, for simplicity.

4.2.4.1 Relative Feature Aggregation Block (RFA)

The relative information among extracted feature maps passed to LoA-SAM for refinement is explicitly captured to aggregate the spatial correlations. The input feature volume h having a size of $C \times H \times W$, where C refers to the number of channels and $H \times W$ denotes the spatial resolution, is mapped by two individual convolutional layers to h_t and h_b , as displayed in Fig. 4.2 (c). The feature volume h_t is compared with its counterpart h_b after being shifted by pixels, hence achieving alignment of values at different positions, which is expressed in the

form:

$$d(q; \Delta_i) = ps(h_t; \Delta_i)(q) - h_b(q), \quad (4.38)$$

where $d(q; \Delta_i)$ is the relative direction vector which delineates the difference of the two feature volumes in the learned linear embedding space, q denotes the spatial position, i.e. $h(q)$ refers to a vector constructed with the values of it at position q across the channel dimension, and $\Delta_i \in \{-L, \dots, 0, \dots, L\}^2$. ps denotes the pixel-shuffle operator which shifts the features spatially by Δ_i :

$$ps(h_t; \Delta_i)(q) = h_t(q + \Delta_i), \quad (4.39)$$

where zero-padding is used for boundary pixels. We select L to be 2 in our experiments, yielding a 5×5 receptive field, to strike a trade-off between model capacity and efficiency. The corresponding relative similarity, dubbed $s(q; \Delta_i)$, is computed by:

$$\begin{aligned} s^*(q; \Delta_i) &= 1 - \text{sim}(ps(h_t; \Delta_i)(q), h_b(q)) \\ s(q; \Delta_i) &= 1 - \frac{s^*(q; \Delta_i)}{\tau(\max_j(s^*(q; \Delta_j)) + \varepsilon)}, \end{aligned} \quad (4.40)$$

where sim denotes the Cosine similarity, the temperature τ is set to 0.25, and $\varepsilon = 10^{-6}$. The feature aggregation $\hat{h}(q)$ is subsequently calculated by combining the direction vector and the associated relative similarity:

$$\begin{aligned} g(q; \Delta_i) &= \text{conv}([d(q; \Delta_i), s(q; \Delta_i)]) \\ \hat{h}(q) &= \text{conv}([g(q; \Delta_1), g(q; \Delta_2), \dots, g(q; \Delta_{(2L+1)^2})]), \end{aligned} \quad (4.41)$$

where conv refers to a convolutional layer and $[]$ denotes the concatenation operation of features along the channel axis.

4.2.4.2 Adaptive Spatial Attention Selection

An intrinsic property of attention-based models is the data-adaptive weights which are predicted from the input sample and allowed to spatially vary [169, 170], having shown great potential to push model performance [171]. We propose an adaptive attention selection method to refine the features extracted from an individual input as well as making the kernel weights sample-adaptive and position-specific.

Instead of sharing convolutional kernels when dealing with different input samples, the spatially varying kernel weights are produced with the feature aggregation coming from the RFA using a tiny U-net, as shown in Fig. 4.2 (b). The size of the output kernel volume, dubbed ω , is $k^2 \times H \times W$ corresponding to k^2 neighboring locations of a specified position q among $H \times W$ candidates. The spatial attention selection leverages an adaptive attention mechanism to promote the representation capacity of basic convolutions of a CNN model, which expresses the refined pixel values in the form of a weighted average of the pixels at neighboring positions with the input-dependent weights of each kernel generated from the aggregated feature information:

$$\tilde{h}(q) = h(q) + \sum_{\delta \in \Delta} \omega(\delta, q) h(q + \delta), \quad (4.42)$$

where $\tilde{h}(q)$ denotes the output features of LoA-SAM with a residual shortcut to facilitate information transmission, δ is the moving footprint, Δ denotes a collection of footprints forming the kernel grid, e.g. $\Delta = \{-1, 0, 1\}^2$ for $k = 3$ as used in our experiments, and $\omega(\delta, q)$ is the generated preference of position $q + \delta$.

Compared to the methods used in [152, 153] which produce the adaptive kernel volume of size $Ck^2 \times H \times W$ and consequently introduce a substantially increased computational overhead, the devised LoA-SAM endows the model with the diversity of attention patterns and the adaptive property with significantly reduced computational costs.

4.2.5 Deep Cascaded Reconstruction Framework

The network architecture used in our experiments is introduced below, which unifies the foregoing model components: the LoA-SAM and DACB modules, to derive a coil-agnostic reconstruction framework for parallel MRI by unfolding a model-based iterative algorithm.

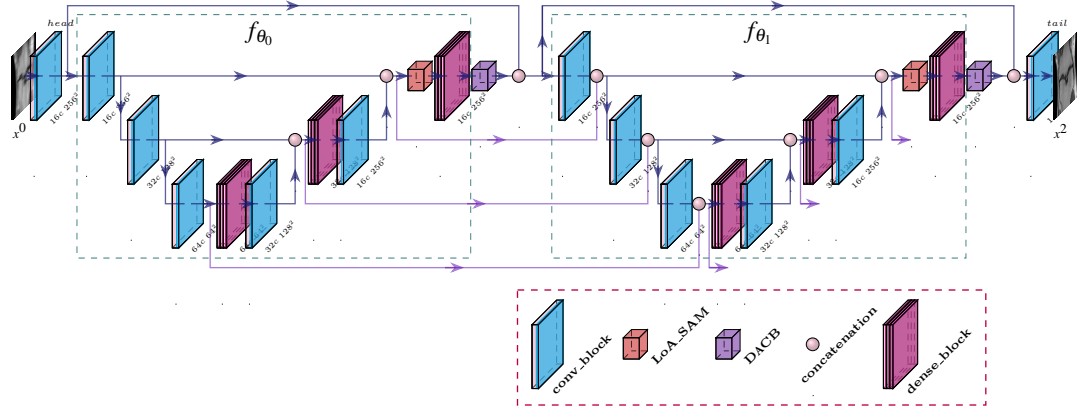


Figure 4.3: Overview of the model architecture. The Root-Sum-of-Squares (RSS) computed with the measurements is first mapped via a head convolutional layer and passed to a network cascade consisting of U-shaped subnetworks with multi-level dense connections and the proposed data aggregation consistency blocks (DACB). For simplicity, only two subnetworks with three transitional levels are displayed. The feature size is indicated by $C N^2$ corresponding to the channel and spatial sizes. For convenience, the input is assumed to have a spatial size of 256×256 .

4.2.5.1 Network Architecture Design

To make full use of the rich representations learned by deep neural networks, a multi-level densely connected architecture is proposed to “reuse” the extracted features at multiple levels via both intra- and inter-shortcuts. To facilitate information propagation between subnetworks and consecutively reuse the features, all previous features from the same level are collected and combined with the current output, which forms the multi-level densely connected structure. Two U-shaped subnetworks with three transitional levels are presented in Fig. 4.3 to demonstrate the network design. Densely connected layers [154], illustrated in Fig. 4.2 (a), are adopted at decoding levels. The LoA-SAM and DACB blocks are embedded at the top decoding levels. Two convolutional layers are used in the DACB as f_1 and f_2 to expand and

reduce feature channels. The features collected from all reconstruction iterations are fused to produce a single-channel outcome as the final reconstruction via a tail convolutional layer. The framework used in experiments is constructed with four subnetworks having four encoding-decoding levels, where the bottom level has $64c$ feature channels with $c = 2$.

4.2.5.2 Unrolled Reconstruction Network Cascade

The reconstruction framework is established based on an iterative algorithm presented in (4.14). The extension of the proposed coil-agnostic data consistency shown in (4.37) can be regarded as a special treatment of DC methods defined in feature space, where the estimated feature volume used to produce an image is corrected with a set of measurements, resorting to a learned coil-invariant mapping. It enables us to transfer the update iteration to feature maps for the purpose of incorporating the dense skip-connections and simplifying the model description.

The illustration of the unrolled reconstruction network cascade is displayed in Fig. 4.4, where the inter-connections between subnetworks are omitted for simplicity. Combining the devised framework components together, the reconstruction iteration is implemented using the following equations with the multi-level skip-connections elided for simplicity:

$$h_{\mathbf{u}}^{k+1} \leftarrow f_{\theta_k}(h_{\mathbf{x}}^k) \quad (4.43a)$$

$$h_{\mathbf{x},\text{cur}}^{k+1} \leftarrow \gamma h_{\mathbf{u}}^{k+1} + (1 - \gamma) f_2^k(\Omega(f_1^k(h_{\mathbf{u}}^{k+1}); \mathbf{DF}, \sum_{i=1}^{n_c} \mathbf{y}_i)) : \text{DACB} \quad (4.43b)$$

$$h_{\mathbf{x}}^{k+1} \leftarrow [h_{\mathbf{x}}^k, h_{\mathbf{x},\text{cur}}^{k+1}] : \text{dense connection}, \quad (4.43c)$$

where $h_{\mathbf{u}}^{k+1}$ refers to the feature prediction coming from a deep network, $h_{\mathbf{x},\text{cur}}^{k+1}$ is the corrected estimate using the DACB, and $h_{\mathbf{x}}^{k+1}$ passed to the successive module collects all previous outputs together via feature concatenation along the channel axis.

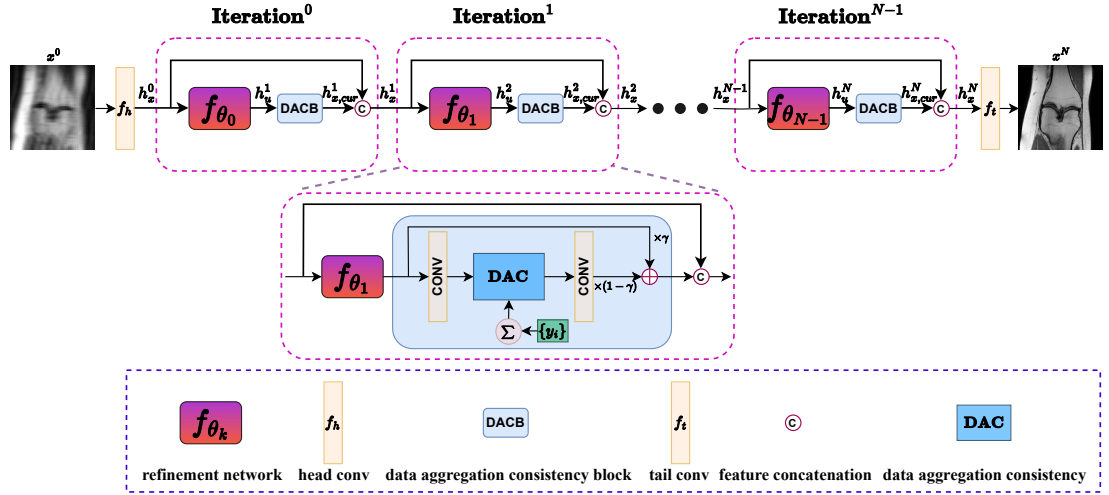


Figure 4.4: Illustration of the unrolled reconstruction network cascade built on a model-based algorithm. Deep neural network modules f_{θ_k} and data aggregation consistency blocks (DACB), illustrated in Fig. 4.1, are interleaved to form a coil-agnostic reconstruction framework, where the inter-connections between subnetworks are omitted, for simplicity. At each iteration, the intermediate prediction coming from the refinement network module f_{θ_k} is passed to the DACB where it is corrected with the multi-coil measurements $\{y_i\}$ by applying the data aggregation consistency (DAC) method, also demonstrated in Fig. 4.1.

The input $\mathbf{x}^0$ to the framework, which is zero-filled, is calculated from the measurements $\{y_i\}$ via the Root-Sum-of-Squares (RSS) algorithm, conforming with the spirit of the reconstruction method without involving the estimation of coil sensitivities:

$$\mathbf{x}^0 = \sqrt{\sum_{i=1}^{n_c} |\mathbf{F}^H \mathbf{y}_i|^2}, \quad (4.44)$$

which gives the initial estimate $h_{\mathbf{x}}^0 = f_h(\mathbf{x}^0)$. The final reconstruction is computed with all collected predictions, giving rise to:

$$\mathbf{x}^N = f_t(h_{\mathbf{x}}^N), \quad (4.45)$$

where N is the number of iterations adopted within the framework and selected to be 4 in our experiments. In accordance with the RSS configuration, $\mathbf{x}^N$ is set to be a single-channel image. For simplicity, two convolutional layers are used for the “head” and “tail” blocks termed f_h and f_t .

4.2.6 Objective Function

The reconstruction framework is trained to minimize the difference between the predicted image and the ground truth. The “fully sampled” k -space data $\{\mathbf{y}_i^{GT}\}$ sensed at multiple coils are combined via the RSS algorithm to produce the reference used to supervise the training procedure in the absence of coil sensitivities:

$$\mathbf{x}^{GT} = \sqrt{\sum_{i=1}^{n_c} |\mathbf{F}^H \mathbf{y}_i^{GT}|^2}. \quad (4.46)$$

The reconstructed image and the ground truth RSS map are represented via a single channel, which enables the calculations of reconstruction errors measured by the L_1 loss and the structural similarity index (SSIM) [58], giving rise to the reconstruction loss shown below,

$$L_{rec} = (1 - \alpha)L_1(\mathbf{x}^N, \mathbf{x}^{GT}) + \alpha L^{SSIM}(\mathbf{x}^N, \mathbf{x}^{GT}), \quad (4.47)$$

where α is practically set to 0.64 to balance the strengths of pixel-wise metric and structure-based evaluation. The least squares GAN (LSGAN) [19] is utilized to encourage sharp and realistic details, leading to the adversarial loss L_{adv} computed by:

$$\begin{aligned} & \min_{\psi} E[(D_{\psi}(\mathbf{x}^{GT}) - 1)^2] + E[(D_{\psi}(\mathbf{x}^N) - 0)^2] \\ & \min_{\theta} E[(D_{\psi}(\mathbf{x}^N) - 1)^2], \end{aligned} \quad (4.48)$$

where θ denotes the collection of the parameters of the devised reconstruction framework and the structure of the discriminator used in CycleGAN [157] is adopted for D_{ψ} parametrized by ψ . The total objective is a weighted combination given by:

$$L = E_{\{\mathbf{x}^N, \mathbf{x}^{GT}\}} [\lambda_{rec} L_{rec} + \lambda_{adv} L_{adv}], \quad (4.49)$$

where we set $\lambda_{rec} = 10$ and $\lambda_{adv} = 0.05$, with a large weight assigned to the reconstruction loss to ensure the success of reconstruction and a small weight assigned to the adversarial loss to prevent training instability.

4.3 Experiments

4.3.1 Multi-Coil MRI Reconstruction Comparisons

4.3.1.1 Implementations

The FastMRI knee dataset [56] is used in experiments to evaluate the proposed reconstruction method, which provides rich structural and textural details with k -space raw data collected by 15 coils. Random masks are adopted in the under-sampling process, where the acceleration rates are respectively set to 8 and 4. The model is trained for 35 epochs using an Adam optimizer [172] with $\beta_1=0.5$, $\beta_2=0.999$, a batch size of 2, and a learning rate of 10^{-5} .

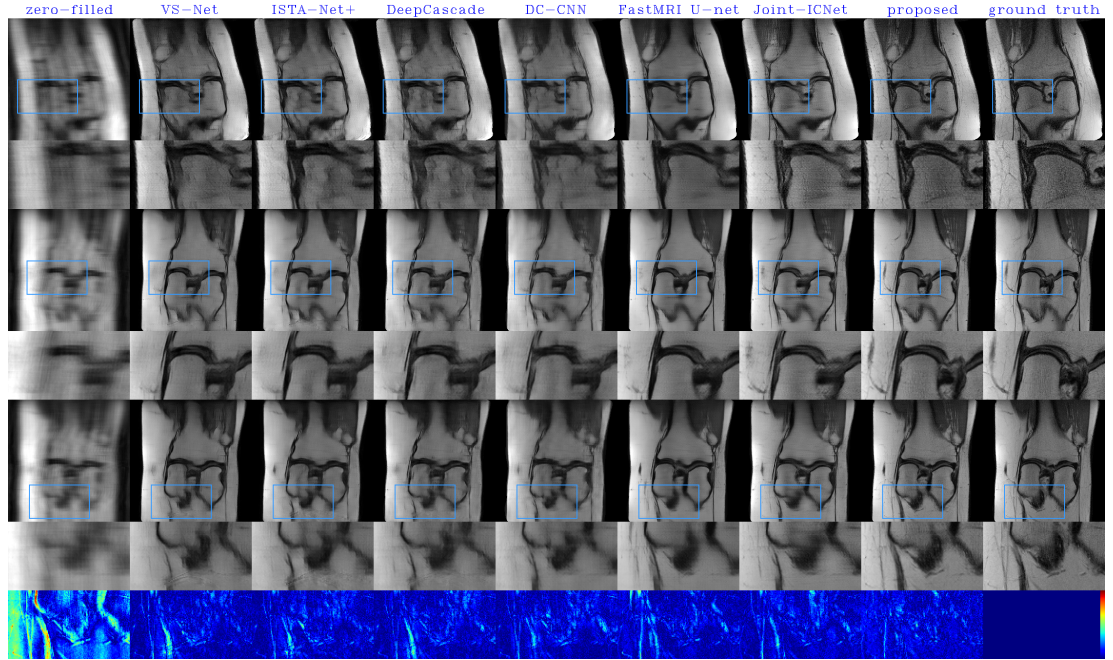


Figure 4.5: Comparison results of $8\times$ accelerated MRI reconstruction.

We compare our method with other prevailing deep learning-based models which are categorized into four groups: 1) an end-to-end reconstruction method, termed FastMRI U-net [56, 100], which directly learns a mapping from the RSS map of the under-sampled data containing heavy aliasing artifacts to the clean image; 2) methods using a model-based sensitivity estimation algorithm including VS-Net [82], ISTA-Net+ [104], and DeepCascade [106, 100], which

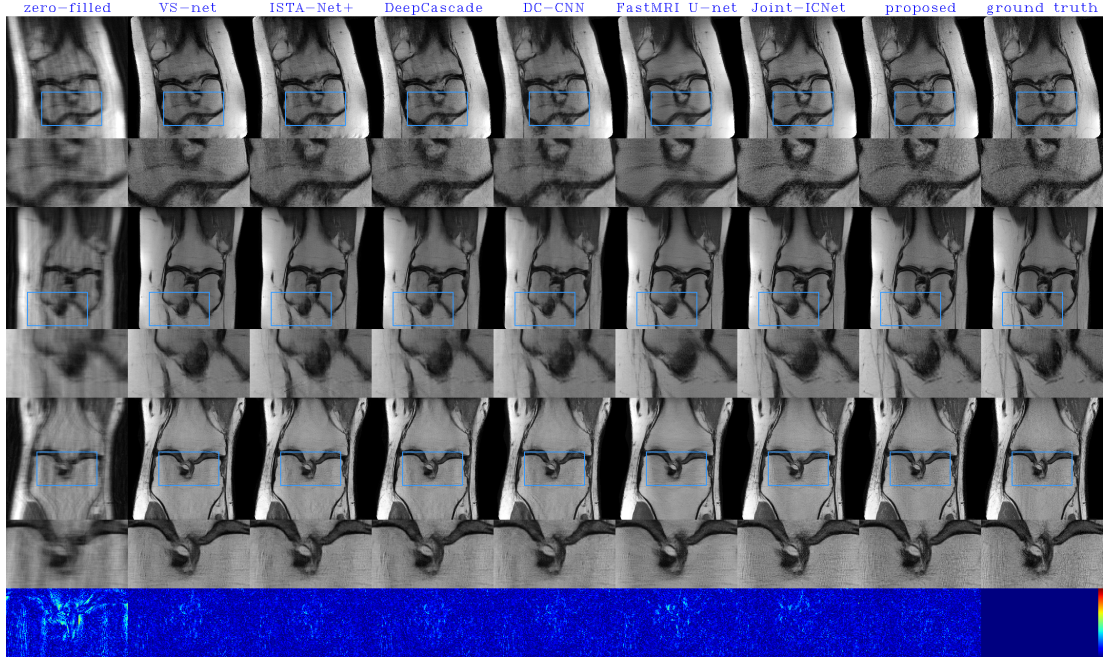


Figure 4.6: Comparison results of $4\times$ accelerated MRI reconstruction.

employ the sensitivity maps pre-computed via ESPIRiT [74] in the forward acquisition process to perform the k -space data constraint; 3) a deep learning-based sensitivity estimation method, dubbed Joint-ICNet [100], which estimates the sensitivity maps using a neural network and jointly updates the reconstructed image and the coil sensitivities in the reconstruction pipeline; and 4) a CNN-based method named DC-CNN [173] which was proposed to simultaneously reconstruct multiple coil frames, hence realizing the target of MR image recovery without using the sensitivity maps. U-net and DeepCascade were also adopted in [100] as comparative approaches. PSNR and SSIM are used as the assessment metrics and the visual quality of reconstructed image is assessed by FID and KID [160].

Table 4.1: Quantitative results of multi-coil knee MRI reconstruction. $4\times$ accelerated reconstructions have the same execution time and parameter size as $8\times$.

	method	PSNR $\uparrow$	SSIM $\uparrow$	FID $\downarrow$	KID $\downarrow$	runtime(s) $\downarrow$	size $\downarrow$
$8\times$	zero-filled	28.75	0.817	379.58	0.459	-	-
	VS-Net [82]	36.18	0.912	141.15	0.078	0.034	1133232
	ISTA-Net+ [104]	36.32	0.911	170.11	0.108	0.189	385613
	DeepCascade [106, 100]	35.82	0.909	146.17	0.081	0.083	453817
	DC-CNN [173]	34.96	0.892	180.47	0.124	0.035	291177
	FastMRI U-net [56, 100]	36.19	0.922	145.58	0.095	0.011	2777993
	Joint-ICNet [100]	35.59	0.921	137.57	0.090	0.283	69103565
	proposed	37.74	0.930	76.08	0.011	0.112	41738102
$4\times$	zero-filled	33.16	0.883	201.20	0.173	-	-
	VS-Net [82]	40.56	0.956	111.34	0.052	-	-
	ISTA-Net+ [104]	40.81	0.956	112.45	0.052	-	-
	DeepCascade [106, 100]	40.57	0.956	109.46	0.050	-	-
	DC-CNN [173]	39.32	0.945	111.32	0.049	-	-
	FastMRI U-net [56, 100]	39.37	0.951	108.52	0.055	-	-
	Joint-ICNet [100]	39.27	0.955	88.69	0.035	-	-
	proposed	41.10	0.963	64.37	0.009	-	-

4.3.1.2 Comparison Results

The reconstruction results generated by different methods are presented in Fig. 4.5 and 4.6, where it shows that the devised reconstruction framework produces more faithful reconstructions with rich and fine-grained structures and textural details, leading to more visually appealing results and demonstrating the superior performance to other comparative methods. The $2\times$ amplified residual maps of the images recovered by our method present fewer errors, validating its ability to provide more accurate reconstructions.

We display the quantitative comparison results in Table 4.1, which demonstrates the consistently improved reconstruction performance of the proposed model in terms of all evaluation metrics ($p\text{-value}\ll 0.05$) at both sampling rates, compared to other approaches. From Table 4.1, we can observe that the proposed method has a comparable execution speed due to the encoding-decoding design and contains fewer parameters than Joint-ICNet, suggesting the efficient design for multi-coil MRI recovery and therefore potentially allowing us to apply the proposed coil-agnostic reconstruction method in real-time scenarios.

In the following, we report the ablation studies conducted to confirm that the superior performance is attributable to the devised model components, and not simply owing to the model size, since the problems associated with the unknown forward process are irrelevant to the parameter size, which we propose to circumvent via a novel data consistency method enforcing the k -space constraint in a coil-agnostic manner.

4.3.2 Ablation Studies on Model Components

4.3.2.1 Framework Configurations

We conduct ablation experiments to validate the efficacy of the proposed model components: multi-level dense connections (ML-dense), LoA-SAM, and DACB, using the knee dataset at $8\times$ acceleration. The configurations of model candidates are provided in Table 4.2, where (E) refers to the candidate with all devised components incorporated in the reconstruction framework. For candidate (A) and (F) with the multi-level dense connections removed, the feature volumes at the convolutional layers, where the skip-connections are expected to occur, are repeated and concatenated along the channel dimension to approximate the parameter size of their counterparts employing ML-dense. Since two convolutional layers are used in the DACB for the control of the channel width, which is also demanded for other candidates, they are preserved as part of the preceding and successive subnetworks to keep the model size for a fair comparison. The U-net adopted in LoA-SAM has a limited number of parameters and is only used for kernel prediction, which enables fair comparisons when removing the devised LoA-SAM with the reconstruction branch containing a vast majority of parameters unchanged.

4.3.2.2 Ablation Results

From the quantitative results shown in Table 4.2, we found that candidate (B) and (E) yield higher performance in terms of the relevant metrics as compared to candidate (A) and (F), showing the improvements made by adopting the multi-level dense connections. Respectively adding LoA-SAM and DACB to candidate (B), leading to model (C) and (D), delivers consid-

Table 4.2: Ablation on model components using $8\times$ accelerated multi-coil knee MRI reconstruction.

method	ML-dense	LoA-SAM	DACB	PSNR $\uparrow$	SSIM $\uparrow$	FID $\downarrow$	KID $\downarrow$
(A)	$\times$	$\times$	$\times$	36.53	0.921	89.83	0.025
(B)	$\checkmark$	$\times$	$\times$	37.06	0.926	85.39	0.024
(C)	$\checkmark$	$\checkmark$	$\times$	37.33	0.926	84.16	0.022
(D)	$\checkmark$	$\times$	$\checkmark$	37.53	0.929	80.14	0.017
(E) ours	$\checkmark$	$\checkmark$	$\checkmark$	37.74	0.930	76.08	0.011
(F)	$\times$	$\checkmark$	$\checkmark$	37.11	0.927	82.13	0.018

erable performance gains, and the final model (E) produces the best results by combining LoA-SAM and DACB together, implying a positive synergy between them in consistently enhancing the reconstruction performance in terms of evaluation metrics. Those observations confirm the effectiveness of the proposed method in pushing model performance.

4.3.3 Ablation Studies on Data Consistency Methods

4.3.3.1 Implementations

The proposed optimization-inspired reconstruction framework adopts the RSS map as input and produces a single-channel image as an attempt to evade the estimation of sensitivity maps required in the forward process of both data acquisition and reconstruction steps, while encouraging a consistent result with the sampled k -space measurements. To elucidate the benefits of the devised DACB and the RSS-based reconstruction pipeline, we compare with three variants using a conventional data consistency method. As shown in Table 4.3, the variant termed (E)-CSM performs data consistency with the pre-calculated sensitivity maps via (4.31), dubbed CSDC, and adopts complex-valued signals as input and output represented by two channels. The complex-valued input is the weighted combination of the observed measurements, which is computed by:

$$\mathbf{x}^0 = \sum_{i=1}^{n_c} \mathbf{S}_i^H \mathbf{F}^H \mathbf{y}_i. \quad (4.50)$$

The extended variant (E)-CSM-a and (E)-CSM-b progressively utilize a single-channel data configuration for the model output and input, as suggested by the proposed method. The evaluation results and reconstructed images of different variants are displayed in Table 4.3 and Fig. 4.7.

Table 4.3: Ablations on data consistency methods and data type at $8\times$ acceleration. The proposed method involves no sensitivity estimation and is therefore free of ME-Arti.

method	input	DC	output	PSNR $\uparrow$	SSIM $\uparrow$	FID $\downarrow$	KID $\downarrow$	ME-Arti
(E)-CSM	complex	CSDC	complex	36.58	0.916	94.56	0.029	strong
(E)-CSM-a	complex	CSDC	single	36.83	0.920	87.46	0.022	weak
(E)-CSM-b	RSS	CSDC	single	36.99	0.925	83.95	0.020	trivial
(E) proposed	RSS	DACB	single	37.74	0.930	76.08	0.011	free

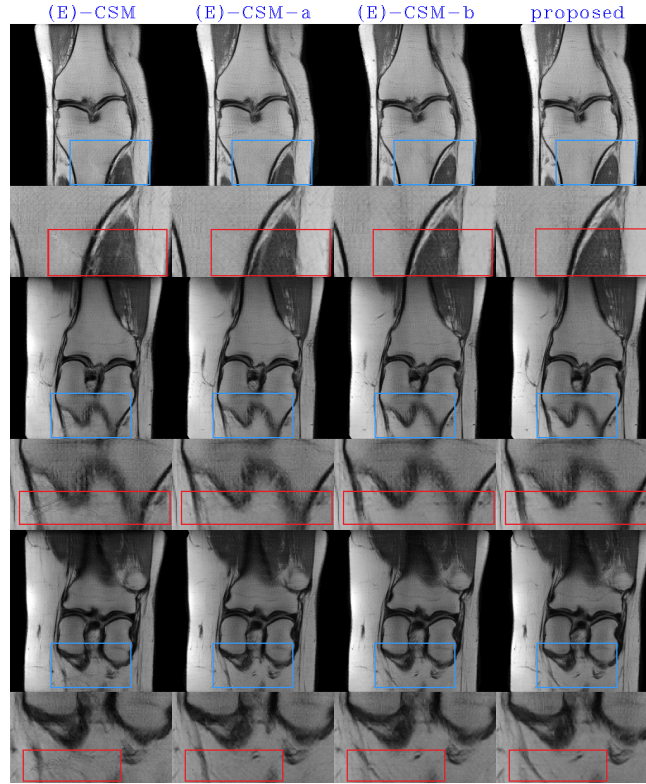


Figure 4.7: Ablation results on data consistency methods and data type at $8\times$ acceleration.

4.3.3.2 Ablation Results

From the results shown in Table 4.3, we found that the data configuration of model input and output considerably affects the reconstruction performance with the increased evaluations while generating a single-channel output and utilizing the RSS map as input. The data aggregation consistency method improves the model performance and yields superior results as compared to variant (E)-CSM-b which applies a conventional DC operation with the estimated sensitivity maps, confirming the efficacy of the DACB.

The reconstruction results of (E)-CSM present strong unwanted artifacts, as displayed in Fig. 4.7, which show to be alleviated in the images produced by variant (E)-CSM-a and -b. We surmise that it is due to the inaccurate estimation of coil sensitivities, which causes irregularities in the complex-valued model input computed with the estimated sensitivity maps. Such errors are eventually propagated to the final outcome of the reconstruction pipeline when it fails to overcome them. The devised framework performs reconstruction without using the sensitivity maps, and thus prevents the errors coming from inaccuracy sensitivity estimation and yields clean images. We demonstrate the sources of such artifacts in the next subsection.

4.3.3.3 Sensitivity Map Estimation Artifacts (ME-Arti)

In Fig. 4.7, we can observe the artifacts in the reconstructed images produced by the ablation variant (E)-CSM, whose source, as demonstrated here, is the errors in the estimation of coil sensitivities. We present an MR image in Fig. 4.8, which is the RSS map of fully sampled signals and shows no artifact. Sensitive maps are pre-computed by a model-based algorithm and used for (E)-CSM in the interleaved DC blocks. A separated display of a single channel, e.g. the one representing the imaginary values, of a sensitive map is provided in Fig. 4.8, where the values of the same sign, e.g. the positive component, form an image clearly showing an irregular shape, which coincides with the artifact displayed in the reconstruction result of (E)-CSM, also see the first sample in Fig. 4.7.

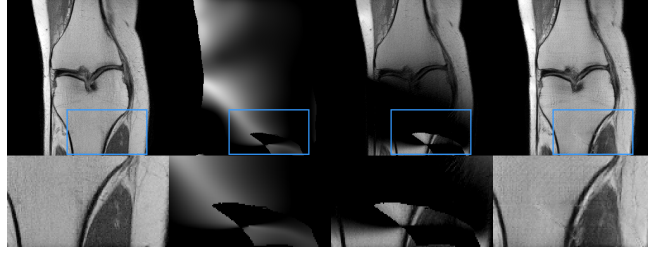


Figure 4.8: Artifacts caused by inaccurate sensitivity map estimation. First) RSS of fully sampled signal, second) positive component of the imaginary part of a sensitivity map, third) positive component of the imaginary part of the weighted average, and last) reconstruction produced by (E)-CSM.

Similarly, the sensitivity-weighted average manifests the irregular pattern when solely visualizing the single component. Those findings demonstrate that the errors of sensitivity estimation can attend to the image channels stealthily in the reconstruction pipeline and cause artifacts in the final outcome when the model fails to overcome them, whereas the devised coil-agnostic framework using the DACB for data consistency enforcement and adopting an RSS-based data configuration circumvents such problems and produces clean images, which shows the adverse impact of inaccurate sensitivity estimation and suggests the usefulness of the proposed method.

4.3.3.4 Ablation on Data Consistency Bottleneck Design

The data consistency method proposed in (4.37) extends the equation in (4.35), which applies the data aggregation consistency purely in image domain, by adopting a residual shortcut to facilitate feature propagation. To demonstrate the impact of the bottleneck design in conventional DC operations and verify the usefulness of the feature-level generalization of the devised coil-agnostic data consistency method, we compare it with the variant, dubbed (E)-ResFree, which implements the DAC method with the skip-connection straddling the DACB, as illustrated in Fig. 4.1 (c), applied to the predicted images, leading to the form expressed in (4.35) and illustrated in Fig. 4.1 (b). We rewrite (4.35) as follows,

$$h' = f_2(\gamma f_1(h) + (1 - \gamma)\Omega(f_1(h); \mathbf{DF}, \sum_{i=1}^{n_c} \mathbf{y}_i)), \quad (4.51)$$

which we denote by ResFree.

Table 4.4: Ablation analysis on residual feature connection in DACB under $8\times$ acceleration using multi-coil knee MRI reconstruction.

method	DC	PSNR $\uparrow$	SSIM $\uparrow$	FID $\downarrow$	KID $\downarrow$
(E) proposed	DACB	37.74	0.930	76.08	0.011
(E)-ResFree	ResFree	37.40	0.928	79.37	0.014

The results in Table 4.4 show reduced PSNR and SSIM scores and concomitantly increased FID and KID values when adopting (4.51) in the reconstruction pipeline, which indicates the negative impact of the bottleneck structure in DC blocks. It confirms that the DACB is capable to alleviate this problem via a natural, simple, and parameter-free modification, as derived in Sect. 4.2.3.4, and improve the reconstruction performance.

4.3.4 Ablation Analysis on Relative Feature Aggregation

To validate the usefulness of the devised relative feature aggregation block used in LoA-SAM to aggregate the relative information among predicted features, we train a variant termed (E)-no-RFA with the RFA removed and instead generating the data-specific kernel weights directly with the feature volume entering the LoA-SAM block. From the results presented in Table 4.5, we can observe decreased PSNR and SSIM values by a small margin and moderately higher FID and KID scores, indicating that the reconstruction performance is slightly affected by the removal of RFA. It shows that the relative information from the RFA performs more favorably than the original input, which improves the capacity of LoA-SAM. Noting that the feature information from the RFA, which is extracted only by utilizing simple mathematical operations, is preferable over the input features, it can be expected to inspire the exploration of more complex aggregation methods for MRI reconstruction.

Table 4.5: Ablation study on relative feature aggregation using $8\times$ accelerated multi-coil knee MRI reconstruction.

method	RFA	PSNR $\uparrow$	SSIM $\uparrow$	FID $\downarrow$	KID $\downarrow$
(E) proposed	$\checkmark$	37.74	0.930	76.08	0.011
(E)-no-RFA	$\times$	37.69	0.929	78.43	0.013

4.3.5 Robustness to Coil Configurations

Instead of estimating the sensitivity maps from the k -space center via model-based algorithms, neural networks, which have shown great flexibility and versatility in vision tasks, can be used to predict the coil sensitivities as part of the reconstruction model, e.g. Joint-ICNet [100]. An alternative approach to solving the forward process problem is to deal with multiple coil frames in parallel. A typical method is DC-CNN [173] which learns to reconstruct multiple views from a collection of measurements and respectively performs data consistency for parallel coil predictions. Such methods can struggle to work properly with different machine configurations among scanning cases, which deters models from being transferred to real-world scenarios. To substantiate this argument and evaluate the model robustness, we introduce the coil configuration discrepancies in inference by randomly permuting the coil frames, i.e. rolling the coil sequence by a randomly sampled shift number. Despite limited diversities of coil orders, the great majority of the FastMRI data follow the same configuration. We compare with Joint-ICNet and DC-CNN which respectively take multi-coil under-sampled images as input to the estimation network and to the reconstruction pipeline.

Table 4.6: Evaluation results of $8\times$ accelerated multi-coil knee MRI reconstruction using random coil permutation.

method	permutation	PSNR $\uparrow$	SSIM $\uparrow$	FID $\downarrow$	KID $\downarrow$
proposed	None/rolled	37.74	0.930	76.08	0.011
Joint-ICNet [100]	None	35.59	0.921	137.57	0.090
Joint-ICNet [100]	rolled	33.18	0.905	143.43	0.092
DC-CNN [173]	None	34.96	0.892	180.47	0.124
DC-CNN [173]	rolled	31.86	0.848	237.91	0.216

The reconstruction results with coil permutation are presented in Fig. 4.9 and Table 4.6. The images produced by Joint-ICNet and DC-CNN are significantly affected by coil rolling, whereas the proposed method is invulnerable to such an attack. In Table 4.6, it is evident that the evaluation performance of both Joint-ICNet and DC-CNN abruptly drops with coil permutation, indicating their sensitivity to machine setting discrepancies. Meanwhile, the number of coils needs to be kept fixed to meet the input requirement, which is another problem violating the deployment of reconstruction pipelines from training to real-world practice. Be-

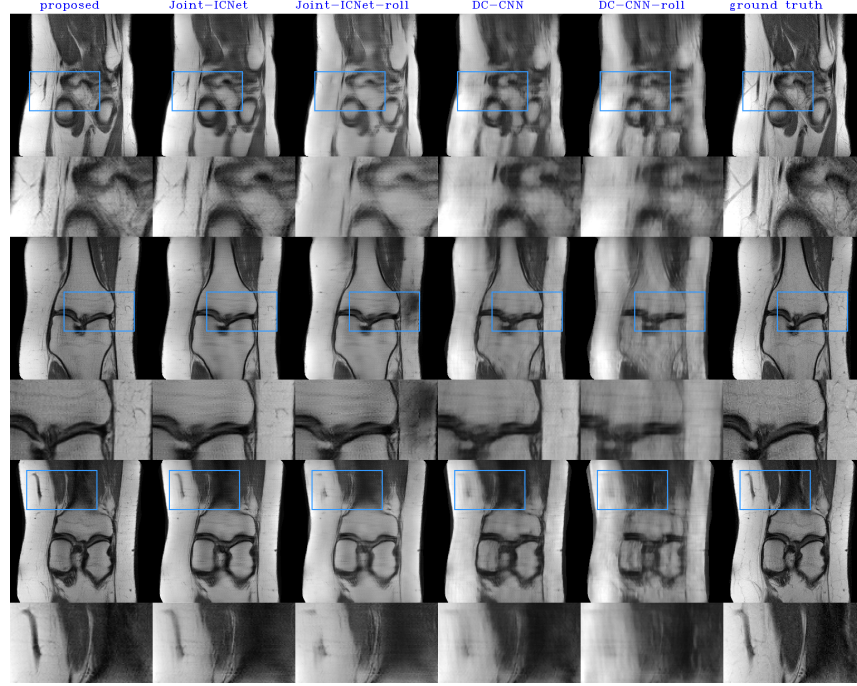


Figure 4.9: Comparisons of $8\times$ accelerated MRI reconstruction with coil deployments randomly permuted.

cause it capitalizes on the linear property of the RSS operator and measurement summation and end-to-end learns a coil-invariant mapping to exhibit data consistency encouragement without involving coil sensitivities, the proposed coil-agnostic reconstruction framework overcomes those issues and works appropriately among different coil deployments, showing a strong invulnerability and robustness capacity.

4.4 Conclusion

In this chapter, a novel reconstruction framework was introduced to handle the unknown forward process problem in parallel MRI reconstruction, which circumvents the coil sensitivity estimation retrospectively demanded by conventional reconstruction methods. A coil-agnostic data consistency block was proposed to model the k -space data constraint and an attention module was designed to capture the spatial local contexts to improve the model performance without tremendous amounts of computational costs associated. The experimental results demonstrated that the proposed framework can provide high-quality reconstructions with a

very low sampling rate and surpass other deep learning-based approaches both qualitatively and quantitatively. In ablations, the proposed coil-agnostic framework showed superior performance over competitive model candidates in accelerated MRI acquisition diets. The future works include applying this method to other anatomical structures and extending it to dynamic MR imaging.

Mask-Guided Motion Feature Incorporated Network for Dynamic MRI Reconstruction

5.1 Introduction

Magnetic resonance imaging (MRI) is known to be a slow medical examine modality due to the long data acquisition time, though it provides a non-invasive and radiation-free imaging tool suitable for a broad range of tissues such as bones and muscles. In dynamic MRI, the demand for high resolution in both spatial and time domains exacerbates the situation and motivates research studies to reconstruct dynamic MR sequences in accelerated scanning diets. To realize a reduction in scan time, the data acquisition process can be accelerated by only collecting a subset of measurements and recovering images needed to diagnose diseases and plan treatments using reconstruction algorithms.

Among MRI reconstruction methods, compressed sensing (CS) methods obtained popularity by showing their ability to recover the signals taken in a sub-Nyquist sampling diet by converting the ill-posed inverse problems into a regularized optimization [36], depending on the sparsity prior of target images assumed to occur in image domain or in some transformed space and achieving a solution via model-based optimization algorithms [138, 139, 45]. Although effective in shortening the acquisition time, the handcrafted prior knowledge can struggle to model the complexity of anatomic structures with rich details which are potentially important

for clinical diagnosis, and the iterative solvers used for optimization problems are often time-consuming, which is against the speed requirement of fast MRI in modern medical imaging. Deep neural networks have gained great success in vision tasks by utilizing the powerful learning and representation capacity of a deep architecture with a large amount of trainable parameters encoding prior information learned from datasets. Many recent works demonstrated promising reconstruction performance by resorting to the deep representations of such models to improve the quality of recovered images [57, 64, 82, 83, 174]. Among those approaches, few works were devoted to dynamic MRI reconstruction which imposes a more challenging topic owing to the need of the recovery of both spatial and temporal high resolutions.

It is a prominent aspect of dynamic MRI reconstruction to exploit the redundancies and informative correlations among a sequence of frames collected at consecutive time steps to recover the underlying dynamic content while resolving the aliasing artifacts appearing in image domain. Recently, several works utilized deep neural networks to recover dynamic MRI sequences from under-sampled signals as a way to accelerate the data acquisition procedure [173]. To jointly leverage the image feature dependencies in spatial and temporal domains, recurrent neural networks (RNN) were used to model time-varying dynamics via the recurrently updated hidden connections allowing for information transmission across consecutive temporal frames [120], learning the complementary and redundant information via the recurrent nature of RNN structures [122, 121]. However, it remains a problem that spatial regions are not equally informative and conducive to improved reconstruction quality. The accumulated history information encoded in the hidden states can be impacted by misalignment and blurring [125, 175]. The recurrent manner of processing image sequences can be difficult to enable parallel computations and possibly results in large memory caching [125, 124].

To make use of the relevant information in adjacent frames, optical flow-based alignment methods warp the reference frames towards the source target to perform sub-pixel motion compensation [175, 176, 136]. The work introduced in [131] combines the adjacent frames with the current target to extend model-based MRI reconstruction methods to dynamic scenarios as a group-wise optimization problem. A neural network was proposed in [123] to estimate the

motion displacement and refine the initial reconstructed MRI sequences by a motion compensation component. However, those methods for dynamic sequence reconstruction can suffer from inevitable errors in flow-based image alignment and motion compensation, which are prone to generate undesirable artifacts [175, 124, 137] due to the difficulty in precisely performing motion offset estimation. The heavy computational costs and large memory consumption are often needed for warping operations in alignment.

In this chapter, we introduce an optimization-inspired deep cascaded framework for dynamic MRI reconstruction, which is derived by unfolding a variable splitting optimization algorithm and has a novel mask-guided motion feature incorporation scheme, dubbed Mask-MFI, to explicitly incorporate the dynamic cues potentially contained in the measurements sampled at consecutive time steps into the reconstruction procedure, showing enhanced ability to capture the dynamic content which benefits from the combination of the motion features extracted by a feature pyramid network and the predictions from reconstruction networks. Attention mechanism proved the usefulness of modelling the correlations among the involved elements in retrieving favorable information to increase the representation capacity, but which introduces huge computational complexity. We extend the efficient design of the Fourier neural operation to dynamic scenes and devise a spatio-temporal Fourier neural block (ST-FNB) to achieve global information exchange among the regions in both spatial and temporal domains without incurring prohibitive computational costs. By experiments, the transferability and generalization capacity of the proposed method are demonstrated when applying it to new unseen data and a variety of models, in addition to showing the superior performance over other competitive deep learning-based methods and verifying the contributions of the devised model components in ablations.

5.2 Methodology

5.2.1 Problem Formulation and Unrolled Reconstruction Pipeline

The data acquisition process of dynamic MRI is accelerated by under-sampling in k -space, obtaining the measurement $\mathbf{y}_i$ corresponding to the i -th coil:

$$\mathbf{y}_i = \mathbf{D}\mathbf{F}\mathbf{S}_i\mathbf{x}, \quad (5.1)$$

where $\mathbf{x}$ denotes the image of the inspected anatomy, $\mathbf{S}_i$ denotes the coil sensitivity map, $\mathbf{F}$ is the Fourier transform matrix, $\mathbf{D}$ refers to the encoding matrix. To provide reconstruction of the target signals from a subset of k -space data, a regularized optimization is formulated to model the under-determined inverse problem as follows,

$$\min_{\mathbf{x}} \sum_{i=1}^{n_c} (\lambda \|\mathbf{D}\mathbf{F}\mathbf{S}_i\mathbf{x} - \mathbf{y}_i\|^2) + R(\mathbf{x}), \quad (5.2)$$

where R is a regularization term, n_c is the number of coils, and λ is a parameter controlling the strength of regularization. The auxiliary variables termed $\mathbf{u}$ and $\mathbf{m}_i$ are employed to solve the minimization problem defined in (5.2) by converting it into the constrained form:

$$\begin{aligned} \min_{\mathbf{x}, \mathbf{u}, \mathbf{m}_i} \sum_{i=1}^{n_c} (\lambda \|\mathbf{D}\mathbf{F}\mathbf{m}_i - \mathbf{y}_i\|^2) + R(\mathbf{u}) \\ s.t. \mathbf{x} = \mathbf{u}, \mathbf{S}_i\mathbf{x} = \mathbf{m}_i. \end{aligned} \quad (5.3)$$

Considering the minimization objective as a function of the involved variables and two Lagrange multipliers α and γ imported to combine the variables together in a single equation, the unconstrained form of the optimization expressed in (5.3) is obtained as follows,

$$\min_{\mathbf{x}, \mathbf{u}, \mathbf{m}_i} \sum_{i=1}^{n_c} (\lambda \|\mathbf{D}\mathbf{F}\mathbf{m}_i - \mathbf{y}_i\|^2) + R(\mathbf{u}) + \alpha \|\mathbf{u} - \mathbf{x}\|^2 + \gamma \sum_{i=1}^{n_c} (\|\mathbf{m}_i - \mathbf{S}_i\mathbf{x}\|^2). \quad (5.4)$$

We utilize a variable splitting method to find a solution to the optimization problem defined in (5.4), by dividing it into the following sub-problems that can be solved using a closed-form expression and alternating minimization over the multiple variables, giving rise to the following expressions:

$$\mathbf{u}^{k+1} \leftarrow \arg \min_{\mathbf{u}} R(\mathbf{u}) + \alpha \|\mathbf{u} - \mathbf{x}^k\|^2 \quad (5.5a)$$

$$\mathbf{m}_i^{k+1} \leftarrow \arg \min_{\mathbf{m}_i} \sum_{i=1}^{n_c} (\lambda \|\mathbf{D}\mathbf{F}\mathbf{m}_i - \mathbf{y}_i\|^2 + \gamma \|\mathbf{m}_i - \mathbf{S}_i \mathbf{x}^k\|^2) \quad (5.5b)$$

$$\mathbf{x}^{k+1} \leftarrow \arg \min_{\mathbf{x}} \sum_{i=1}^{n_c} (\gamma \|\mathbf{m}_i^{k+1} - \mathbf{S}_i \mathbf{x}\|^2) + \alpha \|\mathbf{u}^{k+1} - \mathbf{x}\|^2. \quad (5.5c)$$

The optimization sub-problems linked with variable $\mathbf{m}_i$ and $\mathbf{x}$ have a closed-form solution which can be used to construct a model-based algorithm iteratively solving the reconstruction problem. Leveraging a deep neural network, dubbed f_{θ_k} , to learn the prior knowledge encoded in the regularization term from a dataset of samples, the associated alternating algorithm is expressed as below,

$$\mathbf{u}^{k+1} \leftarrow f_{\theta_k}(\mathbf{x}^k) \quad (5.6a)$$

$$\left. \begin{aligned} \mathbf{m}_i^{k+1} &\leftarrow \overbrace{\mathbf{S}_i \mathbf{x}^k - \eta \mathbf{F}^H (\mathbf{D}\mathbf{F}\mathbf{S}_i \mathbf{x}^k - \mathbf{y}_i)}^{\text{coil projection}} \\ \mathbf{m}^{k+1} &:= \underbrace{\sum_{i=1}^{n_c} \mathbf{S}_i^H \mathbf{m}_i^{k+1}}_{\text{coil combination}} \end{aligned} \right\} \text{DC layer} \quad (5.6b)$$

$$\mathbf{x}^{k+1} \leftarrow \alpha_0 \mathbf{u}^{k+1} + (1 - \alpha_0) \mathbf{m}^{k+1} : \text{coupling layer}, \quad (5.6c)$$

which is executed iteratively to reach a solution, with k numbering the iteration and the parameters satisfying $\eta = \frac{\lambda}{\lambda + \gamma}$ and $\alpha_0 = \frac{\alpha}{\alpha + \gamma}$.

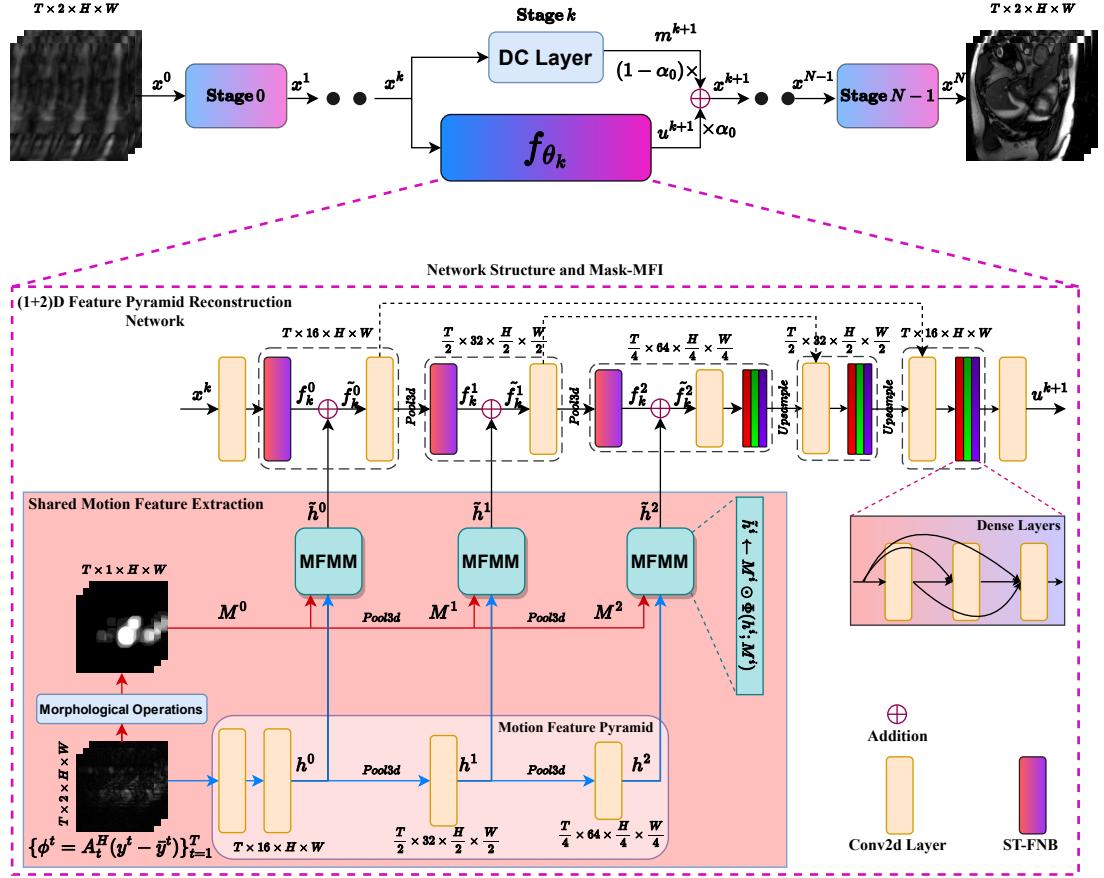


Figure 5.1: Illustration of dynamic MRI reconstruction model. The structure of subnetworks and the mask-guided motion feature incorporation (Mask-MFI) are displayed in the bounding box. The motion features extracted from the under-sampled measurement sequences are incorporated into the reconstruction framework under the guidance of the motion masks generated with the measurement observations to explicitly leverage the dynamic information contained in MRI sequences and improve the reconstruction performance in terms of the recovery of the time-varying content.

The iterative procedure given by (5.6) is unrolled to derive a deep cascaded framework for the reconstruction of dynamic MRI sequences, which takes as input the under-sampled frames collected by parallel coils at different time steps and produces high-quality images by exploiting the learning capacity of network models to provide deep representations of the underlying prior knowledge. The illustration of the devised model is presented in Fig. 5.1, demonstrating the update steps at each iteration including the reconstruction step defined in (5.6a) entailing the k -th network model, the data consistency (DC) layer in (5.6b) aiming to correct the intermediate prediction, and the coupling layer displayed in (5.6c). For stability in training, the

parameters are empirically selected to be $\eta = 0.2$ and $\alpha_0 = 0.2$. In Fig. 5.1, the structure of each subnetwork f_{θ_k} is portrayed in detail, featuring three novel model components: (1+2)D feature pyramid reconstruction network, spatio-temporal Fourier neural block (ST-FNB), and mask-guided motion feature incorporation (Mask-MFI) which we introduce in the following.

5.2.2 (1+2)D Feature Pyramid Reconstruction Network

We devise a deep network architecture to restore MR images with a collection of measurements in dynamic scenes. Inspired by the findings that the low-level high-resolution feature maps coming from a network model have the ability to delineate local structures and textural details of input images and the high-level low-resolution features model more abstract information such as global semantics [177, 24, 178], a (1+2)D feature pyramid reconstruction network, dubbed (1+2)D-FPRN, is adopted as the basic structure of each subnetwork to construct the deep cascaded framework.

As displayed in Fig. 5.1, the (1+2)D-FPRN extracts features from the zero-filled, which is computed by (5.7) as the sensitivity-weighted average of the inverse Fourier transforms of the measurements with unsampled positions filled with zeros, using 3D pooling operation termed *Pool3d* to reduce the feature size in both spatial and temporal domains,

$$\mathbf{x}^0 = \sum_{i=1}^{n_c} \mathbf{S}_i^H \mathbf{F}^H \mathbf{y}_i. \quad (5.7)$$

The feature volumes at different encoding levels are denoted by f_k^i where $i \in \{0, 1, 2\}$ indicates the representations ranging from image resolution to the most abstract bottom level. Owing to the combination of information encoding and 3D pooling operation, the devised network model has the capacity to merge information across a wider range in time space as compared to standard 2D CNNs which only expand the receptive field in spatial domain. Different from 3D convolutional neural networks [173, 107] which entail significantly more parameters than 2D

convolutional layers and abound with shortcomings such as computational costs and excessive memory usage [109, 110], our (1+2)D-FPRN is endowed with the ability to make use of the contextual cues from both spatial and time spaces, while keeping the spirit of efficient model design.

In the decoder of each subnetwork, as displayed in Fig. 5.1, the features coming from the encoder are refined by densely connected layers and up-sampled until the original image resolution by applying interpolation to the spatial and temporal dimensions. Following the ideology of U-shaped networks, skip-connections are utilized to transmit information from the encoder to the decoder. Despite the enlarged receptive range in time space, the (1+2)D-FPRN still has two drawbacks: 1) the dynamic information is not explicitly leveraged, and 2) the exploitation of interdependencies among the extracted features is confined within the local regions, which lacks the modelling of global relations. To address those issues and promote the capacity of the reconstruction framework, a spatial-temporal attention module and a motion feature incorporation mechanism are proposed and introduced below in detail.

5.2.3 Spatio-Temporal Fourier Neural Block (ST-FNB)

As an important constituent aspect in neural network models, attention mechanism was put forward to handle the correlations among involved “tokens” for the purpose of exchanging useful information and enhancing the model performance. Despite achieved advances in vision tasks, attention methods introduce prohibitive computational complexity scaling quadratically with the increased number of tokens when processing high-resolution images. The problem deteriorates with sequences of images passed to the model as input, which makes the application of attention mechanism challenging if not unaffordable. In recent works, local window strategies were proposed to reduce the consumption of computational resources needed to apply attention mechanism, by capturing the interdependencies within non-overlapped local regions partitioned based on a pre-defined window size and realizing cross-window information exchange via circular shift operation [24, 25, 27]. Nevertheless, such methods have several shortcomings: 1) the receptive field is still limited and can cause degraded model performance,

2) a stack of attention layers is needed to enlarge the local area where information communication is enabled, which can be detrimental to model efficiency, and 3) the computational cost remains expensive for some problems involving the process of many high-resolution features, e.g. dynamic sequence reconstruction.

In recent works, the fast Fourier transform (FFT) was utilized as a means to retrieve informative hints from the global region, which converts the heavy quadratic complexity into *log*-linear computational complexity [179, 180] and has proven to work favorably in image tasks, e.g. segmentation, classification, and in-painting [181, 182, 183]. Although effective in practice, the major focus of those methods is on single image tasks, which lacks the ability to model the correlations of a sequence of frames in complementary domains. A novel spatio-temporal Fourier neural block (ST-FNB) is proposed here to simultaneously achieve the retrieval of useful information contained in dynamic MRI sequences in both spatial and temporal domains, by exploiting the Fourier transform to “mix tokens” coming from the global region in a fast and efficient manner. The proposed ST-FNB is embedded in the encoder of each subnetwork, as illustrated in Fig. 5.1, to refine the extracted features which are then decoded to generate the intermediate prediction. The methodology of the Fourier neural attention is introduced below and the description of the details of the ST-FNB ensues.

5.2.3.1 Fourier Neural Attention

Fourier neural attention is proposed to “mix” the tokens involved in the control of the relevant features, which is derived from the extension of attention mechanism to continuous space. Assuming the s -th token of an input tensor X is denoted by $X(s) \in R^{1 \times d}$, e.g. a vector formed by the pixel values at the same spatial position of a feature volume with the batch index omitted for convenience, the kernel summation K is defined by:

$$K(X)(s) = \sum_{\tau=1}^N X(\tau) \kappa(s, \tau), \quad (5.8)$$

where $\kappa(s, \tau) \in \mathbb{R}^{d \times d}$ refers to the attention kernel representing token-specific embedding matrix and N is the number of tokens, admitting $X \in \mathbb{R}^{N \times 1 \times d}$. The continuous expression of attention mechanism can be formulated as the integration computed with the tokens and kernels, given by:

$$K(X)(s) = \int X(\tau) \kappa(s, \tau) d\tau. \quad (5.9)$$

For the special case $\kappa(s, \tau) = \kappa(s - \tau)$, it can be efficiently calculated by resorting to the implementation of the FFT and its inverse:

$$\begin{aligned} K(X)(s) &= \int X(\tau) \kappa(s - \tau) d\tau \\ &= (X * \kappa)(s) \\ &= FFT^{-1}(FFT(X) \otimes FFT(\kappa))(s), \end{aligned} \quad (5.10)$$

where $*$ represents the convolution operation, $\otimes$ refers to matrix multiplication applied to the last two dimensions of the specified tensors whereas the Fourier transform is performed along the first token axis. To enable more flexible and practical implementations of this methodology, the real fast Fourier transform, termed $rFFT$, is utilized to adapt (5.10) to feature space [183]. In general, the Fourier neural attention is expressed as follows,

$$K(X) = rFFT^{-1} \circ \tilde{K} \circ rFFT(X), \quad (5.11)$$

where $\tilde{K}$ represents a mapping parameterized by trainable weights, e.g. convolutional layers [183, 182] and the frequency-domain filter [181], and $\circ$ denotes the composition of operators. Since real-valued signals lead to the conjugate symmetry property in the transformed space, it allows dropping half the pixel values to decrease the memory and computational costs.

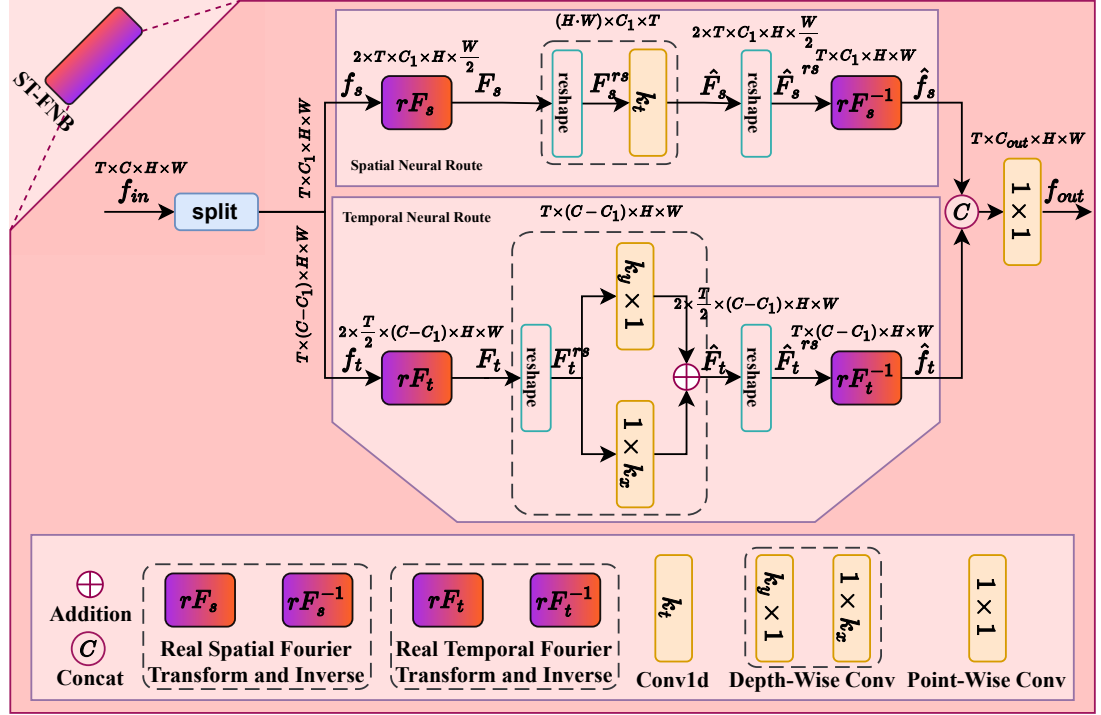


Figure 5.2: Illustration of spatio-temporal Fourier neural block (ST-FNB). The input feature volume is split into two groups which respectively undergo the top spatial neural route and the bottom temporal neural route, where the information exchange among the global region in both spatial and time domains is achieved. The output feature volume is produced via a point-wise convolutional layer with the combination of the outputs coming from the two routes.

5.2.3.2 Spatio-Temporal Dual-domain Modelling

We propose to make use of the efficient design of the Fourier neural attention with the object of retrieving useful cues and achieving information exchange in the global receptive field of the extracted feature representations in both spatial and time domains. Let the input features, dubbed f_{in} , have a size of $T \times C \times H \times W$ corresponding to $[time, channels, height, width]$. The feature volume is divided into two elements along the channel axis before undergoing the Fourier transform respectively applied to the spatial and temporal dimensions, which is, as presented in Fig. 5.2, followed by convolutional layers to handle the relations in complementary dimensions. The input features partitioned into two groups, dubbed f_s and f_t of size $T \times C_1 \times H \times W$ and $T \times (C - C_1) \times H \times W$ where C_1 is a hyperparameter, are respectively passed to two branches, the spatial neural route and the temporal neural route, entailing the operations introduced below.

I) Spatial Neural Route: As demonstrated as Fig. 5.2, the feature group f_s as input to the top route is processed by applying the real spatial Fourier transform, dubbed rF_s , over the spatial dimensions, i.e. $height \times width$, for the purpose of retrieving useful information among the global region. A 1D convolutional layer having a kernel size of k_t is implemented to the transformed and reshaped features F_s^{rs} coming from the spatial Fourier block over the time dimension, mixing the extracted attributes among the temporal frames. The resulting volume, termed $\hat{F}_s$, is then reshaped and mapped by the inverse Fourier transform rF_s^{-1} to the original feature space. The output $\hat{f}_s$ from the spatial neural route has pixel values not only encoding the global spatial correlations but also capturing the local information in time domain.

II) Temporal Neural Route: Along the bottom route, as illustrated in Fig. 5.2, the real temporal Fourier transform, dubbed rF_t and applied to the time axis, and a spatial separable convolutional layer executed over the spatial dimensions with kernel sizes of $k_y \times 1$ and $1 \times k_x$ are used to achieve information exchange in complementary domains. The split group f_t is mapped to the feature volume $\hat{F}_t$ which undergoes the inverse Fourier transform rF_t^{-1} in temporal domain, obtaining the features $\hat{f}_t$ with the global informative cues among the frame sequence mixed.

III) Channel Fusion: To exchange the information coming from the spatial and temporal routes, the outputs from the top and bottom paths are concatenated along the channel axis, forming a feature volume of the same size as the input, and a point-wise convolution is subsequently employed to fuse feature channels, yielding the output f_{out} . This design not only allows mixing the deep representations learned by the two paths, but also enables a difference output size, e.g. $T \times C_{out} \times H \times W$, which enhances the flexibility of the deployment of the ST-FNB.

5.2.3.3 Parameter Count Analysis

We analyze the parameter size of the ST-FNB to showcase the efficiency of the network design. The modelling of the correlations in the spatial and temporal domains leads to the size of the parameters computed by:

$$P_{ST-FNB} = C_1^2 \cdot k_t + (C - C_1) \cdot (k_y + k_x) + C \cdot C_{out}, \quad (5.12)$$

which can be treated as a function of C_1 , k_t , k_y , and k_x , given the channel sizes C and C_{out} of the specified input and output feature volumes. To strike a better trade-off between model ability and efficiency, we upper-bound (5.12) by the parameter count of a depth-wise separable convolution, which is expressed by $P_{sepConv} = C \cdot k'_y \cdot k'_x + C \cdot C_{out}$ with k'_y and k'_x referring to the kernel sizes, giving rise to:

$$\begin{aligned} P_{ST-FNB} &\leq P_{sepConv} \\ \Rightarrow C_1 &\leq \frac{(k_y + k_x) + \sqrt{(k_y + k_x)^2 + 4 \cdot C \cdot k_t \cdot (k'_y \cdot k'_x - k_y - k_x)}}{2k_t}. \end{aligned} \quad (5.13)$$

The relationship depicted in (5.13) enables us to control the parameter size of the ST-FNB by adjusting C_1 in the aforementioned split scheme according to the input channel width C , given a fixed kernel configuration. For instance, setting $(C, C_1) \in \{(16, 3), (32, 6), (64, 8)\}$ with $k_t = k_y = k_x = 3$ satisfies (5.13), which establishes the upper-bound of the parameter count of the ST-FNB using a depth-wise separable convolution with $k'_y = k'_x = 3$.

5.2.4 Mask-Guided Motion Feature Incorporation (Mask-MFI)

We introduce a mask-guided motion feature incorporation (Mask-MFI) scheme, as illustrated in Fig. 5.1, to extract the motion information from the under-sampled measurements acquired at consecutive time steps and incorporate the associated features into the reconstruction iterations to help the model recover the time-varying image features. It is noticeable that dynamic MRI sequences abound with temporal redundancies which can be expressed in two aspects: 1) the stationary or slow-moving background and 2) the dynamic content, also called the foreground, portraying the motion of the organ being imaged.

On one hand, the reconstruction of the background can be formulated as a de-aliasing task which potentially benefits from the correlations with neighboring frames through its static properties, e.g. via the data sharing layer [173]. On the other hand, the time-varying content is encoded in the observations measured at individual time steps, which becomes more challenging to preserve with the increase of acceleration rates due to the lack of relevant dynamic cues. Thanks to the variations in sampling patterns adopted for different frames, which are specifically designed for dynamic MRI, the pixels in k -space can be measured by more than one frame, which enables the extraction of the dynamic hints from a collection of measurement frames sampled at multiple time steps. Motivated by those observations, the Mask-MFI is devised to explicitly utilize the time-dependent observations to assist in restoring the dynamic content. In the following, the constituent components of the Mask-MFI are described in detail.

5.2.4.1 Deep Extraction of Motion Features

D) Temporal Difference Modelling: To extract the motion features and generate the associated motion masks, the weighted encoding average, termed $\bar{\mathbf{y}}^t$ with respect to the t -th time frame, is computed for each temporal frame as a representation for the “average” information conceptually measured at the t -th time step with the encoding matrix $\mathbf{D}_t$, utilizing the data

acquisition prior as follows,

$$\bar{\mathbf{y}}^t(q) = \frac{1}{\sum_{t'} \delta(\mathbf{D}_t, \mathbf{D}_{t'})(q)} \sum_{t'} (\mathbf{D}_t \mathbf{D}_{t'}^H \mathbf{y}^{t'})(q), \quad (5.14)$$

where q denotes the spatial position and $\delta(\mathbf{D}_t, \mathbf{D}_{t'})(q) = 1$ if the sampling patterns spanned by the diagonals of the two encoding matrices $\mathbf{D}_t$ and $\mathbf{D}_{t'}$ are not disjoint and the specified k -space location q belongs to their intersection, and it is zero otherwise. The associated temporal difference map, dubbed ϕ^t , is calculated by subtracting the corresponding weighted encoding average from the t -th observation, obtaining:

$$\phi^t = \sum_{i=1}^{n_c} \mathbf{S}_i^H \mathbf{F}^H (\mathbf{y}_i^t - \bar{\mathbf{y}}_i^t). \quad (5.15)$$

The temporal difference is the sensitivity-weighted combination of the differences between the observation measured at the t -th time step and the average of the dynamic sequence undergoing the time-dependent encoding process, therefore potentially containing the dynamic cues in the form of the changes of image features varying over time.

Different from other approaches which retrieve the relevant information from the neighboring frames without distinguishing the static background and the time-varying objects, the proposed method has the ability to “localize” the regions bearing the underlying dynamics with the background redundancies suppressed. Considering that the moving objects often take a small proportion of the whole spatial area, leveraging the temporal redundancies without imposing the constraint of the forward encoding process corresponding to the individual time frames can unfavorably affect the recovery of the dynamic content, and lead the model to neglect the temporal variations and produce the average of consecutive frames.

II) Motion Feature Generation: It is noteworthy that the resulting $\{\phi^t\}_{t=1}^T$ still suffer from heavy aliasing artifacts. To extract useful information, a feature pyramid network, as demonstrated in Fig. 5.1, is adopted to refine the temporal differences, forming a collection of motion features termed $\{h^i\}_{i=0}^2$ of different levels, where $i = 0$ refers to the top scale and $i = 2$ denotes the bottom level.

III) Motion Mask Generation: We compute a volume of motion masks M^0 from the temporal difference maps via a set of morphological operations, as displayed in Fig. 5.1, in an attempt to roughly indicate the high-variation regions. To construct a pyramid of masks corresponding to the motion features at different scales, the pooling operation is used to decrease the resolution of masks, yielding $\{M^i\}_{i=0}^2$. The paired motion features and masks are shared by all reconstruction iterations to achieve an efficient implementation in terms of computational complexity and parameter count.

5.2.4.2 Motion Feature Modulation and Masking (MFMM)

To enhance the representation capacity of the features extracted to capture the underlying dynamic information, we propose a motion feature modulation and masking method, dubbed MFMM, to modulate the motion features before incorporating them with the features coming from the reconstruction subnetworks, which is implemented under the guidance of the associated motion masks. The MFMM is motivated by the findings in computer vision, revealing that the statistics of feature maps embodies the important attributes of the learned deep representations. The mean of feature maps is shown to be related to the global semantic content whereas the variance encodes the local details, and feature modulation, which performs a supplementary mechanism to feature extraction and normalization, has proven to be effective in improving the model performance in image generation and editing tasks [184, 185, 186]. The involved steps of the MFMM are introduced below.

I) Feature Modulation: The motion features h^i are modulated under the guidance of the generated motion masks M^i , leading to the modulated features termed Φ , which is applied at multiple levels and admits:

$$\Phi(h^i; M^i) = \frac{h^i - \mu(h^i; M^i)}{\sigma(h^i; M^i)}, \quad (5.16)$$

where μ and σ denote the mean and standard deviation operators defined as a function of motion masks:

$$\begin{aligned} \mu(h^i; M^i) &= \frac{1}{|M^i|} \sum_q h^i(q) \odot M^i(q) \\ \sigma(h^i; M^i) &= \frac{\sqrt{\sum_q (h^i(q) - \mu(h^i; M^i))^2 \odot M^i(q)}}{\sqrt{|M^i|}}, \end{aligned} \quad (5.17)$$

and $\odot$ denotes element-wise multiplication, q represents spatial position, and $|M^i|$ is the count of desired pixels in M^i .

II) Feature Masking: We apply the motion masks, which are produced to localize the foreground area where the content changes over time, to the modulated features in a way to specify the desired pixel positions, obtaining the masked features highlighting the moving objects and suppressing the stationary background:

$$\tilde{h}^i = M^i \odot \Phi(h^i; M^i). \quad (5.18)$$

5.2.4.3 Motion Feature Coupling

The features $\tilde{h}^i$ modulated via the MFMM are leveraged to aid the preservation of the dynamic content during the reconstruction procedure of MRI sequences. As illustrated in Fig. 5.1, the features coming from different levels of the reconstruction pyramid at the k -th iteration are coupled with the motion features by:

$$\tilde{f}_k^i = f_k^i \oplus \tilde{h}^i, \quad (5.19)$$

where $\oplus$ refers to element-wise addition. It is worth noting that the generated motion features are shared by all reconstruction iterations to avoid significantly increasing the model parameters and computational overhead.

In other words, the information of the moving content is separately extracted by a pyramid network and coupled with the deep representations produced by the reconstruction pyramids to exploit the dynamic cues explicitly as a means to help capture the time-dependent organ motion, whereas the background, which often keeps static or moving slowly, can be better reconstructed by utilizing the temporal redundancies, e.g. via data sharing [173], recurrent architectures [120, 122], or the proposed ST-FNB.

5.3 Experiments

5.3.1 Implementations

We conduct experiments using the OCMR dynamic dataset [187] where multi-coil k -space data are available to evaluate and compare different methods for dynamic MRI reconstruction. The dataset is randomly divided into training, test, and evaluation groups respectively containing 100, 50, and 33 samples, which were acquired using MAGNETOM scanners with the number of coil sensors ranging from 15 to 38 and provide cardiac cine slices in three different views, potentially raising more challenges to methods for reconstructing dynamic MRI series and making it a benchmark to objectively evaluate reconstruction models for the purpose of expediting clinical translation of those algorithms.

The model is trained for 50 epochs with the fully sampled images as the ground truth (GT) and optimized towards minimizing the objective function where the L_1 and structural similarity index (SSIM) metrics are used to measure the differences between the reconstruction result $\mathbf{x}^N$ and the sensitivity-weighted GT reference $\mathbf{x}^{GT}$:

$$L(\theta) = E_{\{\mathbf{x}^N, \mathbf{x}^{GT}\}} [(1 - \alpha')L_1(\mathbf{x}^N, \mathbf{x}^{GT}) + \alpha'L^{SSIM}(\mathbf{x}^N, \mathbf{x}^{GT})], \quad (5.20)$$

where θ denotes the collection of the parameters of all subnetworks, α' is empirically selected to 0.2 to strike a trade-off between pixel-wise evaluation and structure-based metric, and the number N of reconstruction iterations is set to 5 in our experiments. The coil sensitivity maps are pre-calculated using the ESPIRiT [74] algorithm with the time-averaged k -space center. We compare with other competitive reconstruction approaches including: DL-ESPIRiT [107], DCCNN [173], DIMENSION [108], kt-NEXT [121], CTFNet [122], MODRN [123], LSNet [118], and SLRNet [119], which we briefly summarize in the following:

- DL-ESPIRiT [107] adopts 3D convolutional kernels in a CNN framework to enhance the learning in both spatial and temporal domains for dynamic image recovery.
- DCCNN [173] employs a convolutional neural network for dynamic MRI reconstruction with data sharing layers interleaved to leverage the informative features along the time axis of data sequences.
- DIMENSION [108] combines the k -space frequency learning with the spatial domain learning to improve the prediction of a 3D CNN architecture.
- kt-NEXT [121] proposed to exploit the redundant and complementary information contained in under-sampled MRI series via the recurrence steps of convolutional recurrent neural networks (CRNN) between image and x - f spaces.
- CTFNet [122] performs predictions in both x - t and x - f domains by utilizing two CRNN models with recurrence over iterations and time dimension.
- MODRN [123] adopts a recurrent network for dynamic MRI reconstruction and proposed a motion compensation module to exploit the redundancies in the neighboring frames to enforce motion consistency and refine the initial prediction.
- LSNet [118] utilizes a 3D CNN with a low-rank layer to achieve a reduction in the rank of reconstruction and model the sparsity property of signals.
- SLRNet [119] introduced a learned sparse transform via an ISTA-Net [104] with a soft threshold to reinforce the low-rank constraint for reconstruction problems.

The reconstruction methods are evaluated under $8\times$, $16\times$, and $24\times$ acceleration using the variable density sampling strategy [188] to simulate the accelerated data acquisition, which shows to be an efficient Cartesian sampling scheme for dynamic MRI sequences.

5.3.2 Comparison Results

The quantitative comparisons are reported in Table 5.1 where NRMSE, PSNR, and SSIM are used as evaluation metrics to gauge the reconstruction performance. It is shown that the devised model consistently outperforms other comparison methods in all metrics and presents competitive execution speed with fewer parameters, showcasing the effective and efficient design of the proposed framework.

Table 5.1: Quantitative comparison results evaluated on OCMR dataset. Acceleration factors are respectively $8\times/16\times/24\times$. Running time (s) and parameter size (MB) are same for different acceleration rates.

method	NRMSE↓	PSNR↑	SSIM↑	run-time↓	size↓
proposed	0.0115/0.0146/0.0201	39.28/37.21/34.16	0.939/0.925/0.888	0.039	341019
SLRNet [119]	0.0144/0.0176/0.0223	37.31/35.48/33.31	0.883/0.866/0.838	0.089	865663
LSNet [118]	0.0154/0.0192/0.0276	36.73/34.70/31.56	0.883/0.861/0.817	0.058	605964
MODRN [123]	0.0160/0.0205/0.0251	36.38/34.18/32.30	0.882/0.857/0.826	0.151	5427448
CTFNet [122]	0.0142/0.0176/0.0227	37.50/35.52/33.17	0.888/0.875/0.846	0.112	668921
kt-NEXT [121]	0.0145/0.0180/0.0215	37.37/35.48/33.65	0.885/0.870/0.849	0.040	988955
DIMENSION [108]	0.0139/0.0183/0.0238	37.88/35.33/32.72	0.887/0.869/0.839	0.038	960099
DCCNN [173]	0.0160/0.0200/0.0234	36.47/34.36/32.90	0.882/0.861/0.835	0.026	999448
DL-ESPIRiT [107]	0.0213/0.0246/0.0278	33.72/32.48/31.41	0.850/0.828/0.807	0.956	1526715

The reconstruction results produced by different models are presented in Fig. 5.3, where it is shown that the proposed method is able to achieve more faithful reconstructions at acceleration factors ranging from $8\times$ to $24\times$ with sharper structures, compared to other competing methods. From the associated residual maps, it can be observed that our method produces fewer errors, particularly at aggressive under-sampling rates, showing the improved performance of the devised framework.

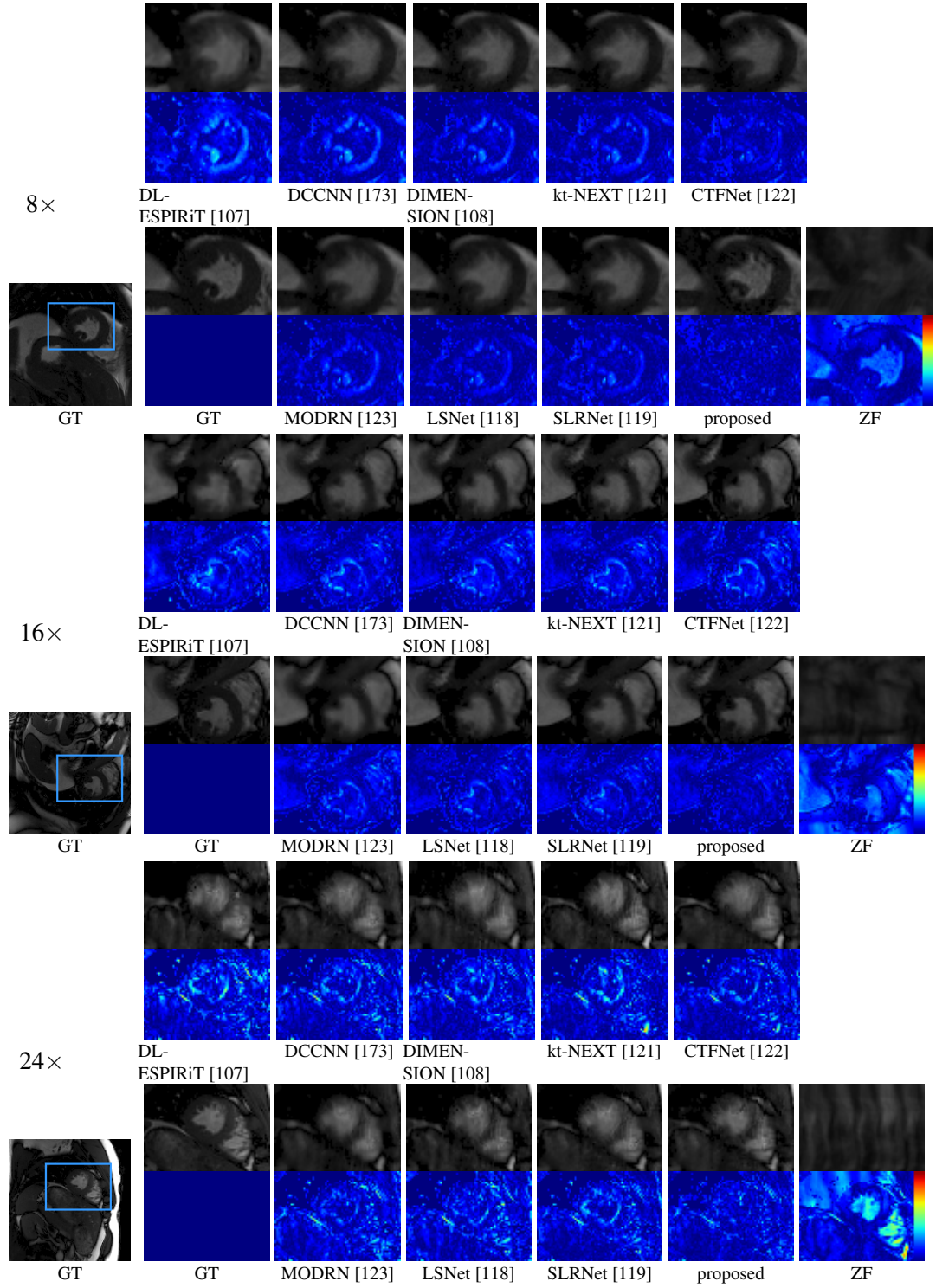


Figure 5.3: Reconstruction comparisons evaluated on OCMR dataset. GT refers to the ground truth image and ZF denotes the zero-filled.

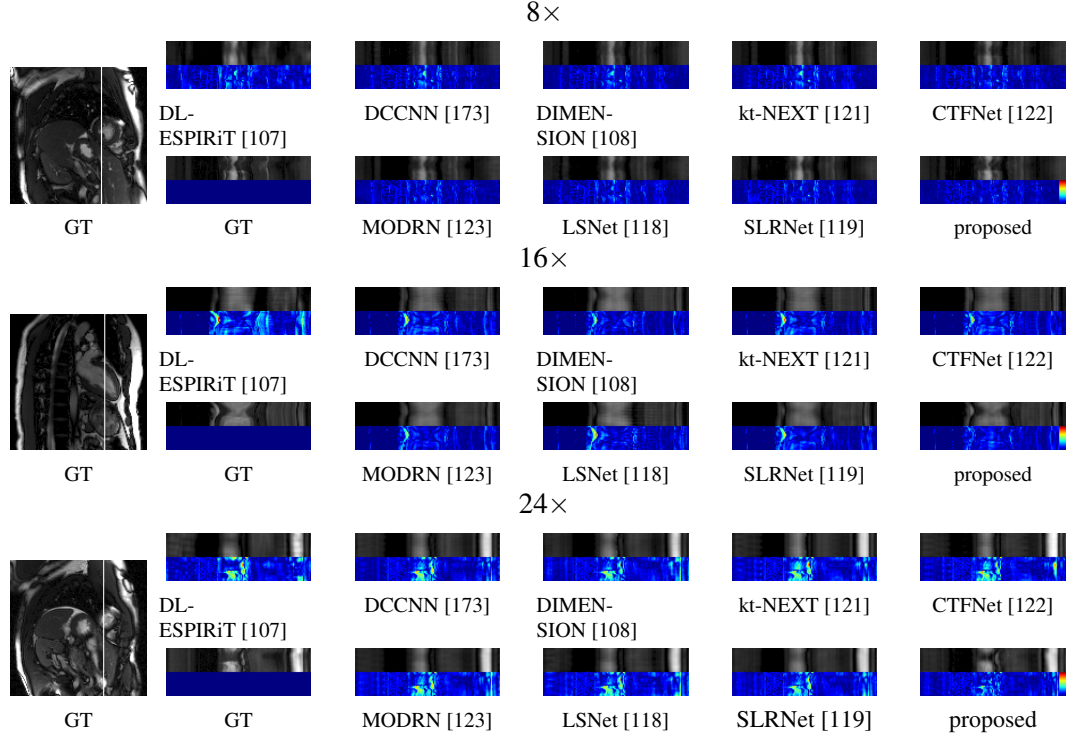


Figure 5.4: Comparison of temporal profiles on OCMR dataset. GT refers to the ground truth image.

In Fig. 5.4, the temporal dynamics is demonstrated by stacking the pixel values on the same column, as indicated in the ground truth, along the time axis, resulting in the temporal profiles indicative of the model capacity to capture and preserve the dynamic attributes in MRI sequences. We observed that the proposed method attains more consistent temporal profiles with the ground truths with fewer errors as displayed in the corresponding residual maps. Under $24\times$ acceleration where very limited cues are available from the observed measurements for the modelling of the dynamic content, our method is still able to capture the subtle motion of the underlying organ, whereas other approaches can struggle to accurately recover the moving content and lead to strong errors in the temporal profiles. Those observations demonstrate the superior performance of our method in terms of the quality of reconstructed images and the ability to preserve the dynamic information.

5.3.3 Ablation on Model Components

We demonstrate the contributions of the devised model components, ST-FNB and Mask-MFI, to improving the network performance in terms of reconstruction accuracy and the recovery of the dynamic content in MRI sequences. To manifest the roles of the proposed components, we evaluate two ablation candidates, termed w/o ST-FNB and w/o Mask-MFI, which are respectively deprived of the ST-FNB and Mask-MFI. The depth-wise separable convolution is used in place of the ST-FNB in candidate w/o ST-FNB in a way to keep the parameter size, pursuant to the criterion displayed in (5.13). The baseline model is derived by removing both the ST-FNB and Mask-MFI. For fair comparisons, the parameters of each candidate are balanced by slightly adjusting the channel size of several convolutional layers, providing approximately the same model size.

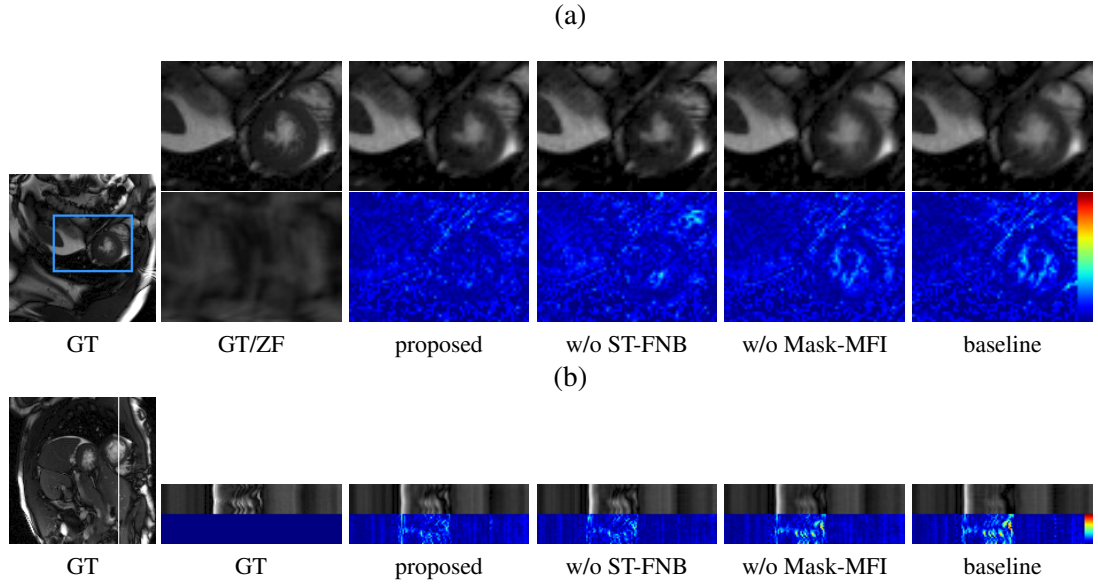


Figure 5.5: Ablation studies on model components evaluated on OCMR dataset under $16\times$ acceleration.

The comparison results of different ablation variants are presented in Fig. 5.5, showing the reconstructed images and the temporal profiles, which are evaluated using the OCMR dataset at $16\times$ acceleration factor. From the results in Fig. 5.5 (a), we can observe a reduction in spatial errors produced by candidate w/o Mask-MFI as compared to baseline, indicating the performance improvement obtained by the ST-FNB. From Fig. 5.5 (b), the comparison

Table 5.2: Ablation studies on model components evaluated on OCMR dataset under $16\times$ acceleration. To provide a fair comparison, we adjust the width of a few convolutional layers in the competing model candidates to balance the parameter size, though it can introduce slightly more parameters.

method	NRMSE↓	PSNR↑	SSIM↑	run-time(s)	size
proposed	0.0146	37.21	0.925	0.039	341019
w/o ST-FNB	0.0152	36.85	0.917	0.029	341068
w/o Mask-MFI	0.0160	36.45	0.902	0.038	342592
baseline	0.0168	35.98	0.891	0.027	342671

between the temporal profiles of candidate w/o ST-FNB and baseline suggests the effectiveness of the Mask-MFI in preserving the dynamic information. Although it is shown in Fig. 5.5 to significantly decrease the errors in the recovery of the dynamic content and therefore generate more faithful reconstructions, adopting the Mask-MFI method seems to cause extra spatial errors while restoring the image features embodying the characteristics of dynamic imaging, as displayed in the residual map of variant w/o ST-FNB in Fig. 5.5 (a). Contrasting with the foregoing observation, the models with the ST-FNB produce fewer spatial errors, e.g. comparing the results of candidate w/o Mask-MFI and baseline, whereas it makes limited contributions to capturing the dynamic characters, see the comparison between candidate w/o Mask-MFI and proposed in Fig. 5.5 (b). Those findings not only verify the efficacy of the devised method, but also indicate a synergistic effect of the two model components that the ST-FNB improves the capacity of the reconstruction framework to alleviate spatial errors and provide more accurate results while the Mask-MFI has a beneficial effect on leveraging the dynamic cues hidden in the input sequence to restore the moving objects.

We present the assessment results in Table 5.2, where the devised model components consistently bring performance improvements and their combination yields the best results in terms of all evaluation metrics, confirming the efficacy and synergy of the proposed method to deliver considerable performance gains. We found that candidate w/o ST-FNB employing the Mask-MFI offers more improvements over the baseline model than candidate w/o Mask-MFI where the ST-FNB is used, which suggests the potential of explicitly extracting and incorporating the motion information as a means to exploit the characteristics of dynamic sequences. From

Table 5.2, it is shown that candidate w/o ST-FNB and baseline offer a faster execution speed, which is due to the removal of the ST-FNB, therefore avoiding the implementation of the Fourier transform in spatial and temporal domains, and suggests further improvement in terms of execution speed by exploring more efficient dual-domain transformations in an attempt to retrieve useful information from the global region.

5.3.4 Ablation on Mask-MFI

The motion feature modulation and masking, termed MFMM and adopted in Mask-MFI as a means to modulate the extracted motion features prior to incorporating with the feature maps from the reconstruction model, is proposed to improve the efficacy of the devised Mask-MFI, which we conduct ablation experiments to verify by respectively removing the modulation and masking operations serving as the major aspects of MFMM, leading to the candidates termed w/o masking and w/o modulation. We also compare against other two feature incorporation approaches. The first one is to simply combine the motion and reconstruction features together by element-wise addition, dubbed addition. The second one utilizes the instance normalization (*IN*) [184], which is often employed to transform the statistics of feature representations for the purpose of encouraging higher performance, to modulate the features, obtaining the aggregation method shown below,

$$\tilde{f}_k^i = f_k^i \oplus IN(h^i), \quad (5.21)$$

which is denoted by IN-addition.

Table 5.3: Ablation on mask-guided motion feature incorporation (Mask-MFI) via OCMR data under $16\times$ acceleration.

method	NRMSE↓	PSNR↑	SSIM↑
proposed	0.0146	37.21	0.925
w/o modulation	0.0157	36.64	0.909
w/o masking	0.0160	36.42	0.909
addition	0.0158	36.52	0.910
IN-addition	0.0155	36.70	0.910

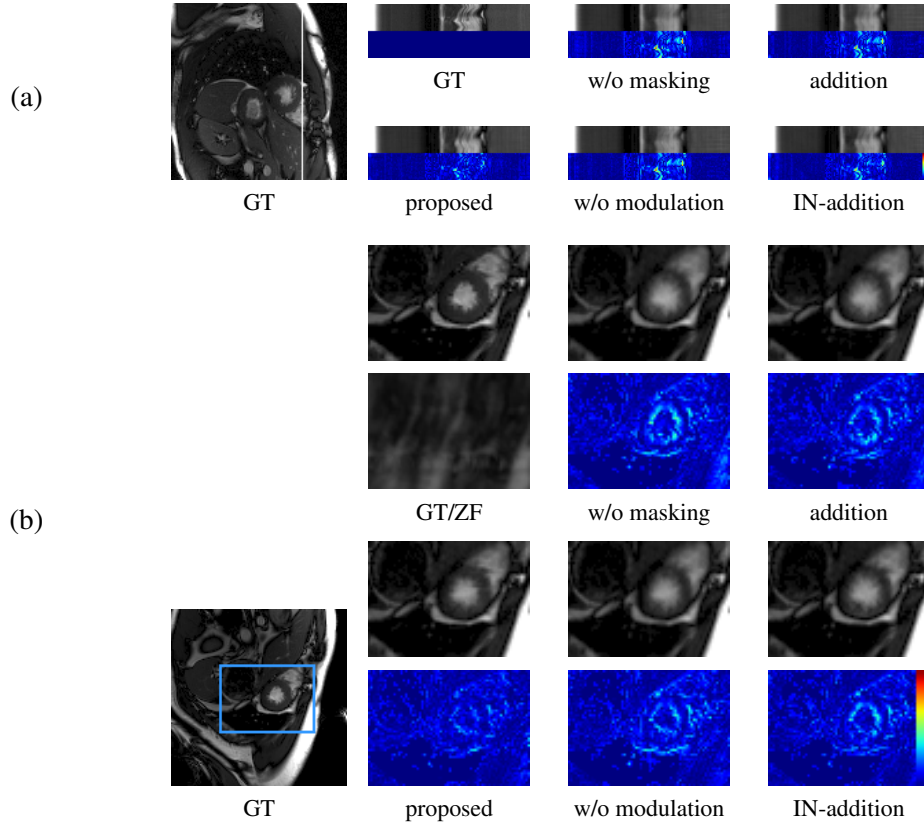


Figure 5.6: Ablation on mask-guided motion feature incorporation (Mask-MFI) using OCMR dataset under $16\times$ acceleration.

The comparison results using different feature modulation approaches are presented in Fig. 5.6, where it is shown that the proposed MFMM can generate more faithful images with fewer spatial errors and more accurately recover the temporal profile. Compared to feature addition, the method IN-addition improves the model capacity to preserve the dynamic content, which implies the usefulness of the modulation of motion features. However, normalizing the feature maps via the statistics computed with pixel values in the global region, as used in IN-addition, can fail to concentrate on the local regions potentially containing the moving objects and lead to inferior performance as shown in Fig. 5.6 (a), compared to the proposed method which focuses on the motion area via the guidance of motion masks. As demonstrated in Fig. 5.6 (b), the proposed method produces fewer reconstruction errors, confirming the enhanced performance by adopting the MFMM approach in the incorporation of motion features.

From the quantitative comparisons presented in Table 5.3, it can be discovered that the motion information exploited via the Mask-MFI method delivers performance gains in terms of evaluation metrics, compared to candidate w/o Mask-MFI in Table 5.2. It is shown that candidate w/o masking produces inferior evaluation results, which is potentially owing to the joint manipulation of the slowly changing background and the foreground featuring the dynamic content which interfere with each other and consequently result in degraded performance. Comparing candidate w/o modulation where the generated masks are used to “localize” the time-varying foreground with candidate addition which adopts neither feature modulation nor masking mechanism, we observed that simply combining the motion features into the reconstruction pipeline fails to obtain considerable performance improvements, whereas applying the instance normalization to the motion features before coupling with those coming from the subnetworks leads to better evaluation scores. It implies the importance and necessity of implementing feature masking to focus on the regions potentially having the underlying dynamic content. The proposed method with both feature modulation and masking attains the best results in all metrics, confirming the effectiveness of the MFMM to improve the efficacy of the Mask-MFI in modeling time-varying dynamics of the anatomy from the under-sampled measurements. It is noteworthy that following the spirit of model simplicity the design of the MFMM only draws on rudimentary mathematical operations which can be efficiently implemented, which allows for further improvement by developing more complex feature modulation and coupling methods.

5.3.5 Transferability of Mask-MFI

We demonstrate the transferability of the proposed mask-guided motion feature incorporation scheme by combining it with other competing models including CTFNet, MODRN, and DCCNN that adopt recurrent or unrolled reconstruction pipelines. To simplify the implementations, each reconstruction stage of those methods, briefly illustrated in Fig. 5.7 (a), are slightly revised to incorporate the motion features extracted by a shallow network with the features generated from the input to the subnetwork by a single convolutional layer before

applying the specified reconstruction step, as delineated in Fig. 5.7 (b). Since the input to the subnetwork of the original reconstruction pipeline can have a distinct shape from that of the features produced by the Mask-MFI, we adjust the channel number of the first layer of each subnetwork and accordingly “reshape” the involved feature volume, for the purpose of meeting the input requirement. The model size is balanced to offer a fair comparison.

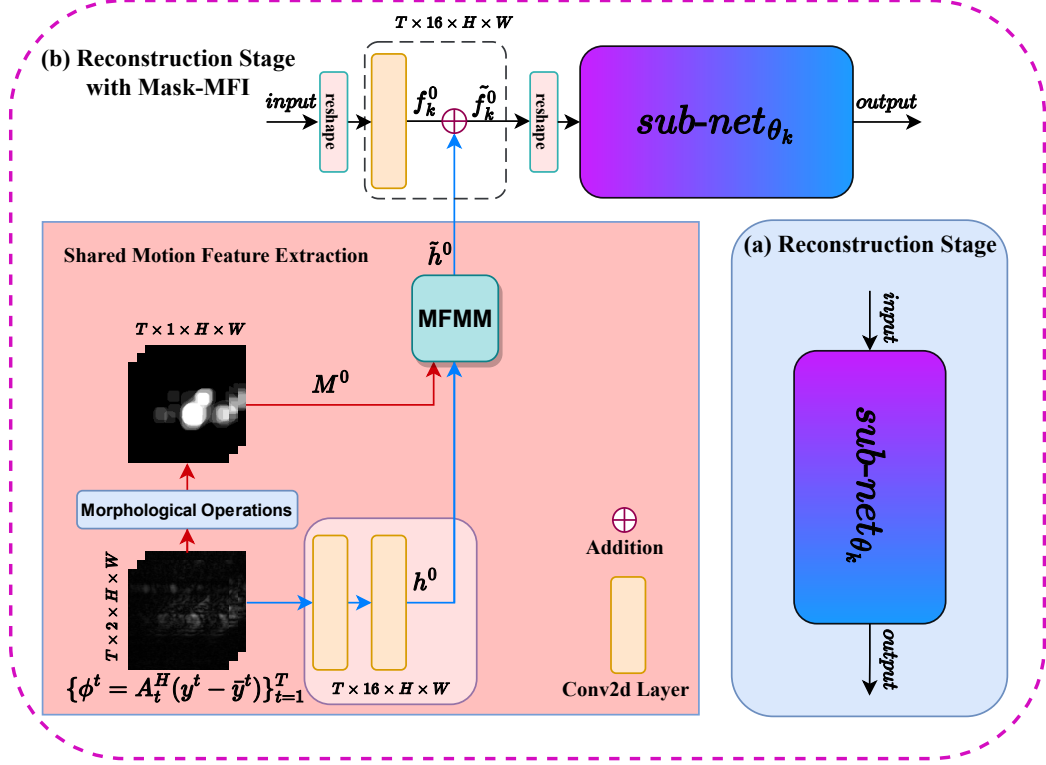


Figure 5.7: Illustration of transferring Mask-MFI to other models. The motion features and masks are generated with the under-sampled MRI sequences and incorporated into each reconstruction iteration by coupling with the features extracted by a convolutional layer at each reconstruction stage.

Table 5.4: Comparison between the models with and without the presence of Mask-MFI evaluated on OCMR dataset under $16\times$ acceleration.

method	NRMSE↓	PSNR↑	SSIM↑
CTFNet [122]	0.0176	35.52	0.875
CTFNet+Mask-MFI	0.0165	36.08	0.877
MODRN [123]	0.0205	34.18	0.857
MODRN+Mask-MFI	0.0174	35.66	0.867
DCCNN [173]	0.0200	34.36	0.861
DCCNN+Mask-MFI	0.0173	35.73	0.868

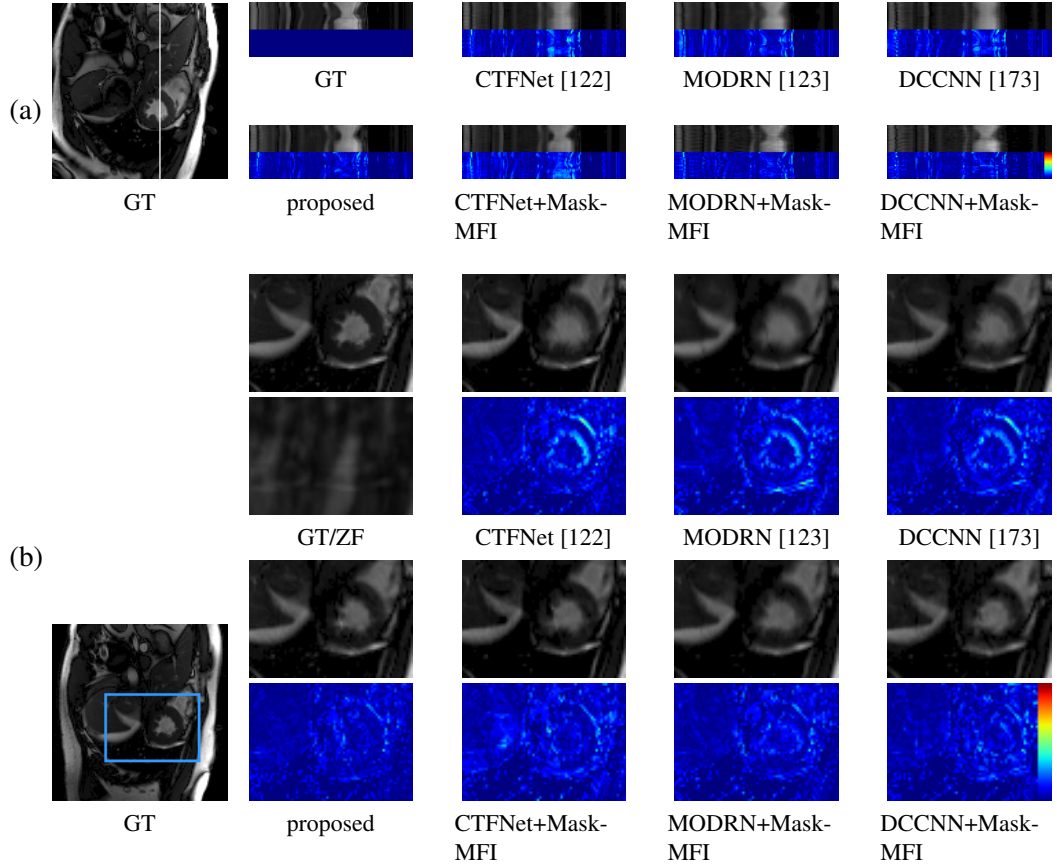


Figure 5.8: Transferability experiment on mask-guided motion feature incorporation (Mask-MFI) using OCMR dataset under $16\times$ acceleration.

The quantitative comparison results of models with and without using the devised Mask-MFI are presented in Table 5.4. It can be observed that the models with the presence of Mask-MFI demonstrate consistently improved reconstruction results in terms of evaluation metrics, which verifies the effectiveness and transferability of the proposed motion feature extraction and incorporation method.

The reconstruction results of models with and without the presence of Mask-MFI are displayed in Fig. 5.8, where it is shown that the dynamic attributes can be better preserved with the motion information explicitly captured and embedded into reconstruction iterations, suggesting the feasibility of the Mask-MFI when applying it to other reconstruction models. Although it improves the performance of CTFNet in terms of the recovery of the dynamic content, as shown in Fig. 5.8 (a), adopting the Mask-MFI leads to extra errors presented in the

temporal profile, which we surmise is due to the recurrent network architecture of CTFNet that is not well suited to propagate the motion information coming from a single forward pass of the Mask-MFI, providing implications for further improvements by employing a propagation-based design for the motion feature incorporation scheme.

The revision for reconstruction models introduced above is simple and straightforward, only aiming to demonstrate the practicality of the Mask-MFI when transferring it to different networks. We hope that it can incentivize the exploration of more useful and effective methods to extract and leverage the dynamic cues hidden in MRI sequences acquired in an acceleration diet.

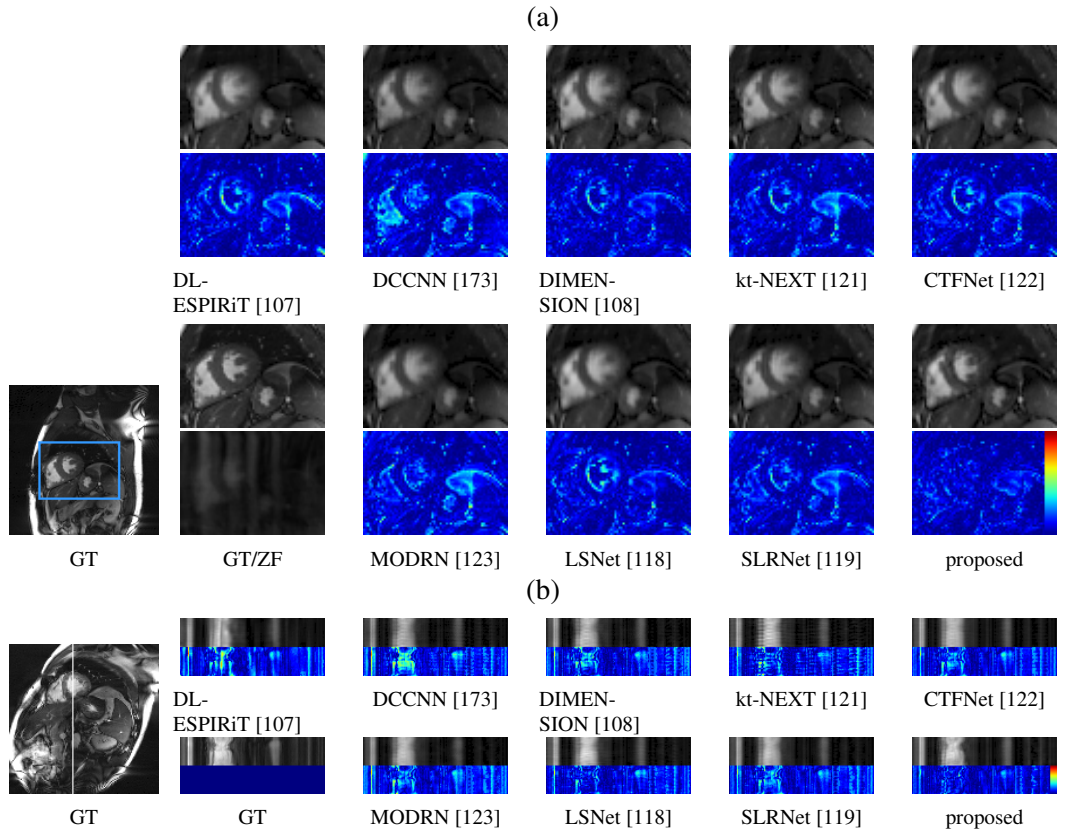
5.3.6 Model Generalization Validation

To evaluate the generalization capacity of reconstruction models, we measure the performance of the models, which are trained on the OCMR dataset, using data from the Harvard data-verse [189] where signals along radial lines were acquired per frame, which can be modelled via the non-uniform fast Fourier transform [162] (NUFFT). We utilize the NUFFT and the variable density sampling method to emulate the k -space under-sampling process. Although radial sampling strategies are also commonly adopted to accelerate MRI acquisition, creating reconstructions requires the use of operations such as the NUFFT and gridding [163], which can potentially propagate interpolation errors having an unfavorable impact on the reconstruction results. The dataset is randomly split into the training, test, and evaluation sets containing 63, 30, and 15 samples. We also report the assessment of the frameworks fine-tuned on the Harvard data for 30 epochs.

The evaluation results of the models with and without fine-tuning are presented in Table 5.5, where the promising results are obtained when measuring the models trained on one dataset with samples from a different dataset. It is also shown that fine-tuning (FT) can consistently deliver performance improvements. We observed that the proposed method surpasses other competing candidates with and without the presence of fine-tuning, indicating the model gener-

Table 5.5: Evaluating models trained on OCMR dataset using Harvard dataverse for testing and fine-tuning (FT) at $16\times$ acceleration factor (w/o FT/w FT).

method	NRMSE↓	PSNR↑	SSIM↑
proposed	0.0256/0.0216	32.37/33.83	0.916/0.923
SLRNet [119]	0.0298/0.0275	30.77/31.67	0.857/0.871
LSNet [118]	0.0294/0.0279	30.93/31.54	0.850/0.879
MODRN [123]	0.0285/0.0271	31.31/31.81	0.836/0.872
CTFNet [122]	0.0265/0.0259	31.99/32.26	0.878/0.900
kt-NEXT [121]	0.0298/0.0270	30.78/31.83	0.852/0.868
DIMENSION [108]	0.0294/0.0270	30.93/31.86	0.870/0.889
DCCNN [173]	0.0293/0.0264	31.18/32.07	0.860/0.883
DL-ESPIRiT [107]	0.0336/0.0324	29.78/30.35	0.841/0.859

**Figure 5.9:** Reconstruction comparisons evaluated on Harvard dataverse with $16\times$ acceleration and fine-tuning.

alization on new datasets. The images produced by fine-tuned models are displayed in Fig. 5.9, where it is shown that our method can produce sharper details with fewer spatial errors and more faithful temporal profiles delineating the organ dynamics, confirming its generalization capacity to restore the under-sampled dynamic sequences.

5.4 Conclusion

In this chapter, an optimization-inspired deep cascaded framework was proposed to reconstruct dynamic MRI sequences from the k -space measurements sampled in accelerated data acquisition diets, with the object of shortening the data sampling process when imaging dynamics of organs and expediting the application of dynamic MR imaging in time-critical practice. We devised a mask-guided motion feature incorporation method, dubbed Mask-MFI, to explicitly capture the motion information of the underlying anatomy being inspected from a subset of measurements, which shows to remarkably enhance the model capacity to preserve the time-varying content and potentially suggests a new direction to model the dynamic content from the under-sampled data in MRI reconstruction.

By experiments, it was demonstrated that the devised framework surpasses other deep learning-based methods in terms of both the reconstruction quality and the recovery of the dynamic content. It was verified that the improved performance is attributable to the devised Mask-MFI and the spatio-temporal Fourier neural block, termed ST-FNB and designed to efficiently retrieve useful information from the global regions in spatial and temporal domains. In addition, the transferability and generalization capacity of the proposed method were validated, showing that our method could generalize well on other types of network pipelines and unseen MRI samples, which allows the application of our method to other frameworks to improve their performance and potentially facilitates the deployment of reconstruction frameworks in real-world applications.

A promising research direction is to speed up the inference by accelerating the execution of the global “mixing” operation adopted in the ST-FNB and explore the potential of this method in real-time applications. It can be an interesting direction to explore other types of sampling strategies for dynamic MRI, e.g. radial sampling patterns and jointly optimizing the data encoding process in a learning-based scenario.

6.1 Summary

The research presented in this thesis concentrates on the reconstruction of MRI images from the k -space data measured under accelerated data acquisition diets aiming to shorten the intrinsically time-consuming sampling process. To provide high-quality reconstructions from highly under-sampled measurements, we developed deep neural network models to recover clean images by exploiting the representation capacity of deep network architectures and their excellent learning ability to generalize on exemplar data.

The attained acceleration rates are closely tied to the achievable quality of reconstructed images. Conventional methods utilize compressed sensing theory to recover the under-sampled signals, which can be difficult to provide high-quality reconstruction due to the limited representation capacity of the sparsity priors assumed by compressed sensing methods. In Chapter 3, we investigated the potential of deep neural networks to achieve high-fidelity reconstruction results at aggressive acceleration rates. An attentive selection generative adversarial network termed ASGAN was introduced to produce images from the under-sampled measurements with fine-grained details by incorporating large-field contextual information integration and attention selection into the reconstruction model. However, as a direct inversion model, it lacks domain-specific knowledge normally expressed in the form of the k -space data condition, which can lead to inconsistent image features and affect the reconstruction quality.

To make use of domain-specific knowledge during the image recovery steps, a novel deep cascaded framework was presented in Chapter 3, which is derived by unrolling an optimization algorithm and enforces the k -space data consistency to pursue more faithful results by means of a devised deep data consistency block. The experiments demonstrated that the proposed framework with an adaptive spatial attention selection module, which was designed to efficiently capture the spatial contextual information of extracted features, yields improved performance in terms of image quality and evaluation metrics. To implement attention mechanism to high-resolution features in an effort to improve the model capacity without causing enormous computational overheads, a region-guided channel-wise attention network was introduced to employ the implicitly learned spatial semantics to guide attention mechanism over feature channels as a means to leverage the rich information contained in the spatial regions of deep representations via a light-weight and efficient design. It was shown that the devised attention method can implicitly capture the spatial semantic information of the anatomical structures to enhance the reconstruction performance and outperform other competing models qualitatively and quantitatively.

In Chapter 4, we introduced the method for parallel MRI reconstruction, with the major focus on solving the unknown forward process conventionally required by reconstruction algorithms and expressed in the form of coil sensitivity maps pre-calculated via model-based methods or jointly estimated by neural networks. A coil-agnostic reconstruction framework was proposed to perform multi-coil MRI reconstruction by leveraging a novel data aggregation consistency block to learn a coil-invariant representation end-to-end and approximately enforce the k -space data consistency without the involvement of coil sensitivities. To improve the model performance, a locality-aware spatial attention module was embedded into the reconstruction network, which endows the model with the capacity to retrieve contextual cues from the extracted features without tremendously increasing the computational costs. It was demonstrated that the proposed coil-agnostic method yields improved reconstruction performance as compared to other competitive approaches relying on the estimation of sensitivity maps, and is robust to different machine configurations, which can potentially expedite the deployment of reconstruction models from training to real-world applications.

In Chapter 5, we introduced our research work focusing on dynamic MRI reconstruction which is often a more challenging task compared to static 2D MRI reconstruction, with the aim of image sequence recovery to detect the motion and dynamics of organs. To provide high-fidelity reconstruction presenting both the spatial features of the parts of human body and time-varying content, a deep cascaded framework built up based on a regularized optimization was proposed to restore dynamic frames from a collection of under-sampled measurements, coming with the ability to exploit the redundant and complementary information from the sequence of temporal frames with the help of a novel mask-guided motion feature incorporation method designed to explicitly incorporate the extracted hints of the dynamic content into the reconstruction procedure. To efficiently achieve information exchange aiming to improve the reconstruction quality, we devised a spatio-temporal Fourier neural block to retrieve useful information from the global field in both spatial and time domains by extending the Fourier neural method to dynamic scenes. It was demonstrated through comprehensive experiments that the proposed framework yields superior performance over state-of-the-art methods for dynamic MRI recovery in terms of image quality and preservation of the organ dynamics.

6.2 Future Works

In this section, we discuss some problems for deep neural network models in compressive MRI reconstruction and provide some suggestions for the future research directions.

6.2.1 MRI Reconstruction Using Diffusion Models

In the past several years, diffusion models [190, 191, 192, 193] have received significant popularity, coming with the outstanding performance in synthesizing photo-realistic and diverse images and showing more stable training as compared to generative adversarial networks which learn the distribution of real data by running an adversarial game with a discriminator. As one typical class of diffusion models, de-noising diffusion probabilistic models reverse a forward diffusion process entailing a series of steps where an image taken from a target distribution

is progressively corrupted by adding small amounts of noise to approach a pre-determined prior, often a Gaussian distribution since it is easy to take samples from. A diffusion model is trained with a large dataset of samples to predict the noise contained in the current estimate and generate a less noisy version following the posterior distribution of the corruption step parametrized with the predicted noise component. The generation process is initialized with a random Gaussian noise image that is passed through the iterative reverse steps to repeatedly remove small portions of noise and synthesize image features ranging from global structures to local details until a reasonable clean result is achieved.

Diffusion models generate images by reversing the corruption process after training a network with exemplar data, which normally needs massive datasets and leads to a resource demanding model. Although it ensures the high-quality and diversity of images synthesized through the reverse process, the de-noising steps applied to remove portions of noise from the intermediate result and explore the search space to direct the random sample towards a mode of the target distribution are iteratively executed to obtain a clean output, often needing several hundred to a few thousand steps to produce images from Gaussian noise, which is inherently slow to sample from and results in a time-consuming generation procedure.

Diffusion models often depend on a large-scale database to obtain high performance in image generation. However, compared to the creation of datasets of natural images which can be comparatively cheap and easy to gather and accumulate, the collection of MR images is expensive and time-consuming, particularly in dynamic imaging scenes, making it prohibitive to train a high-performance model on tremendous amounts of data. Generating images using a trained diffusion model relies on the reverse diffusion process which comprises a series of de-noising steps, leading to a slow sampling procedure and potentially running against the speed demand of MRI reconstruction in some applications where real-time reconstruction is needed to provide the instant information of the anatomy being scanned.

In the machine learning literature, diffusion models have demonstrated a variety of outstanding and impressive applications. Diffusion-based generative frameworks have shown superior performance over the current generative models including generative adversarial networks in terms of image quality and sample diversity [191]. They also made great progress towards a milestone of generating high-fidelity and temporally coherent videos via an extension of the image diffusion network [194]. Latent diffusion models [192] were proposed to enable training with limited computational resources while maintaining the quality of generated images, which expedites the development of diffusion-based image synthesis frameworks [195]. To enable flexible manipulation on diffusion models, DragonDiffusion [196] was developed to perform various editing modes for both synthesized and real images including object moving, resizing, appearance replacement, and content dragging, by utilizing a multiscale design to modify the intermediate representation of the diffusion network. An image editing framework, termed DragDiffusion [197], was introduced to achieve point-based and controllable image editing by leveraging pre-trained diffusion models, which significantly improves the practicality of interactive editing in real-world application. The work introduced in [198] demonstrates that a diffusion-based network can yield improved image super-resolution performance and produce sharp results with plausible details through a sequence of iterative refinement steps. In recent works, diffusion models were employed in data augmentation to address the lack of diversity problem of simple transformations such as flip and rotation and exhibited excellent capacity to modify the high-level image attributes by optimizing the inversion of text-to-image process and generating augmented samples with swapped prompts, which shows to enhance the diversity of data by utilizing a pre-trained stable diffusion model [199, 200]. Those works are indicative of the impressive ability and unprecedented generative performance of diffusion models, compared to other state-of-the-art methods.

With the impressive capacity of diffusion models in image generation, it implies the potential of leveraging diffusion models to improve the quality of MRI reconstruction, by exploiting their coarse-to-fine generation process where abstract image features such as global structures are firstly recovered prior to the modelling of low-level characters such as local textures, therefore yielding significantly improved image quality [191, 193]. However, there

are some limitations of using diffusion models for MRI reconstruction problems. Firstly, the incorporation of the k -space constraint into generation process can impose a challenge when dealing with images with noise progressively removed to achieve a clean result, since the denoising process can have an unfavorable impact on data consistency enforcement and thus restrict the achievable acceleration rate [201]. Secondly, the slow sampling process hinders its implementation in many time-critic applications and struggles to meet the speed demand of modern MRI. Thirdly, it normally requires a large-scale dataset to train a high-performance diffusion model, which can be challenging in medical imaging domain due to the expenses of collecting medical images. Those factors motivate us to explore more effective network architectures and training strategies specifically re-designed for the inverse problems, when harnessing the flexibility and high performance of diffusion models in image generation to the field of image reconstruction in fast MRI scenarios.

6.2.2 Model Learning with Limited Datasets

Although deep learning-based methods have demonstrated the superiority performance over traditional approaches in image recovery by exploiting the powerful representations learned from large datasets where a network model is trained to generalize across exemplar data, it is crucial to collect many samples for model learning to achieve success in inference. More high-quality training samples potentially enable us to develop a higher performance model in the subsequent inference task, which implies more expense on data collection and incredibly increases training costs and time.

Meanwhile, it is normally more expensive to acquire MR images and create large-scale databases for training a network with improved capacity, compared to the collection of natural images. Therefore, how to achieve high-quality reconstruction results with a neural network model established using a limited amount of data forms a challenging, intriguing, and prom-

ising research direction, which can potentially benefit from pre-training with large databases to acquire the ability of feature extraction in general and transferring the learned strong representations to target domain via fine-tuning [202, 203], and learning a reconstruction model from sparse observations using self-supervision strategies [204, 205, 206, 207].

6.2.3 Reconstruction with Cartesian and Non-Cartesian Sampling Strategies

Cartesian under-sampling patterns are often used in accelerated MRI data acquisition procedures due to the simple and tractable design of the encoding process and therefore adopted by many of reconstruction methods in the recovery of clean images. Non-Cartesian data encoding strategies such as radial sampling schemes can also be used as an alternative to Cartesian sampling methods, which have proven effective in capturing informative image components featuring the structures and details of the underlying anatomy. However, adopting non-Cartesian sampling patterns in data acquisition requires the use of the non-uniform fast Fourier transform and gridding [162, 163] to perform mappings between k -space and image domain over the course of reconstruction, which can propagate interpolation errors and have an unfavorable impact on the final reconstruction results [106]. To increase the feasibility and flexibility of devised reconstruction methods and allow for more effective and efficient sampling strategies attempting to improve the quality of recovered images and enable data acquisition schemes under more aggressive acceleration rates, it can be an intriguing direction to explore the potential of reconstruction models which are proposed or re-designed to characteristically take into consideration different types of Cartesian and non-Cartesian under-sampling patterns.

Bibliography

- [1] J. Liu, C. Qin, M. Yaghoobi, High-Fidelity MRI Reconstruction Using Adaptive Spatial Attention Selection and Deep Data Consistency Prior, *IEEE Transactions on Computational Imaging* 9 (2023) 298–313.
- [2] J. Liu, C. Qin, M. Yaghoobi, High-fidelity MRI reconstruction with the densely connected network cascade and feature residual data consistency priors, *Machine Learning for Medical Image Reconstruction* (2022) 34–43.
- [3] J. Liu, C. Qin, M. Yaghoobi, Region-guided channel-wise attention network for accelerated MRI reconstruction, *Machine Learning in Medical Imaging* (2022) 21–31.
- [4] J. Liu, M. Yaghoobi, Fine-grained MRI reconstruction using attentive selection generative adversarial networks, *ICASSP* (2021).
- [5] J. Liu, C. Qin, M. Yaghoobi, Coil-agnostic attention-based network for parallel MRI reconstruction, *ACCV* (2022) 2866–2883.
- [6] J. Liu, C. Qin, M. Yaghoobi, Mask-Guided Deep Network Cascade for Dynamic Cardiac MRI Reconstruction with the Fourier Neural Attention, *Submitted* (2023).
- [7] V. Dumoulin, F. Visin, A guide to convolution arithmetic for deep learning, *arXiv e-prints* (2016) arXiv:1603.07285arXiv:1603.07285.
- [8] A. Krizhevsky, I. Sutskever, G. Hinton, ImageNet classification with deep convolutional neural networks, *Advances in Neural Information Processing Systems* 25 (2012).
- [9] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings* (2015).

- [10] O. Ronneberger, P. Fischer, T. Brox, U-Net: convolutional networks for biomedical image segmentation, *Medical Image Computing and Computer-Assisted Intervention (MICCAI)* 9351 (2015) 234–241.
- [11] S. Hochreiter, J. Schmidhuber, Long short-term memory, *Neural Computation* 9 (8) (1997) 1735–1780.
- [12] J. Chung, C. Gulcehre, K. Cho, Y. Bengio, Empirical evaluation of gated recurrent neural networks on sequence modeling, *NIPS 2014 Workshop on Deep Learning*, December 2014 (2014).
- [13] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using RNN encoder–decoder for statistical machine translation, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2014) 1724–1734.
- [14] X. SHI, Z. Chen, H. Wang, D. Yeung, W. Wong, W. WOO, Convolutional LSTM network: a machine learning approach for precipitation nowcasting, *Advances in Neural Information Processing Systems* 28 (2015).
- [15] N. Ballas, L. Yao, C. Pal, A. Courville, Delving deeper into convolutional networks for learning video representations, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings* (2016).
- [16] I. Goodfellow, A. Pouget, M. Mirza, B. Xu, F. Warde, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, *Advances in Neural Information Processing Systems* 27 (2014) 2672–2680.
- [17] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, A. Courville, Improved training of Wasserstein GANs, *Advances in Neural Information Processing Systems* 30 (2017).
- [18] T. Miyato, T. Kataoka, M. Koyama, Y. Yoshida, Spectral normalization for generative adversarial networks, *International Conference on Learning Representations (ICLR)* (2018).

- [19] X. Mao, Q. Li, H. Xie, R. Lau, Z. Wang, S. Smolley, Least squares generative adversarial networks, 2017 IEEE International Conference on Computer Vision (ICCV) (2017) 2813–2821.
- [20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, *Advances in Neural Information Processing Systems* 30 (2017).
- [21] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-End Object Detection with Transformers, *Computer Vision – ECCV 2020* (2020) 213–229.
- [22] P. Esser, R. Rombach, B. Ommer, Taming Transformers for High-Resolution Image Synthesis, 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020) 12868–12878.
- [23] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: transformers for image recognition at scale, *International Conference on Learning Representations* (2021).
- [24] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: hierarchical vision transformer using shifted windows, *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (2021).
- [25] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, F. Wei, B. Guo, Swin transformer V2: scaling up capacity and resolution, *International Conference on Computer Vision and Pattern Recognition (CVPR)* (2022).
- [26] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, R. Timofte, SwinIR: image restoration using Swin transformer, *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops* (2021) 1833–1844.
- [27] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, H. Hu, Video Swin transformer, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022) 3202–3211.

- [28] Z. Chu, X. and Tian, Y. Wang, B. Zhang, H. Ren, X. Wei, H. Xia, C. Shen, Twins: revisiting the design of spatial attention in vision transformers, *Advances in Neural Information Processing Systems* 34 (2021) 9355–9366.
- [29] N. Kitaev, L. Kaiser, A. Levskaya, Reformer: the efficient transformer, *International Conference on Learning Representations* (2020).
- [30] Y. Tay, D. Bahri, L. Yang, D. Metzler, D. Juan, Sparse Sinkhorn attention, *ICML 2020* (2020).
- [31] M. Zaheer, G. Guruganesh, K. Dubey, J. Ainslie, C. Alberti, S. Ontanon, P. Pham, A. Ravula, Q. Wang, L. Yang, Big bird: transformers for longer sequences, *Advances in Neural Information Processing Systems* 33 (2020).
- [32] H. Peng, N. Pappas, D. Yogatama, R. Schwartz, N. Smith, L. Kong, Random feature attention, *International Conference on Learning Representations* (2021).
- [33] C. Zhu, W. Ping, C. Xiao, M. Shoenybi, T. Goldstein, A. Anandkumar, B. Catanzaro, Long-short transformer: efficient transformers for language and vision, *NeurIPS* (2021).
- [34] D. D.L., Compressed sensing, *IEEE Transactions on Information Theory* 52 (4) (2006) 1289–1306.
- [35] M. Lustig, D. Donoho, J. Pauly, Sparse MRI: the application of compressed sensing for rapid MR imaging, *Magn Reson Med.* 58 (6) (2007) 1182–1195.
- [36] M. Lustig, D. Donoho, J. Santos, J. Pauly, Compressed sensing MRI, *IEEE Signal Processing Magazine* 25 (2) (2008) 72–82.
- [37] J. Ye, S. Tak, Y. Han, H. Park, Projection reconstruction MR imaging using FOCUSS, *Magnetic Resonance in Medicine* 57 (4) (2007) 764–775.
- [38] S. Ramani, J. Fessler, Parallel MR image reconstruction using augmented lagrangian methods, *IEEE Transactions on Medical Imaging* 30 (3) (2011) 694–706.

- [39] S. Ma, W. Yin, Y. Zhang, A. Chakraborty, An efficient algorithm for compressed MR imaging using total variation and wavelets, 2008 IEEE Conference on Computer Vision and Pattern Recognition (2008) 1–8.
- [40] L. Sun, Y. Huang, C. Cai, X. Ding, Compressed sensing MRI using total variation regularization with K-space decomposition, 2017 IEEE International Conference on Image Processing (ICIP) (2017) 3061–3065.
- [41] M. Ehrhardt, M. Betcke, Multicontrast MRI reconstruction with structure-guided total variation, SIAM Journal on Imaging Sciences 9 (3) (2016) 1084–1106.
- [42] M. Panić, D. Jakovetić, D. Vukobratović, V. Crnojević, A. Pižurica, MRI reconstruction using Markov random field and total variation as composite prior, Sensors 20 (11) (2020).
- [43] L. Song, J. Zhang, Q. Wang, MRI reconstruction based on three regularizations: Total variation and two wavelets, Biomedical Signal Processing and Control 30 (2016) 64–69.
- [44] J. Ye, Compressed sensing MRI: a review from signal processing perspective, BMC biomed eng 1 (8 2019).
- [45] S. Ravishankar, Y. Bresler, MR image reconstruction from highly undersampled k-space data by dictionary learning, IEEE Transactions on Medical Imaging 30 (5) (2011) 1028–1041.
- [46] M. Aharon, M. Elad, A. Bruckstein, K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation, IEEE Transactions on Signal Processing 54 (11) (2006) 4311–4322.
- [47] S. Ravishankar, Y. Bresler, Efficient blind compressed sensing using sparsifying transforms with convergence guarantees and application to magnetic resonance imaging, SIAM Journal on Imaging Sciences 8 (4) (2015) 2519–2557.
- [48] H. Jung, J. Ye, E. Kim, Improved k-t BLAST and k-t SENSE using FOCUSS, Phys. Med. Biol. 52 (2007) 3201–3026.

- [49] H. Jung, K. Sung, K. Nayak, E. Kim, J. Ye, k-t FOCUSS: a general compressed sensing framework for high resolution dynamic MRI, *Magn Reson Med* 61(1) (2009) 103–16.
- [50] C. Chen, Y. Li, L. Axel, J. Huang, Real time dynamic MRI with dynamic total variation, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2014* (2014) 138–145.
- [51] C. Chen, Y. Li, L. Axel, J. Huang, Real time dynamic MRI by exploiting spatial and temporal sparsity, *Magnetic Resonance Imaging* 34 (4) (2016) 473–482.
- [52] Y. Saad, *Iterative methods for sparse linear systems*, SIAM (2003).
- [53] N. He, R. Wang, Y. Wang, Dynamic MRI reconstruction exploiting blind compressed sensing combined transform learning regularization, *Neurocomputing* 392 (2020) 160–167.
- [54] S. Lingala, M. Jacob, Blind compressive sensing dynamic MRI, *IEEE Transactions on Medical Imaging* 32 (6) (2013) 1132–1145.
- [55] E. Ahishakiye, M. B. Van Gijzen, J. Tumwiine, R. Wario, J. Obungoloch, A survey on deep learning in medical image reconstruction, *Intelligent Medicine* 1 (3) (2021) 118–127.
- [56] J. Zbontar, F. Knoll, A. Sriram, M. Muckley, M. Bruno, A. Defazio, M. Parente, K. Geras, J. Katsnelson, H. Chandarana, Z. Zhang, M. Drozdal, A. Romero, M. Rabbat, P. Vincent, J. Pinkerton, D. Wang, N. Yakubova, E. Owens, C. Zitnick, M. Recht, D. Sodickson, Y. Lui, FastMRI: an open dataset and benchmarks for accelerated MRI, *CoRR* abs/1811.08839 (2018). [arXiv:1811.08839](https://arxiv.org/abs/1811.08839).
- [57] G. Yang, S. Yu, H. Dong, G. Slabaugh, P. L. Dragotti, X. Ye, F. Liu, S. Arridge, J. Keegan, Y. Guo, D. Firmin, DAGAN: deep de-aliasing generative adversarial networks for fast compressed sensing MRI reconstruction, *IEEE Transactions on Medical Imaging* 37 (6) (2018) 1310–1321.
- [58] H. Zhao, O. Gallo, I. Frosio, J. Kautz, Loss functions for image restoration with neural networks, *IEEE Transactions on Computational Imaging* 3 (1) (2017) 47–57.

- [59] P. Deora, B. Vasudeva, S. Bhattacharya, M. P.P., Structure preserving compressive sensing MRI reconstruction using generative adversarial networks, The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (June 2020).
- [60] K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, C. Loy, Y. Qiao, X. Tang, ESRGAN: enhanced super-resolution generative adversarial networks, The European Conference on Computer Vision Workshops (ECCVW) (September 2018).
- [61] K. Tezcan, C. Baumgartner, R. Luechinger, K. Pruessmann, E. Konukoglu, MR image reconstruction using deep density priors, *IEEE Transactions on Medical Imaging* 38 (7) (2019) 1633–1642. doi:10.1109/TMI.2018.2887072.
- [62] D. Kingma, M. Welling, Auto-encoding variational bayes, 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings (2014).
- [63] D. Rezende, S. Mohamed, D. Wierstra, Stochastic backpropagation and approximate inference in deep generative models, *Proceedings of the 31st International Conference on Machine Learning* 32 (2) (2014) 1278–1286.
- [64] H. Aggarwal, M. Mani, M. Jacob, MoDL: model-based deep learning architecture for inverse problems, *IEEE Transactions on Medical Imaging* 38 (2) (2019) 394–405. doi:10.1109/TMI.2018.2865356.
- [65] J. Cheng, Z. Cui, W. Huang, Z. Ke, L. Ying, H. Wang, Y. Zhu, D. Liang, Learning data consistency and its application to dynamic MR imaging, *IEEE Transactions on Medical Imaging* 40 (11) (2021) 3140–3153.
- [66] Q. Huang, D. Yang, P. Wu, H. Qu, J. Yi, D. Metaxas, MRI reconstruction via cascaded channel-wise attention network, 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019) (2019) 1622–1626doi:10.1109/ISBI.2019.8759423.

- [67] C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, J. Su, This looks like that: deep learning for interpretable image recognition, *Advances in Neural Information Processing Systems* 32 (2019).
- [68] A. Brock, S. De, S. Smith, K. Simonyan, High-performance large-scale image recognition without normalization, *Proceedings of the 38th International Conference on Machine Learning* 139 (2021) 1059–1071.
- [69] M. Hesamian, W. Jia, X. He, P. Kennedy, Deep Learning Techniques for Medical Image Segmentation, *Journal of Digital Imaging* 32 (2019) 582–596.
- [70] D. Sarvamangala, R. Kulkarni, Convolutional neural networks in medical image understanding: a survey, *Evolutionary Intelligence* 15 (2022) 1 – 22.
- [71] M. Yaqub, F. Jinchao, K. Arshid, S. Ahmed, W. Zhang, Nawaz, M.Z., T. Mahmood, Deep learning-based image reconstruction for different medical imaging modalities, *Computational and mathematical methods in medicine* (Jun 2022).
- [72] S. Roth, M. Black, Fields of experts, *International Journal of Computer Vision* 82 (2) (2009) 205–229.
- [73] k. Hammernik, T. Klatzer, E. Kobler, M. Recht, D. Sodickson, T. Pock, F. Knoll, Learning a variational network for reconstruction of accelerated MRI data, *Magnetic Resonance in Medicine* 79 (6) (2017) 3055–3071.
- [74] M. Uecker, P. Lai, M. Murphy, P. Virtue, M. Elad, J. Pauly, S. Vasanawala, M. Lustig, ESPIRiT—an eigenvalue approach to autocalibrating parallel MRI: where SENSE meets GRAPPA, *Magnetic Resonance in Medicine* 71 (3) (2014) 990–1001.
- [75] K. Zhang, W. Zuo, S. Gu, L. Zhang, Learning deep CNN denoiser prior for image restoration, *IEEE Conference on Computer Vision and Pattern Recognition* (2017) 3929–3938.

- [76] J. Rick Chang, C. Li, B. Póczos, B. Vijaya Kumar, A. Sankaranarayanan, One network to solve them all – solving linear inverse problems using deep projection models, *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (Oct 2017).
- [77] J. Schlemper, J. Caballero, J. Hajnal, A. Price, D. Rueckert, A deep cascade of convolutional neural networks for MR image reconstruction, *Information Processing in Medical Imaging* (2017) 647–658.
- [78] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2016).
- [79] M. Mardani, E. Gong, J. Y. Cheng, S. S. Vasanawala, G. Zaharchuk, L. Xing, J. M. Pauly, Deep generative adversarial neural networks for compressive sensing MRI, *IEEE Transactions on Medical Imaging* 38 (1) (2019) 167–179. doi:10.1109/TMI.2018.2858752.
- [80] X. Qiao, J. Du, L. Wang, Z. He, Y. Jia, A model-based deep network for MRI reconstruction using approximate message passing algorithm, *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2020) 1105–1109.
- [81] A. Lahiri, G. Wang, S. Ravishankar, J. Fessler, Blind primed supervised (BLIPS) learning for MR image reconstruction, *IEEE Transactions on Medical Imaging* 40 (11) (2021) 3113–3124.
- [82] J. Duan, J. Schlemper, C. Qin, C. Ouyang, W. Bai, C. Biffi, G. Bello, B. Statton, D. Regan, D. Rueckert, VS-Net: variable splitting network for accelerated parallel MRI reconstruction, *Springer International Publishing* (2019) 713–722.
- [83] Y. Yang, J. Sun, H. Li, Z. Xu, Deep ADMM-Net for compressive sensing MRI, *Advances in Neural Information Processing Systems* 29 (2016).

- [84] R. Ahmad, C. Bouman, T. Buzzard, S. Chan, S. Liu, E. Reehorst, P. Schniter, Plug-and-play methods for magnetic resonance imaging: using denoisers for image recovery, *IEEE Signal Processing Magazine* 37 (1) (2020) 105–116.
- [85] R. Liu, Y. Zhang, S. Cheng, Z. Luo, X. Fan, A deep framework assembling principled modules for CS-MRI: unrolling perspective, convergence behaviors, and practical modeling, *IEEE Transactions on Medical Imaging* 39 (12) (2020) 4150–4163.
- [86] D. Narnhofer, K. Hammernik, F. Knoll, T. Pock, Inverse GANs for accelerated MRI reconstruction, *Wavelets and Sparsity XVIII* 11138 (2019) 111381A.
- [87] M. Griswold, P. Jakob, R. M. Heidemann, M. Nittka, V. Jellus, J. Wang, B. Kiefer, A. Haase, Generalized autocalibrating partially parallel acquisitions (GRAPPA), *Magnetic resonance in medicine* 47 (6) (2002) 1202–1210.
- [88] M. Akcakaya, S. Moeller, S. Weingartner, K. Ugurbil, Scan-specific robust artificial-neural-networks for k-space interpolation (RAKI) reconstruction: Database-free deep learning for fast imaging, *Magnetic Resonance in Medicine* 81 (1) (2019) 439–453.
- [89] C. Zhang, S. Hosseini, S. Weingartner, K. Ugurbil, S. Moeller, M. Akcakay, Optimized fast GPU implementation of robust artificial-neural-networks for k-space interpolation (RAKI) reconstruction, *PLoS ONE* 14 (10) (2019).
- [90] C. Zhang, S. Moeller, O. Demirel, K. Ugurbil, M. Akcakaya, Residual RAKI: a hybrid linear and non-linear approach for scan-specific k-space deep learning, *NeuroImage* 256 (2022) 119248.
- [91] S. Anuroop, Z. Jure, M. Tullie, Z. Lawrence, D. Aaron, D. K.S., GrappaNet: combining parallel imaging with deep learning for multi-coil MRI reconstruction, 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020) 14303–14310.
- [92] B. Zhu, J. Liu, S. Cauley, B. Rosen, M. Rosen, Image reconstruction by domain-transform manifold learning, *Nature* 555 (7697) (2018) 487–492.
- [93] Y. Han, L. Sunwoo, J. Ye, K-space deep learning for accelerated MRI, *IEEE Transactions on Medical Imaging* 39 (2) (2020) 377–386.

- [94] L. Ying, J. Sheng, Joint image reconstruction and sensitivity estimation in SENSE (JSENSE), *Magnetic Resonance in Medicine* 57(6) (2007) 1196–1202.
- [95] A. Sriram, J. Zbontar, T. Murrell, A. Defazio, C. Zitnick, N. Yakubova, F. Knoll, P. Johnson, End-to-end variational networks for accelerated mri reconstruction, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020* (2020) 64–73.
- [96] D. Liang, J. Cheng, Z. Ke, L. Ying, Deep magnetic resonance image reconstruction: inverse problems meet neural networks, *IEEE Signal Processing Magazine* 37 (1) (2020) 141–151.
- [97] F. Knoll, K. Hammernik, C. Zhang, S. Moeller, T. Pock, D. K. Sodickson, M. Akcakaya, Deep-learning methods for parallel magnetic resonance imaging reconstruction: a survey of the current approaches, *IEEE Signal Processing Magazine* 37 (1) (2020) 1281140.
- [98] W. Gan, Y. Hu, C. Eldeniz, J. Liu, Y. Chen, H. An, U. Kamilov, SS-JIRCS: self-supervised joint image reconstruction and coil sensitivity calibration in parallel MRI without ground truth, *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)* (2021) 4031–4039.
- [99] J. Lee, J. Kang, S. Oh, D. Ye, Multi-domain Neumann network with sensitivity maps for parallel MRI reconstruction, *Sensors* 22 (10) (2022).
- [100] Y. Jun, H. Shin, T. Eo, D. Hwang, Joint deep model-based MR image and coil sensitivity reconstruction network (Joint-ICNet) for fast MRI, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021) 5270–5279.
- [101] T. Eo, Y. Jun, T. Kim, J. Jang, H. Lee, D. Hwang, KIKI-net: cross-domain convolutional neural networks for reconstructing undersampled magnetic resonance images, *Magnetic Resonance in Medicine* 80 (5) (2018) 2188–2201.
- [102] M. Arvinte, S. Vishwanath, A. Tewfik, J. Tamir, Deep J-Sense: accelerated MRI reconstruction via unrolled alternating optimization, *ISMRM Annual Conference* (2021).

- [103] C. Qin, D. Rueckert, Artificial Intelligence-Based Image Reconstruction in Cardiac Magnetic Resonance, *Artificial Intelligence in Cardiothoracic Imaging* (2022).
- [104] J. Zhang, B. Ghanem, ISTA-Net: iterative shrinkage-thresholding algorithm inspired deep network for image compressive sensing, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2018) 1828–1837.
- [105] N. Pezzotti, S. Yousefi, M. Elmahdy, J. van Gemert, C. Schülke, M. Doneva, T. Nielsen, S. Kastrýulin, B. F. Lelieveldt, M. P. van Osch, E. de Weerd, M. Staring, An Adaptive Intelligence Algorithm for Undersampled Knee MRI Reconstruction, *arXiv e-prints* (2020) arXiv:2004.07339arXiv:2004.07339.
- [106] J. Schlemper, J. Caballero, J. Hajnal, A. Price, D. Rueckert, A deep cascade of convolutional neural networks for dynamic MR image reconstruction, *IEEE Transactions on Medical Imaging* 37 (2) (2018) 491–503.
- [107] C. Sandino, P. Lai, S. Vasanawala, J. Cheng, Accelerating cardiac cine MRI using a deep learning-based ESPIRiT reconstruction, *Magnetic Resonance in Medicine* 85 (1) (2021) 152–167.
- [108] S. Wang, Z. Ke, H. Cheng, S. Jia, L. Ying, H. Zheng, D. Liang, Dimension: Dynamic mr imaging with both k-space and spatial prior knowledge obtained via multi-supervised network training, *NMR in Biomedicine* 35 (4) (2022) e4131.
- [109] D. Kim, H. Cho, H. Shin, S. Lim, W. Hwang, An efficient three-dimensional convolutional neural network for inferring physical interaction force from video, *Sensors* 19 (16) (2019).
- [110] X. Yang, X. Zhang, Y. Ye, R. Lau, S. Lu, X. Li, X. Huang, Synergistic 2D/3D convolutional neural network for hyperspectral image classification, *Remote Sensing* 12 (12) (2020).
- [111] G. Landi, E. Piccolomini, F. Zama, A total variation-based reconstruction method for dynamic MRI, *Computational and Mathematical Methods in Medicine* 9 (2008) 1–8.

- [112] C. Chen, Y. Li, L. Axel, J. Huang, Real time dynamic MRI with dynamic total variation, *Medical image computing and computer-assisted intervention (MICCAI)* 17 (2014) 138–145.
- [113] F. Knoll, K. Bredies, T. Pock, R. Stollberger, Second order total generalized variation (TGV) for MRI, *Magnetic Resonance in Medicine* 65(2) (2010) 480–491.
- [114] D. Wang, D. Smith, X. Yang, Dynamic MR image reconstruction based on total generalized variation and low-rank decomposition, *Magn Reson Med* 83(6) (2020) 2064–2076.
- [115] J. Zhang, L. Han, J. Sun, Z. Wang, W. Xu, Y. Chu, L. Xia, M. Jiang, Compressed sensing based dynamic MR image reconstruction by using 3D-total generalized variation and tensor decomposition: k-t TGV-TD, *BMC Med Imaging* 22 (2022) 101.
- [116] S. Lingala, Y. Hu, E. DiBella, M. Jacob, Accelerated dynamic MRI exploiting sparsity and low-rank structure: k-t SLR, *IEEE Transactions on Medical Imaging* 30(5) (2011) 1042–54.
- [117] R. Otazo, E. Candes, D. Sodickson, Low-rank plus sparse matrix decomposition for accelerated dynamic MRI with separation of background and dynamic components, *Magn Reson Med* 73(3) (2015) 1125–36.
- [118] W. Huang, Z. Ke, Z. Cui, J. Cheng, Z. Qiu, S. Jia, L. Ying, Y. Zhu, D. Liang, Deep low-rank plus sparse network for dynamic MR imaging, *Medical Image Analysis* 73 (2021) 102190.
- [119] Z. Ke, W. Huang, Z. Cui, J. Cheng, S. Jia, H. Wang, X. Liu, H. Zheng, L. Ying, Y. Zhu, D. Liang, Learned low-rank priors in dynamic MR imaging, *IEEE Transactions on Medical Imaging* 40 (12) (2021) 3698–3710.
- [120] C. Qin, J. Schlemper, J. Caballero, A. Price, J. Hajnal, D. Rueckert, Convolutional recurrent neural networks for dynamic MR image reconstruction, *IEEE Transactions on Medical Imaging* 38 (1) (2019) 280–290.

- [121] C. Qin, J. Schlemper, J. Duan, G. Seegoolam, A. Price, J. Hajnal, D. Rueckert, k-t NEXT: dynamic MR image reconstruction exploiting spatio-temporal correlations, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019* (2019) 505–513.
- [122] C. Qin, J. Duan, K. Hammernik, J. Schlemper, T. Küstner, R. Botnar, C. Prieto, A. Price, J. Hajnal, D. Rueckert, Complementary time-frequency domain networks for dynamic parallel MR image reconstruction, *Magnetic Resonance in Medicine* (Jul. 2021). doi: 10.1002/mrm.28917.
- [123] Q. Huang, D. Yang, H. Qu, J. Yi, P. Wu, D. Metaxas, Dynamic MRI reconstruction with motion-guided network, *Proceedings of The 2nd International Conference on Medical Imaging with Deep Learning* 102 (2019) 275–284.
- [124] T. Isobe, X. Jia, X. Tao, C. Li, R. Li, Y. Shi, J. Mu, H. Lu, Y. Tai, Look back and forth: video super-resolution with explicit temporal difference modeling, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022) 17411–17420.
- [125] M. Suin, A. Rajagopalan, Gated spatio-temporal attention-guided video deblurring, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021) 7802–7811.
- [126] A. Tourapis, O. Au, M. Liou, Predictive motion vector field adaptive search technique (PMVFAST) - enhancing block based motion estimation, *Proceedings of SPIE - The International Society for Optical Engineering* (06 2001).
- [127] H. Jozawa, K. Kamikura, A. Sagata, H. Kotera, H. Watanabe, Two-stage motion compensation using adaptive global MC and local affine MC, *IEEE Transactions on Circuits and Systems for Video Technology* 7 (1) (1997) 75–85.
- [128] Y. Lin, Q. Zhuang, R. Yang, Image reconstruction of dynamic MRI based on adaptive motion estimation, *2012 12th International Conference on Control Automation Robotics and Vision (ICARCV)* (2012) 1586–1590.

- [129] N. Zhao, D. O'Connor, D. Ruan, A. Basarab, K. Sheng, Coupling reconstruction and motion estimation for dynamic MRI through optical flow constraint, SPIE Medical Imaging, Houston, Texas United States, 10/02/18-15/02/18 10574 (2018) (electronic medium).
- [130] N. Zhao, D. O'Connor, A. Basarab, D. Ruan, K. Sheng, Motion compensated dynamic MRI reconstruction with local affine optical flow estimation, IEEE Transactions on Biomedical Engineering 66 (11) (2019) 3050–3059.
- [131] K. Hammernik, J. Pan, D. Rueckert, T. Kustner, Motion-guided physics-based learning for cardiac MRI reconstruction, 2021 55th Asilomar Conference on Signals, Systems, and Computers (2021) 900–907.
- [132] M. Jaderberg, K. Simonyan, A. Zisserman, et al, Spatial transformer networks, Advances in neural information processing systems 28 (2015).
- [133] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Smagt, D. Cremers, T. Brox, FlowNet: learning optical flow With convolutional networks, Proceedings of the IEEE International Conference on Computer Vision (ICCV) (December 2015).
- [134] J. Yu, A. Harley, K. Derpanis, Back to basics: unsupervised learning of optical flow via brightness constancy and motion smoothness, Computer Vision - ECCV 2016 Workshops (2016).
- [135] Q. Huang, Y. Xian, D. Yang, H. Qu, J. Yi, P. Wu, D. Metaxas, Dynamic MRI reconstruction with end-to-end motion-guided network, Medical Image Analysis 68 (2021).
- [136] X. Zhu, H. Hu, S. Lin, J. Dai, Deformable ConvNets V2: more deformable, better results, 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 9300–9308.
- [137] D. P., D. U.K., S. Y.C., Blur more to deblur better: multi-blur2deblur for efficient video deblurring, ArXiv abs/2012.12507 (2020).

- [138] M. Hong, Y. Yu, H. Wang, F. Liu, S. Crozier, Compressed sensing MRI with singular value decomposition-based sparsity basis, *Physics in Medicine and Biology* 56 (19) (2021) 6311–6325.
- [139] S. Lingala, M. Jacob, Blind compressive sensing dynamic MRI, *IEEE Transactions on Medical Imaging* 32 (6) (2013) 1132–1145.
- [140] S. Roohi, D. Zonoobi, A. Kassim, J. Jaremko, Multi-dimensional low rank plus sparse decomposition for reconstruction of under-sampled dynamic MRI, *Pattern Recognition* 63 (2017) 667–679.
- [141] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, W. Shi, Photo-realistic single image super-resolution using a generative adversarial network, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (July 2017).
- [142] D. Lee, J. Yoo, S. Tak, J. Ye, Deep residual learning for accelerated MRI using magnitude and phase networks, *IEEE Trans. Biomed. Eng* 65 (9) (2018) 1985–1995.
- [143] H. Zhang, I. Goodfellow, D. Metaxas, A. Odena, Self-attention generative adversarial networks, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019* 97 (2019) 7354–7363.
- [144] Z. Yuan, M. Jiang, Y. Wang, B. Wei, et al, SARA-GAN: self-attention and relative average discriminator based generative adversarial networks for fast compressed sensing MRI reconstruction, *Frontiers in Neuroinformatics* (2020) 1–12.
- [145] Y. Wu, Y. Ma, J. Liu, J. Du, L. Xing, Self-attention convolutional neural network for improved MR image reconstruction, *Information Sciences* 490 (2019) 317–328.
- [146] G. Li, J. Lv, C. Wang, A modified generative adversarial network using spatial and channel-wise attention for CS-MRI reconstruction, *IEEE Access* 9 (2021) 83185–83198.
- [147] H. Mei, X. Yang, Y. Wang, Y. Liu, S. He, Q. Zhang, X. Wei, R. Lau, Don't hit me! Glass detection in real-world scenes, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2020).

- [148] S. Woo, J. Park, J. Lee, I. Kweon, CBAM: convolutional block attention module, CoRR abs/1807.06521 (2018). [arXiv:1807.06521](#).
- [149] J. Hu, L. Shen, G. Sun, Squeeze-and-Excitation Networks, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018) 7132–7141 [doi:10.1109/CVPR.2018.00745](#).
- [150] H. Tang, D. Xu, Y. Yan, J. Corso, P. Torr, N. Sebe, Multi-channel attention selection GANs for guided image-to-image translation, IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (2022).
- [151] L. Gatys, A. Ecker, M. Bethge, Image style transfer using convolutional neural networks, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016) 2414–2423 [doi:10.1109/CVPR.2016.265](#).
- [152] S. Zhou, J. Zhang, J. Pan, H. Xie, W. Zuo, J. Ren, Spatio-Temporal Filter Adaptive Network for Video Deblurring, Proceedings of the IEEE International Conference on Computer Vision (2019).
- [153] S. Bako, T. Vogels, B. McWilliams, M. Meyer, J. Novák, A. Harvill, P. Sen, T. DeRose, F. Rousselle, Kernel-predicting convolutional networks for denoising Monte Carlo renderings, ACM Transactions on Graphics (TOG) 36 (4) (July 2017).
- [154] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017) 2261–2269.
- [155] T. Quan, D. Nguyen, W. Jeong, Compressed sensing MRI reconstruction using a generative adversarial network with a cyclic loss, IEEE Transactions on Medical Imaging 37 (6) (2018) 1488–1497. [doi:10.1109/TMI.2018.2820120](#).
- [156] B. Sim, G. Oh, S. Lim, C. Jung, J. Ye, Optimal transport driven CycleGAN for unsupervised learning in inverse problems, SIAM Journal on Imaging Sciences 13 (4) (2020) 2281–2306.

-
- [157] J. Zhu, T. Park, P. Isola, A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, 2017 IEEE International Conference on Computer Vision (ICCV) (2017) 2242–2251doi:10.1109/ICCV.2017.244.
- [158] I. Kenji, R. Kuroki, S. Uchida, Explaining convolutional neural networks using Softmax gradient layer-wise relevance propagation, International Conference on Computer Vision Workshop, ICCVW 2019 (2019) 4176–4185.
- [159] E. Shimron, J. Tamir, K. Wang, M. Lustig, Implicit data crimes: machine learning bias arising from misuse of public data, Proceedings of the National Academy of Sciences 119 (13) (2022) e2117203119.
- [160] M. Bińkowski, D. Sutherland, M. Arbel, A. Gretton, Demystifying MMD GANs, International Conference on Learning Representations (2018).
- [161] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L. Chen, MobileNetV2: inverted residuals and linear bottlenecks, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018).
- [162] J. Fessler, B. Sutton, Nonuniform fast Fourier transforms using min-max interpolation, IEEE Trans. Signal Process. 51 (2) (2003) 560–574.
- [163] J. Fessler, On NUFFT-based gridding for non-Cartesian MRI, Journal of magnetic resonance 188 (2) (2007).
- [164] Y. Wang, X. Tao, X. Qi, X. Shen, J. Jia, Image inpainting via generative multi-column convolutional neural networks, Advances in Neural Information Processing Systems (2018) 331–340.
- [165] D. Valsesia, G. Fracastoro, E. Magli, Image denoising with graph-convolutional neural networks, 2019 26th IEEE International Conference on Image Processing (ICIP) (2019).
- [166] D. Valsesia, G. Fracastoro, E. Magli, Deep graph-convolutional image denoising, IEEE Transactions on Image Processing 29 (2020) 8226–8237.

- [167] Y. Guo, C. Wang, H. Zhang, G. Yang, Deep attentive Wasserstein generative adversarial networks for MRI reconstruction with recurrent context-awareness, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020* (2020) 167–177.
- [168] R. Souza, R. Frayne, A hybrid frequency-domain/image-domain deep network for magnetic resonance image reconstruction, 2019 32nd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI) (2019) 257–264.
- [169] Q. Han, Z. Fan, Q. Dai, L. Sun, M. Cheng, J. Liu, J. Wang, On the connection between local attention and dynamic depth-wise convolution, *International Conference on Learning Representations* (2022).
- [170] C. Li, A. Zhou, A. Yao, Omni-dimensional dynamic convolution, *International Conference on Learning Representations* (2022).
- [171] Y. Chen, X. Dai, M. Liu, D. Chen, L. Yuan, Z. Liu, Dynamic convolution: attention over convolution kernels, 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020) 11027–11036.
- [172] D. Kingma, J. Ba, Adam: a method for stochastic optimization, 3rd International Conference on Learning Representations (ICLR) (May 2015).
URL <http://arxiv.org/abs/1412.6980>
- [173] J. Schlemper, J. Duan, C. Ouyang, C. Qin, J. Caballero, J. Hajnal, D. Rueckert, Data consistency networks for (calibration-less) accelerated parallel MR image reconstruction, Abstract from ISMRM 27th Annual Meeting and Exhibition (2019).
- [174] Y. Yang, J. Sun, H. Li, Z. Xu, ADMM-CSNet: a deep learning approach for image compressive sensing, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42 (3) (2020) 521–538. doi:10.1109/TPAMI.2018.2883941.
- [175] X. Wang, K. Chan, K. Yu, C. Dong, C. Loy, EDVR: video restoration with enhanced deformable convolutional networks, *The IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (June 2019).

- [176] Y. Tian, Y. Zhang, Y. Fu, C. Xu, TDAN: temporally-deformable alignment network for video super-resolution, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020).
- [177] T. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature Pyramid Networks for Object Detection, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017) 936–944.
- [178] L. Zhu, F. Lee, J. Cai, H. Yu, Q. Chen, An Improved Feature Pyramid Network for Object Detection, *Neurocomputing* 483 (2022) 127–139.
- [179] L. Chi, B. Jiang, Y. Mu, Fast Fourier convolution, *Advances in Neural Information Processing Systems* 33 (2020) 4479–4488.
- [180] J. Lee-Thorp, J. Ainslie, I. Eckstein, S. Ontanon, FNet: mixing tokens with Fourier transforms, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (2022) 4296–4313.
- [181] Y. Rao, W. Zhao, Z. Zhu, J. Lu, J. Zhou, Global filter networks for image classification, *Advances in Neural Information Processing Systems (NeurIPS)* (2021).
- [182] J. Guibas, M. Mardani, Z. Li, A. Tao, A. Anandkumar, B. Catanzaro, Efficient token mixing for transformers via adaptive Fourier neural operators, *International Conference on Learning Representations* (2021).
- [183] R. Suvorov, E. Logacheva, A. Mashikhin, A. Remizova, A. Ashukha, A. Silvestrov, N. Kong, H. Goka, K. Park, V. Lempitsky, Resolution-robust large mask inpainting with Fourier convolutions, *WACV* (2022) 3172–3182.
- [184] D. Ulyanov, A. Vedaldi, V. Lempitsky, Improved texture networks: maximizing quality and diversity in feed-forward stylization and texture synthesis, 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR (2017) 4105–4113.

- [185] T. Karras, S. Laine, T. Aila, A style-based generator architecture for generative adversarial networks, 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 4396–4405.
- [186] P. Chandran, G. Zoss, P. Gotardo, M. Gross, D. Bradley, Adaptive convolutions for structure-aware style transfer, 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021) 7968–7977.
- [187] C. Chen, Y. Liu, P. Schniter, M. Tong, K. Zareba, O. Simonetti, L. Potter, R. Ahmad, OCMR (v1.0)–open-access multi-coil k-space dataset for cardiovascular magnetic resonance imaging, arXiv e-prints (2020).
- [188] R. Ahmad, H. Xue, S. Giri, Y. Ding, J. Craft, O. Simonetti, Variable density incoherent spatiotemporal acquisition (VISTA) for highly accelerated cardiac MRI, *Magnetic Resonance in Medicine* 74 (5) (2015) 1266–1278.
- [189] H. El-Rewaidy, A. Fahmy, F. Pashakhanloo, X. Cai, S. Kucukseymen, I. Csecs, U. Neisius, H. Haji-Valizadeh, B. Menze, R. Nezafat, Multi-domain convolutional neural network (MD-CNN) for radial reconstruction of dynamic cardiac MRI, *Magnetic Resonance in Medicine* 85 (3) (2021) 1195–1208.
- [190] J. Ho, A. Jain, P. Abbeel, Denoising Diffusion Probabilistic Models, *Advances in Neural Information Processing Systems* 33 (2020) 6840–6851.
- [191] P. Dhariwal, A. Nichol, Diffusion Models Beat GANs on Image Synthesis, *Advances in Neural Information Processing Systems* 34 (2021) 8780–8794.
- [192] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-Resolution Image Synthesis with Latent Diffusion Models, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2022).
- [193] F. Croitoru, V. Hondru, R. Ionescu, M. Shah, Diffusion Models in Vision: A Survey, *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* (2023) 1–20.

- [194] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, D. Fleet, Video Diffusion Models, ICLR Workshop on Deep Generative Models for Highly Structured Data (2022).
- [195] J. Singh, S. Gould, L. Zheng, High-Fidelity Guided Image Synthesis With Latent Diffusion Models, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023) 5997–6006.
- [196] C. Mou, X. Wang, J. Song, Y. Shan, J. Zhang, DragonDiffusion: Enabling Drag-style Manipulation on Diffusion Models (2023).
- [197] Y. Shi, C. Xue, J. Pan, W. Zhang, V. Tan, S. Bai, DragDiffusion: Harnessing Diffusion Models for Interactive Point-based Image Editing, arXiv (2023).
- [198] C. Saharia, J. Ho, W. Chan, T. Salimans, D. Fleet, M. Norouzi, Image Super-Resolution via Iterative Refinement, IEEE Transactions on Pattern Analysis and Machine Intelligence 45 (4) (2023) 4713–4726.
- [199] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. Bermano, G. Chechik, D. Cohen-or, An Image is Worth One Word: Personalizing Text-to-Image Generation Using Textual Inversion, The Eleventh International Conference on Learning Representations (2023).
- [200] B. Trabucco, K. Doherty, M. Gurinas, R. Salakhutdinov, Effective Data Augmentation With Diffusion Models, arXiv (2023).
- [201] H. Chung, B. Sim, J. Ye, Come-Closer-Diffuse-Faster: Accelerating Conditional Diffusion Models for Inverse Problems through Stochastic Contraction, IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, June 18-24, 2022 (2022) 12403–12412.
- [202] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, R. Girshick, Masked Autoencoders Are Scalable Vision Learners, 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022) 15979–15988.

-
- [203] Z. Xie, Z. Zhang, Y. Cao, Y. Lin, J. Bao, Z. Yao, Q. Dai, H. Hu, SimMIM: A Simple Framework for Masked Image Modeling, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022) 9653–9663.
- [204] B. Sim, G. Oh, J. C. Ye, Optimal transport structure of CycleGAN for unsupervised learning for inverse problems, ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2020) 8644–8647.
- [205] Z. Ke, J. Cheng, L. Ying, H. Zheng, Y. Zhu, D. Liang, An unsupervised deep learning method for multi-coil cine MRI, Physics in Medicine Biology 65 (23) (2020) 235041.
- [206] C. Hu, C. Li, H. Wang, Q. Liu, H. Zheng, S. Wang, Self-supervised Learning for MRI Reconstruction with a Parallel Network Training Framework, Medical Image Computing and Computer-Assisted Intervention (MICCAI) Part VI (2021) 382–391.
- [207] B. Yaman, S. Hosseini, M. Akcakaya, Zero-Shot Self-Supervised Learning for MRI Reconstruction, International Conference on Learning Representations (ICLR) (2022).