

UC Davis

UC Davis Electronic Theses and Dissertations

Title

Efficient Channel State Feedback for Advanced Cellular MIMO FDD Systems

Permalink

<https://escholarship.org/uc/item/9fg3g1z5>

Author

Lin, Yu-Chien

Publication Date

2024

Peer reviewed|Thesis/dissertation

Efficient Channel State Feedback for Advanced Cellular MIMO FDD Systems

By

YU-CHIEN LIN
DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY
in

Electrical and Computer Engineering

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Zhi Ding, Chair

Lifeng Lai

Bernard C. Levy

Committee in Charge

2024

Abstract

Massive MIMO technology, leveraging beamforming and precoding, significantly enhances spectrum and energy efficiency when CSI at the transmitter is available. However, the large-scale arrays envisioned for millimeter-wave or terahertz communications drastically increase the downlink CSI training overhead and uplink feedback overhead, necessitating efficient CSI feedback designs in FDD wireless systems. Similar to image compression, after downlink CSI estimation at the UE, downlink CSI is encoded into a low-dimensional codeword stream, transmitted to the base station, and decoded for downlink CSI recovery. Recently, deep learning, particularly autoencoders, has been widely discussed as an efficient CSI feedback approach, outperforming traditional compressive sensing methods. However, several challenges remain unaddressed in learning-based CSI feedback frameworks.

In this dissertation, we address six critical issues in learning-based CSI feedback frameworks. We explore the exploitation of uplink/downlink frequency-division duplexing reciprocity. The energy, delay, and angles of arrival and departure of uplink and downlink channels are highly correlated. Yet, uplink CSI, available at the base station, is seldom used to reduce the uncertainty of downlink CSI recovery. We propose a deep learning CSI feedback framework leveraging this reciprocity, with a redesigned loss function for joint encoding of CSI magnitudes and phases. Evaluation shows superior performance and better utilization of frequency-division duplexing reciprocity compared to previous works.

We also address the reduction of pilot transmission overhead. Given limited pilot resources, it is impractical to transmit pilots for all antenna ports in a massive MIMO system. We introduce a beam-based pilot precoding approach and a deep learning

CSI feedback framework to minimize pilot transmission and CSI feedback overhead from UE. Furthermore, we propose scalable CSI encoding. Existing frameworks often encode and decode full CSI using autoencoders, leading to heavy models and low scalability. Recognizing low correlation between widely spaced antennas, we propose a scalable deep learning CSI feedback framework using a divide-and-conquer principle to encode CSI subarray-by-subarray. This dynamic compression approach achieves significant model size reduction while maintaining downlink CSI recovery performance.

To tackle the reduction of training costs, deep learning models typically require extensive data collection and customization for different channel types. Inspired by the simplicity and generality of JPEG in image compression, we propose a JPEG-based CSI feedback approach, which requires no prior training and adapts to various channels, offering comparable recovery performance to learning-based methods. Additionally, we enhance frequency selective channel performance. Current learning-based models struggle with CSI recovery in frequency selective channels due to sparse pilot placement. We propose an uplink CSI-assisted CSI upsampling module at the base station, compatible with most previous explicit CSI feedback frameworks. Ensuring adherence to standardized feedback, we note that current learning-based CSI feedback methods do not strictly follow standardized CSI feedback, making industry adoption challenging. We propose a lightweight, efficient precoder upsampler as a plug-in module for the base station, enhancing performance in high delay-spread channels.

For prospective researchers, we suggest two directions to bridge artificial intelligence research and cellular communications industries. Addressing channel aging in CSI feedback is crucial due to the time lag between downlink CSI training and downlink data transmission, which may result in outdated precoders. Future research should focus on developing learning-based CSI and precoder prediction to adapt to non-linear

channel variations. Additionally, in practical frequency-division duplexing systems, the base station lacks exact CSI knowledge from UE and relies on fed-back precoders. Researchers should develop precoder-based user scheduling algorithms to avoid interference and maximize system throughput.

Acknowledgement

Completing this dissertation would not have been possible without the support and encouragement of many individuals and organizations.

First and foremost, I would like to express my deepest gratitude to the National Science Foundation (NSF) for their generous financial support with Grants No. 2029027, 20019001, 2002937, and 1824553. These supports provided me with the resources and opportunities to pursue my research and to complete my PhD Degree.

I owe a profound debt of gratitude to my family. To my mother, Hsi-Yun Chang, your sacrifices and love have shaped me into who I am today. To my elder brother, Yu-Nung Lin, thank you for always being there for me, offering your support and understanding during several challenging times. To my father in the heaven, En-Chi Lin, your unwavering belief in me has been my greatest source of strength even if you are not with me. To my best pet and also my little brother in the heaven, Hsiao-Tieh Chang, you are the gift from the God sent to us and I will always remember you and the 12 years you gave us even if you become an angle back to your place.

I am also incredibly thankful to my girlfriend, Yen-Jung Wu. Your patience, love, and understanding have been a beacon of light throughout this journey. You have been my rock, providing me with the emotional support needed to persevere. Thank you for believing in me and for standing by my side through all the ups and downs.

Finally, I would like to extend my gratitude to my friends, colleagues, and mentors who have provided valuable feedback, support, and encouragement throughout this process. To my Ph.D. supervisor, Prof. Zhi Ding, your insights, enthusiasm, and kindness have enriched my research and made this journey a truly rewarding experience. To my Master's supervisor, Prof. Ta-Sung Lee, your teaching and training have transformed me into a well-shaped researcher and engineer.

Thank you all for making this possible.

Contents

Abstract	ii
Acknowledgement	v
1 Introduction	1
1.1 CSI Feedback for MIMO FDD Systems	1
1.2 AI-aided CSI Feedback	2
1.3 Challenges for Learning-based CSI Feedback	3
1.3.1 Inefficient Utilization of FDD Reciprocity	4
1.3.2 Limited Pilot Resources	4
1.3.3 Large Model Size and Low Scalability	5
1.3.4 Efforts for Model Training and Generality	6
1.3.5 Poor Performance in Frequency Selective Channels	6
1.3.6 Deviation from 3GPP Specification	7
1.4 Structure of the article	7
2 System Model and Problem Formulation	9
2.1 Signal Model	9
2.2 Explicit CSI Feedback	11
2.2.1 CSI Estimation via Pilots and Truncation	11
2.2.2 Deep Learning Compression	13
2.3 Implicit CSI Feedback	14

2.3.1	Type I/Type II Precoding	15
2.3.2	eType II Precoding	15
2.3.3	Problem Formulation - Insufficient Feedback Resolution	17
2.4	Overview	17
3	Deep Learning Phase Compression for MIMO CSI Feedback with Limited FDD Channel Reciprocity	19
3.1	Magnitude-aided CSI feedback Framework	20
3.1.1	DualNet-MP	21
3.1.2	Loss Function Redesign	24
3.2	Experimental Evaluations	26
3.2.1	Experiment Setup	26
3.2.2	Different Phase Compression Designs	27
3.2.3	Different Core Layer Designs	28
3.3	Conclusions	29
4	Exploiting Partial FDD Reciprocity for Beam Based Pilot Precoding and CSI Feedback in Deep Learning	32
4.1	BS Precoding for CSI-RS Reduction	33
4.1.1	DL CSI recovery	33
4.1.2	Single-beam BS Precoding and DL CSI recovery	34
4.1.3	Single-beam Precoding and AI-aided DL CSI Recovery	36
4.1.4	Fuzzy-beam Precoding and AI-aided DL CSI Recovery	37
4.2	Encoder-Free CSI Feedback with UL CSI Assistance	39
4.2.1	General Architecture	39
4.2.2	BSdualNet	40
4.2.3	BSdualNet-MN	44

4.3	UL CSI Aided Beam Based Precoding and a Reconfigurable CSI Feed- back Frameworks	46
4.3.1	Frequency Resource Reconfiguration	47
4.3.2	BSdualNet-FR	48
4.4	Experimental Evaluations	50
4.4.1	Experiment Setup	50
4.4.2	Determining significant beam matrix \mathbf{B}_S based on DL and UL CSIs	53
4.4.3	Testing Different Numbers of Available REs	54
4.4.4	Performance for Different Numbers of UEs	55
4.4.5	Different CSI-RS Configurations and Compression Ratios	57
4.4.6	Different Effective Compression Ratio CR_{eff}	58
4.4.7	Complexity: FLOPs and Parameters	61
4.5	Conclusions	62
5	An Efficient and Scalable Deep Learning Framework for Dynamic CSI Feedback under Variable Antenna Ports	63
5.1	Multi-rate CSI Feedback Framework with Flexible Number of Antennas	64
5.1.1	SAB Framework	66
5.1.2	Multi-rate CSI Feedback Framework	67
5.2	Multi-Rate CSI Feedback Framework with Flexible Number of Antenna Ports	69
5.2.1	SAB framework in BD domain	70
5.2.2	DCP Feedback Pruning	70
5.2.3	Local Normalization	72
5.2.4	2D Lightweight Encoder	73
5.3	SAB framework with dynamic CR	74
5.4	Experimental Evaluations	77

5.4.1	Experiment Setup	77
5.4.2	SCENet vs. SCENet+	80
5.4.3	Performance, Complexity and Storage Comparison	80
5.4.4	Testing Different Encoder/Decoder Pairs	82
5.4.5	Testing Different Array Geometries	83
5.4.6	BD SAB Framework in GN and LN approaches	84
5.4.7	DCP feedback pruning	84
5.4.8	2D SAB Framework	85
5.4.9	CSI feedback with dynamic CR	85
5.4.10	Different Noise Powers and CR Selections	86
5.5	Conclusions	87
6	Applying JPEG Compression for Feedback of Massive MIMO Channel State Information	95
6.1	JPEG Image Compression	96
6.2	JPEG-based CSI Feedback Framework	98
6.2.1	Ordering Real/Imaginary CSI	99
6.2.2	Zero Replacement (ZR)	99
6.2.3	Entropy Encoding/Decoding	99
6.3	Experimental Evaluations	102
6.3.1	Experiment Setup	102
6.3.2	Benefit of Huffman Encoding	103
6.3.3	DCT and DFT transformation	103
6.3.4	Testing different channel scenarios	104
6.4	Conclusions	106
7	Physics-Inspired Deep Learning Anti-Aliasing Framework in Efficient Channel State Feedback	108

7.1	Problem Formulation: Aliasing Issue in CSI Feedback	109
7.1.1	DL CSI Preprocessing	109
7.1.2	DL CSI Feedback	110
7.1.3	Aliasing Issue	111
7.2	UL-CSI aided Upsampling with Aliasing Suppression	114
7.2.1	CSI Upsampling with Side Information	114
7.2.2	Multipath Reciprocity	117
7.2.3	UL Masking: UL-Assisted CSI Upsampling with Aliasing Sup- pression	120
7.3	Physic-inspired AI-driven Aliasing Suppression	120
7.3.1	Model Architecture	121
7.3.2	Loss Function Design	125
7.3.3	Limitations and Failure Scenarios	125
7.4	Efficient Channel State Feedback with Aliasing Suppression from Non- uniform Sampling	125
7.4.1	Compressive sensing based CSI upsampling	127
7.4.2	ISTA-Net Framework	127
7.4.3	SRISTA-Net Framework	129
7.4.4	Loss Function Design	130
7.4.5	Initialization	132
7.5	Experimental Evaluations	132
7.5.1	Experiment Setup	132
7.5.2	UL Assisted Bandpass Filter Design for Anti-aliasing	134
7.5.3	SRCSiNet	134
7.5.4	End-to-end CSI Recovery	136
7.5.5	Solving Overfitting problem	137
7.5.6	Temporal Sensitivity of SRISTA-Net	138

7.5.7	Complexity and Storage Requirements	139
7.6	Conclusions	140
8	Plug-in UL-CSI-Assisted Precoder Upsampling Approach in Cellular FDD Systems	142
8.1	Type II/eTypeII based Precoder Upsampling	143
8.1.1	General Architecture	143
8.1.2	Modified Type II/eType II precoding	143
8.1.3	Precoder Upsampler, SRPNet	144
8.2	UL-CSI/SSB Assisted Switch for Low-Complexity Precoder Upsampling	147
8.2.1	PDP-based Switch	147
8.3	Experimental Evaluations	148
8.3.1	Experiment Setup	148
8.3.2	Applying SRPNet to SB-level Type II Precoder	149
8.3.3	Applying SRPNet to SB-level eType II Precoder	150
8.3.4	Applying PDP-based Switch for Complexity Reduction	151
8.4	Conclusions	153
9	Future Works and Conclusions	155
9.1	Channel/Precoder Prediction against Channel Aging	155
9.1.1	Problem Formulation: Future Precoder Prediction	156
9.1.2	Proposed Approaches and Preliminary Results	157
9.2	User Clustering for MU-MIMO	159
9.3	Conclusions	161
A	Appendix	164

Chapter 1

Introduction

In this chapter, we present an introduction to the research topic addressed by this dissertation. First, we briefly review the channel state information (CSI) feedback problem for multiple-input multiple-output (MIMO) frequency-division duplexing (FDD) wireless systems. Then, we concisely discuss the existing CSI feedback frameworks, highlighting the problems in current solutions, including the lack of domain knowledge, model size and scalability, model generality, and low pilot placement density and deviation from the practical CSI feedback mechanisms. Finally, we summarize the structure of this dissertation.

1.1 CSI Feedback for MIMO FDD Systems

Massive MIMO technologies play an important role in improving the spectrum and energy efficiency of 5G and future-generation wireless networks. The power of massive MIMO hinges on accurate DL CSI at the base station or gNodeB (gNB). Without uplink (UL)/DL channel reciprocity assumed in time-division duplexing (TDD) systems, as illustrated in Figure 1.1, an FDD base station typically relies on user equipment (UE) feedback for DL precoder design. The feedback approaches can be broadly categorized into explicit CSI feedback and implicit CSI feedback.

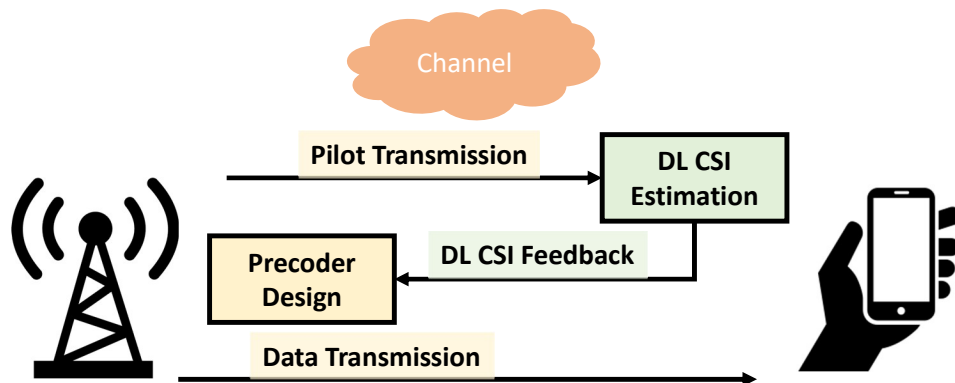


Figure 1.1: Illustration of CSI feedback in FDD system.

Since CSI in most environments has limited delay spread and can be viewed as sparse, CSI feedback by UEs can leverage this low dimensionality for CSI feedback compression. Compressive sensing (CS) [1–4] appears to be a promising solution to recover such sparse signals. However, CS has limitations: [5]:

- (a) CSI recovery is not accurate enough since CSIs are not exactly sparse in any basis,
- (b) the random projection matrix may not fully exploit the structural features of CSIs,
- (c) involvement of iterative approach for signal reconstruction usually is time-consuming while CSI recovery task is somehow time-sensitive.

Thus, both industry and researchers have been dedicated to finding a new effective solution for CSI feedback. The 3rd Generation Partnership Project (3GPP) recently released the features of Release 18 [6], which embraces artificial intelligence (AI) and machine learning (ML) for enhancing CSI feedback (e.g., reducing overhead and improving estimation accuracy).

1.2 AI-aided CSI Feedback

Inspired by the success in image compression, deep neural networks have been widely adopted for CSI feedback frameworks in recent years. To improve feedback efficiency, previous works [5, 7] developed a deep convolutional neural network with an autoen-

coder structure, where the encoder and decoder are deployed at the UE and gNB, respectively, for CSI compression and recovery. Related works and variants [8–11] have demonstrated performance advantages over traditional compressive sensing approaches.

Recent studies have highlighted the importance of exploiting correlated channel information such as UL CSI [12–16], past CSI [17], and CSI of adjacent UEs [18] for improving DL CSI recovery accuracy at base stations. Important physical insights regarding FDD reciprocity, slow changes in the propagation environment over time, and similar propagation conditions within short geographical distances underscore the strong spectral, temporal, and spatial correlations between magnitudes of different CSI in the angle-delay (AD) domain. Since side information from correlated CSI lowers the conditional entropy (uncertainty) of the DL CSI, its effective utilization reduces the encoded feedback payload required from UEs [8, 17].

Recent works have also focused on practical issues such as pilot design, model size, computational complexity, and training strategy. For example, notable progress has been made in recovery performance among recent autoencoder-based CSI feedback frameworks [19] by considering the joint design of pilot and CSI recovery. Lightweight learning models [9, 20] for low-cost UEs have been proposed to ease the broad deployment of such systems. Self-supervised and federated learning strategies have been adopted in some works [21, 22] to address the challenge of scarce data collection. However, fundamental problems of AI-aided CSI feedback still hamper real-world implementation.

1.3 Challenges for Learning-based CSI Feedback

In the following subsections, we describe the problems of existing learning-based CSI feedback frameworks addressed by this article:

1.3.1 Inefficient Utilization of FDD Reciprocity

FDD reciprocity denotes the correlation between UL/DL CSI. Since radiations propagate similarly in both directions, the energy, delay, and angles of arrival/departure (AoAs/AoDs) of multipaths in UL and DL transmission are highly correlated. Thus, there is a high correlation between DL and UL CSI magnitudes in the AD domain. Existing solutions [12, 13, 18] have adopted dual feedback frameworks that separately encode and recover CSI phases from their corresponding magnitudes. These studies use an isolated autoencoder to compress and recover the CSI magnitudes. For phase recovery, significant phases are encoded according to their corresponding magnitudes.

Presently, some magnitude-dependent CSI feedback frameworks train two learning models to encode and recover DL CSI magnitudes and phases, respectively. For example, [18] designed a deep learning model with CSI magnitude and phase autoencoders, training the phase branch with a magnitude-dependent polar-phase (MDPP) loss function to penalize discrepancies in CSI phase estimates with larger magnitudes. However, this does not fully reflect the real MSE loss, thus failing to guarantee minimized MSE.

Both CSI magnitudes and phases depend on the RF propagation environment, including multipath delays, Doppler spread, bandwidth, and scatter distribution. Therefore, CSI magnitude and phase encoding and recovery should be jointly optimized. The structural sparsity of CSI phases and their joint distribution with the CSI magnitude are generally unknown and underexplored.

1.3.2 Limited Pilot Resources

The estimation accuracy of DL CSI at UEs depends on factors such as channel fading properties and reference signal (RS) placement. Beyond feedback overhead, the required resource pilot (i.e., CSI-RS) allocation for CSI estimation grows proportionally with the antenna array size. More resources allocated to CSI-RS improve DL CSI estimation accuracy but degrade spectrum efficiency. In practical systems such as [23],

CSI-RS resources are sparsely allocated on the time-frequency physical resource grid. Only a few studies [19,24] have considered the sparse CSI-RS availability in designing CSI feedback mechanisms.

Our earlier work [24] proposed a deep learning partial CSI feedback framework that reduces RS resource overhead by leveraging temporal CSI correlation. The work in [19] optimizes the DL pilot symbols (i.e., CSI-RS) based on UL CSI without reducing the CSI-RS resources. However, this implementation does not reduce any resource pilot allocation and requires dynamic exchange of optimized pilot symbols between the gNB and the UE, which is incompatible with the current use of predefined CSI-RS.

1.3.3 Large Model Size and Low Scalability

Existing deep learning methods attempt to extract underlying mutual dependency among gNB antennas in massive MIMO configurations by simultaneously feeding CSI of all DL antennas into the learning machine for joint compression. Such large input sizes make it harder to develop low-complexity and lightweight deep learning models. Attempts to reduce encoder model complexity [25–28] have achieved limited success. Additionally, the rigidity of the input/output size of deep learning models necessitates new models for different array sizes.

Inflexible model latent sizes require model retraining for different compression levels. To avoid this, [29] applied transfer learning to reduce training costs for multiple compression levels. A related work [30] designed a multi-rate CSI feedback framework with a matching classification model for selecting a target compression ratio according to the number of channel clusters. However, the physical connection between compressibility and channel cluster number remains unconfirmed.

1.3.4 Efforts for Model Training and Generality

Operators face obstacles in deploying learning-based frameworks in real-world systems, including the cost of training data collection needed for deep learning optimization. CSI data acquisition for massive MIMO requires accurate models and extensive field measurement, posing serious practical challenges [31]. Practical wireless networks are deployed in a wide range of RF environments, requiring multiple deep learning models for different channel scenarios and compression ratios, each trained separately, customized, and selected for a specific scenario.

These practices incur large memory burdens for UEs to store multiple deep learning models, require large training datasets under multiple RF channel scenarios for deep learning optimization, and determine the suitable deep learning model to use. Transfer learning and online learning concepts [32, 33] have moderately reduced training costs. However, the implementation and storage of multiple deep learning models still lead to high costs in hardware and power, especially at the UE side, as channel bandwidth and antenna numbers continue to grow.

1.3.5 Poor Performance in Frequency Selective Channels

In 5G-NR, CSI-RS is used for channel state inference and precoder design instead of full-channel estimation. With low CSI-RS placement density in the frequency domain, the fast variation of CSIs due to large delay spread cannot be captured at the UE side. From the gNB's perspective, recovering full CSI at the subcarrier level may result in strong aliasing effects even if the UE performs perfect CSI feedback. To the best of our knowledge, previous works have barely considered this issue, making it crucial to design a pre- or post-aliasing suppression approach for CSI feedback frameworks.

1.3.6 Deviation from 3GPP Specification

Existing commercialized CSI feedback system is based on implicit feedback instead of explicit CSI feedback. In modern cellular networks such as 5G-NR, UEs feedback precoder matrix indicators (PMI) and layer indicators (LI) for each subband (SB) to select precoders from a codebook at the gNB side. For flat fading channels, it is acceptable to apply a wideband or SB-level precoder to different resource blocks (RBs) since their channels are nearly identical. However, for outdoor scenarios with high delay spread, an SB-level precoder may perform poorly for specific RBs. Although AI-based CSI feedback can perform RB-level CSI compression and recovery with good performance, deploying it in cellular networks is challenging before 3GPP agrees on a new standard. Thus, industry tends to adopt an early-stage solution: a one-sided precoder upsampler deployed only at the gNB, acting as a plug-in module compatible with existing protocols and specifications.

1.4 Structure of the article

In Chapter 2, we describe the signal model and introduce the general CSI feedback problem in FDD MIMO systems. Then, we review how a deep learning-based autoencoder architecture helps reduce uplink overhead and improve CSI estimation accuracy.

In Chapter 3, we tackle the challenge of inefficient utilization of FDD reciprocity. We develop a deep learning-based CSI feedback framework that jointly optimizes magnitude and phase encoding, proposing a new loss function, namely sinusoidal magnitude-adjust phase error (SMAPE), which directly corresponds to the MSE of DL CSI recovery. We also propose novel circularly convolutional neural network (C-CNN) layers to enhance CSI compression efficiency and recovery performance.

In Chapter 4, we address the challenge of limited pilot resources. We develop an efficient and reconfigurable deep learning beam-based CSI feedback framework that

leverages UL/DL angular reciprocity for FDD wireless systems to reduce DL CSI-RS and UL feedback overhead while maintaining DL CSI recovery accuracy.

In Chapter 5, we address the challenge of large model size and low scalability. Utilizing the physical insight that only nearby massive MIMO antennas exhibit non-negligible CSI correlation, we propose a light-weight autoencoder that compresses large-array CSI via a divide-and-conquer principle (DCP). This approach substantially decreases input size and, consequently, the deep learning model size. We also design a novel dynamic-rate CSI feedback framework with a matching classifier to determine the optimal compression level for maximizing codeword efficiency.

In Chapter 6, we address the challenge of model training and generality. We develop a simpler, scalable, and flexible algorithm inspired by the JPEG compression of images. Our zero replacement (ZR) compressive CSI feedback algorithm is universal, accommodating different compression ratios and a wide range of channel scenarios without requiring special training datasets.

In Chapter 7, we address the challenge of poor performance in frequency selective channels. We design a new CSI upsampling module compatible with any explicit CSI feedback frameworks for post-aliasing suppression deployed at the gNB side, exploiting FDD reciprocity to design a bandpass filter (BPF) for aliasing suppression.

In Chapter 8, we address the challenge of specification deviation. We design a plug-in precoder upsampling module deployed at the gNB, which is computationally efficient and compatible with existing Type II and modified enhanced-Type II (eType II) precoders.

In Chapter 9, we summarize the contributions of this article and highlight potential future research directions.

Chapter 2

System Model and Problem Formulation

In this chapter, we delve into the details of system modeling and problem formulation. First, we introduce the general downlink and uplink system signal model. Then, we formulate our CSI feedback problem in FDD wireless systems. Last, we give an big picture illustrating the relationship between the following chapters.

2.1 Signal Model

We consider a single-cell MIMO FDD link where a gNB using an $N_H \times N_V$ uniform planar array (UPA) with $N_b = N_V N_H$ antennas communicates with single-antenna UEs. Note that a UPA becomes a uniform linear array (ULA) when $N_V = 1$ and $N_b = N_H$. In FDD systems, UEs estimate DL CSIs and feedback to the serving base station after encoding and quantization. The gNB then recovers the DL CSI based on this feedback. Following the 3GPP specification [23] and focusing on a specific UE, the DL subband consists of N_{RB} RBs within the bandwidth for both data and pilot transmission. We assume channels within an RB to be under slow, flat, and block fading.

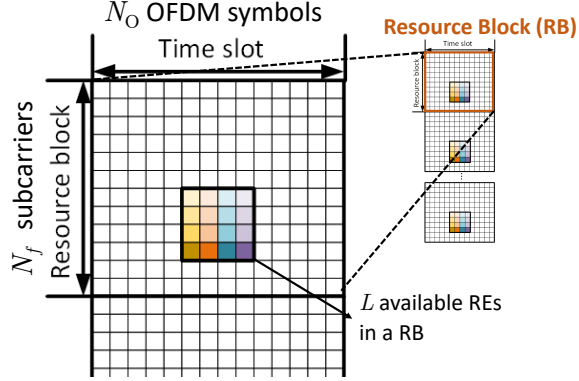


Figure 2.1: Resource block configuration. There are $N_f \times N_O$ time-frequency REs in a RB (N_f subcarriers and N_O OFDM symbols). Pilots are allowed to be placed at those REs in designated region (surrounded by the black frame). There are $L = 16$ available REs for pilot placement in this illustration.

As shown in Figure 2.1, there are $N_f \times N_O$ time-frequency resource elements (REs) in a specific RB (N_f subcarriers and N_O OFDM symbols). In each RB, there are L REs designated for pilot transmission (i.e., CSI-RS placement). We typically adopt a general setting of $L = N_b$ so that the CSI of every antenna port can be estimated in each RB.

The OFDM signal spans N_{RB} DL RBs. The DL signal received from the k th RB is

$$y_{f,\text{DL}} = \mathbf{h}_{f,\text{DL}}^H \mathbf{w}_f x_{f,\text{DL}} + n_{f,\text{DL}},$$

where $(\cdot)^H$ denotes conjugate transpose. Here for the k -th RB, $\mathbf{h}_{f,\text{DL}} \in \mathbb{C}^{N_b \times 1}$ denotes the CSI vector and $\mathbf{w}_f \in \mathbb{C}^{N_b \times 1}$ denotes the corresponding precoding vector¹ whereas $x_{f,\text{DL}} \in \mathbb{C}$ and $n_{f,\text{DL}} \in \mathbb{C}$ denote the DL source signal and additive noise, respectively. With the same antennas, gNB receives UL signal

$$\mathbf{y}_{f,\text{UL}} = \mathbf{h}_{f,\text{UL}} x_{f,\text{UL}} + \mathbf{n}_{f,\text{UL}} \in \mathbb{C}^{N_b \times 1},$$

¹gNB calculates precoding vectors at subcarriers with DL CSI matrix. For example, a maximum-ratio combining precoder $\mathbf{h}_{f,\text{DL}} / \|\mathbf{h}_{f,\text{DL}}\|$ is used for maximizing the receiving gain at UE.

where $\mathbf{h}_{f,\text{UL}} \in \mathbb{C}^{N_b \times 1}$ is the UL channel vector, and the subscript UL denotes the UL signals and noise. Without channel reciprocity in TDD system, $\mathbf{h}_{f,\text{DL}}$ and $\mathbf{h}_{f,\text{UL}}$ are different. Yet, the delays, energy and AoAs/AoDs of multi-paths are highly correlated.

In FDD system, gNB relies on the UE feedback to design precoders to improve the reception power in downlink transmission. The feedback mechanisms can be roughly categorized into two types: explicit and implicit CSI feedback illustrated in Figures 2.2 (a) and (b), respectively. Both architectures are considered in this article. They will be introduced in the following sections.

2.2 Explicit CSI Feedback

The idea of explicit CSI feedback as shown in Figure 2.2 (a) is to compress DL CSI at UE side after channel estimation and recover it at base station side so as to design precoders for downlink transmission. Encoder and decoder are deployed at UE and gNB sides for CSI compression and recovery, respectively.

2.2.1 CSI Estimation via Pilots and Truncation

Assume that UEs and gNB can perfectly estimate DL/UL CSIs, respectively². DL and UL channel vectors can be jointly written as spatial-frequency channel state information (SF-CSI) matrices $\mathbf{H}_{\text{DL}}^{\text{SF}} = [\mathbf{h}_{1,\text{DL}}, \dots, \mathbf{h}_{N_{\text{RB}},\text{DL}}] \in \mathbb{C}^{N_b \times N_{\text{RB}}}$ and $\mathbf{H}_{\text{UL}}^{\text{SF}} = [\mathbf{h}_{1,\text{UL}}, \dots, \mathbf{h}_{N_{\text{RB}},\text{UL}}] \in \mathbb{C}^{N_b \times N_{\text{RB}}}$, respectively. Typically in FDD systems, DL CSI $\mathbf{H}_{\text{DL}}^{\text{SF}}$ is estimated and fed back by UE to gNB. However, the number ($N_b \times N_{\text{RB}}$) of unknowns in $\mathbf{H}_{\text{DL}}^{\text{SF}}$ requires substantial feedback resources in large or massive MIMO systems, consuming excessive bandwidth. To reduce the CSI feedback overhead, we can apply

²Since gNB usually uses high-energy pilots for DL CSI estimation in modern communication systems, recovery loss mainly originates from lossy compression. In such case, there is little difference when using different channel estimation schemes. Furthermore, in the context of CSI, perfect estimation is commonly adopted in most related works.

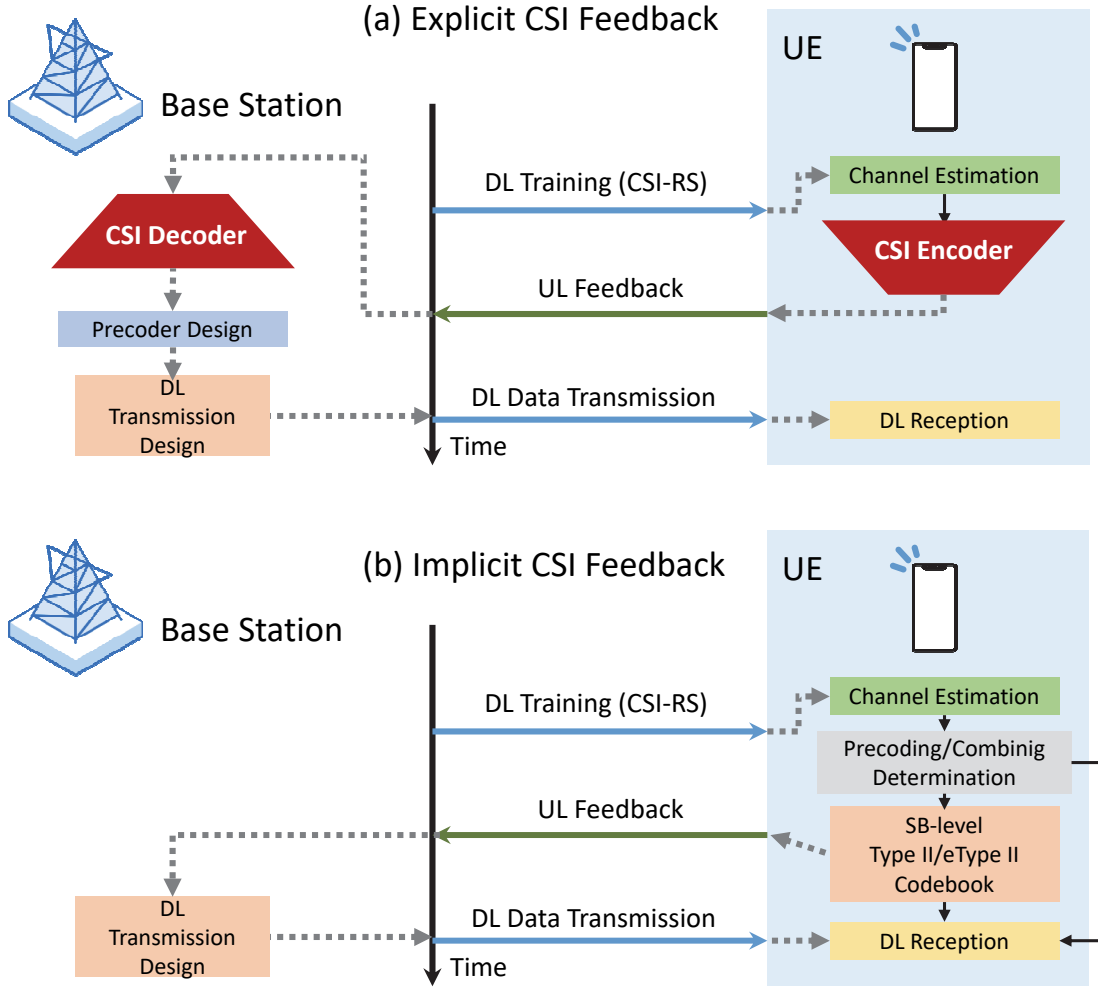


Figure 2.2: Working block diagrams of (a) explicit CSI feedback and (b) implicit CSI feedback in FDD system.

discrete Fourier transform (DFT) $\mathbf{F}_A \in \mathbb{C}^{N_b \times N_b}$ and inverse DFT (IDFT) or discrete cosine transform (DCT) $\mathbf{F}_D \in \mathbb{C}^{N_{RB} \times N_{RB}}$ and on \mathbf{H}^{SF} to generate either spatial-delay (SD) or angle-delay (AD) domain CSI matrices:

$$\mathbf{H}^{\text{AD}} = \mathbf{F}_A^H \mathbf{H}^{\text{SF}} \mathbf{F}_D,$$

$$\mathbf{H}^{\text{SD}} = \mathbf{H}^{\text{SF}} \mathbf{F}_D,$$

which demonstrates sparsity. Note that \mathbf{H}^{SF} denotes either $\mathbf{H}_{\text{UL}}^{\text{SF}}$ or $\mathbf{H}_{\text{UL}}^{\text{SF}}$. Owing to limited multipath delay spread and limited number of scatters, most elements in \mathbf{H}^{SD}

and \mathbf{H}^{AD} are found to be near insignificant, except for the few columns. For simplicity, we shall denote $\mathbf{H}_{\text{DL}}^{\text{SD}}$ and $\mathbf{H}_{\text{DL}}^{\text{AD}}$ as $\tilde{\mathbf{H}}_{\text{DL}}$ in the rest of this article except for cases when ambiguity may arise.

To reduce UL feedback overhead, we exploit the physical multipath delay sparsity of CSI by transforming full DL CSI into delay domain through discrete Fourier transform (DFT) or discrete cosine transform (DCT). We truncate the insignificant near-zero elements in trailing (large) delay indices as follows:

$$\mathbf{H}_{\text{DL}} = \tilde{\mathbf{H}}_{\text{DL}} \underbrace{\begin{bmatrix} \mathbf{I}_{N_t \times N_t} \\ \mathbf{0} \end{bmatrix}}_{\mathbf{T}} \in \mathbb{C}^{N_b \times N_t}, \quad (2.1)$$

where $\mathbf{T} \in \mathbb{C}^{N_{\text{RB}} \times N_t}$ performs delay domain truncation. Note that the design of matrix \mathbf{T} may varies according to transformation \mathbf{F}_D and the CSI properties. Matrix \mathbf{T} in Eq. (2.1) is an example for DCT transformation that drops the last $N_{\text{RB}} - N_t$ columns of $\tilde{\mathbf{H}}$ corresponding to large multipath delays.

2.2.2 Deep Learning Compression

Autoencoder has shown successes in several deep learning frameworks. An encoder at UE compresses its estimated DL CSI for uplink feedback and a decoder at gNB recovers the estimated CSI according to the feedback from UE. Assuming negligible CSI elements at large delays, many have exploited convolutional layers to compress and recover the truncated DL pilot CSI via

$$\text{Encoder: } \mathbf{q} = f_{\text{en}}(\mathbf{H}_{\text{DL}}), \quad (2.2)$$

$$\text{Decoder: } \hat{\mathbf{H}}_{\text{DL}} = f_{\text{de}}(\mathbf{q}). \quad (2.3)$$

The decoder should replace the truncated DL CSI $\hat{\mathbf{H}}_{\text{DL}}$ via zero-padding to transform CSI back from delay domain to estimate DL CSI matrix $\tilde{\mathbf{H}}_{\text{DL}}$ in the subcarrier domain as follows:

$$\hat{\tilde{\mathbf{H}}}_{\text{DL}} = \begin{bmatrix} \hat{\mathbf{H}}_{\text{DL}} & \mathbf{0}_{N_b \times (N_{\text{RB}} - N_t)} \end{bmatrix} \mathbf{F}_D^H. \quad (2.4)$$

Eq. (2.4) is an example for DCT transformation that drops the last $N_{\text{RB}} - N_t$ columns of $\tilde{\mathbf{H}}$ corresponding to large multipath delays. The CSI recovery accuracy can be measured by the normalized mean square error (NMSE) of the full DL CSI:

$$\text{NMSE}(\hat{\tilde{\mathbf{H}}}_{\text{DL}}, \tilde{\mathbf{H}}_{\text{DL}}) = \sum_{d=1}^D \|\hat{\tilde{\mathbf{H}}}_{\text{DL},d} - \tilde{\mathbf{H}}_{\text{DL},d}\|_F^2 / \|\tilde{\mathbf{H}}_{\text{DL},d}\|_F^2,$$

where subscript d denotes the d -th test. When training the autoencoder, the CSI error due to truncation is unavailable. Hence, autoencoder loss function can simply rely on the truncated DL CSI error

$$\text{NMSE}(\hat{\mathbf{H}}_{\text{DL}}, \mathbf{H}_{\text{DL}}) = \sum_{d=1}^D \|\hat{\mathbf{H}}_{\text{DL},d} - \mathbf{H}_{\text{DL},d}\|_F^2 / \|\mathbf{H}_{\text{DL},d}\|_F^2.$$

2.3 Implicit CSI Feedback

The idea of implicit CSI feedback, or codebook-based precoder feedback, is to determine the precoder at UE side from codebooks and feedback to gNB for the following downlink transmission shown in Figure 2.2 (b). There are shared codebooks between gNB and UEs. UEs estimate DL CSI from pilots, and then feedback precoder matrix indicator (PMI) and layer indicator (LI) to gNB. Common codebook-based precoding in modern FDD system include Type I, Type II and eType II [34–36] precoding, which will be introduced below in a high-level manner (for simplicity, we consider one polarization only):

2.3.1 Type I/Type II Precoding

Type I precoding exploits the spatial diversity provided by multiple transmit antennas to enhance communication performance. The UE selects a beam and a co-phase coefficient from an oversampled set of beam directions as the Type I precoder. The Type I precoder for the f -th SB can be expressed as:

$$\mathbf{w}_f = \operatorname{argmax}_{\mathbf{w} \in \Omega_l} \{|\mathbf{h}_{f,\text{DL}}^H \mathbf{w}|\}, f = 1, \dots, N_3. \quad (2.5)$$

Where Ω_l is the codebook containing oversampled beams corresponding to the l -th layer, and N_3 is the number of SBs in BWP. Then the UE acknowledges the selected precoder by feeding back the beam index (i.e., PMI) and the layer l . For the Type II precoder, designed for the multi-user MIMO (MU-MIMO) use case, it provides more flexibility in choosing multiple beams and the degree of freedom to combine the selected beams to match DL CSI. The selected Type II precoder for the f -th SB can be expressed as:

$$\mathbf{w}_f = \sum_{i=1}^L \alpha_{f,i} \mathbf{w}_{f,i} / L. \quad (2.6)$$

where $\mathbf{w}_{f,i}$ is the i -th selected oversampled beam, and $\alpha_{f,i}$ is the complex combining coefficient in the f -th SB. Note that, to reduce the feedback overhead, both Type I and Type II precoders are fed back per SB.

2.3.2 eType II Precoding

The eType II precoder is a more efficient feedback method compared to Type I and Type II precoders. Figure 2.3 reveals the differences between Type II and eType II precoding. In Figure 2.3(a), it can be seen that the UE designs the Type II precoder for each SB independently. Considering the high correlation of the spatial structures of channels for SBs in a BWP, the eType II precoder design allows the gNB to enable UEs to jointly select L wideband beam for all SBs in the entire BWP, which is called

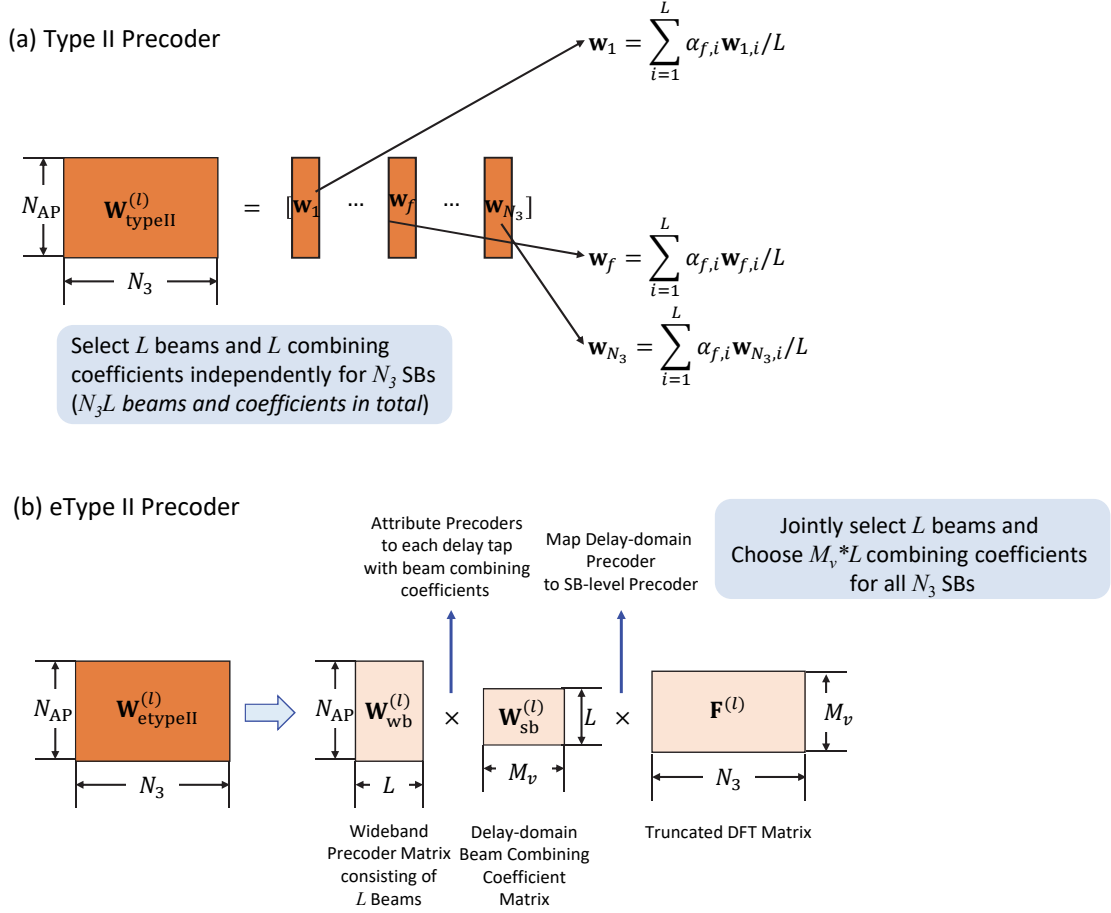


Figure 2.3: The comparison between (a) Type II and (b) eType II Precoder.

spatial compression, and feedback the precoder.

To further reduce the feedback overhead, the eType II precoder performs frequency-domain compression. It first transforms the SB-level precoders into delay-domain ones. According to the principles of radiology and propagation loss, the delay-domain beam combining coefficients are truncated by retaining only the first M_v delay taps. To further compress in the delay domain, due to the sparsity of the truncated delay-domain coefficients, the coefficients can be compressed by a factor of R , feeding back only the significant delay taps and their positions.

2.3.3 Problem Formulation - Insufficient Feedback Resolution

Given the restriction of uplink feedback overhead, all types of codebook-based precoder feedback can only be conducted per SB. For some frequency-selective channels (e.g., outdoor channels with large delay spread), such a low feedback rate cannot fully exploit the channel diversity provided by the precoder. Without modifying the current specification, operators seek a non-linear mapping function $f_{\Theta}(\cdot)$ to upsample the SB-level precoders \mathbf{W}_{SB} to RB-level ones $\hat{\mathbf{W}}_{\text{RB}}$, thereby better exploiting the channel gain via a finer-resolution precoder in the frequency domain. The loss function can be expressed as follows:

$$\Theta = \operatorname{argmax}_{\Theta} \sum_{f=0}^{N_{\text{RB}}} |\mathbf{h}_f^H \hat{\mathbf{W}}_{\text{RB},f}|, \quad (2.7)$$

$$\hat{\mathbf{W}}_{\text{RB}} = f_{\Theta}(\mathbf{W}_{\text{SB}} \in \mathbb{C}^{N_b \times N_3}) \in \mathbb{C}^{N_b \times N_{\text{RB}}}, \quad (2.8)$$

where Θ represents the trainable parameters of the learning-based upsampler, \mathbf{W}_{SB} consists of precoders for N_3 SBs, $\hat{\mathbf{W}}_{\text{RB},f}$ denotes the precoder for f -th RB.

2.4 Overview

Figure 2.4 shows a tree diagram illustrating the features and relationship between the proposed methods in this dissertation. From Sections 3 to 7, we focus on developing efficient compression/recovery algorithms and CSI upsampling approach for explicit CSI feedback considering different important practical needs. The CSI upsampling module proposed in Section 7 can be incorporated with the previous explicit CSI feedback mechanisms and most prior works. For Section 8, we follow the current 3GPP standardized implicit CSI feedback approaches. A plug-in precoder upsampling module, which is compatible with Type II and eType II precoder feedback approaches, was proposed to boost the precoder gain from channels.

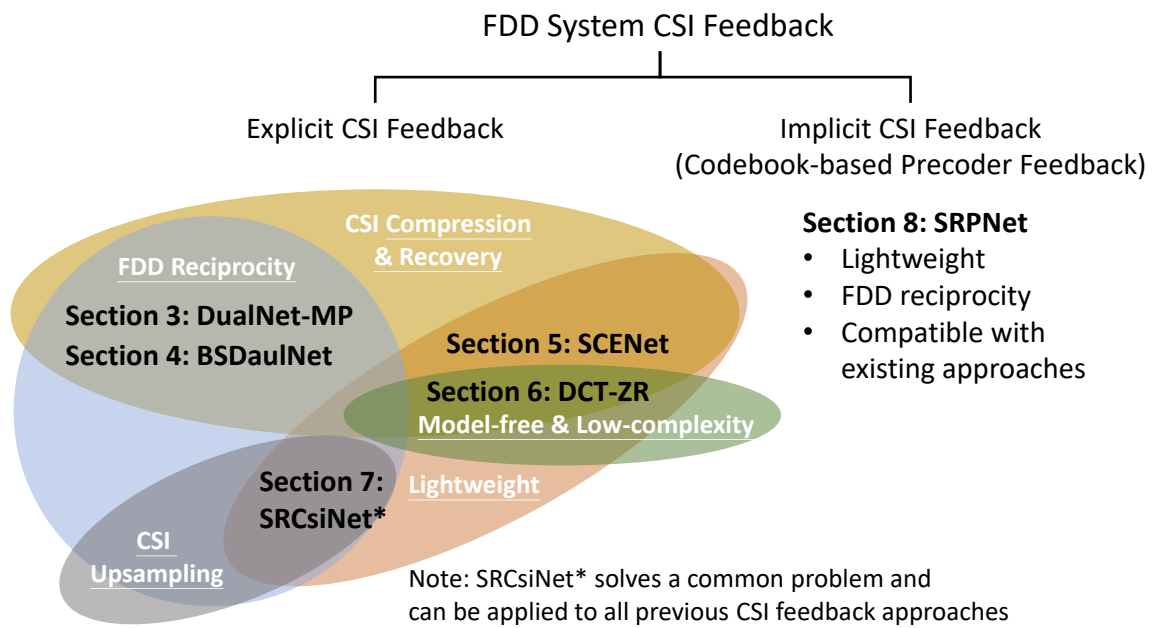


Figure 2.4: The tree diagram of the proposed methods in this dissertation.

Chapter 3

Deep Learning Phase Compression for MIMO CSI Feedback with Limited FDD Channel Reciprocity

Large scale MIMO FDD systems are often hampered by bandwidth required to feedback DL CSI. Previous works [12, 13] have made notable progresses in efficient CSI encoding and recovery by taking advantage of FDD UL/DL reciprocity between their CSI magnitudes. Such framework separately encodes CSI phase and magnitude, which cannot efficiently exploit the correlation between CSI magnitudes and phases. To further enhance feedback efficiency, we propose a new deep learning architecture for joint magnitude and phase encoding based on limited CSI feedback and magnitude-aided information.

In this chapter, we first mention the background of existing DL CSI feedback frameworks with the aids of correlated CSIs locally available at base stations. Then, we propose a new deep learning CSI feedback framework which better exploits UL CSI to jointly encode DL CSI phase and magnitude for better CSI recovery performance. The model architecture and loss function designs are then detailed. Finally, our test results

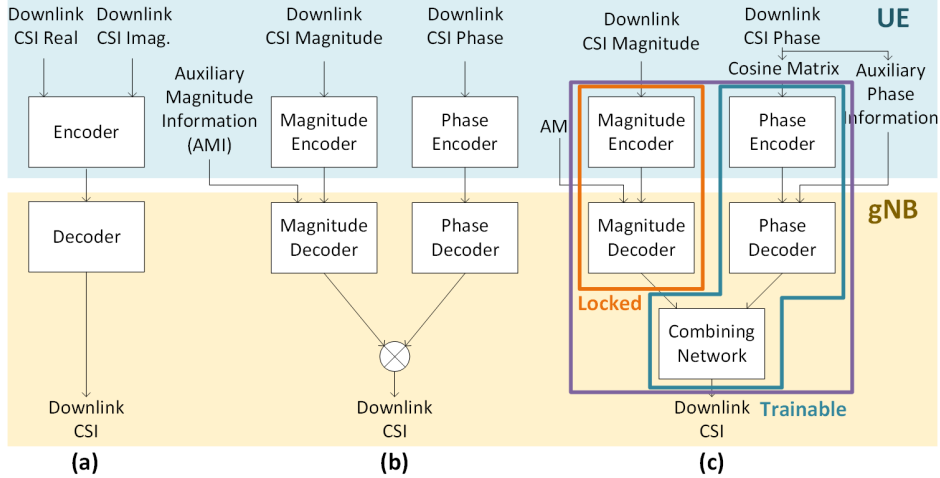


Figure 3.1: General network architecture. (a) Conventional CSI feedback framework, (b) conventional magnitude-aided CSI feedback framework, and (c) proposed magnitude-aided CSI feedback framework.

demonstrate that the newly proposed framework outperforms other SOTAs with or without the aids of UL CSI information. Note that, in this chapter, we represent DL CSI \mathbf{H}_{DL} by \mathbf{H} for simplicity and use \mathbf{H}_{UL} for UL CSI for simplicity.

3.1 Magnitude-aided CSI feedback Framework

Most deep-learning works on CSI compression leverage the success of real-valued deep learning network (DLN) in image processing by separating CSI matrices into real and imaginary parts that are analogous to image files [5, 7, 17], as shown in Figure 3.1(a). Recent studies [13, 17, 18], however, uncovered the benefit of separately encoding magnitudes and phases of \mathbf{H} instead in order to better exploit other correlated CSI magnitudes as auxiliary magnitude information (AMI). Note that, in this chapter, the CSI \mathbf{H} is on AD domain. Such architecture, illustrated in Figure 3.1(b), requires substantially lower feedback overhead for the magnitudes of \mathbf{H} and allocate more feedback resources for phase feedback of \mathbf{H} .

Figure 3.1(c) illustrates our proposed new DLN framework, consisting of magnitude and phase branches. The gNB further contains a combining network to estimate

the full CSI based on results from magnitude and phase decoders. We optimize encoders, decoders, and combining network jointly by minimizing a single loss function for end-to-end learning during offline training. Note that the magnitude branch can be independently optimized. For ease of convergence [37], the training of the DLN has two stages. In stage-1, the CSI magnitude encoder/decoder branch is pre-trained for magnitude recovery. In stage-2, both the CSI phase branch and the combining network are optimized with the help of the magnitude branch, while the parameters of the magnitude branch are fixed.

3.1.1 DualNet-MP

We now present a new DLN called DualNet-MP. As shown in Figure 3.2, DualNet-MP splits each complex CSI matrix into

$$\mathbf{H} = |\mathbf{H}| \odot e^{j\angle\mathbf{H}},$$

where \odot represents Hadamard product. Denote the (m, n) -th entry of \mathbf{H} as $\mathbf{H}_{m,n} = |\mathbf{H}_{m,n}|e^{j\angle\mathbf{H}_{m,n}}$. The magnitude matrix $|\mathbf{H}|$ and consists of entries $|\mathbf{H}_{m,n}|$ and phase matrix $e^{j\angle\mathbf{H}}$ consists of entries $e^{j\angle\mathbf{H}_{m,n}}$.

Similar to [12], we forward the CSI magnitudes to the magnitude encoder network, including four 7×7 circular convolutional layers with 16, 8, 4, and 1 channels and activation functions. Given the circular characteristic of CSI matrices, we introduce circular convolutional layers to replace the traditional linear ones. Subsequently, a fully connected (FC) layer with $\lceil \text{CR}_{\text{MAG}} Q_t N_b \rceil$ elements is connected for dimension reduction after reshaping. CR_{MAG} denotes the magnitude compression ratio. The output of the FC layer is then fed into the quantization module, called the sum-of-sigmoid (SSQ) [12] to generate magnitude codewords for feedback.

At the gNB, a magnitude decoder uses received magnitude codewords and locally

available UL CSI magnitudes¹ as AMI to jointly decode the DL CSI magnitudes. The magnitude branch is first optimized by updating the network parameters Θ_{MAG}

$$\arg \min_{\Theta_{\text{en,MAG}}, \Theta_{\text{de,MAG}}} \left\{ \|\hat{\mathbf{H}} - \mathbf{H}\|_{\text{F}}^2 \right\} \quad (3.1)$$

to minimize the MSE of recovered MIMO CSI magnitude

$$|\hat{\mathbf{H}}| = f_{\text{de,MAG}}(f_{\text{en,MAG}}(|\mathbf{H}|, \Theta_{\text{en,MAG}}), \Theta_{\text{de,MAG}}, \mathbf{H}_{\text{UL}}), \quad (3.2)$$

in which subscripts en, de, UL, and MAG of the $f(\cdot)$ denote the encoder, decoder, UL, and magnitude branch, respectively. Additionally, Θ denotes DLN parameters.

For CSI recovery of MIMO channels, we are only interested in their wrapped phases (i.e., $\angle \mathbf{H}$). There is 1-to-1 relationship between a phase value ϕ and $(\cos(\phi), \text{sign}[\sin(\phi)])$. For these reasons, we propose to form a "cosine" matrix whose entries are cosines of entries from \mathbf{H} denoted by

$$\mathbf{Cos} = \cos(\angle \mathbf{H}). \quad (3.3a)$$

Denote entry $\mathbf{A}_{m,n} = \text{sign}[\sin[\angle(\mathbf{H}_{m,n})]]$. We further form a sign matrix

$$\mathbf{A} = [\mathbf{A}_{m,n}]. \quad (3.3b)$$

Thus $(\mathbf{Cos}, \mathbf{A})$ uniquely determines $\angle \mathbf{H}$.

Since \mathbf{Cos} matrix is real, we can adopt a phase encoder similar to the magnitude encoder. Let CR_{PHA} denotes the phase compression ratio. Each \mathbf{Cos} generates a $\lceil \text{CR}_{\text{PHA}} Q_t N_b \rceil$ -element codeword. Our DLN uses tanh activation function in each circular convolutional layer of the phase encoder to capture the underlying features of significant phases associated with large magnitudes. Upon completion of encoder training, the UE processes each CSI \mathbf{H} , and feeds back the CSI magnitude codeword $f_{\text{en,MAG}}(|\mathbf{H}|, \Theta_{\text{en,MAG}})$, the phase codeword $f_{\text{en,PHA}}(\mathbf{Cos}, \Theta_{\text{en,PHA}})$ and the sign matrix \mathbf{A} to gNB.

At the gNB receiver, the phase codeword $f_{\text{en,PHA}}(\mathbf{Cos}, \Theta_{\text{en,PHA}})$ and the feedback

¹UL CSI is estimated at the gNB and assumed to be perfectly estimated.

sign matrix \mathbf{A} are sent to the phase decoder with the tanh activation function as the last layer to constrain the entries of DL CSI cosine matrix $\widehat{\mathbf{C}\text{os}}$ within $[-1, 1]$. The magnitude codeword and side information are used by the magnitude decoder to obtain an estimated CSI magnitude matrix $|\widehat{\mathbf{H}}|$. Based on the relationship $\sin(\phi) = \text{sign}(\sin[\phi])\sqrt{1 - \cos^2(\phi)}$, we form

$$\widehat{\mathbf{S}\text{in}} = \mathbf{A} \odot (\mathbf{1} - \widehat{\mathbf{C}\text{os}} \odot \widehat{\mathbf{C}\text{os}})^{1/2}.$$

Therefore, we can directly generate a preliminary CSI estimate

$$\widehat{\mathbf{H}} = \left[|\widehat{\mathbf{H}}| \odot \widehat{\mathbf{C}\text{os}}, |\widehat{\mathbf{H}}| \odot \widehat{\mathbf{S}\text{in}} \right]$$

from locally available $\widehat{\mathbf{C}\text{os}}, \mathbf{A}, |\widehat{\mathbf{H}}|$. The combining network is trainable and can include two residual blocks containing four circular convolutional layers to refine the DL CSI matrix.

For end-to-end optimization, we apply the following training criterion herein

$$\underset{\Theta_{\text{en,PHA}}, \Theta_{\text{de,PHA}}, \Theta_{\text{C}}}{\text{minimize}} \left\{ \|\widehat{\mathbf{H}} - \mathbf{H}\|_{\text{F}}^2 \right\}, \quad (3.4)$$

to optimize the parameters $\Theta_{\text{en,PHA}}$ of phase encoder $f_{\text{en,PHA}}$ and parameters $\Theta_{\text{de,PHA}}$ of phase decoder $f_{\text{de,PHA}}$ to generate an estimate

$$\widehat{\mathbf{C}\text{os}} = f_{\text{de,PHA}}(f_{\text{en,PHA}}(\mathbf{C}\text{os}, \Theta_{\text{en,PHA}}), \Theta_{\text{de,PHA}}). \quad (3.5)$$

Using the same loss function (3.4), we also train the combining network f_{C} by optimizing parameters Θ_{C} to generate

$$\widehat{\mathbf{H}} = f_{\text{C}}(|\widehat{\mathbf{H}}|, \widehat{\mathbf{C}\text{os}}, \mathbf{A}, \Theta_{\text{C}}). \quad (3.6)$$

Since the training of the magnitude learning branch can be decoupled, our framework optimizes the entire architecture by minimizing the overall CSI MSE of (3.4). It is possible, however, to also partially incorporate the MSE of (3.4) to further refine the

magnitude DLN branch by adopting a slower learning rate.

3.1.2 Loss Function Redesign

Considering the MSE loss function, it may be intuitive to simply rewrite the loss function as follows:

$$\begin{aligned}
 \text{MSE}_0 &= \text{MSE}_{\text{CSI}}(|\widehat{\mathbf{H}}|, \angle\widehat{\mathbf{H}}) = \|\mathbf{H} - \widehat{\mathbf{H}}\|_{\text{F}}^2 \\
 &= \||\mathbf{H}| \odot \cos(\angle\mathbf{H}) - |\widehat{\mathbf{H}}| \odot \cos(\angle\widehat{\mathbf{H}})\|_{\text{F}}^2 \\
 &\quad + \||\mathbf{H}| \odot \sin(\angle\mathbf{H}) - |\widehat{\mathbf{H}}| \odot \sin(\angle\widehat{\mathbf{H}})\|_{\text{F}}^2.
 \end{aligned} \tag{3.7}$$

This means that $|\mathbf{H}|$ and $\angle\mathbf{H}$ are used as encoder network input variables whereas their estimates are the decoder network output variables. However, the presence of infinitely many and shallow local minima of sinusoidal functions $\sin(\cdot)$ and $\cos(\cdot)$ often lead to training difficulties [38]. To overcome this problem, the authors in [18] recently proposed a weighted MDPP loss function

$$\text{MSE}_{\text{MDPP}} = \||\angle\mathbf{H} - \angle\widehat{\mathbf{H}}| \odot |\mathbf{H}|\|_{\text{F}}^2 \tag{3.8}$$

which still uses $|\mathbf{H}|$ and $\angle\mathbf{H}$ as input and output variables. where $\angle\mathbf{H}$ and $\angle\widehat{\mathbf{H}}$ denote the true and estimated phases, respectively. By weighting the original phase discrepancy with the true CSI magnitude, this new loss function helps capture the underlying features of the critical phases associated with CSI coefficients with dominant magnitudes. However, the loss function is not equivalent to our final goal for minimizing MSE of DL CSI. We now propose a reparameterization of the same MSE loss function during training. Instead of changing the loss function, we can overcome the training problem of directly parameterization in Eq. (3.7). Instead, recognizing that only the wrapped phases of $\angle\mathbf{H}$ are of interest, we replace $\angle\mathbf{H}$ with \mathbf{Cos} and \mathbf{A} via the following

reparameterization:

$$\begin{aligned}
\text{MSE}_{\text{SMAPE}}(\widehat{\mathbf{H}}, \widehat{\mathbf{C}}\mathbf{os}, \mathbf{A}) &= \|\mathbf{H} - \widehat{\mathbf{H}}\|_{\text{F}}^2 \\
&= \|\mathbf{H} \odot \mathbf{C}\mathbf{os} - |\widehat{\mathbf{H}}| \odot \widehat{\mathbf{C}}\mathbf{os}\|_{\text{F}}^2 \\
&\quad + \|\mathbf{H} \odot \mathbf{S}\mathbf{in} - |\widehat{\mathbf{H}}| \odot \widehat{\mathbf{S}}\mathbf{in}\|_{\text{F}}^2,
\end{aligned} \tag{3.9}$$

where we have used the sign matrix \mathbf{A} feedback to generate

$$\mathbf{S}\mathbf{in} = \mathbf{A} \odot (\mathbf{1} - \mathbf{C}\mathbf{os} \odot \mathbf{C}\mathbf{os})^{1/2} \tag{3.10a}$$

$$\widehat{\mathbf{S}}\mathbf{in} = \mathbf{A} \odot (\mathbf{1} - \widehat{\mathbf{C}}\mathbf{os} \odot \widehat{\mathbf{C}}\mathbf{os})^{1/2}. \tag{3.10b}$$

This formulation saves about half the bandwidth by sending the sign matrix \mathbf{A} without encoding matrix $\mathbf{S}\mathbf{in}$.

Moreover, the sparsity of \mathbf{H} means that we only need to feed back partial entries of \mathbf{A} associated with a swath of entries with dominant magnitudes. If we define a reduction ratio R_s to further reduce feedback overhead². The total phase feedback overhead (in bits) is summarized as follows:

$$B_{\text{SMAPE}} = \text{CR}_{\text{PHA}}(K_{\text{PHA}}N_tN_b + R_sN_tN_b)(\text{bits}), \tag{3.11}$$

where K_{PHA} denotes the number of encoding bits for each entry of the compressed cosine matrix $f_{\text{en,PHA}}(\mathbf{C}\mathbf{os}, \Theta_{\text{en,PHA}})$.

To summarize our training strategy of DualNet-MP, we use Eq. (3.1) as the loss function during the first training stage. In the second training stage, we used Eqs. (3.9) as the loss function to build an end-to-end learning architecture.

²Usually, the reconstruction performance can remain approximately the same even if the sign ratio R_s is less than 0.25 due to the sparsity.

3.2 Experimental Evaluations

3.2.1 Experiment Setup

In our experiments, we let the UL and DL bandwidths be 20 MHz and the subcarrier number be $N_f = 1024$. We consider both indoor and outdoor cases. We place the gNB with a height of 20 m at the center of a circular cell coverage with a radius of 20 m for indoor and 200 m for outdoor. The number of gNB ULA antennas is $N_b = 32$ whereas each UE has a single antenna. A half-wavelength inter-antenna spacing is considered. For each trained model, the number of epochs and batch size were set to 1,000 and 200, respectively. We generate two datasets consisting of 100,000 random channels for both indoor and outdoor cases from two different channel models. 60,000 and 20,000 random channels are for training and validation. The remaining 20,000 random channels are test data for performance evaluation.

In the first dataset (indoor), we used the industry-model COST 2100 [39] to generate indoor channels at 5.1-GHz UL and 5.3-GHz DL. We generate a second dataset (outdoor) using the QuaDRiGa method, described in 3GPP TR 38.901 [40]. For the outdoor dataset, We consider the urban microcell (UMi) scenario at 2 and 2.1 GHz of UL and DL bands, respectively, without line-of-sight (LOS) paths. The number of cluster paths was set as 13. For more detailed data generation settings, please refer to the preprint version [37]

In the following section, we evaluate the performance of CSI recovery by adopting the proposed optimization method and encoder/decoder architecture. Thus, we trained DualNet-MP with the same core network design for magnitude recovery. However, we test different methods to reconstruct the CSI phases for two phase compression ratios of $CR_{\text{PHA}} = 1/8$ and $1/16$ ³:

- SMAPE: the network architecture follows DualNet-MP. The sign ratio R_s varies

³All alternate approaches consume 1.2 and 0.625 bits/phase entry

between $[0.25, 0.125]$ and we use $K_{\text{PHA}} = 8$ bits for both $\text{CR}_{\text{PHA}} = \{1/8, 1/16\}$.

- MDPQ [13]: the design assigns $[0, 0, 0, 3, 7]$ and $[0, 0, 0, 0, 5]$ bits for $\text{CR}_{\text{PHA}} = [1/8, 1/16]$, respectively, to encode the CSI phases corresponding to $[0, 0.5, 0.7, 0.8, 0.9]$ of the cumulative distribution of CSI magnitude.
- MSE_0 : instead of cosine, CSI phases are fed directly to the phase encoder. Both cosine and sine functions are appended as the final layer of the phase decoder. The loss function for phase reconstruction is given by Eq.(3.7). We set K_{PHA} to 8 bits.
- MDPP [18]: we reuse the loss function Eq.(3.8) with the same network architecture. We set K_{PHA} to 10 bits.

Detailed setting about the alternatives can be found in [37].

3.2.2 Different Phase Compression Designs

To demonstrate the superiority of the proposed SMAPE loss function, we applied different phase reconstruction approaches to DualNet-MP for different phase compression ratios CR_{PHA} . Figures 3.3 (a) and (b) show the NMSE performance of different approaches under indoor and outdoor scenarios, respectively, at different compression ratios. As expected, DaulNet-MP encounters training difficulties when using the simple loss function MSE_0 . By adopting MDPP loss functions, DualNet-MP performs much better than the simple loss function Loss_0 . Although DualNet-MP appears to be better when using MDPQ instead of MDPP, encoding bit-assignment require careful tuning to achieve a satisfactory result. Finally, DualNet-MP based on the proposed SMAPE loss function achieves 4-dB performance improvement in terms of NMSE reduction for $\text{CR}_{\text{PHA}}=1/8$ at outdoor and 7-dB improvement for $\text{CR}_{\text{PHA}}=1/8$ at outdoor.

3.2.3 Different Core Layer Designs

To investigate the appropriate core layer designs of DualNet-MP in order to efficiently extract the underlying features of CSI phases, we provide a performance evaluation using FC, linear convolutional, and circular convolutional layers, respectively, for the core network. Denoted respectively as DNN, CNN and C-CNN, these networks adopted SSQ [12] and binary-level quantization (BLQ) as the quantization module at the encoder. Denote that the DNN design follows the recent work [18]. CNN and C-CNN design follow the DualNet-MP without and with utilization of circular convolutional layers, respectively. We consider the phase compression ratio of $CR_{\text{PHA}} = 1/8$. For SSQ, we assign $K_{\text{PHA}} = 8$ bits for each codeword. That is, there are $CR_{\text{PHA}}Q_tN_b = 128$ 8-bit codewords sent to the gNB. In contrast, there are $K_{\text{PHA}}CR_{\text{PHA}}Q_tN_b = 1024$ 1-bit codewords when applying BLQ.

Figures 3.4.(a) and (b) show the NMSE performance for the considered core layer designs. For both indoor and outdoor scenarios, DualNet-MP demonstrates superiority when adopting SSQ and C-CNN, which can be attributed to two possible reasons. Firstly, unlike BLQ, SSQ is differentiable such that it is easier to train. Secondly, there are many structural and circular features of CSI phases in the angle-delay domain that can be extracted better with the proposed structural changes.

In terms of storage and complexity of the proposed architecture, we note that C-CNN with only 826K parameters is considerably simpler than DNN requiring 11.6M parameters, whereas required floating point operations are comparable. As a result, we find the proposed new DualNet-MP architecture that combines SSQ and C-CNN delivers both performance advantages and cost benefits.

3.3 Conclusions

This work presents a new deep-learning framework for large scale CSI estimation that leverages feedback compression and auxiliary CSI magnitude information in FDD systems. Utilizing strong domain knowledge in deep-learning for CSI estimation to overcome known training issues, our new framework provides a novel loss function to enable efficient end-to-end learning and improves CSI recovery performance. We further exploit the circular characteristics of the underlying CSI in DA domain to propose an innovative circular convolution neural network (C-CNN). Our test results reveal significant improvement of overall CSI recovery performance for both indoor and outdoor scenarios and complexity reduction in comparison with a number of published alternative deep-learning compression designs for MIMO CSI feedback.

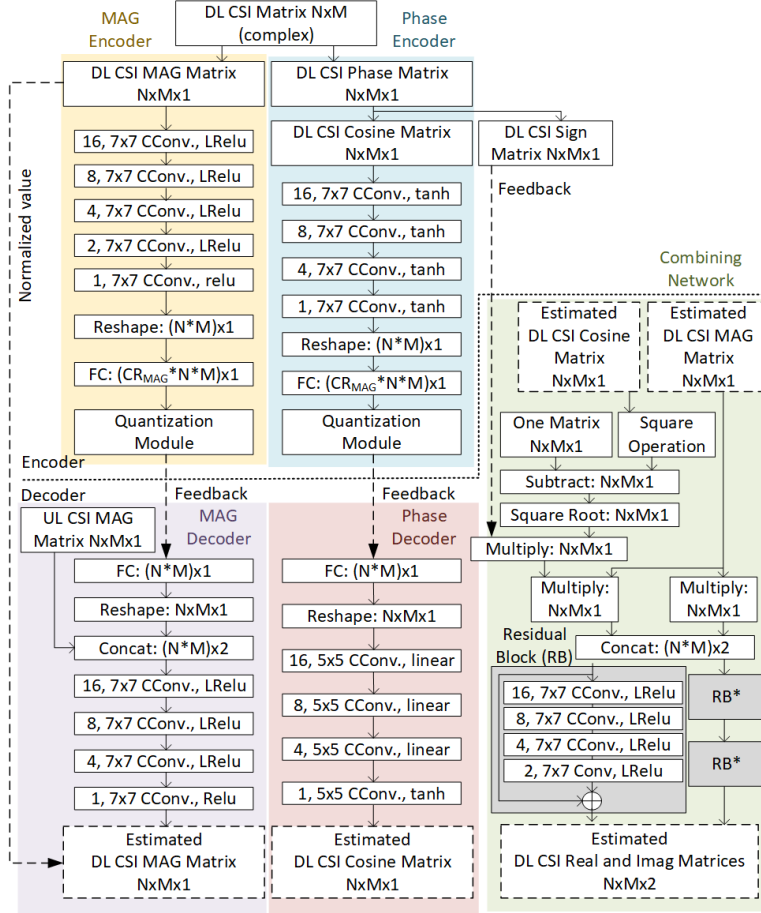


Figure 3.2: Network architecture of DualNet-MAG-PHA.

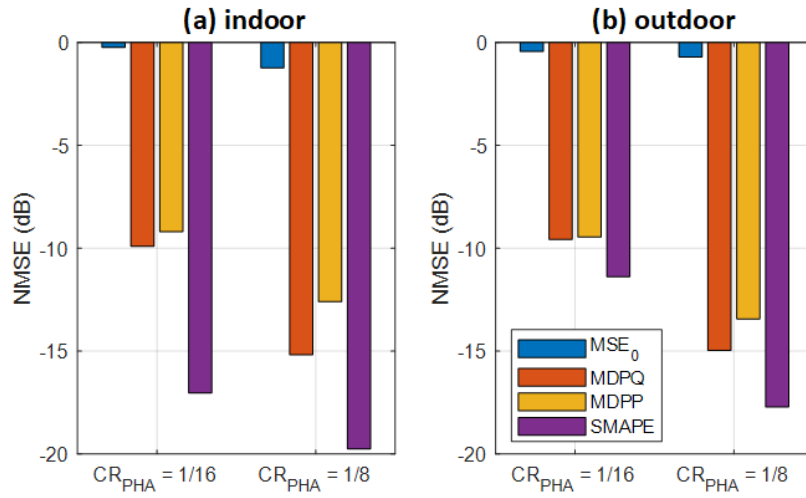


Figure 3.3: NMSE performance for different loss functions in (a) indoor and (b) outdoor scenarios.

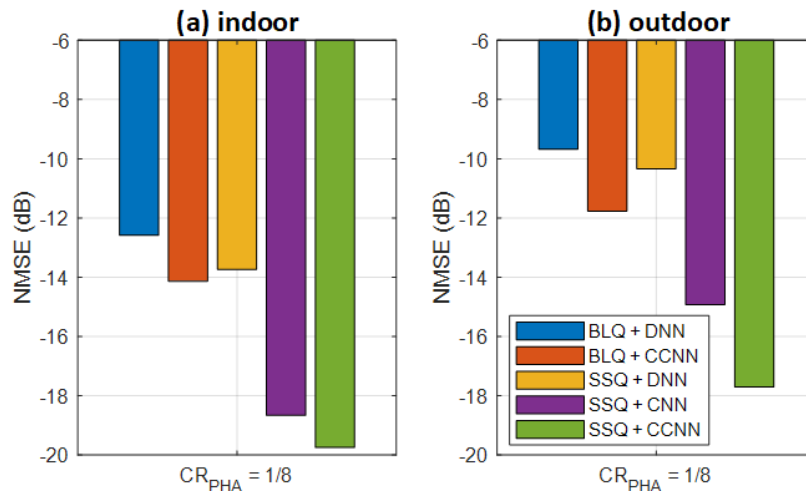


Figure 3.4: NMSE performance for different core layer designs in (a) indoor and (b) outdoor scenarios.

Chapter 4

Exploiting Partial FDD Reciprocity for Beam Based Pilot Precoding and CSI Feedback in Deep Learning

Massive MIMO systems can achieve high spectrum and energy efficiency in DL based on accurate estimate of CSI. Existing works have developed learning-based DL CSI estimation that lowers UL feedback overhead. One often overlooked problem is the limited number of DL pilots available for CSI estimation. One proposed solution leverages temporal CSI coherence by utilizing past CSI estimates and only sending CSI-RS for partial arrays to preserve CSI recovery performance. Exploiting CSI correlations, FDD channel reciprocity is helpful to base stations with direct access to UL CSI. In this work, we propose a new learning-based feedback architecture and a reconfigurable CSI-RS placement scheme to reduce DL CSI training overhead and to improve encoding efficiency of CSI feedback.

In this chapter, we first describe a beam-space (BS) precoding and CSI feedback scheme for pilot reduction in Section 4.1. In Section 4.2, we proposed a neural network, BSdualNet₀, to recover low-dimensional BS CSIs according to inter-beam correlation.

Then, we further formulate a criterion to find the *optimal beam merging matrix* for compact representation of DL CSIs and the *approximation mapping function* to recover full DL CSIs. In Section 4.3, we propose two neural networks, BSdualNet and BSdualNet-MN, to find the beam merging matrix and the approximation function. In Section 4.4, to further reduce DL CSI training and UL feedback overhead, we propose a CSI feedback framework with a reconfigurable CSI-RS placement by exploiting the sparsity and smoothness in beam and frequency domains, respectively. In Section 4.5, the numerical results demonstrate the proposed frameworks provide better CSI recovery performance while maintaining a descent complexity and storage requirement advantages under different compression ratios.

4.1 BS Precoding for CSI-RS Reduction

In FDD systems, UEs estimate DL CSIs and feedback to the serving base station after quantization. Then, gNB recovers the DL CSI based on the feedback. For brevity, we use $\tilde{\mathbf{h}}_{\text{DL}}$ and $\hat{\mathbf{h}}_{\text{DL}}$ in the following section to represent the estimated CSI obtained at UE and gNB, respectively. Focusing on a specific UE, the DL subband consists of κ RBs within the bandwidth for CSI-RS placement. We assume channels within an RB to be under slow, flat and block fading. As shown in Figure 4.1, there are $N_f \times N_o$ time-frequency REs in a specific RB (N_f subcarriers and N_o OFDM symbols). Since the same processing procedures are applied for every RB, without loss of generality, we only discuss the processing in a single RB in this section.

4.1.1 DL CSI recovery

Given that the gNB assigns $L = N_b$ REs for DL CSI training for N_b antennas, the received signal vector $\mathbf{y}_{\text{DL}} \in \mathbb{C}^{N_b \times 1}$ at UE can be expressed as

$$\mathbf{y}_{\text{DL}} = \mathbf{S}_{\text{DL}, N_b} \cdot \mathbf{h}_{\text{DL}} + \mathbf{n}_{\text{DL}}, \quad (4.1)$$

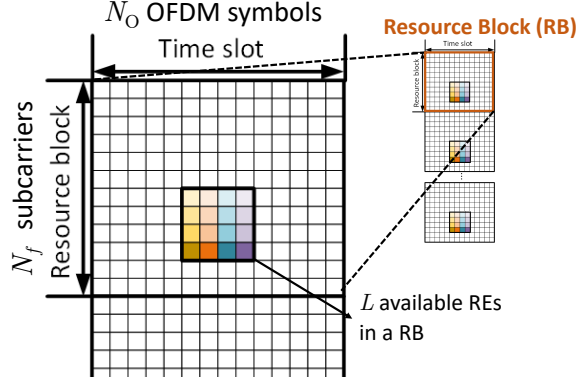


Figure 4.1: Resource block configuration. There are $N_f \times N_O$ time-frequency REs in a RB (N_f subcarriers and N_O OFDM symbols). Pilots are allowed to be placed at those REs in designated region (surrounded by the black frame). There are $L = 16$ available REs for pilot placement in this illustration.

where $\mathbf{h}_{\text{DL}} = \text{vec}(\mathbf{H}_{\text{DL}}) \in \mathbb{C}^{N_b \times 1}$ denotes the DL CSI vector whereas $\mathbf{S}_{\text{DL}, N_b} = \text{diag}(\mathbf{s}_{\text{DL}}) \in \mathbb{C}^{N_b \times N_b}$ denotes the CSI-RS training symbol matrix which is diagonal matrix with diagonal entries of training symbols $s_{\text{DL}}^{(n)} \neq 0, n = 1, \dots, N_b$. $\mathbf{n}_{\text{DL}} \in \mathbb{C}^{N_b \times 1}$ denotes the additive noise. $\mathbf{H}_{\text{DL}} \in \mathbb{C}^{N_H \times N_V}$ denotes the DL CSI matrix before reshaping. With the assumption of perfect channel estimation, from known training symbols in $\mathbf{S}_{\text{DL}, N_b}$, the UE can estimate its DL CSI for feedback to gNB via $\tilde{\mathbf{h}}_{\text{DL}} \approx \mathbf{h}_{\text{DL}}$.

4.1.2 Single-beam BS Precoding and DL CSI recovery

Existing wireless systems [23, 41] have applied beamforming/precoding techniques to CSI-RS symbols for beam selection, DL CSI estimation, or resistance to attenuation in high frequencies. According to [42], we can find N_b orthogonal beams to construct an unitary “orthogonal beam matrix (OBM)” $\mathbf{B} = [\mathbf{b}^{(1)} \mathbf{b}^{(2)} \dots \mathbf{b}^{(N_b)}]$. As shown in Figure 4.2.A, applying the OBM to the CSI-RS matrix $\mathbf{S}_{\text{DL}, N_b}$ in the digital beamforming module, the UE receives signals at different REs:

$$\mathbf{y}_{\text{DL}} = \mathbf{S}_{\text{DL}, N_b} \mathbf{B}^T \mathbf{h}_{\text{DL}} + \mathbf{n}_{\text{DL}}. \quad (4.2)$$

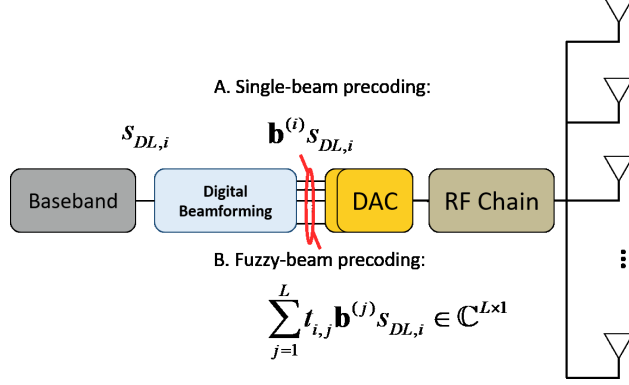


Figure 4.2: Signal processing flow for BS precoding

From the orthogonality of the OBM, DL CSI can be reconstructed at the gNB from the BS quantized UE feedback $\bar{\mathbf{g}}_{\text{B}} = Q(\tilde{\mathbf{h}}_{\text{BS,DL}} \approx \mathbf{B}^T \mathbf{h}_{\text{DL}}) \in \mathbb{C}^{N_b}$ via $\hat{\mathbf{h}}_{\text{DL}} = \mathbf{B}^* \bar{\mathbf{g}}_{\text{B}} \in \mathbb{C}^{N_b}$ where $Q(\cdot)$ denotes a differentiable quantizer. More details about $Q(\cdot)$ can be found in [12].

Given the angular sparsity of DL CSIs, especially for DL CSIs in LOS scenarios, the beam space (BS) DL CSI $\mathbf{h}_{\text{BS,DL}} (= \mathbf{B}^T \mathbf{h}_{\text{DL}})$ can be assumed as a L -sparse vector and thus DL CSI \mathbf{h}_{DL} can be approximated according to the most significant L ($L < N_b$) beams as follows:

$$\hat{\mathbf{h}}_{\text{DL}} = \mathbf{B}_{\text{S}}^* \bar{\mathbf{g}}_{\text{B,S}} \quad (4.3)$$

where $\mathbf{B}_{\text{S}} \in \mathbb{C}^{N_b \times L}$ and $\bar{\mathbf{g}}_{\text{B,S}} \in \mathbb{C}^{L \times 1}$ respectively denote the significant beam matrix consisting of the steering vectors of the most significant L orthogonal beams, and the corresponding quantized beam responses¹. Relying on L significant beams, the gNB only need to assign L ($< N_b$) REs for CSI-RS in DL to reduce UL feedback. We denote the heuristic DL CSI approximation approaches as BS-DL and BS-UL when using DL CSI and UL CSI to obtain the significant beam matrix \mathbf{B}_{S} , respectively. Note that BS-DL is an ideal approach with the assumption that we already have perfect DL CSI.

Typically, the S significant beams could be found through beam training or direction

¹According to [43], in propagation channels with low angular spread, only few significant beams contribute to most DL CSI energy in beam domain. This is also shown in Table 4.1

finding [43,44] by utilizing additional bandwidth and power resources. Fortunately, the FDD UL/DL reciprocity in magnitudes of angular CSI [13] can help gNB implement this beam selection process by relying the available UL CSI at gNB.

4.1.3 Single-beam Precoding and AI-aided DL CSI Recovery

To find the optimal serving beam in the beam management task, the work [43] efficiently derived a fine-resolution beam response map by only scanning few "eigen-beams" with the help of deep-learning network according to the correlation between vertically and horizontally adjacent beam responses. Reference [43] shows that it is possible to recover the full BS CSI via few significant beam responses. Equally important is the fact that UL CSI magnitudes are highly correlated to DL CSI magnitude in beam space so that gNB can find significant DL beams based on locally available UL CSIs.

Leveraging these insights, we first develop a heuristic CSI feedback framework, *BSdualNet₀*. As shown in Figure 3, the *BSdualNet₀* consists of three phases:

- **UL-CSI aided significant beam selection:** the gNB selects L beams containing the largest UL beam response magnitudes (i.e., $|\mathbf{H}_{\text{BS,UL}}|$) as significant beams. Next, the gNB applies the L significant beams to training symbols on L REs for CSI-RS transmission to UEs. We denote the index set of these beams as Ω_{B} .
- **Beam response feedback:** the UE estimates the beam responses $\tilde{\mathbf{h}}_{\text{BS,DL}}$ via well known channel estimation methods for direct encoding and feedback quantized beam responses $\bar{\mathbf{g}}_{\text{FB}}$ to the gNB.
- **Beam response refinement:** the gNB first generates a sparse map \mathbf{M} filled with quantized beam responses $\bar{\mathbf{g}}_{\text{FB}}$ and zeros as initial BS DL CSI estimate according to the index set of the selected beams Ω_{B} . The sparse map \mathbf{M} and local UL

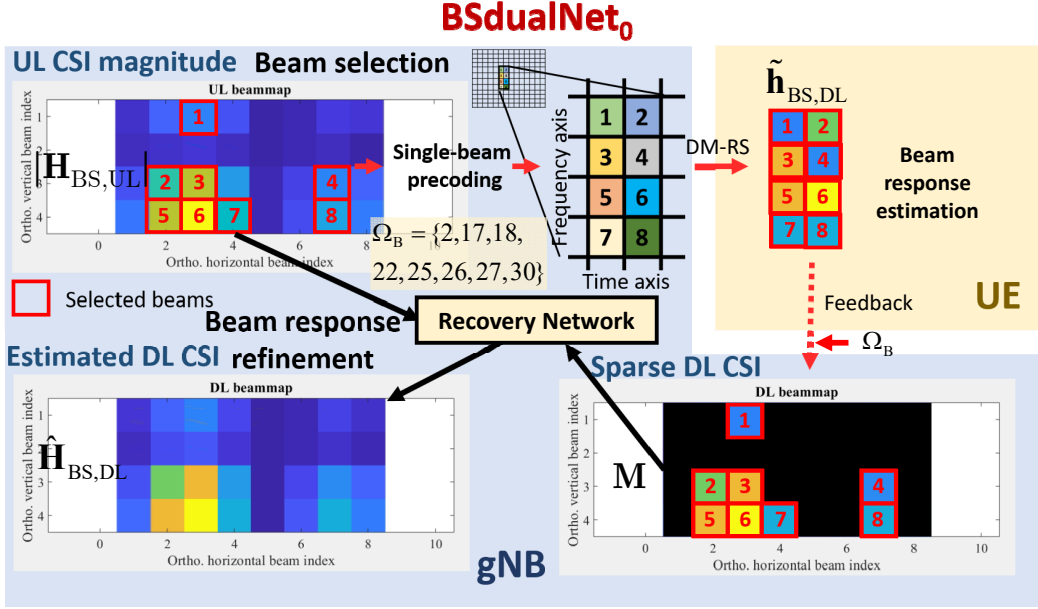


Figure 4.3: Illustration of BSdualNet₀. In this illustration, we consider a gNB equipped with an 8×4 UPA communicating with a single-antenna UE. The number of REs, L , is set to 8. This means that 8 significant beams are chosen according to BS UL CSI $|\mathbf{H}_{UL}|$ for precoding the CSI-RS and recorded in a chosen beam index set Ω_B . The UE obtains the BS DL CSI estimate $\tilde{\mathbf{h}}_{BS,DL}$ and feeds it back to the serving gNB after CSI quantization. Next, the gNB forms a sparse map \mathbf{M} according to the beam response feedback $\tilde{\mathbf{g}}_{FB}$ and the chosen beam index set Ω_B , and recovers the full BS DL CSI according to the sparse map \mathbf{M} and BS UL CSI magnitudes $|\mathbf{H}_{UL}|$.

CSI magnitudes $|\mathbf{H}_{UL}|$ form inputs to a deep learning network for estimating the missing elements in the sparse map for DL CSI refinement. The CNN generates refined DL beam domain CSI $\hat{\mathbf{H}}_{DL}$.

In this BS DL CSI recovery framework, the gNB assigns L orthogonal beams to L REs and recovers the full BS DL CSI based on the feedback of the L beam responses from UEs via correlation between adjacent beam responses.

4.1.4 Fuzzy-beam Precoding and AI-aided DL CSI Recovery

We also develop a BS DL CSI recovery framework which assigns all orthogonal beams to L REs ($L < N_b$). Instead of utilizing a single beam for each RE (Figure 4.2.A), as shown in Figure 4.2.B, a combination of weighted beams is applied. Let us denote an

$L \times N_b$ beam merging matrix

$$\mathbf{T} = \begin{bmatrix} \mathbf{t}_1^T \\ \mathbf{t}_2^T \\ \vdots \\ \mathbf{t}_L^T \end{bmatrix} \in \mathbb{C}^{L \times N_b}, \quad \mathbf{t}_i = \begin{bmatrix} t_{1,i} \\ \vdots \\ t_{N_b,i} \end{bmatrix}. \quad (4.4)$$

The received signal vector at UE is expressed as

$$\begin{aligned} \mathbf{y}_{\text{DL}} &= \begin{bmatrix} \sum_{i=1}^{N_b} t_{i,1} \mathbf{h}_{\text{DL}}^T \mathbf{b}^{(i)} s_{\text{DL}}^{(1)} \\ \sum_{i=0}^{N_b-1} t_{i,2} \mathbf{h}_{\text{DL}}^T \mathbf{b}^{(i)} s_{\text{DL}}^{(2)} \\ \vdots \\ \sum_{i=0}^{N_b-1} t_{i,L} \mathbf{h}_{\text{DL}}^T \mathbf{b}^{(i)} s_{\text{DL}}^{(L)} \end{bmatrix} + \mathbf{n}_{\text{DL}} = \begin{bmatrix} \mathbf{h}_{\text{DL}}^T \mathbf{B} \mathbf{t}_1^* s_{\text{DL}}^{(1)} \\ \mathbf{h}_{\text{DL}}^T \mathbf{B} \mathbf{t}_2^* s_{\text{DL}}^{(2)} \\ \vdots \\ \mathbf{h}_{\text{DL}}^T \mathbf{B} \mathbf{t}_L^* s_{\text{DL}}^{(L)} \end{bmatrix} + \mathbf{n}_{\text{DL}} \\ &= \mathbf{S}_{\text{DL},L} \mathbf{T} \mathbf{B}^T \mathbf{h}_{\text{DL}} + \mathbf{n}_{\text{DL}} = \mathbf{S}_{\text{DL},L} \mathbf{T} \mathbf{h}_{\text{BS,DL}} + \mathbf{n}_{\text{DL}}, \end{aligned} \quad (4.5)$$

where \mathbf{T} is used to reduce the required REs and to find a compact representation of DL CSI. $\mathbf{h}_{\text{BS,DL}} = \mathbf{B}^T \mathbf{h}_{\text{DL}}$ denotes the DL CSI vector in beam domain. Since the recovery loss mainly attributes to the quantization and compression error instead of CSI estimation discrepancy, we adopt a common assumption without loss of generality in our benchmarks [5,9,14,17] that UEs provide perfect channel estimation for simplicity. The raw and quantized response vectors of the merged beam responses are denoted by $\mathbf{g}_{\text{FB}} = \mathbf{T} \mathbf{h}_{\text{BS,DL}}$ and $\bar{\mathbf{g}}_{\text{FB}} = Q(\mathbf{g}_{\text{FB}})$, respectively.

Our goal is to find a beam merging matrix $\mathbf{T} \in \mathbb{C}^{N_b \times L}$ and a mapping function f_{re} for recovering the DL CSI based on the quantized feedback vector via the principle of

$$\arg \min_{\mathbf{T}, \Omega_{\text{re}}} \|\mathbf{B}^* f_{\text{re}}(Q(\mathbf{T} \mathbf{h}_{\text{BS,DL}})) - \mathbf{h}_{\text{DL}}\|_2^2 \quad (4.6)$$

where Ω_{re} denotes the deep learning model parameters to be optimized. Following this principle, the detailed design and architecture of an UL CSI-aided feedback framework

for DL CSI estimation will follow in the next section. Since the encoding function $Q(\cdot)$ [12] and each layer of the combining network are differentiable, the deep model parameters Ω_{re} can be optimized through backpropagation.

4.2 Encoder-Free CSI Feedback with UL CSI Assistance

In last section, to optimize the CSI recovery discrepancy, an efficient beam merging matrix and a recovery mapping function need to be found. In this section, we start with the general architecture of the two proposed frameworks (*BSdualNet*, *BSdualNet-MN*) which characterize different designs of the mapping function f_{re} . Both exploit UL/DL reciprocity to design the beam merging matrix \mathbf{T} for dimension reduction but utilize different recovery schemes. Next we introduce detailed learning model objectives and design principle. Note that, existing learning-based frameworks often treat CSI compression as a black box with the help of DNN encoders deployed on the UEs, thereby imposing heavy memory and computation burden on low cost UEs. In fact, they require an even more complex DNN decoder to resolve the low-dimensional feedback. Instead, our new framework unloads CSI compression efforts of UEs by utilizing available beamforming hardware at gNB to lowers the required REs for CSI-RS of DL MIMO channels and reduces UL feedback overhead.

4.2.1 General Architecture

Consider a wireless communication system with L REs assigned in each RB for CSI-RS placement. For CSI feedback reduction, we first design a beam merging matrix \mathbf{T} to match N_b orthogonal beams with different weights to the L REs that carry CSI-RS. We use a beam merging network that use UL CSI magnitudes in beam domain as inputs. Owing to the high correlation between magnitudes of UL and DL CSIs in

beam domain, the beam merging network learn to assign suitable weights to orthogonal beams according to the corresponding BS UL CSI magnitudes $|\mathbf{B}^T \mathbf{h}_{\text{UL}}|$ that are locally available at gNB.

Next, we apply the beam merging matrix \mathbf{T} to L CSI-RS symbols on the L REs. The linear mapping matrix \mathbf{T} instead of a general or non-linear mapping function $f : \mathbb{C}^{N_b} \rightarrow \mathbb{C}^L$ for pilot dimension reduction provides the advantage of simpler implementation and easier decoupling of CSI-RS symbols. Consequently, *the effective channels at UEs after CSI estimation would be the weighted sum of beam responses as estimate of the full CSI at downlink*. Obtaining effective channels, the UE simply quantizes and feeds back the channel information to the gNB. The gNB recovers DL CSI by sending the quantized feedback and the known beam merging matrix \mathbf{T} into the proposed deep learning decoder network. For simplicity, Figure 4.4 shows the general architecture of the proposed CSI feedback framework for a single UE, though the same principle applies for multiple UEs.

Unlike previous works, our new framework does not require any encoder at UE to store and compress full DL CSI. This is beneficial to UE devices with limited computation, storage, and/or power resources. Moreover, we reduce the DL overhead of CSI-RS and provide higher spectrum efficiency while previous frameworks require REs proportional to its transmit antennas. The number of required REs for DL CSI training in our framework heavily depends on the sparsity in beam domain. Since the beam sparsity increases with larger array, this would bring more benefits in reducing DL CSI overhead when considering a large-scale transmit antenna array.

4.2.2 BSdualNet

Figure 4.5 shows the proposed CSI feedback framework, BSdualNet, in multi-user scenarios (i.e., N UEs). BSdualNet consists of three learning networks at gNB which serve on distinct objectives:

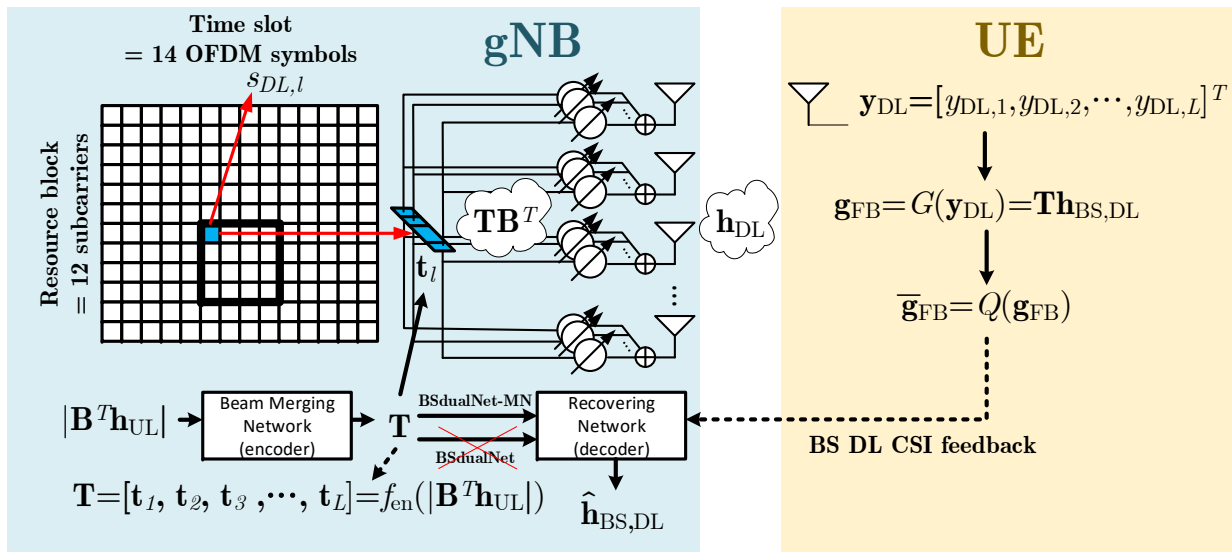


Figure 4.4: General architecture of the proposed BS CSI feedback framework. (Each of the small grids is a RE. The region covered by the bold black frame is the designated place for RS replacement. Thus, in this example, the available number of REs, L , is 16.)

- **Beam merging network:** it designs an matrix \mathbf{T} which is applied to DL CSI training for reducing the required REs and UL feedback overhead while maintaining accurate CSI recovery. With the aids of partial FDD UL/DL reciprocity, the beam merging matrix \mathbf{T} transforms effective BS CSI at UEs into a compressive representation.
- **Recovery network:** it estimates the full BS CSI according to the quantized estimated beam responses from UEs.
- **Combing network:** it refines the magnitudes of DL BS CSI by using the known magnitudes of UL BS CSI based on partial FDD UL/DL reciprocity.

As shown in Figure 4.6, we aggregate and reshape the magnitudes of BS UL CSIs of each UE $\mathbf{H}_{\text{BS,UL}}^{(i)} \in \mathbb{C}^{N_H \times N_V} = \text{reshape}(\mathbf{h}_{\text{BS,UL}}^{(i)}), i = 1, \dots, N$, into a tensor $|\mathcal{H}_{\text{BS,UL}}| \in \mathbb{C}^{N_H \times N_V \times N}$, which is sent to the beam merging network at gNB. The beam merging deep learning network (Figure 4.6) consists of four 3×3 circular convolutional layers with 16, 8, 4, and 2 channels, respectively, to learn the importance of different

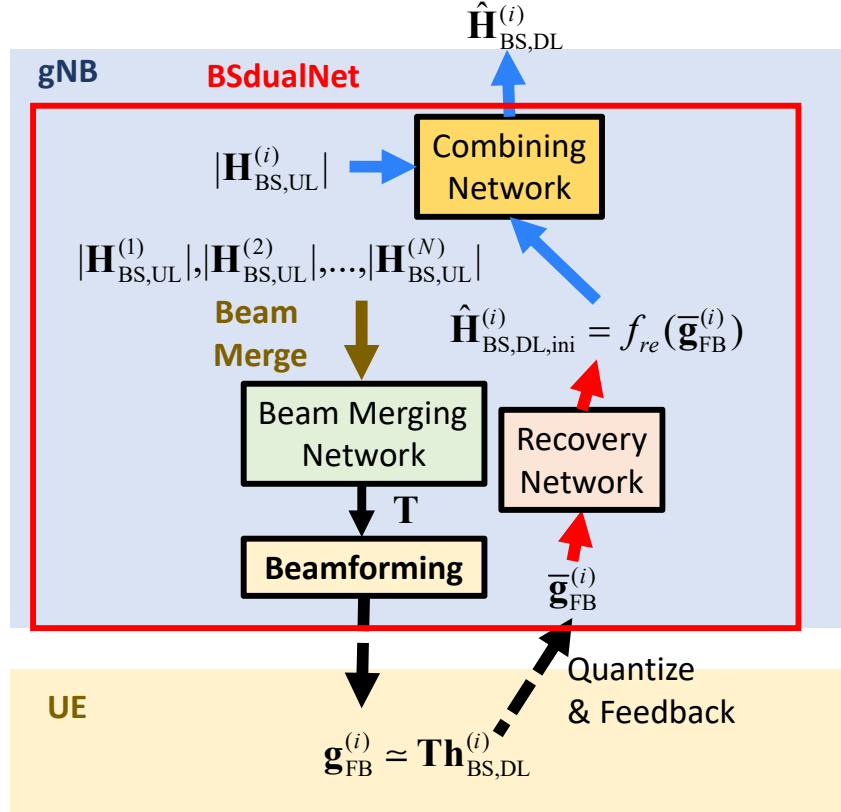


Figure 4.5: Block Diagram of BSdualNet. gNB designs beam merging matrix \mathbf{T} via the beam merging network, f_{bm} , according to locally available UL CSIs $\mathbf{H}_{\text{BS,UL}}^{(i)}$, $i = 1, \dots, N$ of N UEs and apply it to DL CSI estimation. UE estimates its effective BS CSI $\mathbf{g}_{\text{FB}}^{(i)}$ and feed it back to gNB after quantization. The gNB first obtains initial estimate of full BS DL CSI $\hat{\mathbf{H}}_{\text{BS,DL,ini}}^{(i)}$ by the recovery network f_{re} according to the quantized BS CSI. Then, it is refined as the final BS DL CSI $\hat{\mathbf{H}}_{\text{BS,DL}}^{(i)}$ via the combining network f_c with the knowledge of BS UL CSI magnitudes.

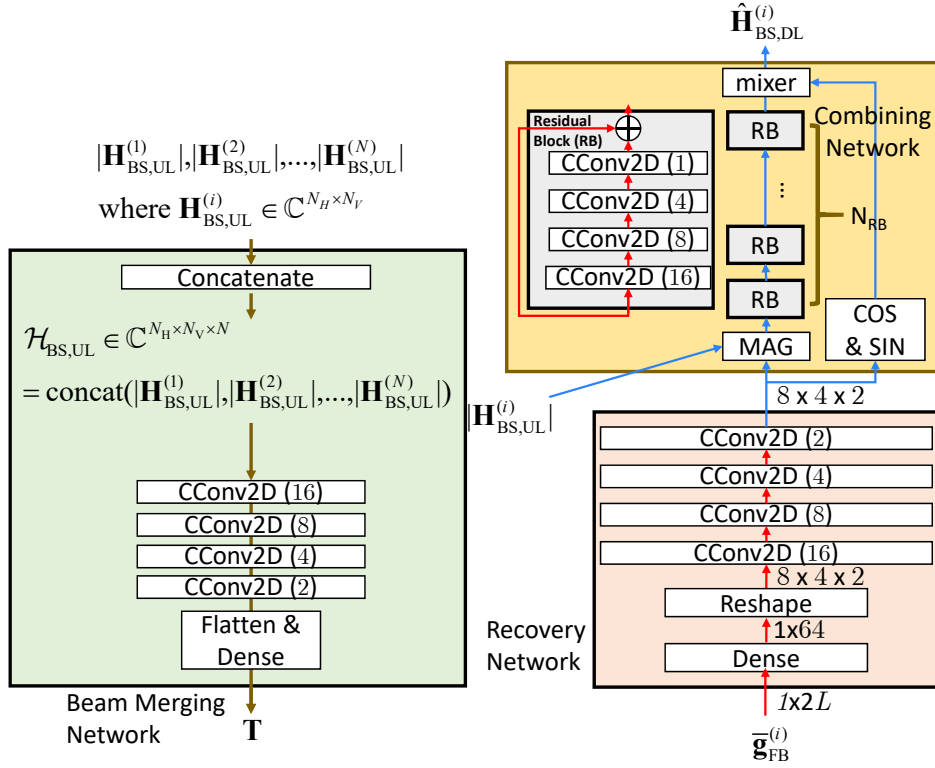


Figure 4.6: Network design of BSdualNet.

orthogonal beams according to the spatial structures of UL beam domain CSI magnitudes. Given the circular characteristic of BS CSI matrices, we introduce *circular convolutional layers* to replace traditional convolution. Subsequently, a fully connected (FC) layer with $2N_bL$ elements is included to generate desired dimension after reshaping (Recall that \mathbf{T} is a complex matrix with size of $N_b \times L$). After CSI estimation at UEs, the gNB receives the N copies of quantized feedbacks from N UEs and obtains quantized feedbacks $\bar{\mathbf{g}}_{\text{FB}}^{(i)} \in \mathbb{C}^{2L}$, $i = 1, 2, \dots, N$.

Now we focus on the network at gNB. For the i -th UE, we forward the received feedback $\bar{\mathbf{g}}_{\text{FB}}^{(i)}$ to a FC layer with $2N_b$ elements. After reshaping the feedback data into a matrix of size $N_H \times N_V \times 2$, we use four 3×3 circular convolutional layers with 16, 8, 4, and 2 channels and activation functions to generate initial BS DL CSI estimate $\hat{\mathbf{H}}_{\text{BS,DL,ini}}^{(i)} = f_{\text{re}}(\bar{\mathbf{g}}_{\text{FB}}^{(i)}) \in \mathbb{C}^{N_H \times N_V}$. Next, the gNB forwards the initial BS DL CSI estimate $\hat{\mathbf{H}}_{\text{BS,DL,ini}}^{(i)}$ together with the corresponding BS UL CSI magnitudes

$|\mathbf{H}_{\text{BS,UL}}^{(i)}|$ to the combining network for final DL CSI estimation $\widehat{\mathbf{H}}_{\text{BS,DL}}^{(i)}$. The combining network f_c uses N_B residual blocks, each block contains the same design of circular convolutional layers and activation functions as the network for DL CSI recovery.

Since all layers and quantization function are differentiable, the BSdualNet is optimized via backpropagation and gradient descent to update the network parameters Θ_{bm} , Θ_{re} and Θ_c of non-linear beam merging, recovery, and combining networks f_{bm} , f_{re} and f_c :

$$\arg \min_{\Theta_{\text{bm}}, \Theta_{\text{re}}, \Theta_c} \left\{ \sum_{i=0}^{N-1} \|\widehat{\mathbf{h}}_{\text{BS,DL}}^{(i)} - \mathbf{h}_{\text{BS,DL}}^{(i)}\|_2^2 \right\}, \quad (4.7)$$

where $\widehat{\mathbf{h}}_{\text{BS,DL}}^{(i)} = \text{vec}(\widehat{\mathbf{H}}_{\text{BS,DL}}^{(i)})$ and $\mathbf{h}_{\text{BS,DL}}^{(i)} = \text{vec}(\mathbf{H}_{\text{BS,DL}}^{(i)})$ denote the vectorized estimated and original BS DL CSIs where $\widehat{\mathbf{H}}_{\text{BS,DL}}^{(i)} = f_c(f_{\text{re}}(\bar{\mathbf{g}}_{\text{FB}}^{(i)}, |\mathbf{H}_{\text{BS,UL}}^{(i)}|)$. $\bar{\mathbf{g}}_{\text{FB}}^{(i)} = Q(\mathbf{T}\mathbf{h}_{\text{BS,DL}}^{(i)})$ denotes the quantized BS DL CSI and the beam merging matrix is given by $\mathbf{T} = f_{\text{bm}}(|\mathbf{H}_{\text{BS,UL}}^{(1)}|, |\mathbf{H}_{\text{BS,UL}}^{(2)}|, \dots, |\mathbf{H}_{\text{BS,UL}}^{(N)}|)$. Note that the superscript (i) denotes the UE index. $\mathbf{H}_{\text{BS,DL}}^{(i)}$ and $\mathbf{H}_{\text{BS,UL}}^{(i)} \in \mathbb{C}^{N_H \times N_V}$ denote original BS DL and UL CSIs at the i -th UE.

4.2.3 BSdualNet-MN

In BSdualNet, the beam merging network provides a beam merging matrix \mathbf{T} to generate an efficient representation of the convoluted responses of all orthogonal beams. Although \mathbf{T} is optimized for the ease of decoupling individual beam responses, the decoder remains a blackbox such that the information within \mathbf{T} may not be fully exploited due to its indirect use. In this section, we would redesign the decoder by directly using the beam merging matrix \mathbf{T} to achieve better architectural interpretability and performance improvement.

Unlike the previous works that split the deployment of CSI encoder and decoder at UEs and gNB, respectively, *our gNB knows the exact encoding and decoding processes* in our framework. Thus, we can exploit the locally known beam merging matrix \mathbf{T} to

decode the feedback more efficiently. To this end, we reformulate the problem of DL CSI recovery for $\widehat{\mathbf{h}}_{\text{BS,DL}}^{(i)}, i = 0, \dots, N - 1$ by seeking a minimum-norm solution to an under-determined linear system

$$\mathbf{y}_{\text{DL}}^{(i)} = \mathbf{T}\mathbf{h}_{\text{BS,DL}}^{(i)} + \mathbf{n}_{\text{DL}}^{(i)}, i = 0, \dots, N - 1.$$

As seen from Figure 4.7, the output of the recovery network can be expressed as follows:

$$f_{\text{re}}(\bar{\mathbf{g}}_{\text{FB}}^{(i)}) = \mathbf{T}^H(\mathbf{T}\mathbf{T}^H)^{-1}\bar{\mathbf{g}}_{\text{FB}}^{(i)}, \quad (4.8)$$

Clearly, the minimum norm solution depends on matrix \mathbf{T} . Assuming perfect quantization and zero noise, we can approximate the decoder (See Appendix) of Eq. (4.8) as

$$\begin{aligned} f_{\text{re}}(\tilde{\mathbf{g}}_{\text{FB}}^{(i)}) &= \widehat{\mathbf{h}}_{\text{BS,DL}}^{(i)} \approx \mathbf{T}^H(\mathbf{T}\mathbf{T}^H)^{-1}\mathbf{T}\mathbf{h}_{\text{BS,DL}}^{(i)}, \\ &= \underbrace{\sum_{i=1}^L \mathbf{v}_i \mathbf{v}_i^H}_{\tilde{\mathbf{I}}} \mathbf{h}_{\text{BS,DL}}^{(i)} = \tilde{\mathbf{I}} \cdot \mathbf{h}_{\text{BS,DL}}^{(i)}, \end{aligned} \quad (4.9)$$

where $\mathbf{v}_i, i = 1, 2, \dots, N_b$ are right singular vectors of \mathbf{T} . Since $\text{Trace}(\tilde{\mathbf{I}}) = L$, $\mathbf{h}_{\text{BS,DL}}^{(i)}$ cannot be fully recovered by only relying on the diagonal entries of $\tilde{\mathbf{I}}$. If strong spatial correlation exists in the beam domain, we will need a recovery matrix $\tilde{\mathbf{I}}$ with larger off-diagonal entries, representing the correlation between beams. Given the FDD UL/DL reciprocity in beam domain, by capturing the correlation between adjacent beam response magnitudes of UL CSI, it would be more reasonable to define a merging matrix \mathbf{T} which contains well-behaved right singular vectors such that $\sum_{i=0}^{N-1} \|\tilde{\mathbf{I}}\mathbf{h}_{\text{BS,DL}}^{(i)} - \mathbf{h}_{\text{BS,DL}}^{(i)}\|_2^2$ can be minimized.

With the same design of the beam merging network in BSdualNet, the recovery network in BSdualNet-MN simply includes a series of matrix products. Thus, BSdualNet-MN is not only more interpretable, its computational complexity and required model

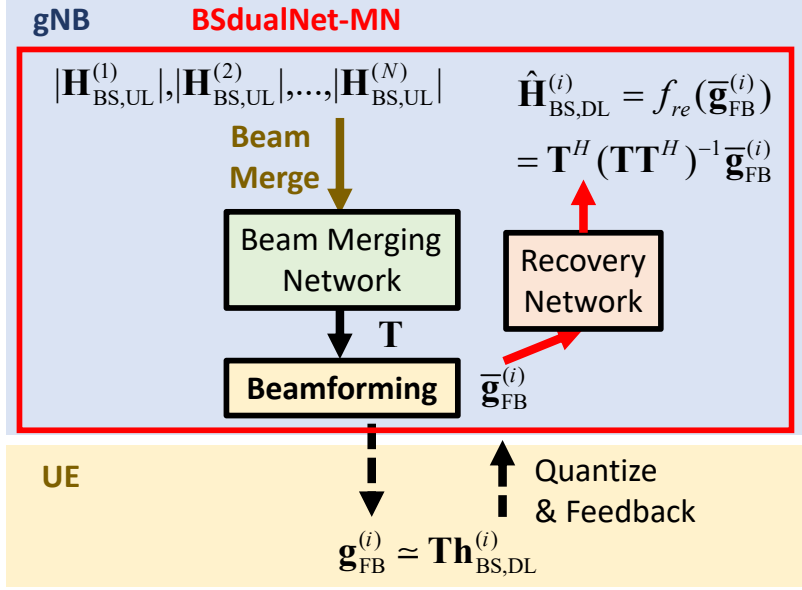


Figure 4.7: Block Diagram of BSdualNet-MN. The decoder design of BSdualNet-MN is different from BSdualNet. It directly applies the beam merging matrix \mathbf{T} to recover full BS DL CSI $\hat{\mathbf{H}}_{\text{BS,DL}}^{(i)}$.

memory are also lower.

4.3 UL CSI Aided Beam Based Precoding and a Reconfigurable CSI Feedback Frameworks

Generally, the aforementioned methods perform better with high sparsity CSI in beam domain. Yet, such spatial sparsity may not hold for CSI of every propagation channels. For example, indoor propagation channels tend to exhibit rich multi-paths with high angular spreads. This could lessen spatial sparsity and degrade recovery accuracy of DL CSI. Interestingly, however, such channels are alternatively characterized by large coherence bandwidth because of the dominance of low-delay paths dominate[45]. This means that for such channels, it is not necessary to have high CSI-RS density in frequency domain.

In this section, a reconfigurable CSI feedback framework will be described as a

more flexible solution to reduce the number of pilots by selecting frequency reduction (FR) and beam reduction (BR) ratios. Instead of regarding feedback of each RB independently, as discussed in the signal model of Section II, we exploit the large coherence bandwidth and consider a joint UL feedback for a total of K RBs. By leveraging spectral coherence, we can further reduce the UL feedback overhead by applying an autoencoder network. In what follows, we elaborate on the reconfiguration of CSI-RS placement and the design of a learning-based CSI feedback framework, BSdualNet-FR.

4.3.1 Frequency Resource Reconfiguration

In modern wireless protocols, there are designated resource regions for CSI-RS placement [23]. Compatible with existing RS configurations, we can reduce the CSI-RS placement density along the frequency domain by a frequency reduction factor FR by placing pilots only at RB indices $k = 1, 1 + FR, 1 + 2FR, \dots, 1 + (K/FR - 1)FR$ as shown in Figure 4.8. We can also further reduce the required REs by a beam reduction factor of $BR(= \text{round}(N_b/L))$ by applying beam merging matrix \mathbf{T} designed by using a three-dimensional (3-D) beam merging network with 3-D convolutional kernels as shown in Figures 4.9 and 4.10. Jointly, the total REs for CSI-RS placement can be reduced by a factor of $BR \cdot FR$. Thus, the total number of pilot REs becomes $N_b K / (BR \cdot FR)^2$.

The DL received signal vector $\mathbf{y}_{\text{DL}}^{(i,k)} \in \mathbb{C}^{L \times 1}$ at the i -th UE in the k -th RB can be expressed as

$$\mathbf{y}_{\text{DL}}^{(i,k)} = \mathbf{S}_{\text{DL},L}^{(k)} \mathbf{T} \mathbf{h}_{\text{BS,DL}}^{(i,k)} + \mathbf{n}_{\text{DL}}^{(k)}, \quad (4.10)$$

where the superscript (i, k) denotes the UE and RB indexes, respectively. Following Section II, UE- i estimates beam response vectors $\mathbf{g}_{\text{FB}}^{(i,k)}$, $k = 1, 1 + FR, \dots, 1 + (K/FR -$

²In previous proposed frameworks, the total REs for CSI-RS placement are reduced by a factor of BR . The total number of pilot REs for K RBs is $N_b K / BR$.

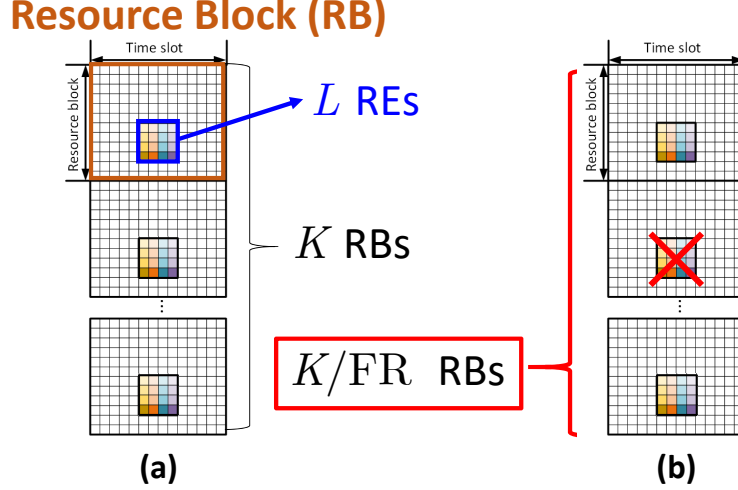


Figure 4.8: Illustration of pilot placement (a) before and (b) after reduction in frequency domain. Note that the color grids represent the designated REs in one of the pilot placement configurations defined in 5G specification [23]. The largest allowable L is 32 and FR can be 1 or 2 in legitimate pilot placement configurations. For example, $FR = 2$ means that one of every 2 consecutive RBs in the assigned bandwidth for CSI-RS is used for pilot placement. In this work, we assume FR can be any positive integer.

1) FR as a beam response matrix

$$\mathbf{G}_{\text{FB}}^{(i)} = \left[\mathbf{g}_{\text{FB}}^{(i,1)}, \mathbf{g}_{\text{FB}}^{(i,1+FR)}, \dots, \mathbf{g}_{\text{FB}}^{(i,1+(K/FR-1)FR)} \right] \in \mathbb{C}^{L \times K/FR}$$

where the estimates $\mathbf{g}_{\text{FB}}^{(i,k)} = \mathbf{T}\mathbf{h}_{\text{BS,DL}}^{(i,k)} \in \mathbb{C}^L$ are based on pilots reduced by FR .

4.3.2 BSdualNet-FR

For further reduction of UL feedback overhead, we compress the beam responses $\mathbf{G}_{\text{FB}}^{(i)}$ by implementing a frequency compression module (FCM) similar to an autoencoder. The FCM consists of an encoder at UE and decoder at gNB for CSI compression and recovery, respectively. The encoder consists of four 3×3 circular convolutional layers with 16, 8, 4 and 2 channels. Subsequently, an FC layer with $\lceil 2LK/(\text{CR} \cdot \text{FR}) \rceil$ elements accounts for dimension reduction by a factor of $\text{CR}_{\text{eff}} = \text{BR} \cdot \text{FR} \cdot \text{CR}$ after

reshaping. CR_{eff} and CR respectively denote the effective and feedback compression ratios. The FC layer output is sent to a quantization module which uses a trainable soft quantization function as proposed in [12] to generate feedback codewords.

At the gNB, the codewords from different UEs are forwarded to the FMC decoder network to recover their respective DL CSIs. The decoder first expands the dimension of the codewords to their original size of $2N_bK$. Reshaped into a size of $N_b \times K \times 2$, a codeword enters four 3×3 circular convolutional layers with 16, 8, 4 and 2 channels to generate the FCM output. Note that the dimensions in both the frequency and beam domains are already the same as our target output in this stage. The FCM output serves as an initial DL CSI estimate $\widehat{\mathbf{H}}_{\text{BS,DL,ini}}^{(i)} \in \mathbb{C}^{N_b \times K}$ which is used to calculate the first loss.

$$\text{loss}_1 = \sum_{i=0}^{N-1} \|\widehat{\mathbf{H}}_{\text{BS,DL,ini}}^{(i)} - \mathbf{H}_{\text{BS,DL}}^{(i)}\|_{\text{F}}^2, \quad (4.11)$$

$$\mathbf{H}_{\text{BS,DL}}^{(i)} = \begin{bmatrix} \mathbf{h}_{\text{BS,DL}}^{(i,1)} & \mathbf{h}_{\text{BS,DL}}^{(i,2)} & \cdots & \mathbf{h}_{\text{BS,DL}}^{(i,K)} \end{bmatrix} \quad (4.12)$$

$$\widehat{\mathbf{H}}_{\text{BS,DL,ini}}^{(i)} = f_{\text{FMC,de}}(f_{\text{FMC,en}}(\mathbf{G}_{\text{FB}}^{(i)})). \quad (4.13)$$

Next, the combining network refines the initial estimate with the help of UL CSI magnitudes. The combining network first split the magnitude and the phase of the initial estimate before sending the initial estimate magnitudes and the UL CSI magnitudes into five residual blocks which are constructed by a shortcut and four circular convolutional layers with 16, 8, 4, 2 and 1 channels and activation functions for magnitude refinement. From there, the refined magnitudes of DL CSI and their corresponding phases form the final output $\widehat{\mathbf{H}}_{\text{BS,DL}}^{(i)} \in \mathbb{C}^{N_b \times K}$ to determine the second loss function.

$$\text{loss}_2 = \sum_{i=0}^{N-1} \|\widehat{\mathbf{H}}_{\text{BS,DL}}^{(i)} - \mathbf{H}_{\text{BS,DL}}^{(i)}\|_{\text{F}}^2, \quad (4.14)$$

$$\mathbf{H}_{\text{BS,UL}}^{(i)} = \begin{bmatrix} \mathbf{h}_{\text{BS,UL}}^{(i,1)} & \mathbf{h}_{\text{BS,UL}}^{(i,2)} & \cdots & \mathbf{h}_{\text{BS,UL}}^{(i,K)} \end{bmatrix} \quad (4.15)$$

$$\widehat{\mathbf{H}}_{\text{BS,DL}}^{(i)} = f_c(\widehat{\mathbf{H}}_{\text{BS,DL,ini}}^{(i)}, |\mathbf{H}_{\text{BS,UL}}^{(i)}|), \quad (4.16)$$

The BSdualNet-FR is optimized by updating the network parameters Θ_{bm} , $\Theta_{\text{FMC,en}}$, $\Theta_{\text{FMC,de}}$ and Θ_c of the non-linear 3-D beam merging, FMC encoder/decoder, and combining networks f_{bm} , $f_{\text{FMC,en}}$, $f_{\text{FMC,de}}$ and f_c :

$$\arg \min_{\Theta_{\text{bm}}, \Theta_{\text{FMC,en}}, \Theta_{\text{FMC,de}}, \Theta_c} \{\alpha \cdot \text{loss}_1 + (1 - \alpha) \cdot \text{loss}_2\}$$

where hyperparameter α adjusts the weighting. Note that both loss_1 and loss_2 are differentiable.

Note that the deep learning network contains many hyperbolic tangent activation functions and a soft quantization function which could lead to the gradient vanishing problem for parameters in those layers. To mitigate this problem, we suggest a two-stage training scheme for optimizing the proposed framework. In the first stage, we train the model by setting $\alpha = 1$ for N_{first} epochs, freezing the combining network and focusing on finding the best beam merging matrix and encoding/decoding networks. In the second stage, we change $\alpha = 0.1$ and focus on refining the final estimates with the aid of UL CSI magnitudes. Using the elbow method [46], we found that $N_{\text{first}} = 30$ is usually sufficient to obtain a good tradeoff.

4.4 Experimental Evaluations

4.4.1 Experiment Setup

In our numerical test, we consider both indoor and outdoor cases. Using channel model software, we position a gNB of height equal to 20 m at the center of a circular cell with a radius of 30 m for indoor and 200 m for outdoor environment. We equip the

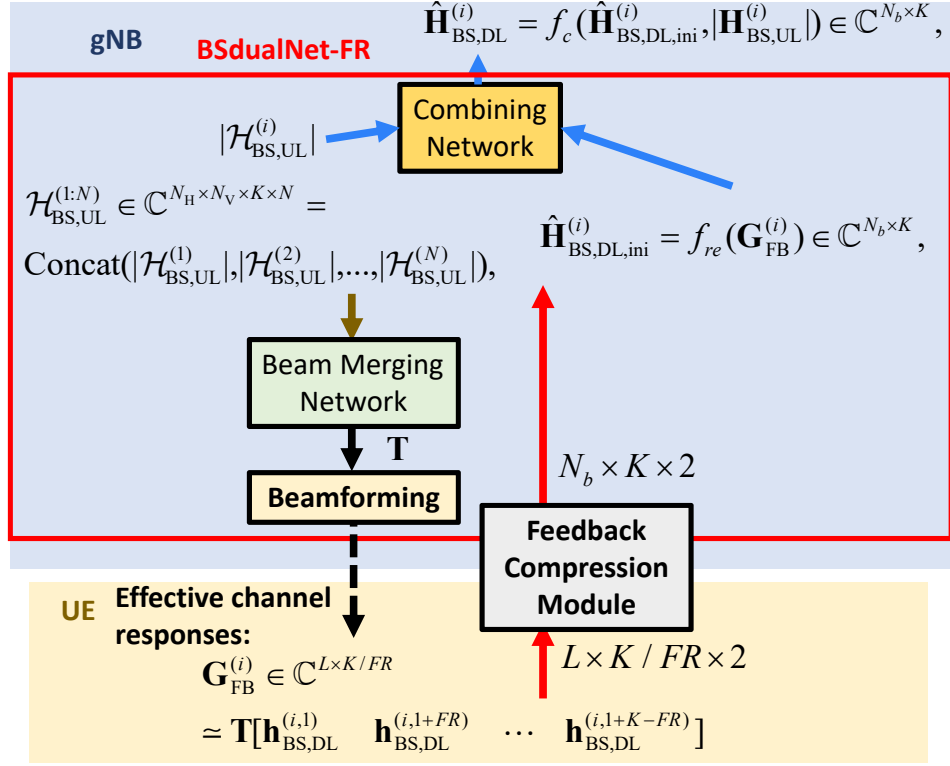


Figure 4.9: Block Diagram of BSdualNet-FR. The gNB designs beam merging matrix \mathbf{T} via the beam merging network, f_{bm} , according to locally available UL CSIs $\mathcal{H}_{\text{BS,UL}}^{(i)}$, $i = 1, \dots, N$ of N UEs over K RBs and applies it to DL CSI estimation. The i -th UE estimates its effective BS CSI $\mathbf{G}_{\text{FB}}^{(i)}$ and feeds it back to gNB after compression by the FMC encoder, $f_{\text{FCM,en}}$, and quantization. The gNB first obtains initial estimate $\hat{\mathbf{H}}_{\text{BS,DL,ini}}^{(i)}$ of full BS DL CSI by the FCM decoder, $f_{\text{FCM,de}}$, and refines it with the knowledge of BS UL CSI magnitudes via the combining network f_c . Note that $\mathbf{H}_{\text{BS,UL}}^{(i)} \in \mathbb{C}^{N_b \times K} = \text{reshape}(\mathcal{H}_{\text{BS,UL}}^{(i)} \in \mathbb{C}^{N_H \times N_V \times K})$.

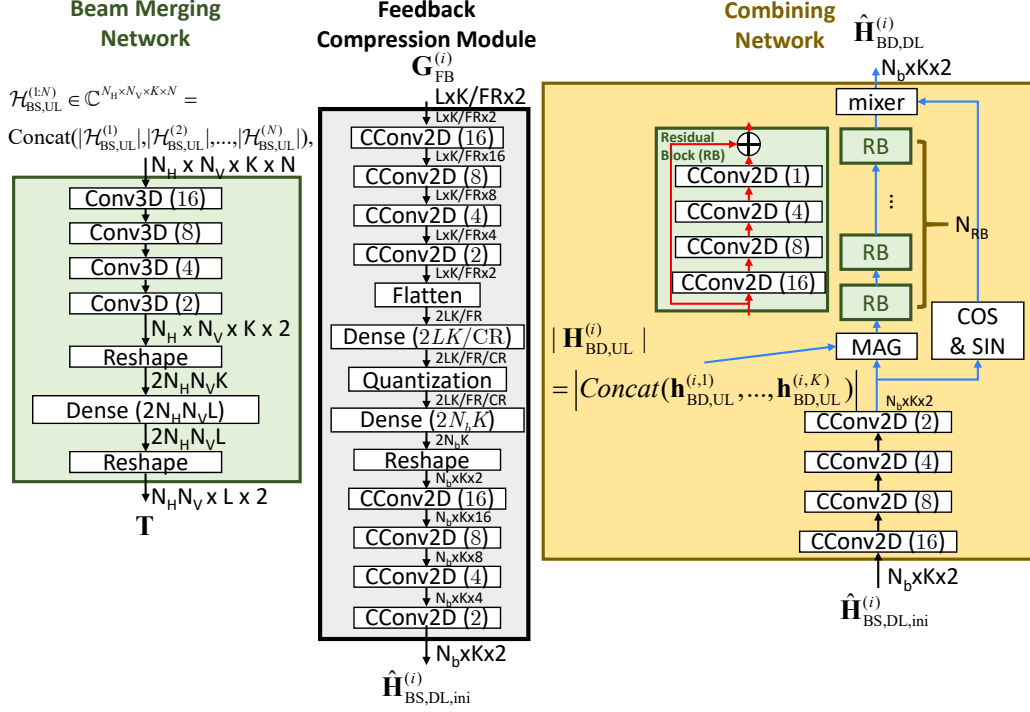


Figure 4.10: Network design of BSdualNet-FR.

gNB with a $8 \times 4(N_H \times N_V)$ UPA for communication with single-antenna UEs. UPA elements have half-wavelength uniform spacing. The number of residual blocks in the combining network is set to $N_B = 5$ throughout.

For our proposed model and other competing models, we set the number of epochs to 300 and 1500, respectively. We use batch size of 200. For our model, we start with learning rate of 0.001 before switching to 10^{-4} after the 100-th epoch. Using channel simulators, we generate several indoor and outdoor datasets, each containing 100,000 random channels. We use one seventh of these channels as test data for performance evaluation. The remaining channels are split into 2/3 and 1/3 for training and validation, respectively. For both indoor and outdoor, we use the QuaDRiGa simulator [40] using the scenario features given in *3GPP TR 38.901 Indoor* and *3GPP TR 38.901 UMa* at 5.1-GHz and 5.3-GHz, and 300 and 330 MHz of UL and DL with LOS paths, respectively. For both scenarios, 1024 subcarriers with a 15K-Hz spacing are considered for each subband. Here, we assume UEs are capable of perfect channel estimation. We

set antenna type to *omni*. We use normalized MSE as the performance metric

$$\frac{1}{ND} \sum_{d=1}^D \sum_{n=1}^N \|\widehat{\mathbf{H}}_{\text{BS,DL},d}^{(i)} - \mathbf{H}_{\text{BS,DL},d}^{(i)}\|_{\text{F}}^2 \|\mathbf{H}_{\text{BS,DL},d}^{(i)}\|_{\text{F}}^2, \quad (4.17)$$

where the number D and subscript d denote the total number and index of channel realizations, respectively.

4.4.2 Determining significant beam matrix \mathbf{B}_S based on DL and UL CSIs

Figure 4.11 illustrates the recovery performance of DL CSI by determining precoding matrix \mathbf{B}_S which consists of the L significant beams selected according to CSI magnitudes in UL and DL beam domains, respectively. The modest difference in terms of CSI estimation error demonstrates the high correlation between CSI magnitudes in UL and DL beam domains (partial FDD reciprocity). Specifically, the L dominant beams of UL and DL channels are highly correlated. Good CSI recovery performance requires sufficient number of beams L or REs for CSI-RS.

To evaluate the beam sparsity for different array geometries, Table I demonstrates the average numbers and ratios of significant beams to recover 90% of total CSI energy for UPA with different antenna numbers. We see that larger array and lower angular spread (outdoor channels) lead to a higher beam sparsity. The proposed framework exploits beam sparsity to allocate a small number of required REs while maintaining recovery performance. Namely, more REs are saved for DL CSI training for large-scale arrays and channels with low angular spread. This shows the practical potential of such feedback framework in communications systems with large-scale arrays.

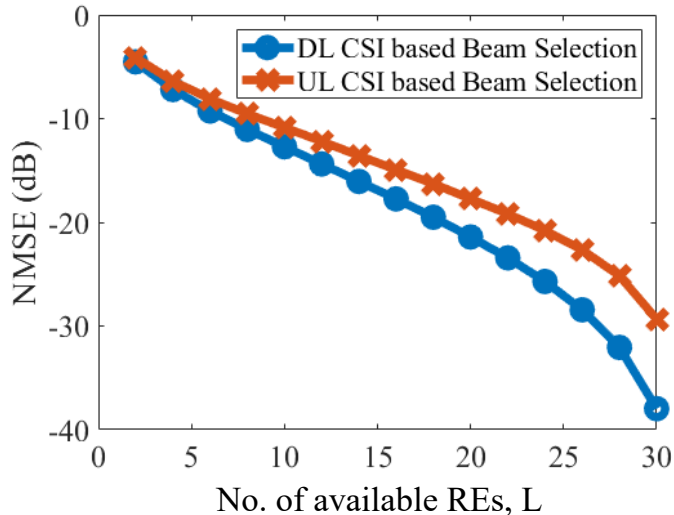


Figure 4.11: Normalized mean square error (NMSE) of the recovered results obtained by beam selection according to UL/DL CSI magnitudes. (This experiment is based on simulated outdoor UMa channels generated by QuaDRiGa channel simulator [40].)

Table 4.1: Beam sparsity evaluation for different array geometries. Higher sparsity means lower ratio of beams required to achieve 90% of total energy.

Indoor				
No. of antennas ($N_H \times N_V$)	32 (8×4)	64 (16×4)	128 (16×8)	256 (32×8)
No. of beams ($> 90\%$)	11.688	19.43	31.45	55.398
Ratio of beams ($> 90\%$)	0.36	0.30	0.25	0.22
Outdoor				
No. of antennas ($N_H \times N_V$)	32 (8×4)	64 (16×4)	128 (16×8)	256 (32×8)
No. of beams ($> 90\%$)	8.23	13.3	18.71	32.9
Ratio of beams ($> 90\%$)	0.26	0.21	0.15	0.13

4.4.3 Testing Different Numbers of Available REs

We evaluate the performance of CSI recovery by adopting the proposed encoder-free CSI feedback frameworks, BSdualNet₀, BSdualNet and BSdualNet-MN. To test the efficacy without considering quantization, we first compare BSdualNet₀ with two heuris-

tic approaches (denoted as BS-UL and BS-DL) that recover DL CSIs according to L beam responses where the beams are selected according to the UL and DL CSI magnitudes, respectively. Note that BS-UL should serve as the lower bound of BSdualNet₀ since BSdualNet₀ is equivalent to refine the result of BS-UL with an additional combining network.

Figures 4.12 (a) and (b) provide the NMSE performance for different number of available REs L in an RB for BSdualNet₀, BS-UL and BS-DL in both indoor and outdoor scenarios, respectively. The results show that BSdualNet₀ delivers better performance than BS-UL and also BS-DL in outdoor scenario owing to the high spatial correlation in beam domain. Because of the high angle spread induced by the more complex multi-path environment in indoor scenarios, the combining network in BSdualNet₀ only marginally improve the recovery performance.

Figures 4.13 (a) and (b) illustrate the NMSE performance for different number L of REs within a RB for BSdualNet₀, BSdualNet and BSdualNet-MN for both indoor and outdoor channels, respectively. We can observe the benefits of the beam merging matrix \mathbf{T} especially in outdoor cases. Furthermore, instead of using a convolution-layer based combining network, changing the combining function as a minimum-norm solution yields a significant performance improvement in both indoor and outdoor scenarios. Since minimum-norm solution directly uses the beam merging matrix \mathbf{T} , it becomes more efficient to decouple the superposition of weighted beam responses by minimizing the MSE of DL CSIs.

4.4.4 Performance for Different Numbers of UEs

Similar to our beam merging matrix \mathbf{T} , measurement matrix in compressive sensing based frameworks [47, 48] also functions to shrink the dimension of original data and derive a better representation for their sparsity that can be easier to recover. To demonstrate the relative performance of the proposed frameworks, we also compare

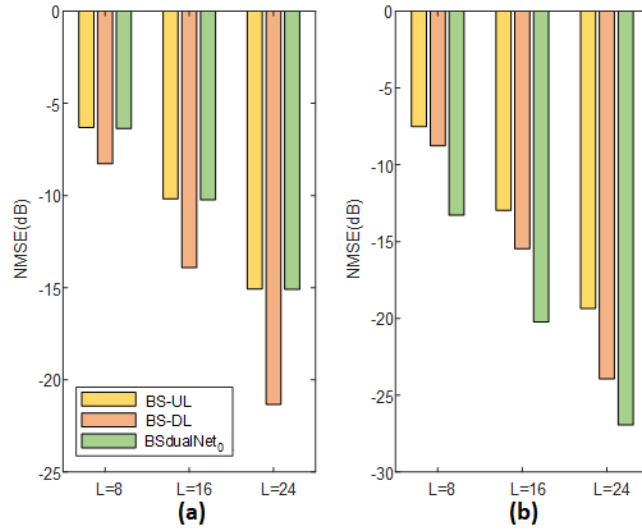


Figure 4.12: NMSE performance of BS-UL, BS-DL, and BSdualNet₀ for different REs L in (a) indoor, (b) outdoor scenarios.

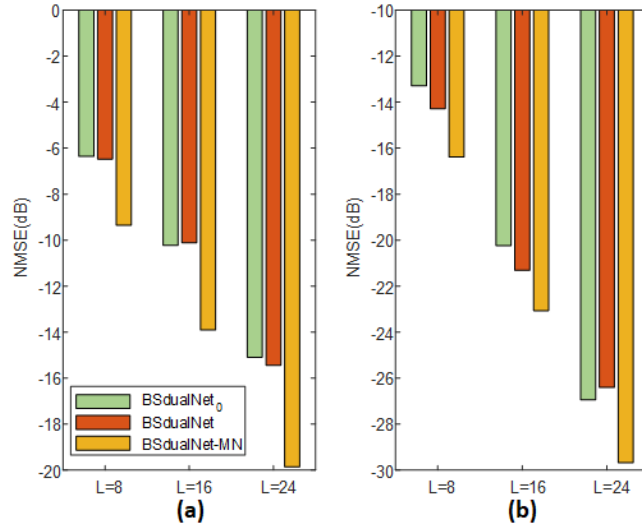


Figure 4.13: NMSE performance of BSdualNet₀, BSdualNet, and BSdualNet-MN for different REs L in (a) indoor, (b) outdoor scenarios.

with two successful compressive approaches ISTA [47] and ISTA-Net [48]:

- **Iterative Shrinkage-Thresholding Algorithm (ISTA)**: Its regularization parameter and maximum iteration number are set to 0.5 and 3000, respectively.
- **ISTA-Net**: The phase and epoch numbers are set to 5 and 1000, respectively.

Figures 4.14 (a) and (b) provide the NMSE performance comparison for different numbers of UEs N for $L = 8$ REs in a RB for BSdualNet, BSdualNet-MN, ISTA and ISTA-Net and under indoor and outdoor scenarios, respectively. From the results, we observe the clear performance degradation for BSdualNet and BSdualNet-MN as UE number grows. This is intuitive since it is difficult to find an optimum beam merging matrix for all active UEs. Fortunately, for most cases, the performance degradation tends to saturate after the UE number exceeds a certain number typically less than 10 for BSdualNet-MN.

Our tests show that both BSdualNet and BSdualNet-MN deliver better performance over ISTA and ISTA-Net under different UE numbers. Our heuristic insight is that measurement matrix in ISTA and ISTA-Net is unknown at recovery whereas the beam merging matrix is designed by the gNB and can be explicitly utilized by the recovery decoders of BSdualNet and BSdualNet-MN.

4.4.5 Different CSI-RS Configurations and Compression Ratios

We consider a 5.76 MHz subband (i.e., 32 RBs each of bandwidth 180K-Hz). Each codeword element uses 8 quantization bits. To comprehensively evaluate BSdualNet-FR, The two tables in Figure 4.15 and Figure 4.16 provide the NMSE performance of BSdualNet-FR against different CSI-RS configurations and compression ratios in outdoor and indoor scenarios, respectively. We apply the same background color on results with the same pilot and feedback overhead reduction ratios.

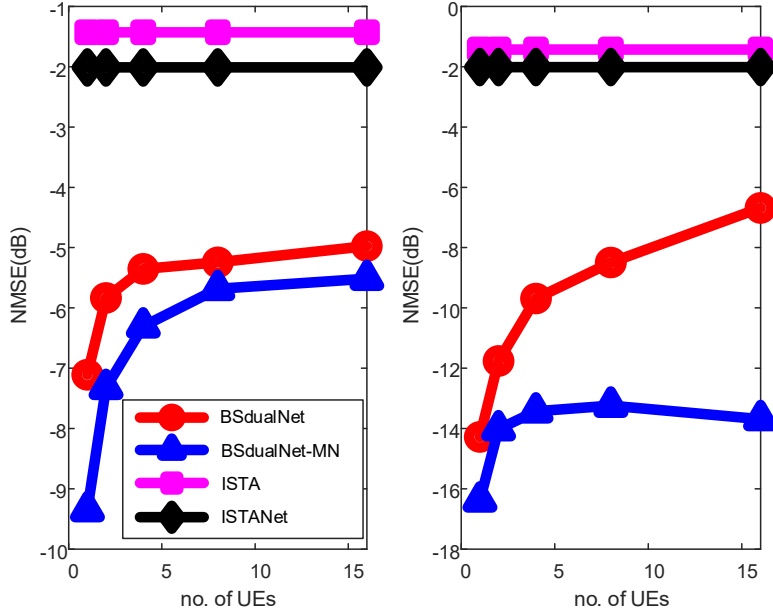


Figure 4.14: NMSE performance for different number of UEs N when $L = 8$ in (left) indoor, (right) outdoor scenarios.

Since outdoor channels generally exhibit stronger sparsity and larger delay spread respectively in beam and delay domains, we observe a slight performance degradation with BR increase as opposed to FR increase. Importantly, for $BR = 4$, there is a clear performance loss even when using the same pilot and feedback overhead reduction ratio. Despite the channel sparsity, with the use of half-wavelength antenna spacing (i.e., Nyquist sampling in spatial domain), the overly aggressive compression in beam domain cause too much information loss to recovery at the gNB. For indoor channels, we observe a slight performance degradation when increasing FR instead of BR because of larger angular and shorter delay spread of indoor CSI.

4.4.6 Different Effective Compression Ratio CR_{eff}

As benchmarks, we also compare BSdualNet-FR with CsiNet [5], CRNet[9], CsiNet-Pro [17] and another successful method DualNet-MP [14]. The newly proposed DualNet-MP also exploits FDD reciprocity by incorporating the UL CSI magnitude as side

CR = 2	BR=1 (L=32)	BR=2 (L=16)	BR=4 (L=16)
FR=2	-34.45	-10.16	-6.66
FR=4	<u>-34.5</u>	-10.2	
FR=8	<u>-27.2</u>		

CR = 4	BR = 1 (L=32)	BR = 2 (L=16)	BR = 4 (L=8)
FR = 1	<u>-34.63</u>	-10	-6.66
FR = 2	-34.06	-10.12	-6.85
FR = 4	-27.03	-10.25	
FR = 8	<u>-17.41</u>		

CR = 8	BR=1 (L=32)	BR=2 (L=16)	BR=4 (L=8)
FR=1	-33.46	-10.14	-6.75
FR=2	-26	-10.15	
FR=4	-17.31		





CR _{eff} = 4		CR _{eff} = 16	
CR _{eff} = 8		CR _{eff} = 32	

Figure 4.15: NMSE performance of BSdualNet-FR for different CSI-RS placement configurations in indoor scenarios. (The results with the same effective compression ratio are denoted as the same color. The best performance at the same effective compression ratio is denoted by bold fonts with underline.)

CR = 2	BR=1 (L=32)	BR=2 (L=16)	BR=4 (L=16)
FR=2	-18.41	-15.21	-12.28
FR=4	-15.73	-12.94	
FR=8	-10.78		

CR = 4	BR = 1 (L=32)	BR = 2 (L=16)	BR = 4 (L=8)
FR = 1	<u>-19.78</u>	-15.43	-10.89
FR = 2	-16.08	-13.19	-9.98
FR = 4	-13.18	-10.79	
FR = 8	-9.642		

CR = 8	BR=1 (L=32)	BR=2 (L=16)	BR=4 (L=8)
FR=1	<u>-16.54</u>	<u>-13.3</u>	-10.07
FR=2	-13.28	<u>-10.95</u>	
FR=4	-9.30		





CR _{eff} = 4		CR _{eff} = 16	
CR _{eff} = 8		CR _{eff} = 32	

Figure 4.16: NMSE performance of BSdualNet-FR for different CSI-RS placement configurations in outdoor scenarios. (The results with the same effective compression ratio are denoted as the same color. The best performance at the same effective compression ratio is denoted by bold fonts with underline.)

information at CSI decoder of gNB. Table II presents the comparison of NMSE for CsiNet, CRNet, CsiNet-Pro, DualNet-MP and BSdualNet-FR with different values of effective compression ratio CR_{eff} in indoor and outdoor cases. Benefiting from the UL CSI magnitudes, both BSdualNet-FR and DualNet-MP can outperform CsiNet, CRNet and CsiNet-Pro in most cases. Interesting, better utilization of UL CSI by BSdualNet-FR provides better performance than DualNet-MP. Although the performance gain becomes less impressive for higher CR_{eff} , it is practically important to note the additional benefit of the BSdualNet-FR framework in reducing REs for DL CSI-RS by a factor of $BR \cdot FR$, which enables gNB to reconfigure the CSI-RS placement to enhance DL spectrum efficiency.

Table 4.2: NMSE performance of different CSI feedback frameworks at different CR_{eff} .

CR_{eff}	CsiNet		CRNet		CsiNet-Pro		DualNet-MP		BSdualNet-FR	
	Indoor	Outdoor	Indoor	Outdoor	Indoor	Outdoor	Indoor	Outdoor	Indoor	Outdoor
4	-17.1	-11.3	-18.1	-12	-24.2	-13	-27.3	-19.1	-34.6 (FR = 1, BR = 1)	-19.8 (FR = 1, BR = 1)
8	-16.7	-10.4	-17.6	-11.8	-20.8	-12.5	-20.9	-16.4	-34.5 (FR = 4, BR = 1)	-16.5 (FR = 1, BR = 1)
16	-16.4	-10	-17.3	-10.5	-14.4	-11.8	-20.2	-13.3	(FR = 8, BR = 1)	-13.3 (FR = 1, BR = 2)
32	-13.4	-8.9	-14.3	-9.1	-13.2	-8.6	-16.8	-11	-17.4 (FR = 8, BR = 1)	-11 (FR = 2, BR = 2)

To demonstrate the benefits of DL spectrum efficiency, we use achievable rate as another performance metric. We allow gNB to choose MRC precoder $\mathbf{w} = \frac{\hat{\mathbf{h}}_{\text{DL}}}{\|\hat{\mathbf{h}}_{\text{DL}}\|_2}$ for maximizing DL transmission gain. According to 5G NR specification, we assume 32 REs (for 32 antenna ports) among all 168 REs in each RB for CSI-RS transmission and we adopt a frequency reduction rate FR to lower CSI-RS placement density. We can define the achievable rate in each RB as follows:

$$R = \gamma \cdot E\left[\log_2\left(1 + \frac{|\hat{\mathbf{h}}_{\text{DL}}^H \mathbf{h}_{\text{DL}}|}{\|\hat{\mathbf{h}}_{\text{DL}}\|_2 \|\mathbf{h}_{\text{DL}}\|_2}\right)\right] (\text{bit/s/Hz}), \quad (4.18)$$

where $\hat{\mathbf{h}}_{\text{DL}}$ and \mathbf{h}_{DL} respectively denote the estimated and original DL CSIs. N_0 denotes

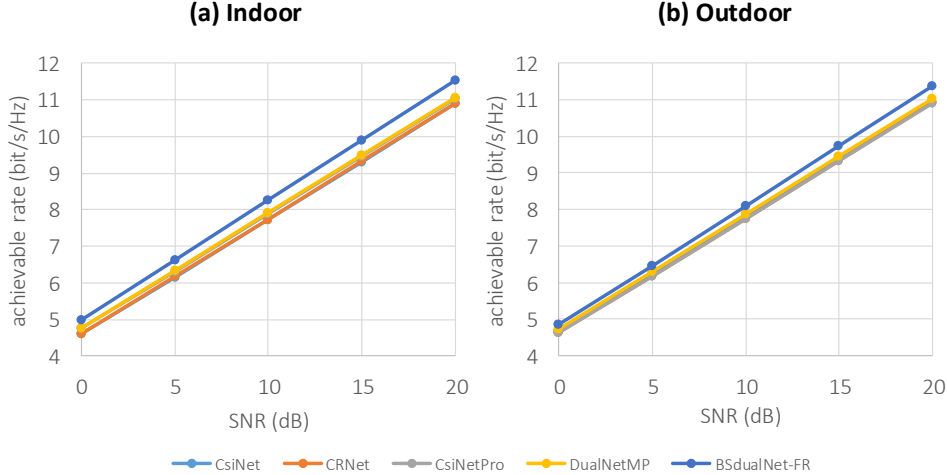


Figure 4.17: DL achievable rate under different SNRs for (a) indoor and (b) outdoor scenarios. Note that we consider a CSI-RS sparsity $P = 4$ and $CR_{\text{eff}} = 32$ ($FR = 8, BR = 1, CR = 4$ is for indoor channels whereas $FR = 2, BR = 2, CR = 8$ is for outdoor channels) in the test results.

the ambient noise level. The quantity

$$\gamma = \frac{K \cdot 168 \cdot P - K \cdot 32 / (FR \cdot BR)}{K \cdot 168 \cdot P} = \frac{168P - 32 / (FR \cdot BR)}{168P}$$

denotes the effective ratio of REs being used for data transmission, where P is the sparsity of CSI-RS placement in terms of slots³. Figure 4.17 shows the achievable rate of all alternatives under different signal-to-noise ratios (SNRs) for both indoor and outdoor scenarios. We observe that BSdualNet outperforms other compared approaches in terms of DL achievable rate although its NMSE performance may not always prevail. This is due to the effect of saving REs for DL signaling to avoid bandwidth waste for data transmission.

4.4.7 Complexity: FLOPs and Parameters

Most UEs have stronger memory, computation, and power constraints. The system design favors light-weight and simpler encoders for deployment at UEs. In comparison

³One out of every P slots is assigned for CSI-RS placement.

Table 4.3: Comparison of parameters (PARAs) and FLOPs at encoder.

	CsiNet		CRNet		CsiNet-Pro		DualNet-MP		BSdualNet-FR	
CR _{eff}	PARAs	FLOPs	PARAs	FLOPs	PARAs	FLOPs	PARAs	FLOPs	PARAs	FLOPs
4	2.8M	1.1M	2.8M	1.2M	4.3M	11.1M	3.8 M	19.2M	$\frac{7.6M}{(FR*BR)}$	$\frac{11.1M}{(FR*BR)}$
8	1.4M	0.56M	1.4M	0.68M	3.8M	10.56M	1.9M	18.9M	$\frac{3.8M}{(FR*BR)}$	$\frac{10.6M}{(FR*BR)}$
16	0.7M	300K	0.7M	420K	1.9M	10.3M	980 K	18.8M	$\frac{1.9M}{(FR*BR)}$	$\frac{10.3M}{(FR*BR)}$
32	350K	170K	350K	290K	950K	10.2M	490 K	18.7M	$\frac{950K}{(FR*BR)}$	$\frac{10.2M}{(FR*BR)}$

with the baseline CsiNet Pro and DualNet, Table 4.3 shows dimension reduction in frequency and beam domains and smaller input size of our encoder/decoder architecture. BSdualNet-FR provides significant reduction in terms of FLOPs and the number of model parameters. Similarly, if the total reduction factor $FR \cdot BR \geq 4$, BSdualNet-FR shows lower storage requirement than those light-weight models CsiNet and CRNet.

4.5 Conclusions

This work presents a new deep learning framework for CSI estimation in massive MIMO downlink. Leveraging UL CSI estimate to reduce its CSI-RS resources, the gNB designs a beam merging matrix based on UL channel magnitude information to transform DL CSI observation at UEs into a lower dimensional representation that is easier for feedback and recovery. We further develop an efficient minimum-norm CSI recovery network to improve recovery accuracy. Our new framework does not deploy training deep learning models at UEs, thereby lowering UE complexity and power consumption. We achieve further reduction of DL CSI training and feedback overhead, by introducing a reconfigurable CSI-RS placement. Test results demonstrate significant improvement of CSI recovery accuracy and reduction of both DL CSI training and UL feedback overheads.

Chapter 5

An Efficient and Scalable Deep Learning Framework for Dynamic CSI Feedback under Variable Antenna Ports

Existing deep learning architectures for downlink CSI feedback and recovery show promising improvement of UE feedback efficiency and eNB/gNB CSI recovery accuracy. One notable weakness of current deep learning architectures lies in their rigidity when customized and trained according to a preset number of antenna ports for a given compression ratio. To develop flexible learning models for different antenna port numbers and compression levels, this work proposes a novel scalable deep learning framework that accommodates different numbers of antenna ports and achieves dynamic feedback compression. It further reduces computation and memory complexity by allowing UEs to feedback segmented DL CSI. We showcase a multi-rate successive convolution encoder with under 500 parameters. Furthermore, based on the multi-rate architecture, we propose to optimize feedback efficiency by selecting segment-dependent

compression levels.

In this chapter, in Section 5.1, we first developed a multi-rate light-weight subarray-based (SAB) CSI feedback framework with flexible number of antennas via a DCP. Then, in Section 5.2, following the same principle, we proposed a SAB framework for flexible number of antenna ports. For further uplink feedback overhead reduction and better recovery performance, we introduced DCP feedback pruning scheme and local normalization, respectively. In Section 5.3, we design a dynamic CR CSI feedback framework to adaptively encode CSI according to its significance. In Section, 5.4, test results demonstrate superior performance, good scalability, and high efficiency for both indoor and outdoor channels. Finally, in Section 5.5, we summarize the proposed light-weight SAB framework and its future research directions. Note that, in this chapter, we represent DL CSI \mathbf{H}_{DL} by \mathbf{H} for simplicity.

5.1 Multi-rate CSI Feedback Framework with Flexible Number of Antennas

There have been notable progresses in terms of recovery performance among the recent autoencoder-based CSI feedback frameworks [14, 18–20]. Since UEs often have limited resources [20], an important consideration is the computation complexity and storage needed by the CSI encoder at the UE. Unfortunately, naïve use of autoencoders from image compression for CSI compression requires direct input of full CSI matrix \mathbf{H} as a 2D “image” to deep learning networks for feature extraction. The inevitably large input size necessitates large autoencoder learning models at both UE and gNB, thereby making it highly challenging to effectively reduce model complexity and storage need.

This raises an question: is it necessary to *simultaneously* feed full DL CSI matrix into the model for encoding CSI features across all ports? The answer may vary. In

application when antenna configuration avoids spatial aliasing ¹ (e.g., half-wavelength antenna spacings), CSI correlation across the multiple antenna ports tends to be weak and negligible. Thus, it may be unnecessary to import CSIs across many antennas of the same MIMO configuration to the UE encoder for compression and feedback.

We can gain some insights from the following test results. Figures 5.1 (a) and (b) show the correlation between different antennas and the statistics at different delay taps for different antennas. It is apparent that correlation between antennas is weak and, in fact, CSI statistics at different delay taps even for different antennas appears similar. This recognition motivates us to propose to apply a common and smaller deep-learning model to encode and decode the DL CSI across large number of antenna ports when distinct antennas serve as multiple activated ports.

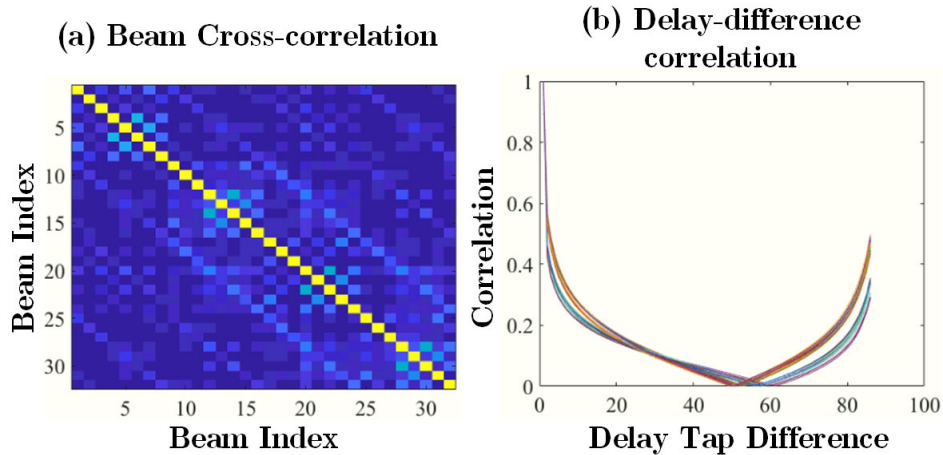


Figure 5.1: (a) Cross-correlation between different beams (we consider a 8×4 orthogonal beam set), and (b) correlation versus various delay tap difference (we consider CSIs of 32 antennas denoted by curves with different colors). The low cross-correlation between beams and the high similarity of these curves in delay domain imply the possibility to compress and recovery CSI antenna-by-antenna.

¹As a rule of thumb, CSI of antennas spaced more than one wavelength apart are nearly independent.

5.1.1 SAB Framework

Previous works such as [5, 7, 13, 14] send full CSI matrix like an image as encoder input for compression. Such 2-D CSI structure in antenna and delay domains is akin to a natural 2-D image. However, from the preliminary results of Figures 5.1 (a) and (b), the inter-antenna independence and similar statistics of delay profile of different antennas motivate a simpler subarray based (SAB) CSI encoding and decoding framework. In this section, we propose an SAB framework which divides a full DL CSI into non-overlapping several subarray pieces before their individual compression and gNB recovery.

We first define a new quantity, *subarray width*, as the spatial domain width of the new framework input. Let subarray width be K to capture K consecutive antenna ports among the N_b rows of the CSI matrix that exhibit correlation [20]. We concatenate real and imaginary parts of the full DL CSI matrix $\mathbf{H} = [\mathbf{h}_1 \ \mathbf{h}_2 \ \dots \ \mathbf{h}_{N_b}]^T$ (Note that we adopt a DL CSI in antenna-delay domain in this section) in an interleaving manner as an augmented real-value full DL CSI matrix $\mathbf{H}_{\text{aug}} = [\text{Real}(\mathbf{h}_1) \ \text{Imag}(\mathbf{h}_1) \ \text{Real}(\mathbf{h}_2) \ \dots \ \text{Imag}(\mathbf{h}_{N_b})]^T$ of size $2N_b \times N_t$ before partitioning the $2N_b$ rows to form $2N_b/K$ matrices of size $K \times N_t$ as follows:

$$\mathbf{H}^{(i)} = \mathbf{H}_{\text{aug}}(Ki + 1 : Ki + K, :), \quad i = 0, 1, \dots, N_b/K - 1. \quad (5.1)$$

We train a common autoencoder for each of the K subarray CSIs. Each subarray matrix \mathbf{H}_i enters the common encoder $\mathbf{q}^{(i)} = f_{\text{en}}(\mathbf{H}^{(i)})$ at UE for compression and feedback. At the gNB, the decoder $\hat{\mathbf{H}}^{(i)} = f_{\text{de}}(\mathbf{q}^{(i)})$ recovers the subarray CSI before stacking them back into the full DL CSI matrix

$$\hat{\mathbf{H}}_{\text{aug}} = \left[\hat{\mathbf{H}}^{(1)}; \hat{\mathbf{H}}^{(2)}; \dots; \hat{\mathbf{H}}^{(N_b/K)} \right] \quad (5.2)$$

By extracting rows at the odd and even indexes, we can obtain the estimate of the full DL CSI matrix $\hat{\mathbf{H}}$.

5.1.2 Multi-rate CSI Feedback Framework

In practical applications, physical environment affects the MIMO CSI characteristics including its sparsity and entropy. Therefore, the degree to which an MIMO CSI can be compressed in a deep learning framework would vary with physical environment. Without knowing the actual CSI a priori, multiple encoder-decoder pairs may have to be deployed at UEs and gNB to achieve the required accuracy and feedback compression. Training multiple encoders would lead to higher memory use to store the models and possibly higher complexity to test the outcomes of different compression models (i.e., ratios).

To this problem, the authors of [25] proposed a multi-rate CSI framework as illustrated in Figure 5.2. Its encoder of [25] can generate 4 different output arrays of 4 distinct compression ratios. The parameters of all layers in its encoder are common except for a final fully-connected (FC) layer. This framework of [25] reduces the total number of encoder parameters by enforcing convolutional layers for different compression ratios to remain the same so as to generate similar features. Only the final layer decides the encoder output for feedback at different compression ratios.

In this chapter, we consider a similar architecture but proposing a new encoder design with fully convolutional layers and the proposed SAB framework. We name the new architecture “successive convolutional encoding network (SCENet)” whose model complexity can be significantly tamed while preserving good recovery performance. To achieve a good tradeoff between performance and model complexity, we focus on complexity reduction at the encoder for low cost UEs. For the UE encoder, we introduce a fully-convolutional down-sizing block (FCDS) to lower the input size by half. The FCDS block consists of 1×7 , 1×5 and 1×3 convolutional layers with 2 channels, respectively. Note that the stride lengths are all 1 except for the final horizontal stride in the last convolutional layer which is of length 2 to drop the input size by half. Figure 5.3 shows an example of a CSI feedback framework using S FCDS blocks for dealing

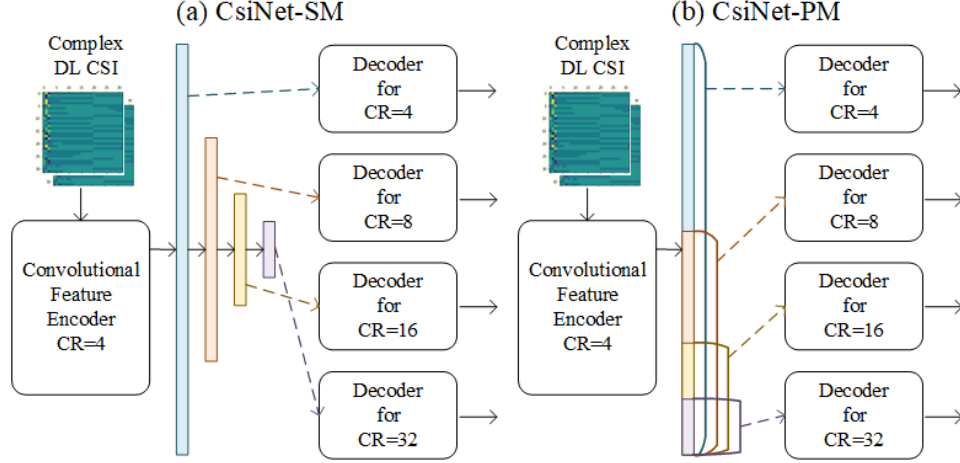


Figure 5.2: Illustration of previous multi-rate CSI feedback frameworks, CsiNet-SM and CsiNet-PM. The encoders share model parameters at different compression ratios except for FC layers, which contribute the majority of model complexity.

with 4 compression ratios ($S=4$ throughout this chapter). Specifically, the output of i -th block with size of $K \cdot N_t/2^i$ represents codewords with compression ratio = $2^i, i = 1, \dots, S$.

Since gNBs are less resource constrained, individual CSI decoder is designed for each compression ratio. For the i -th decoder, the codeword is first fed to a $K \cdot N_t$ FC layer, a 1×3 convolutional layer and activation function after reshaping for initial estimation. An ensuing RefineBlock [25] provides refinement. RefineBlock uses a residual structure and consists of three 1×3 convolutional layers with 16, 8 and 1 channels, respectively. The RefineBlock is followed by a $K \cdot N_t$ FC layer for generating real/imaginary CSI estimates. To further improve recovery accuracy, we provide another SCNnet, called SCENet+ by adding an additional FC layer at the end of each FCDS block which provides extra non-linearity at the same output size.

The parameters of the SCENet are optimized according

$$\Omega_{en}, \Omega_{de} = \arg \min \sum_{d=1}^D \sum_{s=1}^{S=4} W_s \cdot \|\mathbf{H}_d - \hat{\mathbf{H}}_{d,s}\|_F^2, \quad (5.3)$$

$$\hat{\mathbf{H}}_{d,1}, \hat{\mathbf{H}}_{d,2}, \hat{\mathbf{H}}_{d,3}, \hat{\mathbf{H}}_{d,4} = f_{de}(f_{en}(\mathbf{H}_d)), \quad (5.4)$$

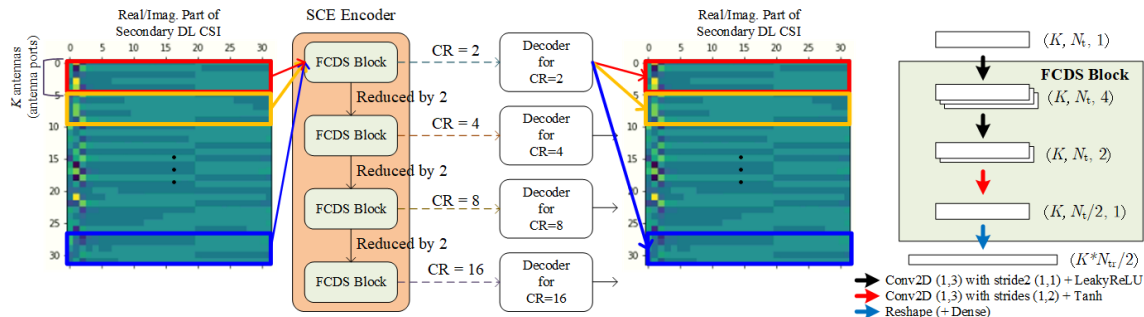


Figure 5.3: SAB Framework and SCE Network Architecture. Input data are first split into real and imaginary and separated into subarray matrices. These matrices are fed to the SCE network and recovered in parallel. Note that, at encoder, after each FCDS block, the total size of input is reduced by half. The fully convoluted FCDS blocks share parameters. We also provide another alternative encoder (SCENet+ encoder) where a FC layer is attached at the end of each FCDS block for enhancing performance.

where subscript s denotes the outcome from the s -th compression ratio and Ω_{en} , Ω_{de} denote the trainable parameters of encoder f_{en} and decoder f_{de} . D is the training data size. In [25], hyper-parameters $\{W_1, W_2, W_3, W_4\}$ were chosen as $\{30/39, 6/39, 2/39, 1/39\}$.

5.2 Multi-Rate CSI Feedback Framework with Flexible Number of Antenna Ports

The proposed SAB framework can effectively reduce the model size and computational complexity. However, the uplink feedback overhead is not lower with this framework. To reduce feedback information, we observe that CSI in beam domain (i.e., angular domain) appears to be sparse. For instance, outdoor propagation channels usually characterized with its low angular spread. If we transform CSI matrices from antenna (i.e. spatial) domain to beam domain before compression and recovery with the proposed SAB framework, we may require fewer or even no codewords for those subarray CSIs with negligibly low energy. With this motivation, we propose a *DCP feedback pruning* mechanism to further reduce the uplink information for CSI feedback and the computational complexity of encoding/decoding at UE and gNB, respectively.

5.2.1 SAB framework in BD domain

We represent the full CSI matrix in antenna-delay domain as

$$\mathbf{H}_{\text{AP}} = \mathbf{M} \cdot \mathbf{H} \quad (5.5)$$

where $\mathbf{M} \in \mathbb{C}^{N_b \times N_b}$ is an orthogonal transformation matrix transforming from antenna to antenna port (AP) domain. Without loss of generality, we can have a DL BD CSI matrix \mathbf{H}_B by designing an orthogonal beam matrix $\mathbf{M} = \mathbf{B}$ which be found via the mechanism in [42]. Following the same preprocessing in the previous section, we first concatenate real and imaginary parts of CSIs as an augmented DL BD matrix $\mathbf{H}_{B,\text{aug}}$ and divide the augmented DL BD matrix into $2N_b/K$ subarray CSI matrices of the same size $K \times N_t$ given below

$$\mathbf{H}_B^{(i)} = \mathbf{P}^{(i)} \mathbf{H}_{B,\text{aug}}, \quad \forall i = 1, 2, \dots, N_b/K. \quad (5.6)$$

Thus, the parameters of the SCEnet are optimized according to criterion:

$$\Omega_{en}, \Omega_{de} = \arg \min \sum_d^D \sum_s^{S=4} W_s \cdot \|\mathbf{H}_d - \mathbf{B}^H \hat{\mathbf{H}}_{B,d,s}\|_F^2, \quad (5.7)$$

$$\hat{\mathbf{H}}_{B,d,1}, \hat{\mathbf{H}}_{B,d,2}, \hat{\mathbf{H}}_{B,d,3}, \hat{\mathbf{H}}_{B,d,4} = f_{de}(f_{en}(\mathbf{H}_{B,i})). \quad (5.8)$$

5.2.2 DCP Feedback Pruning

Due to small angular spread, outdoor CSIs in beam domain are usually sparse in angular domain. To take advantage of this physical property, we propose a DCP feedback pruning method to exploit the beam sparsity to further reduce the uplink feedback overhead and encoding/decoding computations by skipping feedback of those insignificant subarray CSI matrices of negligibly low Frobenius norm.

To evaluate whether a subarray CSI matrix is insignificant, we measure its relative energy ratio

$$R_{E,i} = \|\mathbf{H}_{B,i}\|_F^2 / \|\mathbf{H}_{B,\text{aug}}\|_F^2. \quad (5.9)$$

Subarray CSI matrices with energy ratio below a predefined threshold T are regarded as insignificant and are ignored at the UE encoder. Importantly, UEs need to transmit extra information bits to indicate insignificant subarray to the gNB during feedback.

To minimize the information bits, as illustrated in Figure 5.4, we suggest that UE could utilize a prefix bit indicating whether to send a *zero-skipping request* to base station. As depicted in Figure 5.5, the additional bit is appended before the bit stream of each subarray CSI matrix as a prefix which is decoded first at gNB to avoid the subsequent CSI recovery for the insignificant subarray CSI matrix. For subarray CSI matrix with energy ratio $R_E \geq T$, UE encodes the CSI matrix and the codeword feedback on uplink to gNB with the indicator bit = 1. Otherwise, UE sends zero uplink feedback with indicator bit = 0. Alternatively, a $2N_b/K$ bitmap can lead or trail the CSI codeword feedback as indicators to the decoder. The gNB examines these indicator bits to decide whether to decode the corresponding subarray CSI codeword or to zeropad the corresponding subarray CSI before moving onto the next subarray CSI.

By doing so, a larger threshold T tends to skip more encoding/decoding process, use less uplink bandwidth for feedback, but possibly cause performance degradation due to the zero-skipping process. Thus, the selection of threshold T becomes a trade-off between the amount of uplink feedback overhead and recovery performance. Fortunately, due to the sparsity in angular domain, we can effectively reduce uplink feedback bandwidth and computations while not sacrificing too much recovery performance in general, especially for channels with low angular spreads.

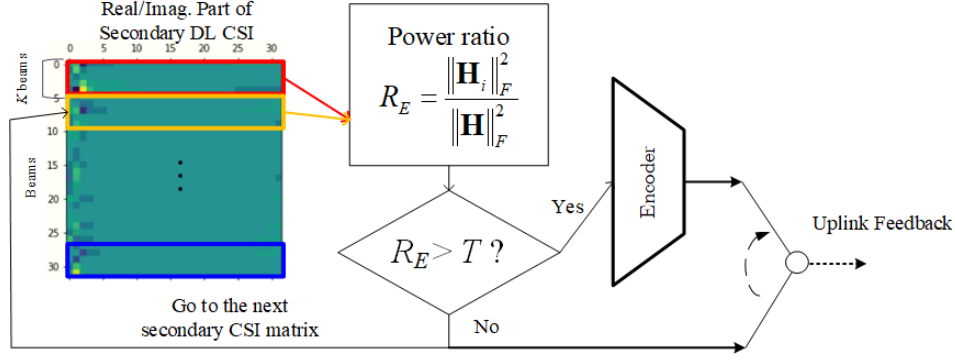


Figure 5.4: Illustration of DCP Feedback Pruning. If the energy ratio of the i -th DL BD subarray CSI matrix is less than the predetermined threshold T , UE skips encoding and send only one bit to tell base station to fill zeros in the corresponding region of the DL BD subarray CSI matrix. Otherwise, UE operates SAB framework normally.

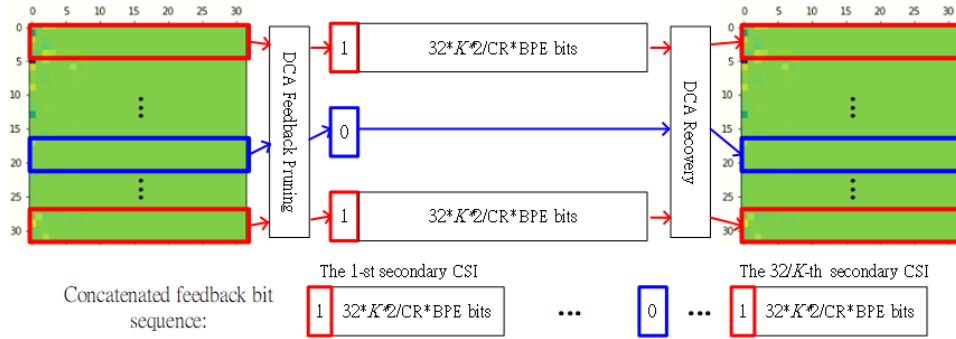


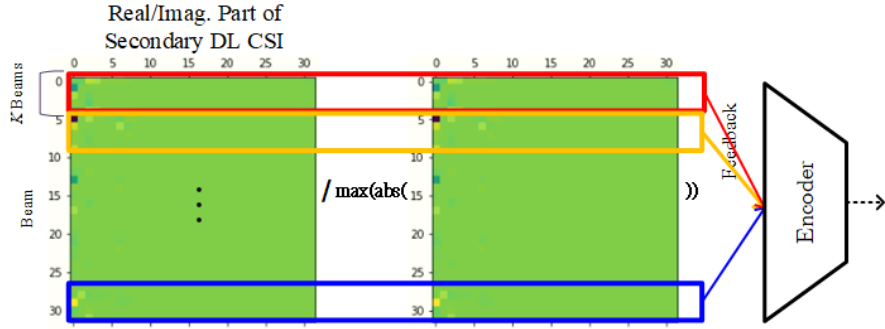
Figure 5.5: DCP feedback pruning block diagram and ordered feedback bit sequence.

5.2.3 Local Normalization

Recall that one assumption for SAB framework is the similar statistics of delay profile of different antennas. After transforming CSI from antenna to beam domain, although the relative delay profile is still similar for different beams, CSI energy concentrates in a few specific angles (directions) for in most propagation with low angular multipath spreads. As a result, CSI recovery may degrade because of training bias in which deep learning model endeavor to recover those stronger subarray CSI matrices better. This may lead to very poor recovery performance for subarray CSI matrices of modest energy. To tackle this problem, as depicted in Figure 5.6(b), we let UE normalize each encoded subarray CSI matrix individually and encode the normalization factor as a

feedback to gNB.

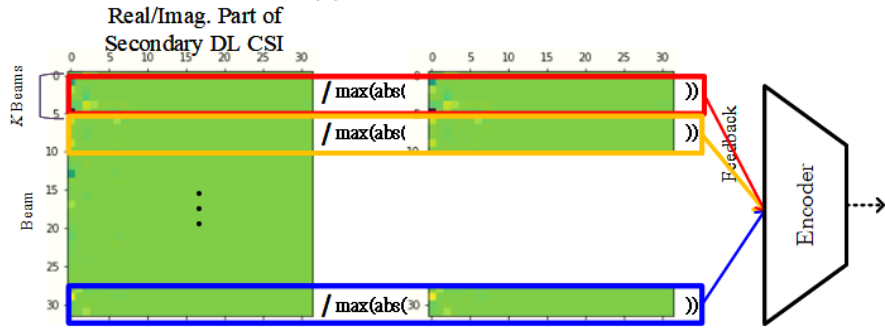
(a) Global Normalization



Feedback:

1. $32/K$ times of codewords with length of $32*K/CR$
2. A global normalization element among $32/K$ secondary CSI matrices

(b) Local Normalization



Feedback:

1. $32/K$ times of codewords with length of $32*K/CR$
2. **$32/K$ normalization elements of $32/K$ secondary CSI matrices**

Figure 5.6: Illustrations of (a) global normalization and (b) local normalization.

5.2.4 2D Lightweight Encoder

In this section, we proposed a SAB framework in BD domain along with subarray row feedback and pruning to further reduce uplink feedback overhead by taking advantages of its beam domain sparsity. In fact, sparsity is also observed in the delay domain. A natural extension is develop a two-dimensional (2D) SAB framework as illustrated in Figure 5.7 along with feedback pruning method to skip near-zero CSI matrix blocks

for reducing uplink feedback bandwidth.

However, overly aggressive model reduction as such requires the CSI energy to be not only similarly distributed in the delay domain across antenna ports but also similarly distributed in spatial domain for each delay. Such property has not been experimentally verified. Therefore, although a 2D lightweight encoder admits a low complexity autoencoder structure, we must carefully weigh the complexity-accuracy tradeoff of such efforts.

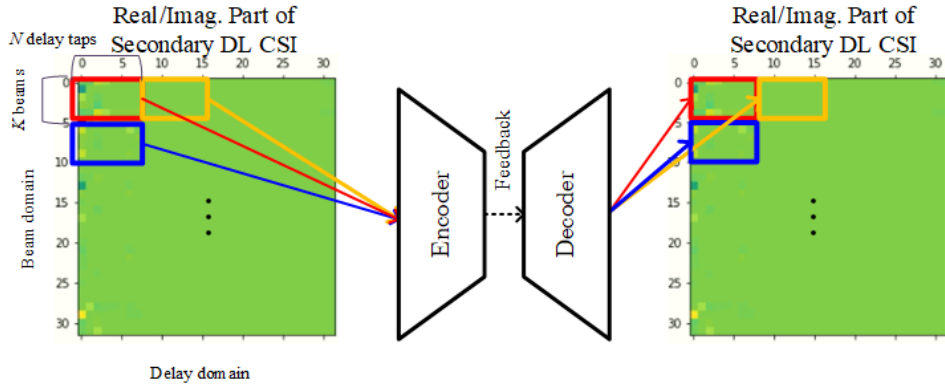


Figure 5.7: Illustrations of 2D SAB framework (N and K are the numbers of delay taps and beams being considered in a single subarray CSI matrix, respectively).

5.3 SAB framework with dynamic CR

The proposed SAB feedback switches on/off the encoding of CSI subarrays for achieving a higher effective compression ratio. We can utilize more feedback resource on high-energy subarray CSI matrices especially for channels with sparse distribution in beam or angular domain such as outdoor channels. Yet, instead of using a hard decision to determine whether to feedback or skip the encoding/recovery process, the multi-rate architecture motivates a softer decision approach. Here, we propose a dynamic CR CSI feedback framework which compresses subarray CSI matrices in a full DL CSI using dynamic CR by a energy-based CR selector according to their significance (i.e.,

normalized subarray CSI energy) to maximize the codeword efficiency (CE) of CSI feedback.

To define the efficiency of codeword, we measure the expected capacity provided by each codeword. With orthogonal multiple access, when using estimated full CSI $\widehat{\mathbf{H}} = [\widehat{\mathbf{H}}_1; \widehat{\mathbf{H}}_2; \dots; \widehat{\mathbf{H}}_{2N_b/K}]$ as a maximum-ratio combining (MRC) precoder for DL transmission at gNB, the expected capacity for the i -th subarray CSI matrix in DL transmission can be reasonably set as

$$C_i = \log_2(1 + \text{SNR}_i) \quad (5.10)$$

$$\text{SNR}_i = \frac{\|(\widehat{\mathbf{H}}_i)^* \widetilde{\mathbf{H}}_i / \|\widehat{\mathbf{H}}_i\|_{\text{F}}\|_{\text{F}}^2\|_{\text{F}}^2}{K \cdot N_f \cdot P_N} \quad (5.11)$$

where $\|(\widehat{\mathbf{H}}_i)^* \widetilde{\mathbf{H}}_i / \|\widehat{\mathbf{H}}_i\|_{\text{F}}\|_{\text{F}}^2\|_{\text{F}}^2 / (K \cdot N_f)$ and P_N denote the average signal and noise power, respectively, over N_f subcarriers and K antenna ports. $\widetilde{\mathbf{H}}_i$ denotes true subarray CSI matrix.

Let the sum length of uplink feedback codeword $\mathbf{q} = [\mathbf{q}_1; \mathbf{q}_2; \dots; \mathbf{q}_{2N_b/K}]$ from UE to gNB be $L = \sum_{i=1}^{2N_b/K} L_i$. We can define the average CE as

$$\text{CE} = \sum_{i=1}^{2N_b/K} \frac{C_i}{L_i} \frac{K}{2N_b} (\text{bits/s/Hz/codeword}). \quad (5.12)$$

This metric measures the contribution of each codeword to the eventual end-to-end CSI feedback performance.

Take the multi-rate CSI feedback framework, DCnet, as an example, it provides four distinct lengths of codewords (corresponding to four compression ratios) for different compressing/recovery quality. To achieve the best performance, we should compress CSI with the least compressive codewords and vice versa. There always exists a trade-off between uplink feedback cost and recovery performance. Yet, although there is no best choice of compression ratio, the most efficient one exists.

By dividing a full-size CSI matrix into several subarray CSI matrices, we discover that only a fraction of subarray CSI matrices dominate in terms of energy. That is, if we could recover those subarray CSI matrices well, we will have a high-quality CSI recovery even if other subarray CSI matrices are recovered with large errors. Hence, to improve feedback efficiency, we should utilize more resources (i.e., $CR = 2$) on subarray CSI matrices with larger significance (i.e., higher energy) and less resources ($CR = 16$) on those with less significance. We first evaluate the significance of the i -th subarray CSI matrix for each data sample according to its normalized CSI energy $R_{E,i}$ defined in (5.9).

We design a energy-based CR selector which selects CR according to the normalized energy of subarray CSI matrices. The CR determined by the CR selector for the i -th subarray CSI matrix is given by

$$CR_i = \begin{cases} 2 & a_0 \leq R_{E,i} < a_1 \\ 4 & a_1 \leq R_{E,i} < a_2 \\ 8 & a_2 \leq R_{E,i} < a_3 \\ 16 & a_3 \leq R_{E,i} \leq a_4 \end{cases} \quad (5.13)$$

As illustrated in the Figure 5.8, there are five anchor points $\mathbf{a} = [a_0 = 1, a_1, a_2, a_3, a_4 = 0]$ where $1 \geq a_1 \geq a_2 \geq a_3 \geq 0$ and a_1, a_2, a_3 are trainable. If we optimize the three anchor points by maximizing CE in Eq. 5.12, since the nominator does not grow proportionally as the denominator increases, we will have a trivial CR selector, which always suggests adopting the largest CR to achieve the highest codeword efficiency. Unfortunately, this induces a fairness problem since the CR selector tends to secure CE and ignore those CSI estimates with extremely poor performance. Those cases should be considered as recovery failure. Thus, using a standard step function $u(\cdot)$, we

define the mean outage capacity as

$$\text{CE} = E \left\{ \sum_{i=1}^{2N_b/K} \frac{C_i}{L_i} \cdot u(\text{NMSE}_i - T_{\text{out}}) \right\} \quad (5.14)$$

We define an outage threshold T_{out} to reject cases when gNB totally fails to estimate DL CSI. In the training stage, as a rule of thumb, a typical value of T_{out} is set as -10 dB.

In this chapter, we provide a heuristic training strategy for searching optimal points by following Alg. 6.1. Note that, since we consider four possible CRs, we need extra two-bit information for each subarray CSI matrix in the uplink feedback to gNB for correctly identifying the correct decoder of the corresponding CR as shown in Figure 5.9.

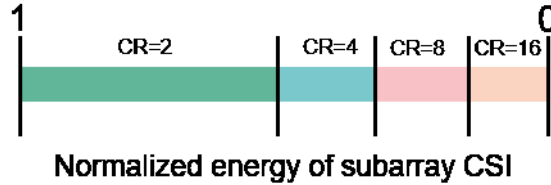


Figure 5.8: Five anchor points of normalized energy of subarray CSI matrix for CR decision. Note that we only need to train the three anchor points a_1, a_2, a_3 for separating the operating regions of four CRs.

5.4 Experimental Evaluations

5.4.1 Experiment Setup

In our experiments, we consider both indoor and outdoor cases. Using channel model software [40], we place a gNB of height equal to 20 m at the center of a circular cell with a radius of 30 m for indoor and 200 m for outdoor environment. The gNB equipped with a $8 \times 4(N_H \times N_V)$ UPA for communicates with single-antenna UEs. UPA elements

Algorithm 5.1 Multi-point linear searching algorithm

Require: $\mathbf{a} = [1, 0, 0, 0, 0]$, N_{ter} , N , $\text{CE}_f = 0$, $\Omega = \{1, 2, 3\}$

Ensure: $\mathbf{a} = [1, a_1, a_2, a_3, 0]$, CE_f

```
for  $i = 1 : 1 : N_{\text{ter}}$  do
   $j \leftarrow \text{mod}(i, \text{length}(\Omega)) + 1$ 
   $\mathbf{v}_f \leftarrow [a_j - \frac{a_{j-1} - a_j}{(N/2)+1}; \dots; a_j - (N/2) \frac{a_{j-1} - a_j}{(N/2)+1}]$ 
   $\mathbf{v}_b \leftarrow [a_j + \frac{a_j - a_{j+1}}{(N/2)+1}; \dots; a_j + (N/2) \frac{a_j - a_{j+1}}{(N/2)+1}]$ 
   $\mathbf{v} \leftarrow [\mathbf{v}_f; \mathbf{v}_b]$ 
   $\mathbf{a}_{\text{old}} \leftarrow \mathbf{a}$ 
  flag  $\leftarrow$  False
  for  $k = 1 : 1 : N$  do
     $a_{\Omega_j} \leftarrow \mathbf{v}[k]$ 
    Evaluate CE according to  $\mathbf{a}$ 
    if  $\text{CE} > \text{CE}_f$  then
       $\text{CE}_f \leftarrow \text{CE}$ 
      flag  $\leftarrow$  True
    end if
  end for
  if  $|a_2 - a_1| < 0.005$  then
     $\Omega = \{[1, 2], 3\}$ 
  else if  $|a_3 - a_2| < 0.005$  then
     $\Omega = \{[1], [2, 3]\}$ 
  else if  $|a_2 - a_1| < 0.005$  and  $|a_3 - a_2| < 0.005$  then
     $\Omega = \{[1, 2, 3]\}$ 
  end if
  if flag = False then
     $\mathbf{a} \leftarrow \mathbf{a}_{\text{old}}$ 
  end if
end for
```

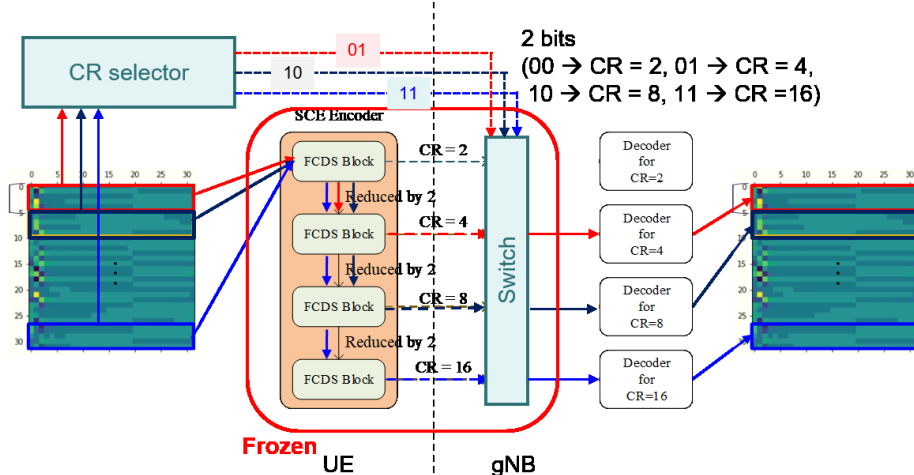


Figure 5.9: UE sends extra two bits for each subarray CSI matrix to indicate the adopted CR. gNB selects the corresponding decoder according to the extra information.

have half-wavelength uniform spacing.

For our proposed model and other competing models, we set the number of epochs to 1000. We use batch size of 200. For our model, we start with learning rate of 0.001 before switching to 5×10^{-4} after 300 epochs. Using the channel simulator, we generate several indoor and outdoor datasets, each containing 100,000 random channels. One seventh of these channels is test data for performance evaluation. Two and one thirds of the remaining are for training and validation. For both indoor and outdoor, we use the QuaDRiGa simulator [40] using the scenario features given in *3GPP TR 38.901 Indoor* and *3GPP TR 38.901 UMa* at 5.1-GHz and 5.3-GHz, and 300 and 330 MHz of UL and DL with LOS paths, respectively. To accurately assess recovery accuracy, we assume UEs are capable of exact CSI estimation. For each data channel, we consider $N_f = 1024$ subcarriers with 15K-Hz spacing and place $M_f = 86$ pilots with downsampling ratio $DR_f = 12$ as illustrated in the Figure 5.10. We set antenna type to *omni*. We use NMSE Eq. 2.2.2 as the performance metric.

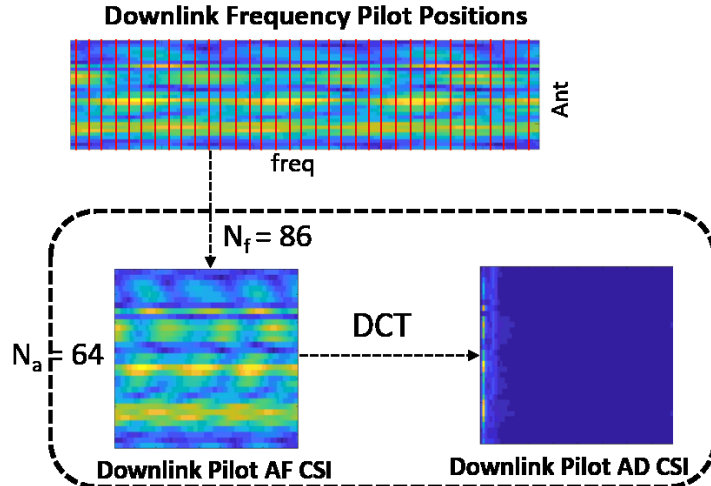


Figure 5.10: Pilot placement illustration (Note that the red lines indicate the time-frequency resources to be placed pilot symbols. AF and AD stand for antenna-frequency and antenna-delay domains, respectively).

5.4.2 SCENet vs. SCENet+

Figures 5.11 (a) and (b) summarize NMSE performance for the two proposed models at different compression ratios in indoor and outdoor scenarios, respectively. We observe the benefits of the extra FC layer at encoder for low compression ratios. Considering the negligible error improvement in linear scale, SCENet and SCENet+ achieve similar performance. Yet, SCENet+ has more flexible coding rate owing to the use of FC layers. For brevity, we use SCENet+ as our benchmark in the rest of this section.

5.4.3 Performance, Complexity and Storage Comparison

For comparison, besides the proposed models SCENet and SCENet+, we also include two recent multi-rate CSI feedback alternatives which **take full DL CSI as model input** and are listed below:

- **CsiNet-SM** [25]: Figure 5.2 (a) shows its general architecture. Note that we accommodate the model for desired compression ratios by adjusting the size of FC layers.

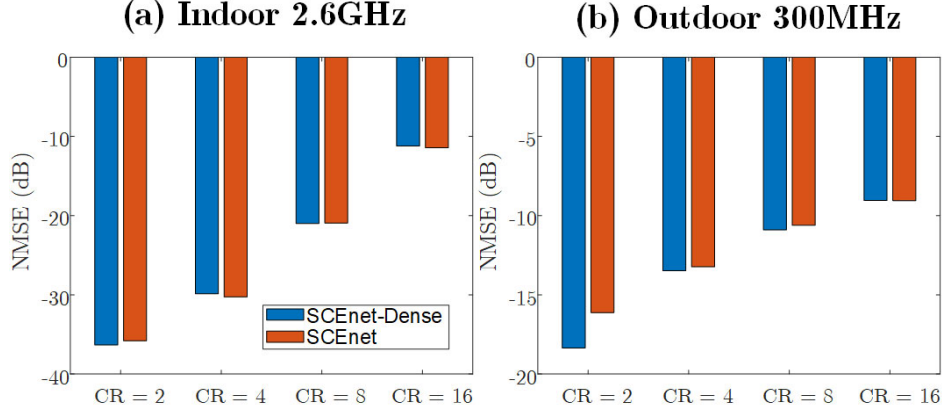


Figure 5.11: NMSE performance at different compression ratios for SCENet and SCENet+ in indoor and outdoor scenarios.

- **CsiNet-PM** [25]: Figure 5.2 (b) shows the general architecture. Note that CsiNet-PM is a more compact model than CsiNet-SM but suffers slight performance degradation in general.

Note that the proposed models adopt a similar decoder as the alternatives in comparison with required accommodations such as reduced sizes of FC layers and one dimensional convolutional filter size (i.e., (1,3), (1,5) and (1,7)).

Most UEs have strict memory, computation and power constraints, thereby favoring light-weight and simpler encoders for deployment. Figures 5.12 (a) and (b) model size of encoder and decoder, respectively, for SCENet, SCENet+, CsiNet-SM, and CsiNet-PM. Table 5.1 reveals computation complexity of encoder and decoder for alternatives in comparison. Table 5.2 shows the NMSE performance at different compression ratios and subarray width (K) for SCENet+, CsiNet-SM and CsiNet-PM including both indoor and outdoor scenarios. We observe that SCENet+ with $K = 64$ generally outperforms CsiNet-SM and CsiNet-PM and requires less FLOP number and storage at UE side. Leveraging the SAB framework of smaller subarray width K , we enjoy much lower complexity and storage with slight performance degradation. The selection of $K = 2$ yields an acceptable recovery performance and delivers several orders of encoder

Table 5.1: Floating-point operation of alternatives in comparison. Compression ratio is 8 for calculating decoder's FLOP numbers.

	SCENet	SCENet+	CsiNet-SM	CsiNet-PM
Encoder FLOPs	1.16M	1.4M	4.3M	2.2M
Decoder FLOPs (K=2)	10.7M		49.4M	
Decoder FLOPs (K=4)	12M			
Decoder FLOPs (K=8)	14.75M			

and decoder size reduction as well². Moreover, SCENet+ becomes scalable and can be a universal CSI feedback framework which can be applied to CSI feedback with various numbers of antenna ports (according to the 3GPP specification, 2, 4, 8, 16, 32 are possible antenna port number).

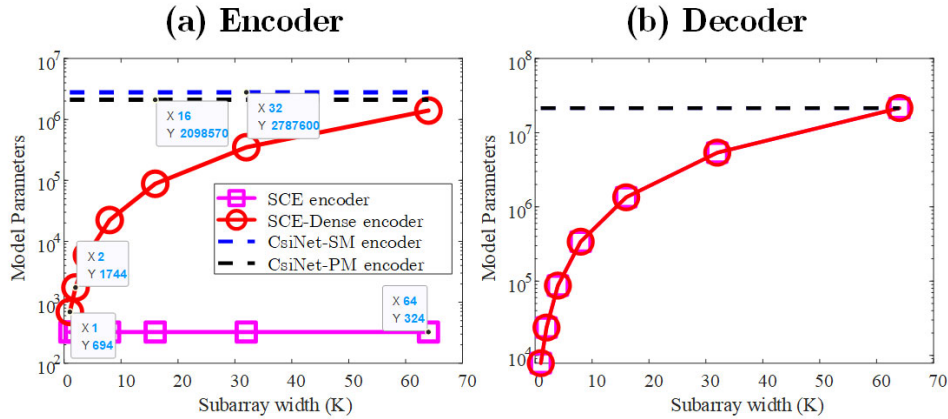


Figure 5.12: (a) Encoder and (b) decoder model size comparison of SCENet, SCENet+, CsiNet-SM and CsiNet-PM.

5.4.4 Testing Different Encoder/Decoder Pairs

To show the efficacy of SAB framework, Figure 5.13 shows the NMSE performance at different compression ratios and three encoder/decoder pairs: 1) SAB encoder plus

²The major model size reduction is attributed to smaller input size. However, smaller input size does not simplify the computation complexity by the same order. Although FLOP number grows proportionally with input size, the encoder is applied to multiple subarray CSI matrices.

Table 5.2: NMSE performance of the CsiNet-SM and SCENet for different selections of subarray width (K).

CR	Scen.	SCENet+				CsiNet -SM	CsiNet -PM
		K=2	K=4	K=8	K=64		
2	Ind.	-39.2	-38.8	-36.7	-39.6	-29.7	-29.8
	Out.	-17.8	-16.3	-16.1	-19.8	-18.9	-18.8
4	Ind.	-31.7	-32.0	-31.9	-31.5	-26.0	-25.9
	Out.	-13.6	-13.3	-12.6	-14.7	-15.3	-14.5
8	Ind.	-20.7	-21.8	-22.2	-24.3	-20.3	-19.1
	Out.	-11.5	-11.0	-10.6	-12.7	-12.3	-11.2
16	Ind.	-12.8	-12.3	-11.9	-15.4	-13.0	-12.0
	Out.	-10.3	-9.7	-9.5	-11.5	-10.2	-9.2

SAB decoder 2) SAB encoder plus pooling decoder 3) full-size encoder and decoder. We consider a subarray width of 2 for SAB encoder and decoder. Pooling decoder consists of 32 copies of SAB decoder and is followed by 2 residual blocks with 3×3 convolutional layers with 16, 8, 1 channels for pooling purpose. A full-size encoder and decoder are the SAB ones with $K = 64$. With respect to limited correlation between antennas, we can observe that the SAB encoder/decoder pair only causes slight performance degradation while requiring much less storage and computational burdens for UEs and base stations.

5.4.5 Testing Different Array Geometries

To show the scalability of SCENet+, Figure 5.14 shows the NMSE performance at different compression ratios and array geometries (8-element ULA, 16, 32-element UPAs) in indoor and outdoor scenarios. The results show no obvious performance difference for arrays of different sizes. This demonstrates the scalability of the proposed SAB framework.

5.4.6 BD SAB Framework in GN and LN approaches

The sparsity of CSI matrix in beam domain allows DCP feedback pruning for further uplink feedback reduction. On the other hand, it may cause power imbalance across antenna ports and performance degradation. Fortunately, this problem could be mitigated by LN.

Figure 5.15 shows the NMSE performance by applying GN and LN to SCEnet+ when $K = 2$ and 4 in indoor and outdoor channels. We observe better performance by selecting a smaller subarray width K because of limited correlation between adjacent beams. Additionally, we also see that performance improvement, especially for outdoor scenario, is achieved by utilizing LN approach. Since outdoor channels characterize with its low angular spread, this causes severe power imbalance problem over different subarray BD CSI matrices when using GN approach. The experiment results show that LN can effectively alleviate power imbalance problem. Note that LN is adopted in the following results.

5.4.7 DCP feedback pruning

In DCP feedback pruning, only subarray CSI matrices with energy ratio larger than T are encoded and fed back. The remaining are fed back to gNB with a bit "zero" as illustrated in Figure 5.5. For a better understanding, we define a metric, called *pruning ratio*, to be the ratio of the number of encoded subarray CSIs to all. Note that a larger T can increase pruning ratio but cause performance degradation.

Figures 5.16 and 5.17 show the NMSE performance under different pruning ratios in indoor and outdoor scenarios, respectively. The results suggest that the degradation of 20% pruning (pruning ratio = 0.2) is acceptable. Although low compression ratios appear to exhibit more severe performance loss in logarithm-scale, the actual discrepancy in MSE is quite small. From Figure 5.17, we can observe that pruning exhibits more advantages in outdoor case. It is because its high sparsity in beam domain gives

rise to many near-zero subarray CSI matrices which can be skipped with little CSI distortion.

5.4.8 2D SAB Framework

Figure 5.18 shows the NMSE performance versus compression ratios for different settings of N under subarray width $K = 2$. We find that 2D SAB framework with a small N degrades less when increasing pruning ratio. However, due to the low sparsity for each subarray CSI matrix, the 2D SAB framework with a small N performs worse than that with a large N . Note that the 2D SAB framework with $N = 32$ is equivalent to the original SAB framework operating in BD domain. Performance degradation due to a small N can be attributed to the aforementioned two factors: 1) incompatibility with the requirement of similar delay profile and 2) trade-off between sparsity and recovery performance. Yet, since the number of model parameters are nearly proportional to the input size squared, a smaller size of inputs in 2D SAB framework could further significantly reduce the model size of both encoder and decoder. However, the current model size using $K = 2$ is already under 1000 parameters, an extraordinarily small number for deep learning models. Further reduction of encoder model size appears to be less critical. However, since SAB framework can compress and recover in parallel, if the designer has strict computation time constraint, a 2D SAB framework may be a viable choice.

5.4.9 CSI feedback with dynamic CR

To show the benefits of the dynamic CR CSI feedback, we compare the recovery performance and codeword efficiency of the SAB CSI feedback frameworks with fixed and dynamic CRs. Since compression ratio cannot be perfectly controlled in dynamic CR

CSI feedback, we define an effective CR below for fair comparison

$$\text{CR}_{\text{eff}} = \frac{1}{D} \sum_{d=1}^D \frac{2N_t N_b}{\sum_i^{2N_b/K} (L_{d,i} + L_{\text{CR}})}. \quad (5.15)$$

L_{CR} denotes the prefix codeword length to indicate adopted CR (i.e., 2 bits), which is equivalent to $2/B$ codeword elements. B denotes the quantization bits used for each codeword element. The beam-domain sparsity in outdoor channels reduces the cost of uplink feedback with minor performance loss via DCP feedback pruning. Furthermore, by properly assigning CRs to subarray CSIs, we can achieve performance improvement and codeword efficiency.

We consider four possible CRs ($= 2, 4, 16, \infty$), where $CR = \infty$ denotes the case of DCP feedback pruning. We define an outage CSI estimate when its NMSE is higher than a predetermined $T_{\text{out}} = -5$ dB, rendering the CSI recovery unusable. We use an outage threshold $T_{\text{out}} = -10$ dB and $P_N = 0.01$ for training anchor points. Figure 5.19 shows the average outage probability and codeword efficiency in outdoor scenario. The optimal anchor points are located at $\mathbf{a} = [1, 0.018, 0.018, 0, 0]$. This result reveals that two CRs (i.e., $CR = 2, 16$) is sufficient to maximize codeword efficiency. This further suggests that DCP feedback pruning is relatively inefficient owing to oversimplifying the low-energy subarray CSIs. Moreover, the SAB framework via dynamic CR feedback (effective CR is 5.9) can achieve comparable outage probability against a fixed low $CR = 2$ (requiring the most resources and achieving the best recovery).

5.4.10 Different Noise Powers and CR Selections

From the previous results, we know that $CR = \infty$ is unused in dynamic CR. Therefore, we attempt an additional combination of CRs [2, 4, 8, 16]. Table 5.3 shows the optimal points trained with different choices of $P_N = [0.01, 0.0001, 1e - 7]$ and CR sets (i.e., [2, 4, 8, 16] and [2, 4, 16, ∞]) to maximize codeword efficiency. The results show that the

Table 5.3: The resulting five anchor points.

Noise Power	4 CRs	a_0	a_1	a_2	a_3	a_4
$P_N = 0.01$	[2,4,16, ∞]	1	0.018	0.018	0	0
	[2,4,8,16]	1	0.018	0.018	0.018	0
$P_N = 0.0001$	[2,4,16, ∞]	1	0.014	0.014	0	0
	[2,4,8,16]	1	0.014	0.014	0.014	0
$P_N = 1e - 7$	[2,4,16, ∞]	1	0.012	0.012	0	0
	[2,4,8,16]	1	0.012	0.012	0.012	0

optimal anchor points are insensitive to P_N and continue to suggest that we only need two CRs (CR = 2 and 16) for maximizing codeword efficiency. We conclude that the most efficient strategy is to use the lowest CR to secure those subarray CSIs with high significance and keep the codeword stream as compact as possible for subarray CSIs with low energy. Also, we only need 1-bit information for acknowledging the adopted CR to gNB.

Figure 5.20 shows the NMSE performance and outage probability via fixed CR and dynamic CR feedback. The anchor points shown in Table 5.3 are trained with different noise powers. The results show that dynamic CR manner not only improves the outage probability but also leads to better recovery performance than fixed CR for outdoor channels.

5.5 Conclusions

This work proposes a lightweight deep-learning architecture for encoding and feeding back downlink CSI in massive MIMO wireless systems. This new CSI feedback framework flexibly accommodates different numbers of antenna ports in use and also requires lower computational and storage hardware at resource constrained UEs. By developing a SAB CSI feedback framework, a common encoder allows encoding of subarray CSI matrices separately. We further develop a dynamic encoding principle to flexibly compress subarray CSI matrices by applying dynamic compression ratios according to

their significance. The new framework includes a channel-based CR selector at UE for determining CRs to achieve the maximum of codeword efficiency. Numerical results show the proposed framework generally outperforms the SOTAs, CsiNet-SM and CsiNet-PM. In summary, the proposed SAB framework heralds a simple and systematic CSI feedback manner with higher flexibility, and scalability while requiring lower storage and computational complexity.

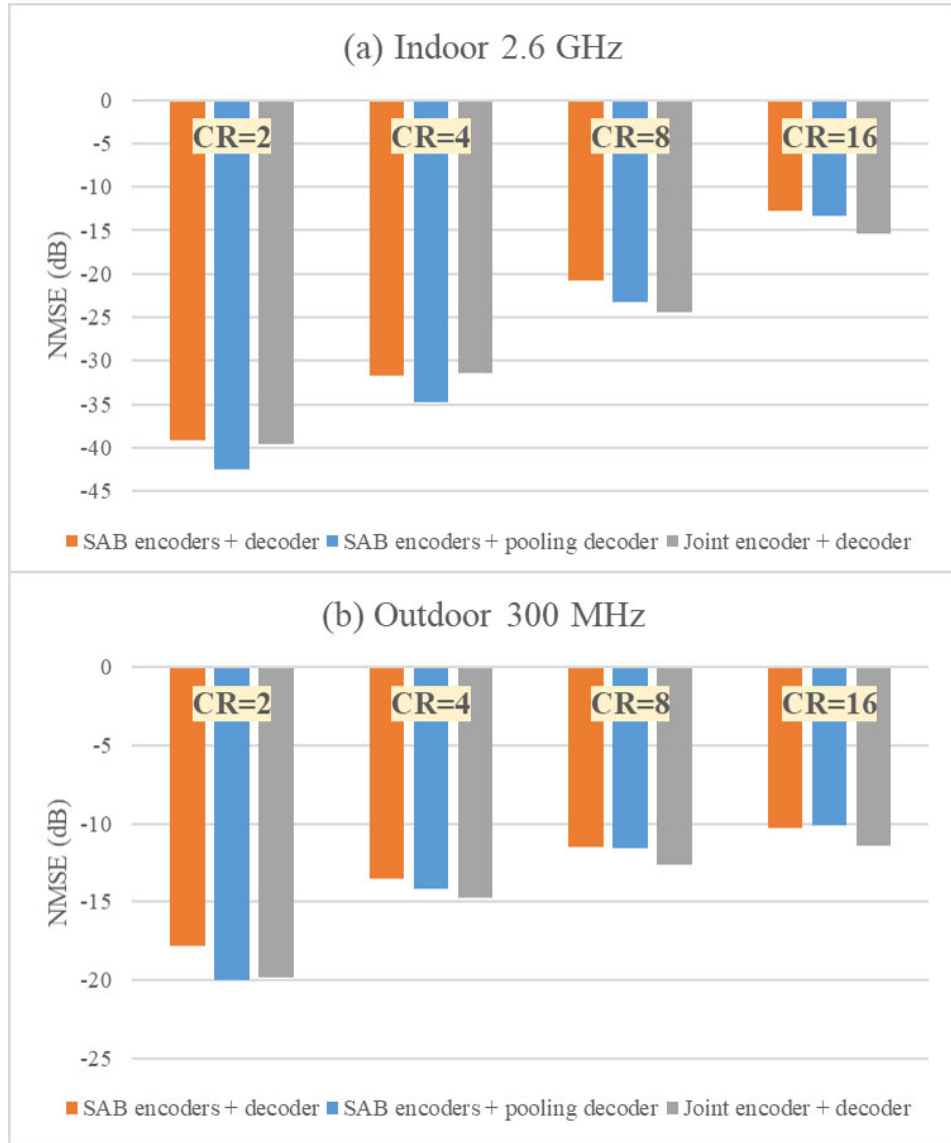


Figure 5.13: NMSE performance versus compression ratios with different encoder/decoder pairs in (a) indoor and (b) outdoor scenarios.

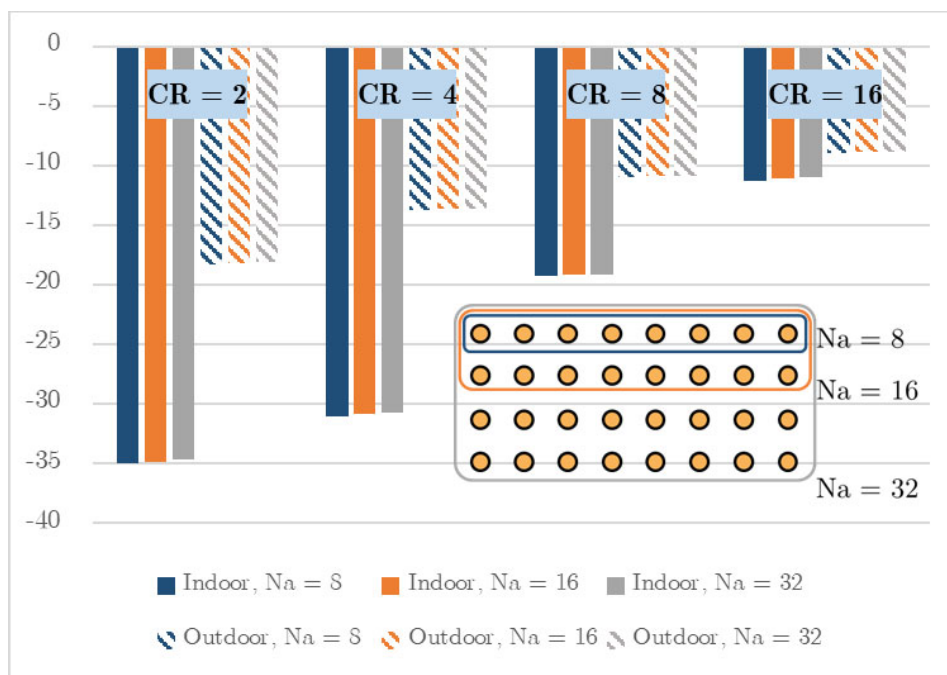


Figure 5.14: NMSE performance of SCENet+ for arrays with different array geometries. We consider 8-element ULA and 16- and 32-element UPAs.

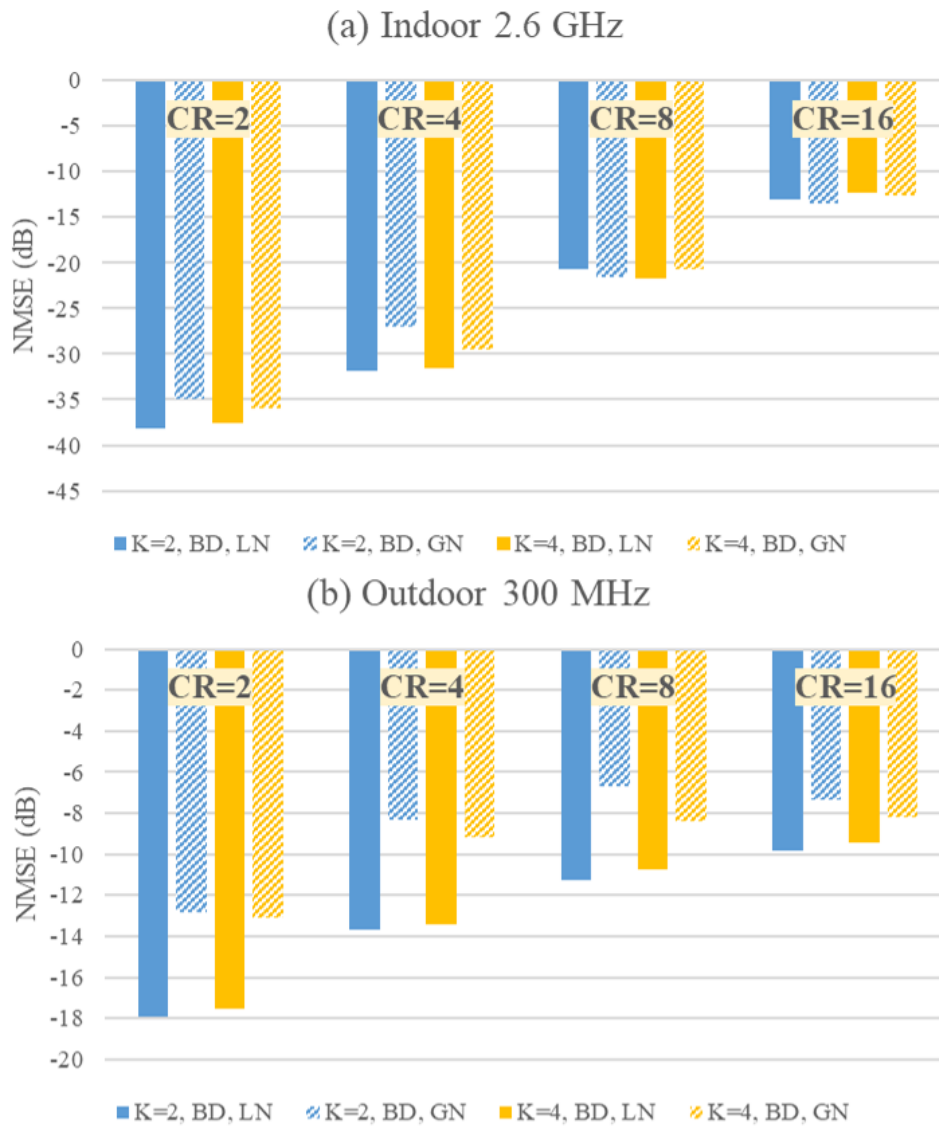


Figure 5.15: NMSE performance versus compression ratios with or without local normalization (LN) in (a) indoor and (b) outdoor scenarios.

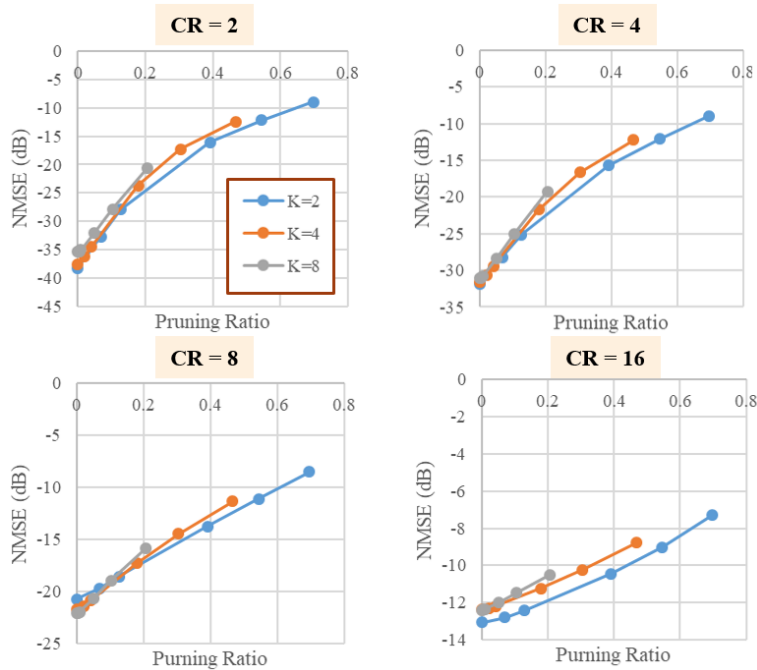


Figure 5.16: NMSE performance versus pruning ratio for different selections of subarray width K in indoor scenario.

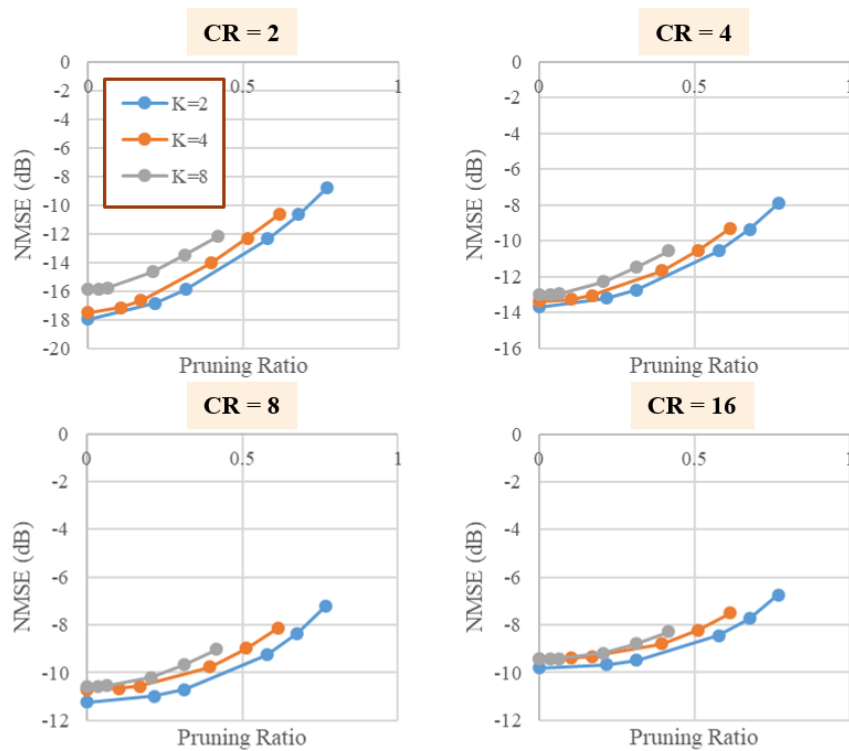


Figure 5.17: NMSE performance versus pruning ratio for different selections of subarray width K in outdoor scenario

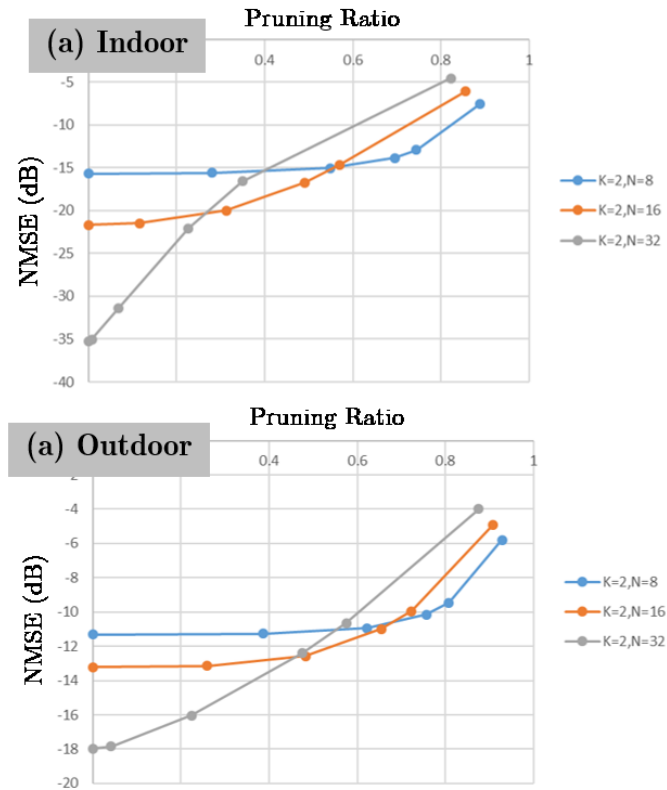


Figure 5.18: NMSE performance versus pruning ratio for different N in (a) indoor and (b) outdoor scenarios.

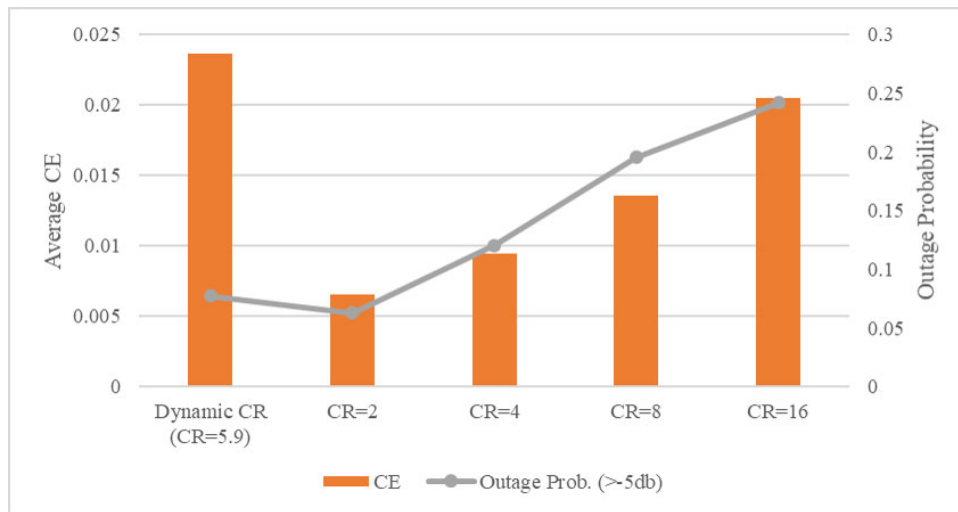


Figure 5.19: Average CE and outage probability for dynamic CR and fixed CR CSI feedback framework in outdoor scenario.

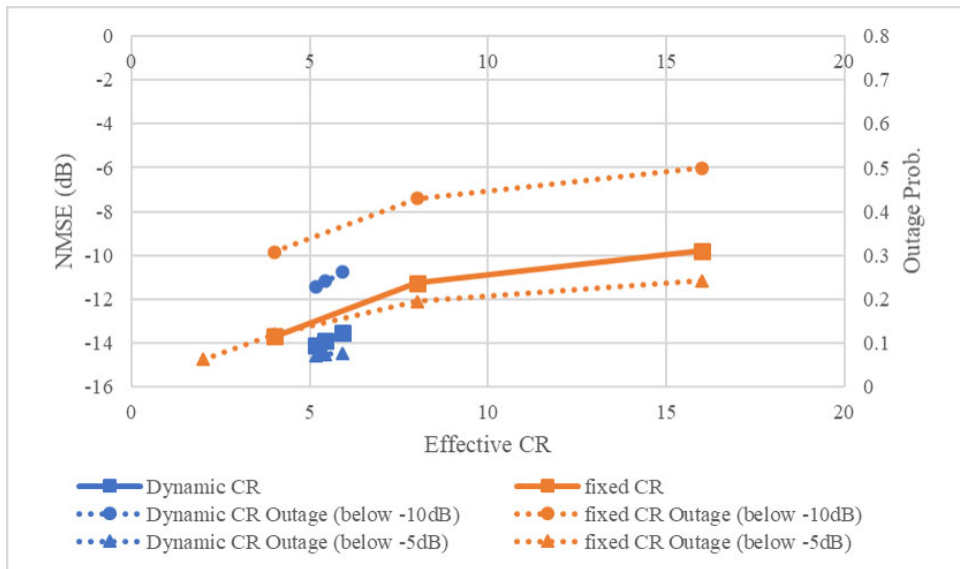


Figure 5.20: NMSE performance and outage probability for dynamic CR and fixed CR CSI feedback framework in outdoor scenario.

Chapter 6

Applying JPEG Compression for Feedback of Massive MIMO Channel State Information

DL CSI acquisition plays a vital role in massive MIMO FDD systems. To improve spectrum and energy efficiency, deep learning architectures for UE-side CSI feedback and basestation-side recovery show notable improvement in feedback efficiency and CSI recovery accuracy. However, deep learning based approaches often manifest practical inflexibility since DL models are customized and trained for typical RF channel environments at specific compression level. The simplicity and success of image compression algorithms for a wide range of images motivates us to investigate their adoption for CSI compression and recovery. This work proposes a model-free JPEG architecture for compressive CSI feedback that easily accommodates a variety of channel types and compression ratios. We present a lossless entropy encoder to further lower CSI feedback bandwidth. Test results of the proposed simple algorithm demonstrate very competitive CSI recovery accuracy and feedback efficiency for various propagation channels against DL models.

In this chapter, we first introduce JPEG image compression in Section 6.1. Then we reveal the proposed JPEG-based CSI feedback framework in Section 6.2. In Section 6.3, we demonstrate the simulation results as compared to the learning-based SOTAs. Finally, we give conclusion and future work in Section 6.4.

6.1 JPEG Image Compression

JPEG is one of the most successful image lossy compression methods. JPEG achieves efficient compression by focusing on visually sensitive components. We now briefly summarize JPEG compression to ease the presentation of our proposed CSI feedback framework inspired thereof.

Color Transformation and DCT Transformation

JPEG first splits an image for encoding into small square blocks $\widetilde{\mathbf{M}}_{\text{RGB}} \in \mathbb{U}_8^{8 \times 8 \times 3}$ for separate compression and recovery, where \mathbb{U}_8 denotes a 8-bit unsigned integer set from 0 to 255. Each block $\widetilde{\mathbf{M}}_{\text{RGB}}$ is transformed into blocks of colors $\widetilde{\mathbf{M}}_{\text{YBR}} \in \mathbb{U}_8^{8 \times 8 \times 3}$ (including luminance $\widetilde{\mathbf{M}}_{\text{Y}} \in \mathbb{U}_8^{8 \times 8 \times 1}$, blue component $\widetilde{\mathbf{M}}_{\text{B}} \in \mathbb{U}_8^{8 \times 8 \times 1}$ and red component $\widetilde{\mathbf{M}}_{\text{R}} \in \mathbb{U}_8^{8 \times 8 \times 1}$). Here we only use the 8×8 brightness block $\widetilde{\mathbf{M}}_{\text{Y}}$ as an example for brevity. Then, we apply 2D DCT-II, termed as DCT in the following for simplicity, to the brightness block $\widetilde{\mathbf{M}}_{\text{Y}} \in \mathbb{U}_8^{8 \times 8}$ given below

$$\mathbf{M}_{\text{Y}} = \mathbf{T}^H \widetilde{\mathbf{M}}_{\text{Y}} \mathbf{T} \quad (6.1)$$

where $\mathbf{T} \in \mathbb{R}^{8 \times 8}$ is a DCT transformation matrix.

Quantization and Dequantization

After 2D-DCT transformation, the upper-left area captures the low-frequency components that dominate the image. According to human vision sensitivity to different

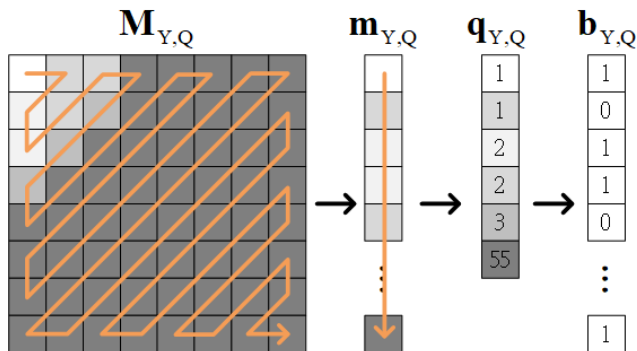


Figure 6.1: Example of entropy encoding in JPEG compression.

components, higher precision is given to low-frequency components via element-wise division by a quantization matrix \mathbf{Q} to yield a transformed luminance block \mathbf{M}_Y :

$$\mathbf{M}_{Y,Q} = \lfloor \mathbf{M}_Y \oslash \mathbf{Q} \rfloor, \quad (6.2)$$

where $\lfloor \cdot \rfloor$ denotes rounding down. The exact quantization matrices \mathbf{Q} can be found in [49].

Entropy Encoding and Decoding

$\mathbf{M}_{Y,Q}$ is rearranged as a vector $\mathbf{m}_{Y,Q} \in \mathbb{U}_8^{64}$ in a zigzag order and encoded as a codeword stream $\mathbf{q}_{Y,Q} \in \mathbb{U}_8^{64}$ via run-length encoding (RLE) algorithm. From the distribution of $\mathbf{M}_{Y,Q}$, zigzag reordering increases likelihood of repetitive zeros. Due to high efficiency of RLE with data that contain many repetitive segments, RLE provides good compression for $\mathbf{m}_{Y,Q}$. Huffman encoding [49] can transform the codeword stream $\mathbf{q}_{Y,Q}$ into a bit stream $\mathbf{b}_{Y,Q}$. Figure 6.1 demonstrates an example of entropy encoding in JPEG compression.

6.2 JPEG-based CSI Feedback Framework

Recent works on DL-based CSI feedback frameworks have shown successes in UE compression of CSI and recovery by BSs. However, one remaining problem is the rigidity of DL frameworks that need to customize multiple DL models for different channel scenarios and multiple compression levels. This shortcoming makes it difficult to apply DL models flexibly at different compression levels and for practical gNBs which often may consist of different numbers of MIMO antennas. The need for UEs to store multiple pre-trained DL models is detrimental to widespread deployment, especially at low cost UEs. Furthermore, training different DL configurations targeting different compression ratios, different numbers of antennas, and various channel scenarios would require suitable channel models and large amount of training data for each channel scenario.

We aim to develop a simple, scalable, and general compressive CSI feedback algorithm. Note that 2D CSI matrices are similar image data. In particular, CSI exhibits certain delay and angular sparsity that can be revealed by DFT/DCT transformations. On the other hand, CSI data also are different from images. For example, due to multipath propagation, CSI energy may spread over various directions, leading to distinct patterns from that of visual images. Furthermore, unlike human vision perception, a small amount of high-frequency CSI discrepancy may further lead to notable performance loss by MIMO precoders. Since JPEG may not be directly applicable to CSI compression, we apply domain knowledge and develop a model-free JPEG-based compressive CSI feedback framework that accommodates different propagation scenarios, compression ratios, pilot numbers, and antenna array configurations.

6.2.1 Ordering Real/Imaginary CSI

JPEG compression is scalable to different image sizes for ease of hardware design. However, careless use of JPEG on CSI matrices may obscure sparsity and lead to compression degradation. To this end, we separate each full BD domain CSI matrix $\mathbf{H} \in \mathbb{C}^{N_b \times N_t}$ for compression into two real-value CSI matrices as $\mathbf{H}_R = \text{Real}(\mathbf{H}) \in \mathbb{R}^{N_b \times N_t}$ and $\mathbf{H}_I = \text{Imag}(\mathbf{H}) \in \mathbb{R}^{N_b \times N_t}$ as images to be processed in parallel.

6.2.2 Zero Replacement (ZR)

An fixed quantization matrix design for CSI compression is intractable since the CSI energy may come from any directions. To efficiently truncate insignificant information, we need a dynamic sampling approach responsive to changing CSI energy distributions. We represent CSI input matrix \mathbf{M} (e.g., \mathbf{H}_R or \mathbf{H}_I) as (1) a sequence \mathbf{m} and (2) an indicator matrix \mathbf{I} . We down-sample the input matrix \mathbf{M} by a factor CR (compression ratio) to obtain \mathbf{m} consisting only of the top $(1/CR)\%$ elements ranked by magnitude. The indicator matrix $\mathbf{I} \in \{0, 1\}^{N_b \times N_t}$ contains only 1's and 0's corresponding to sampled and discarded elements, respectively.

6.2.3 Entropy Encoding/Decoding

Figure 6.2 shows the process flow of the proposed CSI entropy encoding. For effective binary representations, we express the downsampled sequence \mathbf{m} as bit stream Ω_m by μ -law companding entropy encoding with Q bits. We propose a modified RLE (mRLE) to reduce transmission overhead of the indicator matrix \mathbf{I} . Following the pseudo code of mRLE shown in (Alg. 6.1), we obtain a symbol list Ω_S marking the numbers of consecutive zeros between ones in a back-and-forth scanning pattern shown in Figure 6.2. The final entry EOS in the symbol list Ω_S denotes "end of symbols" and also means that no more ones in the remaining sequence.

Algorithm 6.1 Modified RLE

Require: $\Omega_S = \{\}, \mathbf{I}$
Ensure: Ω_S
 $\mathbf{i} \leftarrow \text{vec}(\mathbf{I})$
 $N_0 \leftarrow 0$
for $k = 1 : 1 : \text{length}(\mathbf{i})$ **do**
 if $\mathbf{i}(k) = 1$ **then**
 $\Omega_S \leftarrow \{\Omega_S, N_0\}$
 $N_0 \leftarrow 0$
else
 $N_0 \leftarrow N_0 + 1$
end if
end for
if $N_0 \geq 1$ **then**
 $\Omega_S \leftarrow \{\Omega_S, \text{EOS}\}$
end if

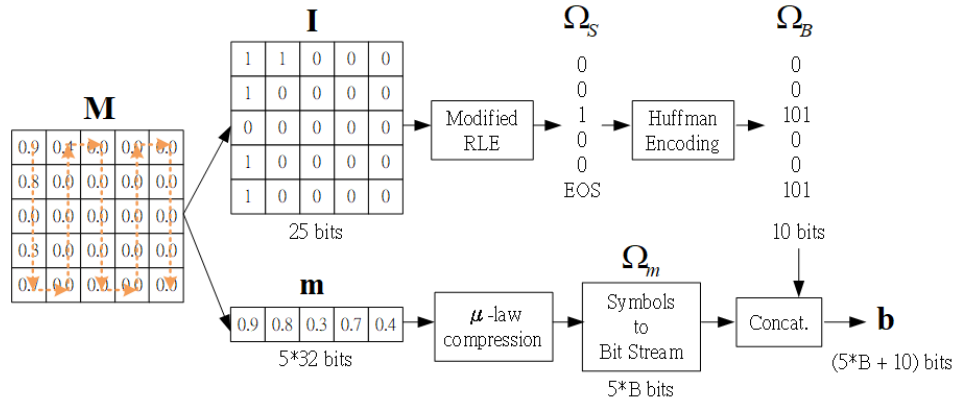


Figure 6.2: Example of CSI entropy encoding (Note that the numbers 0.0 represent negligible elements in matrix \mathbf{M}).

We next transform the symbol list Ω_S into a bit stream Ω_B by a modified Huffman coding. Owing to the back-and-forth scanning pattern and CSI energy distribution in BD domain, as shown in the symbol histogram (Figure 6.3), cases with small numbers of consecutive zeros between ones are the majority. To lower transmission cost, we design a modified Huffman coding Table 6.1. Note that we use three bits (101) to represent *EOS*. The more frequently used symbols are presented in a small-size bit stream. Prefix bit streams are designed for unambiguous matching. By reversing the Huffman coding (decoding) and ZR, we could recover an estimate of input matrix $\widehat{\mathbf{M}}$.

Table 6.1: Modified Huffman coding table

Size	Number of zeros					Bit stream
0	0					0
1	1	EOS				10
1	2	3				110
2	4	5	6	7		1110
3	8	9	...	14	15	11110
4	16	17	...	30	31	111110
5	32	33	...	62	63	1111110
6	64	65	...	126	127	11111110
7	128	129	...	254	255	111111110
8	256	257	...	510	511	1111111110
9	512	513	...	1022	1023	11111111110

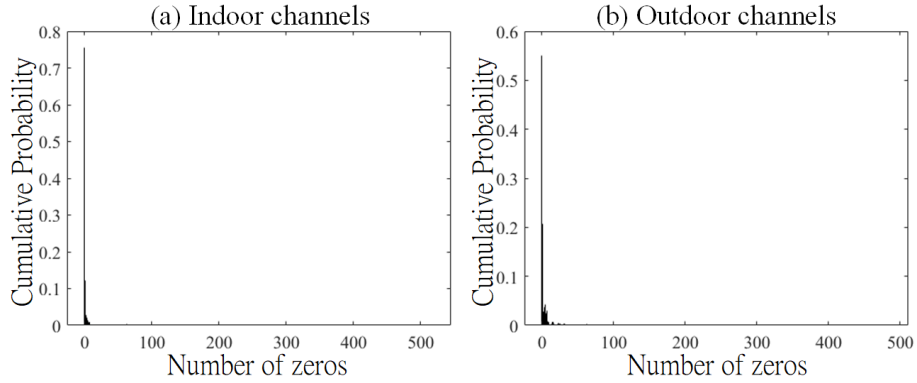


Figure 6.3: Accumulative ratio of symbols in symbol list Ω_S for (a) indoor channels at 5 GHz and (b) outdoor channels at 300 MHz generated by QuaDRiGa channel simulator.

6.3 Experimental Evaluations

6.3.1 Experiment Setup

In our tests, we consider both indoor and outdoor channels. We generate CSIs using widely used channel model softwares [39, 40]. Our configuration features a gNB of height equal to 20 m at the center of a circular cell with 30 m radius for indoor setting and 200 m radius for outdoor environment. We consider gNB with a 8×4 UPA and gNB with a 32-element ULA serving single-antenna UEs in QuaDRiGa and COST2100 simulators, respectively. Both UPA and ULA elements have half-wavelength uniform spacing.

For both indoor and outdoor channels, we utilize QuaDRiGa and COST2100 simulators [39, 40] using scenario features given in 3GPP TR 38.901 *Indoor (QuaDRiGa)* and *IndoorHall 5GHz(COST2100)* at 5.1 and 5.3 GHz, and in 3GPP TR 38.901 *UMa (QuaDRiGa)* and *SemiUrban 300MHz (COST2100)* at 300 and 330 MHz of uplink and downlink with LOS paths, respectively. For each data channel, we consider $N_f = 1024$ subcarriers with 15K-Hz spacing and place $M_f = 86$ pilots with downsampling ratio $DR_f = 12$. We set antenna type to “**omni**”. To accurately assess recovery accuracy, we assume UEs have accurate CSI estimates. We use NMSE as the performance metric.

We compare our framework with multiple DL-based models, CsiNet[5], CRNet[9], CsiNetPro[17]. We apply DCT and DFT transformations in our framework, termed as DCT-ZR and DFT-ZR respectively. For DL models, we set the number of epochs to 1000 and use batch size of 200. We start with learning rate of 0.001 before switching to 5×10^{-4} after 300 epochs. Using channel simulators, we generate several indoor and outdoor datasets, each containing 100,000 random channels. We use one seventh of these channels as test data for performance evaluation. The remaining channels are split into 2/3 and 1/3 for training and validation, respectively.

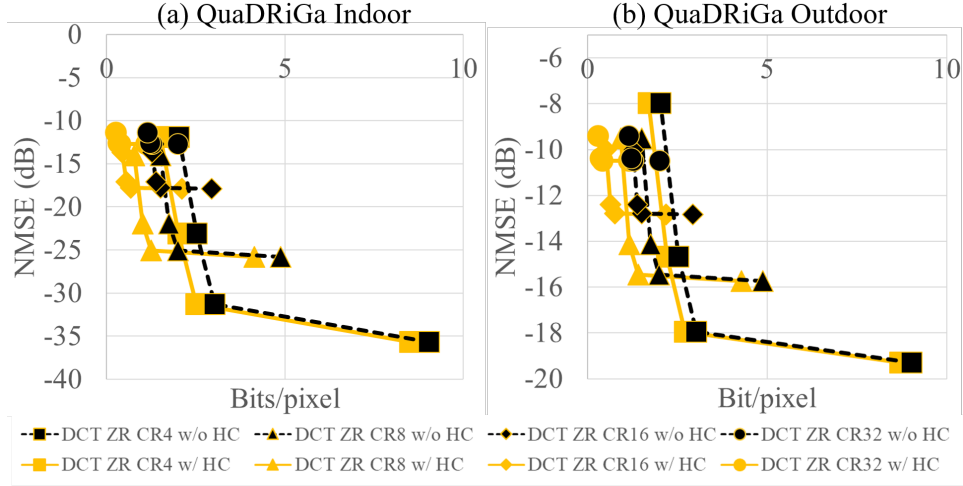


Figure 6.4: NMSE performance versus BPP for DCT ZR with and without Huffman coding under different compression ratios (four anchor points of each curve corresponds to the use of quantization bits $Q = 4, 6, 8, 32$).

6.3.2 Benefit of Huffman Encoding

Without using Huffman encoding, transmission overhead of the 32×32 indicator matrix \mathbf{I} is 1024 bit and hence we need at least 1 bit per pixel (BPP)¹. Fig 6.4 shows the NMSE performance with and without Huffman coding under different compression ratios and quantization levels. We observe that DCT-ZR breaks through the theoretical boundary of $\text{BPP} = 1$ by leveraging Huffman codes.

6.3.3 DCT and DFT transformation

DCT transformation provides very compact image representation and yields good compression efficiency in classic JPEG application. On the other hand, DFT can effectively transform complex periodic features into low-dimensional subspaces in CSI compression. Figure 6.5 shows the NMSE performance under different compression ratios at various quantization levels for DCT-ZR and DFT-ZR in indoor and outdoor channels generated by QuaDRiGa and COST2100. We see that DCT transformation deliv-

¹We need at least one bit to show whether the corresponding element is transmitted or not for each pixel (element) even if we choose $CR = \infty$.

ers better recovery performance for most scenarios (QuaDRiGa UMa Outdoor at 300 MHz, QuaDRiGa and COST2100 indoor at 5 GHz). Meanwhile, DFT performs better in COST2100 outdoor channels. Figure 6.6 shows the power delay profile of different channels under test. We discover that DCT performs less efficiently on COST outdoor channels that give large DS. Unlike DFT, the low-frequency DCT sidelobes folds back to superimpose instead of wrapping around. This allows channels with low DS at the center of low frequency to provide more sparse representation via DCT than DFT transformation. Since our framework is not rigidly frozen to a specific transformation, it allows UE to chose the best transformation based on the sparsity of results from multiple transformations. Specifically, UE only need 1-bit information for indicating the selected transformation between DFT/DCT to the serving gNB for better recovery performance.

6.3.4 Testing different channel scenarios

Training based DL model optimization for different channel scenarios and compression ratios elevates the difficulty for broad practical deployment of DL-based compressive MIMO CSI feedback in FDD communications systems. On the other hand, our proposed DCT-ZR approach is directly applicable to different channel scenarios, different antenna sizes, and various compression ratios.

Figure 6.7 compares the NMSE performance of DCT ZR, CsiNet, CsiNetPro, CR-Net at different compression ratios and quantization levels in QuaDRiGa channels. For QuaDRiGa channels, DCT-ZR exhibits superior recovery performance at higher feedback rate while delivering performance comparable to DL-based benchmarks when considering smaller feedback bandwidth.

To illustrate the scalability and flexibility of the proposed framework, we test different propagation channels. Figure 6.8 shows the NMSE results for DCT-ZR, CsiNet, CsiNetPro, and CRNet when $BPP = 1$ and 2 , respectively for both QuaDRiGa and

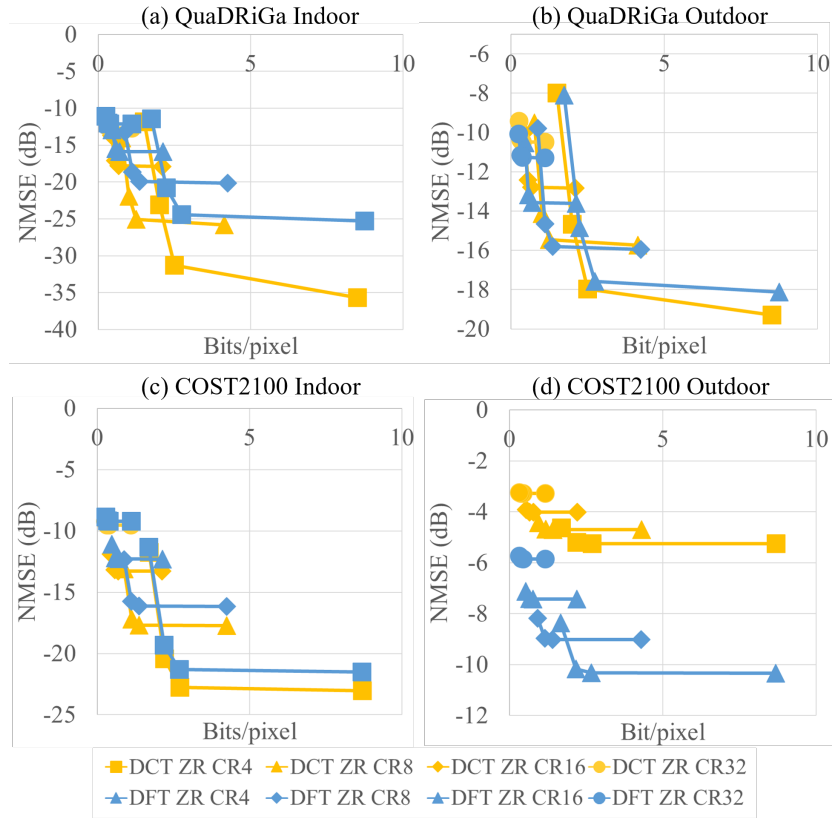


Figure 6.5: NMSE performance versus BPP for DCT and DFT ZR under different compression ratios (four anchor points of each curve corresponds to the use of quantization bits $Q = 4, 6, 8, 32$).

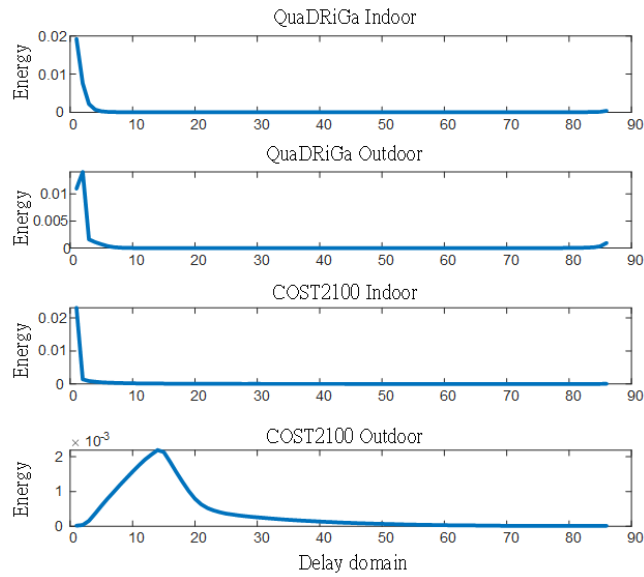


Figure 6.6: Power delay profile for indoor and outdoor channels generated from QuaDRiGa and COST2100 simulators.

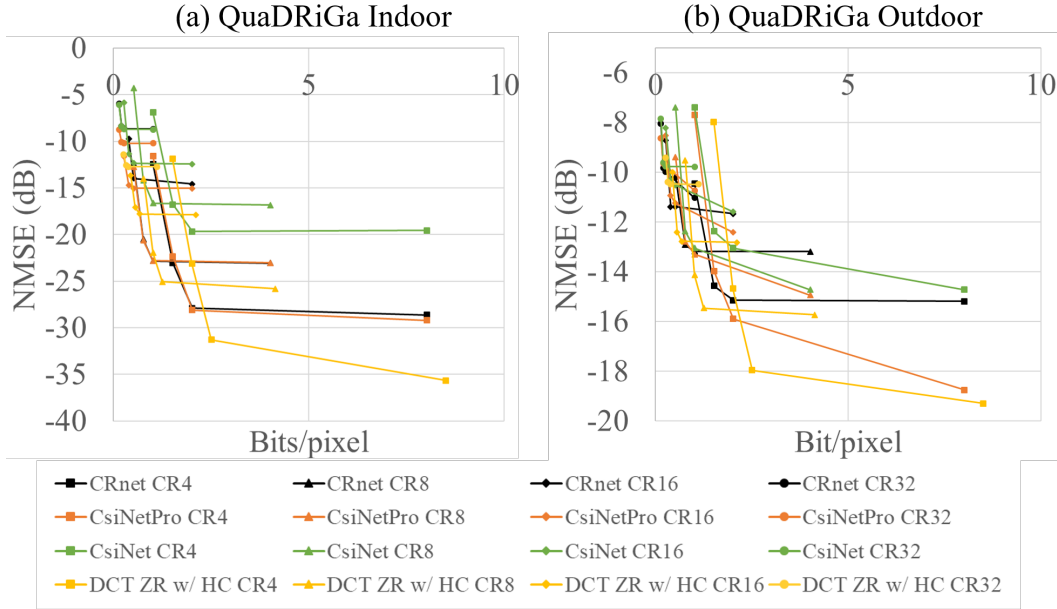


Figure 6.7: NMSE performance versus BPP for CsiNet, CRNet, CsiNetPro, DCT-ZR under different compression ratios (four anchor points of each curve corresponds to the use of quantization bits $Q = 4, 6, 8, 32$).

COST2100 channels. DCT-ZR continues to outperform other DL-based compressive feedback models, except in the case of COST2100 indoor channels. Nevertheless the simple DCT-ZR algorithm still delivers highly competitive performance in comparison to the DL-models dedicated for each specific channel and are optimized through training with large size datasets, each with as many as 57,000 channels. These comparative results confirm that the proposed simple, scalable, and flexible DCT-ZR algorithm is highly effective in compressive CSI feedback and recovery of massive MIMO channel information.

6.4 Conclusions

We propose a low complexity model-free CSI feedback framework for encoding and recovering downlink CSI in massive FDD MIMO wireless systems. Inspired by the success of image compression, this framework exploits sparsity of CSI matrices in massive

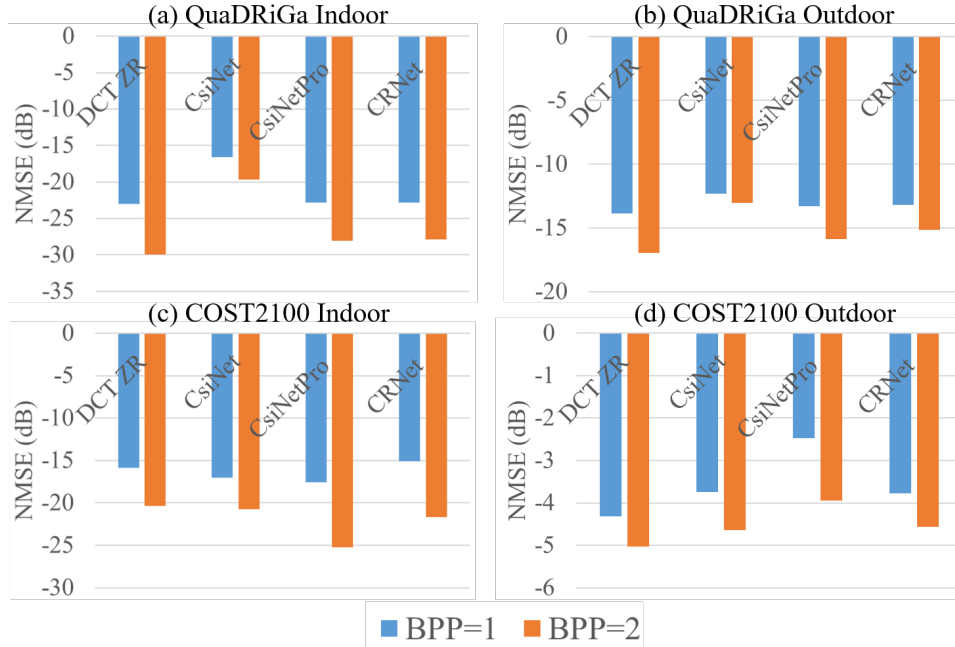


Figure 6.8: NMSE performance versus BPP for CsiNet, CRNet, CsiNetPro, DCT-ZR under different compression ratios (four anchor points of each curve corresponds to the use of quantization bits $Q = 4, 6, 8, 32$).

MIMO wireless systems to design a JPEG-inspired compressive feedback mechanism. Free from customized training for different propagation channels at various compression ratios, this modified feedback framework neither requires high volumes of training data nor needs to configure multiple DL models for different RF channel environments and/or different compression ratios. Unlike DL approaches, this new ZR model can be directly applied to new and unseen channel scenarios without pre-training or customization. This flexible and scalable framework is simple to implement and amenable to broad deployment in practical massive MIMO wireless systems. Numerical results demonstrate DCT-ZR performance to be competitive with complex state-of-the-art DL models such as CsiNet, CRNet and CsiNet-Pro in most tested propagation channels. To highlight, this new JPEG-based framework heralds a simple and easy-to-deploy CSI feedback approach that do not require large dataset and can be rapidly deployed without any prior training.

Chapter 7

Physics-Inspired Deep Learning

Anti-Aliasing Framework in

Efficient Channel State Feedback

Acquiring downlink channel state information (CSI) at the base station is vital for optimizing performance in massive Multiple input multiple output (MIMO) Frequency-Division Duplexing (FDD) systems. While deep learning architectures have been successful in facilitating UE-side CSI feedback and gNB-side recovery, the undersampling issue prior to CSI feedback is often overlooked. This issue, which arises from low density pilot placement in current standards, results in significant aliasing effects in outdoor channels and consequently limits CSI recovery performance. The main objective of this work is to solve this issue by introducing a new CSI upsampling framework at the gNB as a post-processing solution to address the gaps caused by undersampling. Leveraging the physical principles of discrete Fourier transform shifting theorem and multipath reciprocity, our framework effectively uses uplink CSI to mitigate aliasing effects. We further develop a learning-based method that integrates the proposed algorithm with the Iterative Shrinkage-Thresholding Algorithm Net (ISTA-Net) architecture, enhanc-

ing our approach for non-uniform sampling recovery. Our numerical results show that both our rule-based and deep learning upsampling methods significantly outperform traditional interpolation techniques and current state-of-the-art approaches by 8-13 dB and 2-10 dB, respectively, in terms of normalized mean square error.

In this chapter, in Section 7.1, we first described the aliasing issue in explicit CSI feedback. Then, in Section 7.2, to tackle aliasing issue, we proposed a UL-CSI-aided CSI upsampling method with exploitation of FDD multipath reciprocity. In Section 7.3, we proposed an AI-driven CSI upsampling approach, SRCsiNet, which elegantly guides each part of the NN with desired functions. In Section 7.4, we further proposed an advanced version of SRCsiNet, which is called SRISTANet, which take the advantages of the two networks ISTANet and SRISTANet to perform CSI upsampling from non-uniform sampled CSI. In Section 7.5, test results demonstrate superior performance, good scalability, high efficiency of SRCsiNet and SRISTANet for high outdoor channels. We also show the downsides of SRISTANet and the suggestions when applying to the practical system. Finally, we give conclusion in Section 7.6.

7.1 Problem Formulation: Aliasing Issue in CSI Feedback

7.1.1 DL CSI Preprocessing

We consider a single-cell MIMO FDD link where a gNB with N_b antennas serves a plurality of single-antenna UEs. Following 3GPP technical specifications, sparse pilot symbols (i.e., CSI-RS) are uniformly distributed in frequency domain for DL channel acquisition. Assuming each subband contains N_f subcarriers with a spacing of Δf and a pilot spacing of D_{RS} subcarriers, adjacent CSI-RSs are separated by $D_{RS} \cdot \Delta f$ Hz. We denote $\mathbf{h}_i \in \mathbb{C}^{M_f \times 1}$ as CSI-RS DL CSI of the i -th antenna at gNB at M_f pilot

positions. Let the superscript $(\cdot)^H$ denote the conjugate transpose. By collecting CSI of each gNB, a pilot sampled DL CSI matrix \mathbf{H}_{RS} relates to the full DL CSI matrix $\mathbf{H} \in \mathbb{C}^{N_b \times N_f}$ via

$$\mathbf{H}_{\text{RS}} = \mathbf{H}\mathbf{Q}_{D_{\text{RS}}} = \begin{bmatrix} \mathbf{h}_1 & \mathbf{h}_2 & \cdots & \mathbf{h}_{N_b} \end{bmatrix}^H \in \mathbb{C}^{N_b \times M_f},$$

where $\mathbf{Q}_{D_{\text{RS}}} = [\mathbf{e}_1, \mathbf{e}_{1+D_{\text{RS}}}, \dots, \mathbf{e}_{1+(M_f-1)D_{\text{RS}}}] \in \mathbb{C}^{N_f \times M_f}$ is a downsampling matrix with pilot rate D_{RS} with $\mathbf{e}_i \in \mathbb{C}^{N_f}$ being the i -th column vector of an identity matrix of size N_f .

7.1.2 DL CSI Feedback

Autoencoder has shown success in CSI compression. An encoder at UE compresses its estimated DL CSI based on reference signals for UL feedback and a decoder at gNB recovers the CSI according to the feedback from UE. Before compression and after recovery, some works [5, 50] may or may not transform CSI into the domain with sparse features as pre-processing, which usually only pose slight impact. Many have exploited convolutional and fully connected layers to compress and recover the DL pilot CSI via

$$\text{Encoder: } \mathbf{q} = f_{\text{en}}(\mathbf{H}_{\text{RS}} + \mathbf{N}),$$

$$\text{Decoder: } \hat{\mathbf{H}}_{\text{RS}} = f_{\text{de}}(\mathbf{q}).$$

We note that the size of the codeword $\mathbf{q} \in \mathbb{C}^{\frac{N_b M_f}{CR}}$ for the UL feedback is determined by a specific compression ratio CR . We can evaluate the feedback loss by the NMSE of the pilot DL CSI:

$$Loss_{\text{FB}}(\hat{\mathbf{H}}_{\text{RS}}, \mathbf{H}_{\text{RS}}) = \sum_{d=1}^D \frac{\|\hat{\mathbf{H}}_{\text{RS},d} - \mathbf{H}_{\text{RS},d}\|_F^2}{\|\mathbf{H}_{\text{RS},d}\|_F^2},$$

where subscript d denotes the d -th random test.

7.1.3 Aliasing Issue

Fig. 7.1 demonstrates the block diagram of a practical explicit CSI feedback framework. Since UE can only acquire \mathbf{H}_{RS} , gNB needs to upsample $\hat{\mathbf{H}}_{RS}$ the actual full DL CSI, denoted as \mathbf{H} after CSI encoding and decoding for precoder design. Our primary interest shifts towards the total discrepancy between the actual full DL CSI, denoted as \mathbf{H} , and the estimated full DL CSI, denoted as $\hat{\mathbf{H}}$. The discrepancy is given as follows:

$$Loss = \text{NMSE}(\hat{\mathbf{H}}, \mathbf{H}) = \sum_{d=1}^D \frac{\|\hat{\mathbf{H}}_d - \mathbf{H}_d\|_F^2}{\|\mathbf{H}_d\|_F^2},$$

$$\hat{\mathbf{H}} = f_{\uparrow}(f_{\text{de}}(f_{\text{en}}(\mathbf{H}_{RS} + \mathbf{N}))),$$

where $f_{\uparrow}(\cdot)$ is the upsampling operation and $\hat{\mathbf{H}} \in \mathbb{C}^{N_b \times N_f}$ is the estimated DL CSI after upsampling/interpolation.

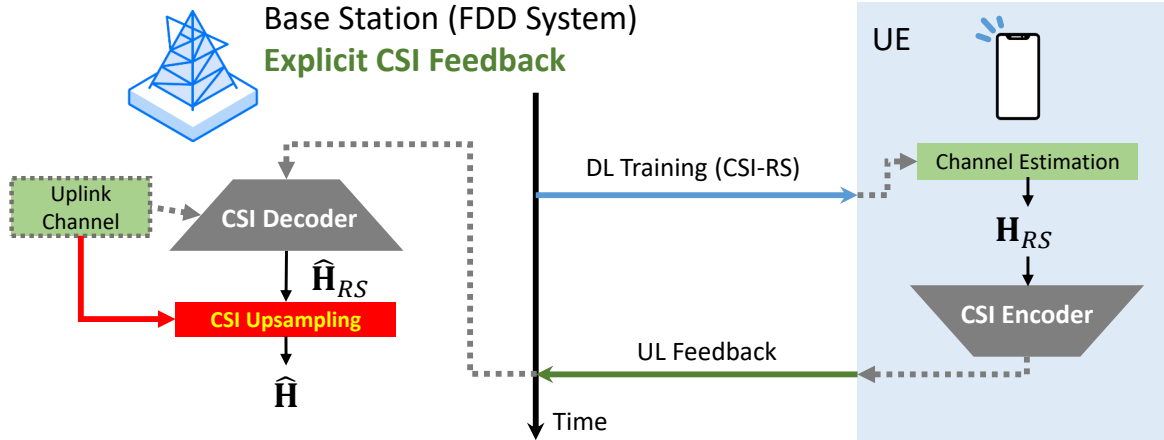


Figure 7.1: Block diagram of a CSI feedback framework. To the best of our knowledge, all previous works neglected the necessity to upsample from RS CSI to full DL CSI or they assumed that UE is able to acquire full DL CSI, which is not practical. This work aims to design a CSI upsampler that leverages uplink channels and side information against the aliasing issue.

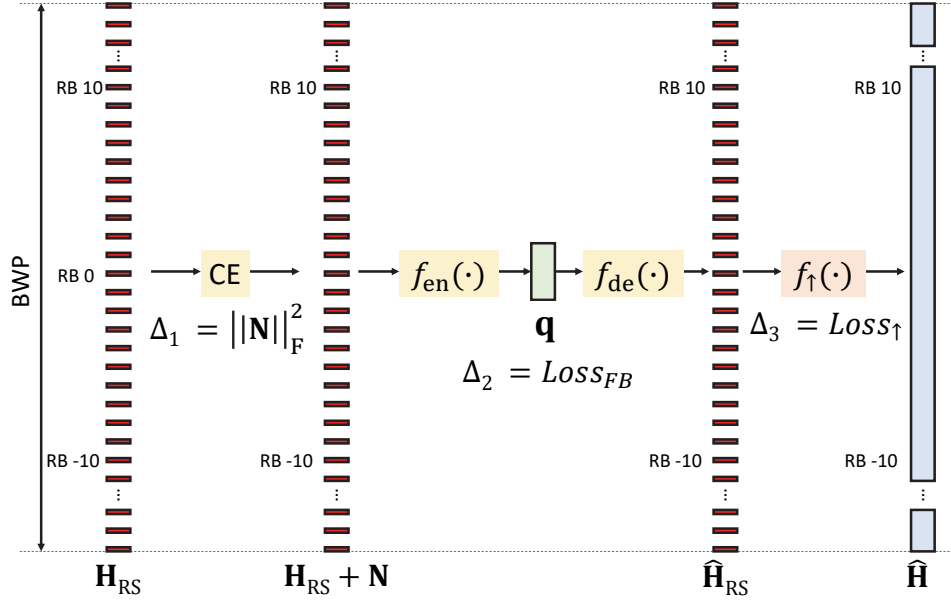


Figure 7.2: Illustration of the total discrepancy related to the losses at different stages. Δ_1 , Δ_2 and Δ_3 denote the distortions from channel estimation at UE side, feedback from UE to gNB, and upsampling, respectively.

As shown in Figure 7.2, the total discrepancy in recovering the full DL CSI, denoted as $Loss$, arises from three main factors: channel estimation (CE) noise \mathbf{N} , feedback loss $Loss_{FB}$, and upsampling/interpolation loss $Loss_{\uparrow}$. The CE loss, resulting from imperfect CE at the UE side, has been effectively addressed by rule-based methods like Least Square (LS) and MMSE estimation [51], as well as advanced learning-based denoising networks [52, 53]. Feedback loss, due to limited CSI feedback, has been extensively explored in existing CSI feedback frameworks [5, 13]. However, there has been less focus on upsampling loss. This loss occurs when interpolating full DL CSIs from a limited number of known estimated pilot DL CSIs. While feedback loss $Loss_{FB}$ is typically predominant in indoor propagation channels, the insufficient density of current CSI-RS placements means that upsampling loss $Loss_{\uparrow}$ becomes a significant challenge in recovering DL CSIs with large delay spread (i.e., fast-varying in frequency domain).

Prior research often assumes adequate pilot density in the frequency domain for all

types of channels. However, the density of pilot placement in CSI-RS, as specified in cellular network standards [23], falls short for outdoor scenarios, particularly for channels with a high delay spread. This leads to a significant issue: the CSI-RS DL CSI matrix, \mathbf{H}_{RS} , may experience aliasing due to downsampling, rendering it impossible to accurately recover the full DL CSI, \mathbf{H} . Let us define the pilot sample rate in frequency as S_F and the maximum delay tap as Δt_{max} seconds. If $\frac{1}{2S_F} \leq \Delta t_{\text{max}}$, the channels captured from CSI-RS are considered to be aliased signals. Generally, recovering aliased signals (i.e., aliased downsampled (DS) CSI) to their original form (i.e., full CSI) is not feasible.

To give some realistic examples, based on the highest density of placement of CSI-RS, which is per 12 subcarriers with a spacing of 15 kHz, the frequency sampling interval is 180 kHz. According to the Nyquist theorem, the maximum measurable delay is half the inverse of the frequency interval, i.e., $\frac{1}{2 \cdot 180 \text{ kHz}} = 2.778$ microseconds. Consequently, any path with a delay greater than the maximum measurable delay will wrap around into the low delay region (i.e., so-called aliasing effect). Specifically, once the delay spread exceeds 1.4 microseconds, aliasing effects are inevitable regardless of the mean excess delay. If the mean excess delay is significant, aliasing can also occur even if the delay spread is less than 1000 nanoseconds. In practical field tests [54, 55], some research findings corroborate our points by demonstrating that, in the sub-6 GHz band, the delay spread of some measured channels can exceed 1000 nanoseconds. Additionally, according to the 3D channel model study for 5G NR [56], the delay spread of about 20% of NLoS Urban Macro channels is greater than 1 microsecond, as illustrated in Fig. 7.3.

However, if the DS signals satisfy certain constraints, we may recover the full CSI with aids of side information, which will be introduced in the following sections. Pre-

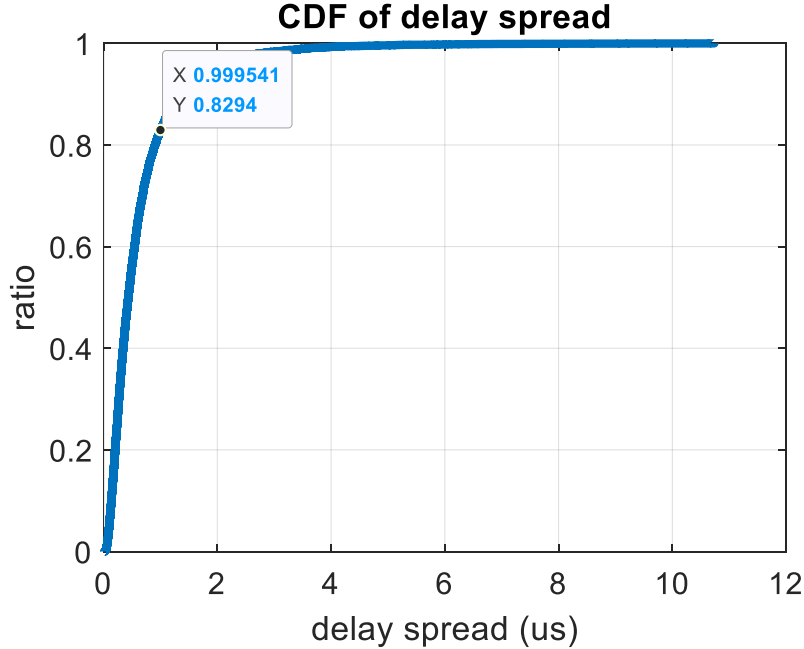


Figure 7.3: The empirical delay spread CDF of 10000 NLoS Urban Macro channel realizations according to 3D channel model of 5G NR.

vious studies often assume an overly idealistic approach to upsampling/interpolation, which can be a *critical operation* in channels with a large delay spread, and results in a bottleneck in reducing the total discrepancy¹. To enhance the overall performance, our focus should shift to improving this critical operation rather than the other two.

7.2 UL-CSI aided Upsampling with Aliasing Suppression

7.2.1 CSI Upsampling with Side Information

For an arbitrary channel $\mathbf{H} \in \mathbb{C}^{N_b \times N_f}$ in frequency domain and its DS version $\mathbf{H}_{\text{RS}} = \mathbf{H}\mathbf{Q}_{D_{\text{RS}}} \in \mathbb{C}^{N_b \times M_f}$ by a factor of D_{RS} . If we upsample the \mathbf{H}_{RS} by inserting $D_{\text{RS}} - 1$

¹As the three operations (estimation, feedback, and interpolation) are sequential, the one causing the largest loss becomes the bottleneck in reducing the total discrepancy. This operation is termed the critical operation.

zeros between any two consecutive samples along frequency domain, we have

$$\mathbf{H}_{\text{DS}}[:, j] = \begin{cases} \mathbf{H}[:, j], & \forall j \in \Psi_{\text{RS}}, \\ \mathbf{0}, & \forall j \notin \Psi_{\text{RS}}, \end{cases} \quad (7.1)$$

where $\Psi_{\text{RS}} = \{0, D_{\text{RS}}, \dots, (M_f - 1)D_{\text{RS}}\}$ is a downsampling index set. Note that \mathbf{H}_{DS} consists of the entries of \mathbf{H}_{RS} at frequencies with pilots and zeros elsewhere. By DFT/IDFT transformation, the full and DS DL CSI in beam-delay (BD) domain can be obtained as follows:

$$\mathbf{H}_{\text{BD}} = \mathbf{F}_{\text{AB}} \mathbf{H} \mathbf{F}_{\text{FD}} \in \mathbb{C}^{N_b \times N_f},$$

$$\mathbf{H}_{\text{DS, BD}} = \mathbf{F}_{\text{AB}} \mathbf{H}_{\text{DS}} \mathbf{F}_{\text{FD}} \in \mathbb{C}^{N_b \times N_f}, \quad (7.2)$$

where $\mathbf{F}_{\text{AB}} \in \mathbb{C}^{N_b \times N_b}$ and $\mathbf{F}_{\text{FD}} \in \mathbb{C}^{N_f \times N_f}$ are DFT and IDFT transformation matrices, respectively. The subscripts AB and FD denote the transformation from antenna/frequency to beam/delay domains, respectively. Note that we use subscript BD, AD, AF to denote CSI in beam-delay, angle-delay, and angle-frequency domains, respectively. We use no subscript to denote CSI in the original domain which is antenna-frequency domain.

Given the *DFT shifting theorem*[57], after IDFT transformation, we have the following relationship between the full and DS DL CSIs:

$$\mathbf{H}_{\text{DS, BD}}[i, j] = \begin{cases} \frac{\mathbf{H}_{\text{BD}}[i, j] + \mathbf{H}_{\text{BD}}[i, j + M_f] + \dots + \mathbf{H}_{\text{BD}}[i, j + M_f(D_{\text{RS}} - 1)]}{D_{\text{RS}}}, & \forall 0 \leq j < M_f \\ \mathbf{H}_{\text{DS, BD}}[i, \text{mod}(j, M_f)], & \text{otherwise} \end{cases} \quad (7.3)$$

Note that $\mathbf{H}_{\text{DS, BD}}$ is periodic in the delay domain with a period of $M_f = N_f/D_{\text{RS}}$. If

$\mathbf{H}_{\text{BD}}[i, j] \neq 0$ for any $j > M_f$, we can say that the aliasing effect occurs and it cannot be recovered to the original version \mathbf{H} in general cases since we can only measure $\mathbf{H}_{\text{DS, BD}}$, the sum of the multipaths. However, since $\mathbf{H}_{\text{DS, BD}}$ is periodic in the delay domain with a period of M_f , which matches the wrapped-around effect due to downsampling, the IDFT transformation *unwraps* the delay bins of \mathbf{H}_{BD} to the original delay positions. Thus, \mathbf{H} can be recovered if $\mathbf{H}_{\text{BD}}[i, j]$ in the delay domain satisfies the two requirements shown below:

- **Bin Isolation Property:** for any non-zero $\mathbf{H}_{\text{DS, BD}}[i, j]$ in Eq.(7.3), only one from the D_{RS} aliased copies $\mathbf{H}_{\text{BD}}[i, j], \mathbf{H}_{\text{BD}}[i, j + N_f/D_{\text{RS}}], \dots, \mathbf{H}_{\text{BD}}[i, j + N_f(D_{\text{RS}} - 1)/D_{\text{RS}}]$ is non-zero. Namely, the delay bins (i.e., $\mathbf{H}_{\text{BD}}[i, j], j > M_f$) and the low-delay bin (i.e., $\mathbf{H}_{\text{BD}}[i, j], j \leq M_f$) are isolated after wrapped-around in its DS version. If the bin isolation property holds, each non-zero DS signal $\mathbf{H}_{\text{DS, BD}}[i, j]$ in delay domain maps to a scaled unique delay bin in the original signal (i.e., $\mathbf{H}_{\text{DS, BD}}[i, j] = \mathbf{H}_{\text{BD}}[i, n_k]/D_{\text{RS}}$). Note that n_k can only be $j, j + M_f, \dots$, or $j + (D_{\text{RS}} - 1)M_f$.
- **Knowledge of bin locations:** we have the perfect knowledge map $\Phi \in \mathbb{C}^{N_b \times N_f}$ with ones at the positions with non-zero values in the the full CSI matrix $\mathbf{H}_{\text{BD}}[i, j]$ and zeros elsewhere.

Figure 7.4 shows a simple illustration for the single antenna case with the intermediate results of the proposed CSI upsampling approach using the bin location information. If the full CSI matrix \mathbf{H}_{BD} satisfies the above two requirements, \mathbf{H}_{BD} can be ideally obtained by

$$\widehat{\mathbf{H}}_{\text{BD}} = D_{\text{RS}} \Phi \circ \mathbf{H}_{\text{DS, BD}} \approx \mathbf{H}_{\text{BD}}.$$

Note that \circ denotes the operation of the element-wise product. Φ acts like a bandpass filter in BD domain. Although the two requirements are ideal, they lead us to a rationale to deal with aliasing problems. That is, to deal with sparse signals,

we can suppress aliasing peaks with the knowledge of the non-zero bin locations as a *bandpass filter*. In practice, DL CSI is somehow sparse so that a *quasi-bin isolation property* can hold. As for the knowledge of bin locations of DL CSI, we can estimate it according to UL CSI at base stations.

7.2.2 Multipath Reciprocity

Typically, acquiring the exact delay bin location information without the original DL CSI, denoted as \mathbf{H}_{BD} , is challenging. However, in communications systems, the DL CSI \mathbf{H}_{BD} is often closely correlated with the UL CSI, which is readily available at base stations, especially in terms of magnitudes in the BD domain. Although DL and UL CSIs do not exhibit full correlation in FDD wireless systems, as illustrated in Figure 7.5, they often share similar large-scale multipath geometries. This multipath reciprocity results in comparable delay and angle profiles, a finding supported by field tests and mathematical analysis [58, 59]. Therefore, UL CSI in the BD domain is typically considered a reliable estimate for the AD profiles of DL CSI. Owing to the relatively high pilot placement density in UL CSI, there are no aliasing effects, allowing for the design of a bandpass filter to mitigate aliasing effects in DL CSIs.

In modern communication systems, as depicted in Figure 7.6, the pilot placement density in the frequency domain of the Sounding Reference Signal (SRS) is much higher (every two subcarriers) compared to that of CSI-RS (every 12 subcarriers). Consequently, the maximum non-aliasing delay (i.e., measurable delay) of UL CSI is approximately six times greater than that of DL CSI, virtually eliminating aliasing effects in UL CSIs. Based on the principle of multipath reciprocity, this work proposes designing the bandpass filter Φ using UL CSI information.

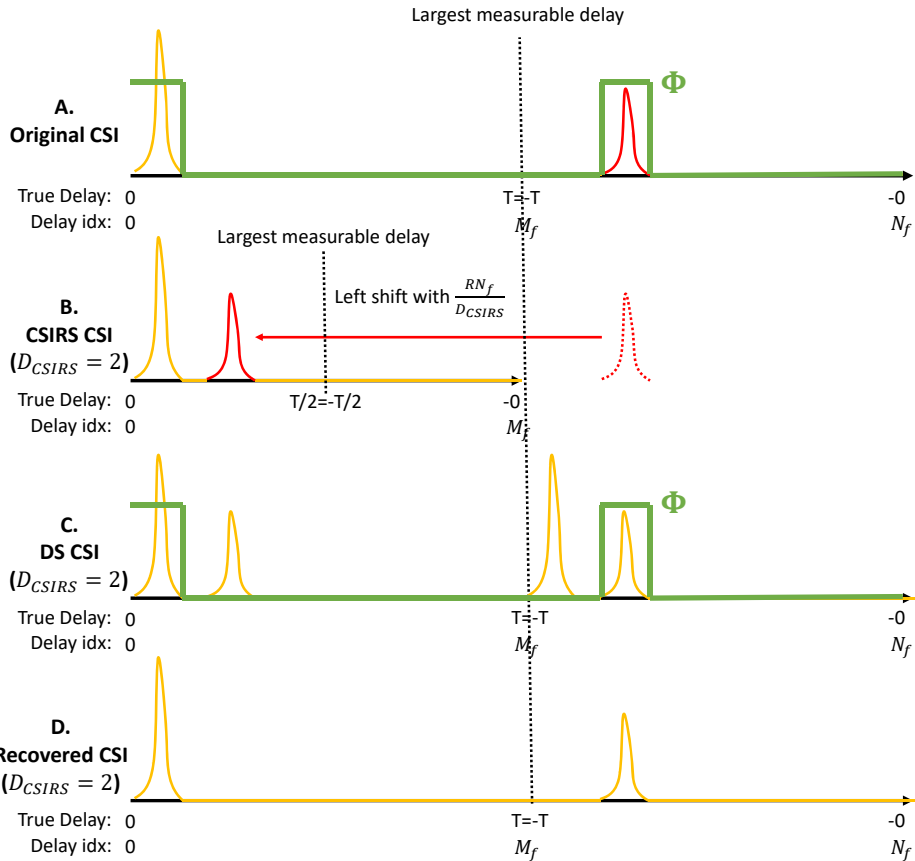


Figure 7.4: Illustration of CSI upsampling with side information. (A) shows the original CSI magnitude in delay domain. (B) demonstrates the CSIRS CSI magnitude in delay domain when $D_{RS} = 2$. We can find that the high negative delay peak wraps around ($R = 1$) into the low delay region, leading aliasing effect. (C) shows the DS CSI magnitude in delay domain by inserting zero inbetween samples of CSIRS CSI in frequency domain. The green curve represents an ideal binary bandpass filter Φ to be the side information. (D) is the resulting DL CSI magnitude in delay domain after applying the binary bandpass filter Φ .

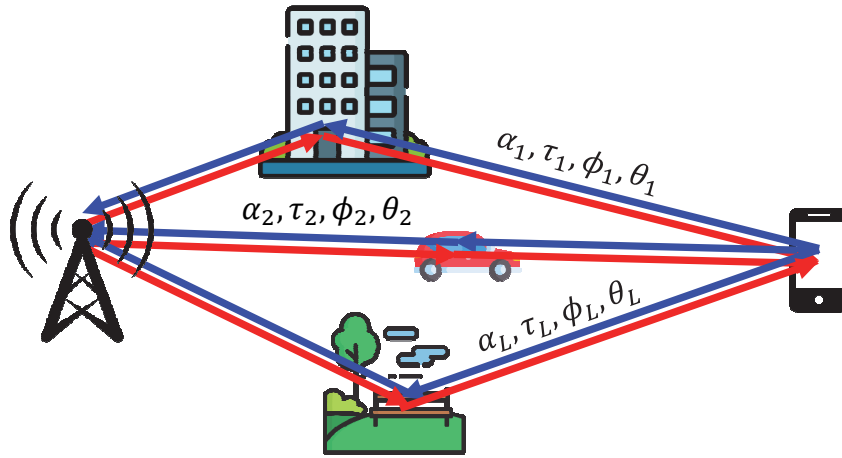


Figure 7.5: Illustration of multipath reciprocity between UL and DL propagation channels.

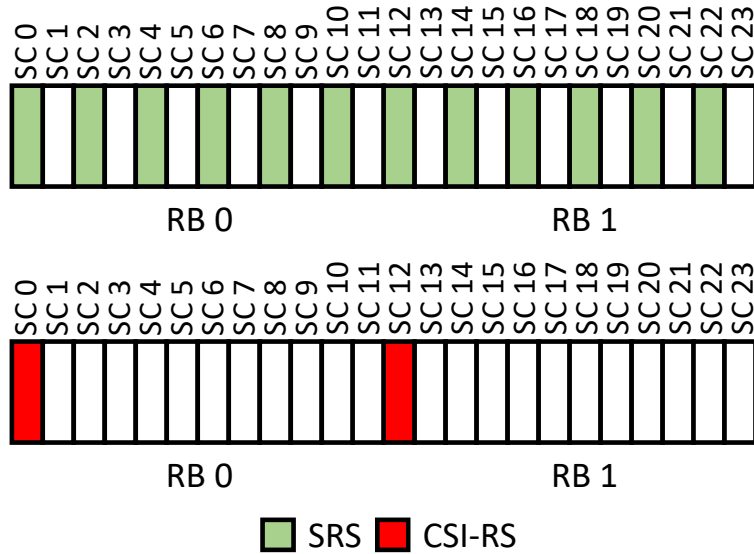


Figure 7.6: Comparison of SRS and CSI-RS placement density.

7.2.3 UL Masking: UL-Assisted CSI Upsampling with Aliasing Suppression

Assume that we have perfect UL CSI \mathbf{H}_{UL} . According to the multipath reciprocity between UL and DL CSIs, we can design a two-dimensional bandpass filter based on the UL CSI magnitude in BD domains as follows:

$$\Phi_{\text{UL}}[i, j] = \begin{cases} 0, & |\mathbf{H}_{\text{UL,BD}}[i, j]| < T, \\ 1, & |\mathbf{H}_{\text{UL,BD}}[i, j]| \geq T, \end{cases}$$

$$\mathbf{H}_{\text{UL,BD}} = \mathbf{F}_{\text{AB}}\mathbf{H}_{\text{UL}}\mathbf{F}_{\text{FD}} \in \mathbb{C}^{N_b \times N_f},$$

where we set $T = R \cdot \sqrt{P}$ and P is the average power of $\mathbf{H}_{\text{UL,BD}}$. We next can estimate the BD domain DL CSI by

$$\hat{\mathbf{H}}_{\text{BD}} = \Phi_{\text{UL}} \circ \mathbf{H}_{\text{DS,BD}}.$$

Due to the multipath reciprocity, the filter can effectively suppress the aliased copies as long as we design a proper threshold T which determines the pass band in delay and angle domains. However, it is challenging to find a reasonable threshold T for all CSIs.

7.3 Physic-inspired AI-driven Aliasing Suppression

Previous works [11, 18, 19] have been successfully applied to in CSI compression and recovery. Enough pilot sampling rate was usually assumed. In fact, following the 3GPP 5G NR standard [23], UEs estimate the CSI-RS channels and send channel state feedback. However, the frequency density of CSI-RS is not sufficient to capture the fast channel variation along the frequency domain. Even if a perfect CSI feedback is achieved, the aliasing loss due to downsampling is theoretically not possible to recover.

7.3.1 Model Architecture

There are plenty of successful network architecture which can enhance image details while maintaining visual fidelity after SR operation. In a sense of information theory, the model learns prior information from the training data to fill the information gap between the target and desired images. There are lots of common features in images such as facial features, colors textures, edges and shapes. For example, as long as the deep learning model can recognize a specific patch as a face, it can largely lower the uncertainty to upsample the LR images since there exists nothing else except facial features. However, unlike SR task in computer vision, the details of CSIs are random and difficult to learn as prior information stored in the deep learning model. To fill the information gap, we propose to utilize UL CSI information by exploiting multipath reciprocity against aliasing effects due to an insufficient pilot sampling rate.

This section introduces a general learning framework designed to effectively up-sample LR tensors into SR equivalents. This process is akin to the SR challenge in computer vision, where numerous successful networks [60–62] have been developed to enhance image details while preserving visual fidelity after SR operation. From the perspective of information theory, the model employs prior knowledge obtained from training data to fill the gap between actual and desired images. Certain image features, including facial characteristics, colors, textures, edges, and shapes, are common across various images. These features are retained as prior knowledge within the model, ready to be utilized as necessary to aid in image processing tasks. For instance, if a deep learning model identifies a particular segment as part of a face, it significantly reduces the uncertainty involved in upscaling LR images, since the expected features are confined to those associated with faces.

However, unlike the SR task in computer vision, the intricacies of CSI are random and challenging to learn as pre-existing information within a deep learning model. To

overcome this information gap, we propose leveraging UL CSI data, exploiting the principle of multipath reciprocity to counteract the aliasing effects stemming from an inadequate pilot sampling rate. Figure 7.7 gives a high-level understanding of the proposed architecture. This framework is designed to be deployed at base stations and consists of three modules: a) non-aliasing selection map generation, b) true peak recovery, and c) CSI attention and refinement which are described in detail as follows:

True Peak Recovery

This module aims to upsample LR DL CSIs by inserting zeros and transform them into the beam and delay domains. By doing so, we can have a DL CSI map in BD domain which is periodic in delay domain. According to the DFT shifting invariance property, we can map the aliasing delay bins to its original positions by inserting $D - 1$ zeros in between samples. On the other hand, this will also lead to more false peaks in the repetition map at the false delay positions. To implement, we basically follow Eqs. (7.1) and (7.2) to generate the desired repetition map $\mathbf{H}_{\text{BD,DS}}$. We describe these operations as a linear function $f_{\text{TPR}}(\cdot)$ such that $\mathbf{H}_{\text{BD,DS}} = f_{\text{TPR}}(\mathbf{H}_{\text{RS}})$.

Non-aliasing Selection Map Generation (Bandpass Filter Design)

This module aims to generate a bandpass filter in the BD domain which can suppress aliasing peaks at wrong delay positions. Regarding the multipath reciprocity, we can reply on UL CSI to infer where the true peaks are. Instead of using a rule-based approach mentioned in the previous section, we adopt a neural network to design a bandpass filter. We first transform the HR UL CSI into BD domain as $\mathbf{H}_{\text{BD,UL}}$ with the same size of the matrix $\mathbf{H}_{\text{BD,DS}}$ to be filtered. We then feed $\mathbf{H}_{\text{BD,UL}}$ into three convolutional layers with two ReLU activations at the outputs of the first two convolutional layers. We then utilize a sigmoid function as the last activation function to output the bandpass filter Φ_{UL} since it perfectly matches the soft filtering purpose (i.e.,

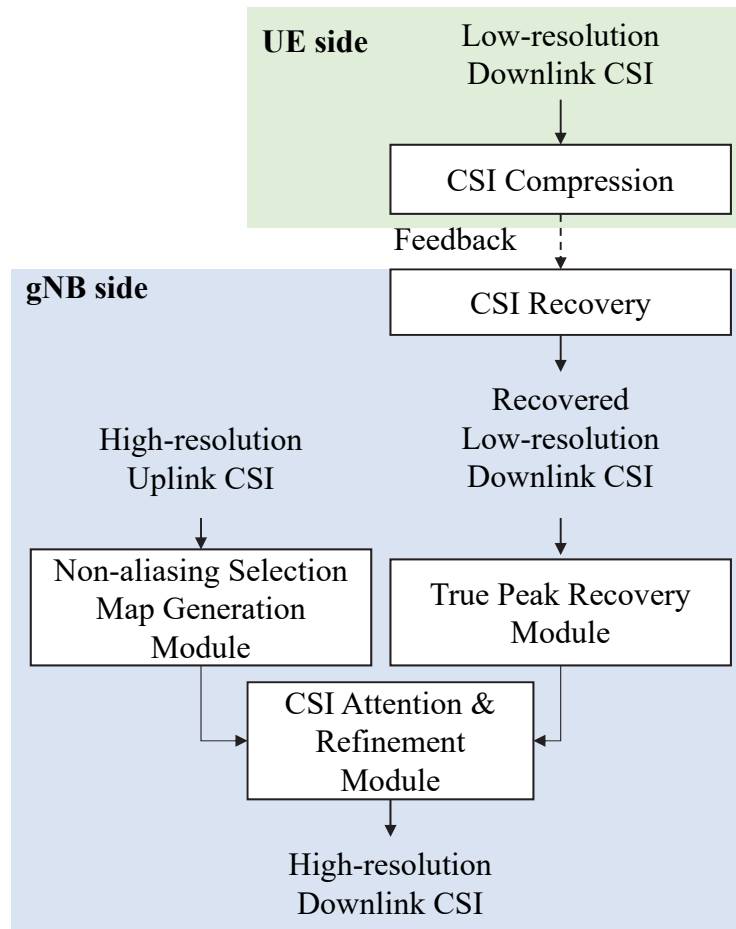


Figure 7.7: General architecture of the proposed physic-inspired AI-driven aliasing suppression framework. This framework consists of two parts. The first part is CSI compression and recovery which are deployed at UE and base station sides, respectively. The other part is the SR operation for the LR CSIs.

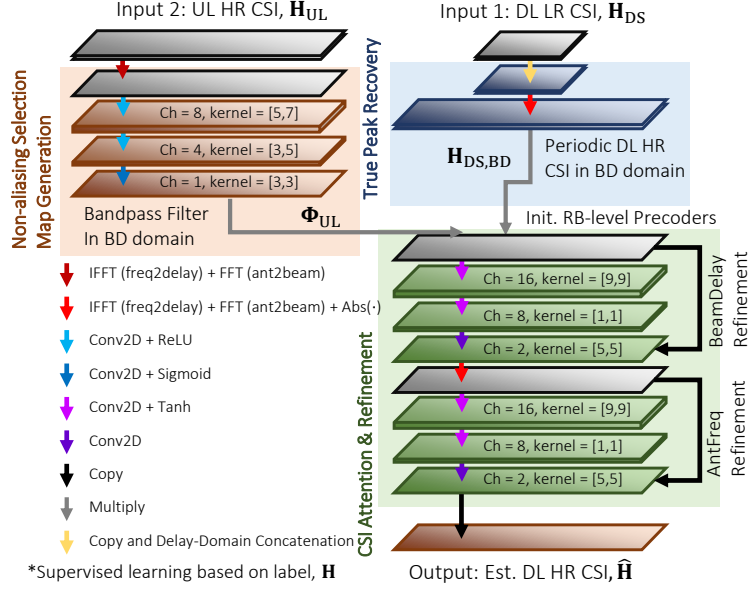


Figure 7.8: Network architecture of SRCsiNet. It consists of three modules: 1) Non-aliasing selection map generation, 2) True peak generation and 3) CSI attention and refinement.

model cannot only yield zeros to suppress aliasing delay positions and ones elsewhere, but also yield values between 0 and 1 to represent the model uncertainty and provide flexibility). We called it as Bandpass Filter Design (BFD) Block. For brevity, we can express the output of the branch of the model as

$$\Phi_{UL} = f_{BFD}(\mathbf{H}_{UL}). \quad (7.4)$$

CSI Attention and Refinement

This module aims to filter out the aliasing peaks and do refinement to generate the final DL CSI estimates which can be expressed as $\hat{\mathbf{H}} = f_{AR}(\Phi_{UL} \circ \mathbf{H}_{BD,DS})$. The function $f_{AR}(\cdot)$ aims to further refine and smooth the filtered result, which may have some artifacts due to the imperfect bandpass filter Φ_{UL} and the overlapped delay bins in $\mathbf{H}_{BD,DS}$. We apply two residual blocks with SRCNN block [60] as the backbone to refine the estimate first in BD domain and then in AF domain.

7.3.2 Loss Function Design

This network aims to minimize the upsampling loss $Loss_{\uparrow}$ which is defined as

$$\begin{aligned} Loss_{\uparrow}(\Theta_{\text{BFD}}, \Theta_{\text{AR}}) &= \frac{1}{D} \sum_d^D \|\hat{\mathbf{H}}_d - \mathbf{H}_d\|_F^2, \\ &= \frac{1}{D} \sum_d^D \|f_{\text{AR}}(\Phi_{\text{UL},d} \circ \mathbf{H}_{\text{BD,DS}}) - \mathbf{H}_d\|_F^2, \\ &= \frac{1}{D} \sum_d^D \|f_{\text{AR}}(f_{\text{BFD}}(\mathbf{H}_{\text{UL},d}) \circ \mathbf{H}_{\text{BD,DS}}) - \mathbf{H}_d\|_F^2, \end{aligned}$$

where Θ_{BFD} and Θ_{AR} are trainable parameters of the functions $f_{\text{BFD}}(\cdot)$ and $f_{\text{AR}}(\cdot)$, respectively.

7.3.3 Limitations and Failure Scenarios

As mentioned in Section 7.2.1, the full DL CSI \mathbf{H} can be recovered if \mathbf{H}_{BD} satisfies two requirements: bin sparsity and knowledge of bin locations. Low sparsity tends to cause overlapped delay bins, which cannot be separated. Even if channel sparsity is high but the magnitude correlation is low, the proposed approaches would generate a poor-quality mask that cannot correctly mitigate aliasing delay bins. Yet, considering the propagation model and path reciprocity, the two requirements are true for most cases. If the two requirements are not met, in fact, there is little else we can do from the point of view of information theory.

7.4 Efficient Channel State Feedback with Aliasing Suppression from Non-uniform Sampling

The true delay position information can significantly improve the CSI recovery for high-delay scenarios. In the perspective of information theory, if we can increase the

mutual information between the input and the desired output, we can further improve the CSI recovery accuracy.

According to the 3GPP 5G-NR standards [23], the primary and secondary synchronization signals (PSS and SSS) play crucial roles in cell identification and frame synchronization, appearing periodically every 25 subframes (approximately 25ms) and spanning 64-128 subcarriers in bandwidth. Beyond these primary functions, as depicted in Figure 7.9, UEs can also utilize PSS and SSS to estimate DL CSI, treating these signals as *virtual pilots* for DL CSI acquisition. Furthermore, the Physical Broadcast Channel (PBCH), instrumental for broadcasting system information and aiding UEs in network access, also contributes to DL CSI estimation by UEs, acting as additional virtual pilots. This dense placement of virtual pilots (SSS, PSS, and PBCH) aids in detecting multipath effects with large delays, which CSI-RS might miss, despite the mismatch in bandwidth coverage with the bandwidth part (BWP) designated for UEs.

In an ideal scenario, combining the channels from sparse uniform pilots (CSI-RS) with those from dense virtual pilots would enable us to harness the strengths of both pilot types, leading to more accurate CSI recovery. However, the effectiveness of our proposed architecture, SRCsiNet, hinges on maintaining a uniform sampling relationship between input and output to exploit the inversion discrete Fourier transform (IDFT) shifting invariance property.

This section will introduce the integration of a compressive sensing-based deep learning model into SRCsiNet, to address the challenges posed by a nonuniform pilot setup while effectively employing a bandpass filter. We will begin by outlining the compressive sensing-based CSI upsampling method, followed by an introduction to a novel framework, SRISTA-Net.

7.4.1 Compressive sensing based CSI upsampling

As illustrated in Figure 7.9, considering the extra subcarrier-level DL CSIs, we can express the non-uniform pilot DL CSI, termed as LR DL CSI for simplicity, as

$$\mathbf{H}_{\text{LR}}[i, j] = \begin{cases} \mathbf{H}[i, j], \forall j \in \Psi_P, \\ 0, \forall j \notin \Psi_P, \end{cases} \quad (7.5)$$

where $\Psi_P = \Psi_{\text{RS}} \cup \Psi_{\text{ex}}$ is the union of Ψ_{RS} and $\Psi_{\text{ex}} = \{I, I + 1, \dots, I + P - 1\}$ with I being the smallest subcarrier index in SSS, PSS or PBCH. Ψ_{ex} is the index set of consecutive pilots with size of P . We can reformulate the LR DL CSI based on the full AD DL CSI as

$$\begin{aligned} \mathbf{H}_{\text{LR}} &= \mathbf{H}\mathbf{I}[:, \Phi_P] = \mathbf{H}\mathbf{F}_{\text{FD}}\mathbf{F}_{\text{FD}}^H\mathbf{I}[:, \Phi_P] \\ &= \mathbf{H}_{\text{AD}}\mathbf{F}_{\text{FD}}^H\mathbf{I}[:, \Phi_P] = \mathbf{H}_{\text{AD}}\tilde{\mathbf{F}}_{\text{DF}}, \end{aligned} \quad (7.6)$$

where $\tilde{\mathbf{F}}_{\text{FD}} = \mathbf{F}_{\text{FD}}[:, \Phi_P] \in \mathbb{C}^{N_f \times |\Phi_P|}$ is the trimmed DFT transformation matrix.

Mathematically, the goal of compressive sensing reconstruction is to infer the original signal $\mathbf{x} \in \mathbb{C}^N$ from a low-dimensional measurement $\mathbf{y} = \Phi\mathbf{x} \in \mathbb{C}^M$, where $M \ll N$. By transposing Eq.(7.6), we have an exact projection of the problem of interest to a compressive sensing reconstruction problem (i.e., $\mathbf{y} = \mathbf{H}_{\text{LR}}[i, :]^T$, $\Phi = \tilde{\mathbf{F}}_{\text{FD}}^T$, $\mathbf{x} = \mathbf{H}_{\text{AD}}[i, :]^T$ where $i = 1, \dots, N_b$). This inversion is typically an ill-posed problem. However, it can be solved by compressive sensing reconstruction since the sparsity of the original CSIs regularizes the possible outputs.

7.4.2 ISTA-Net Framework

Previous works have proposed a deep unfolding approach called ISTA-Net [48]. The basic idea of ISTA-Net is to map the previous ISTA [47] approach updating steps to a deep learning network. This architecture consists of a fixed number of phases, each of

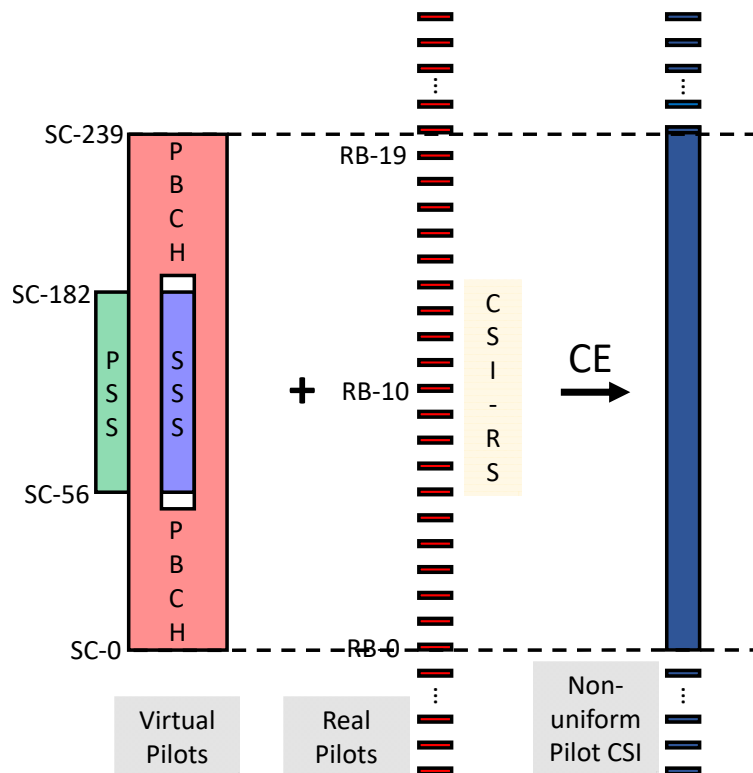


Figure 7.9: Illustration of virtual pilots (i.e., PBCH, SSS and PSS) and non-uniform pilot DL CSI. With the sparse uniform pilots (CSI-RS) and the dense virtual pilots, we can have an effective non-uniform DL CSI.

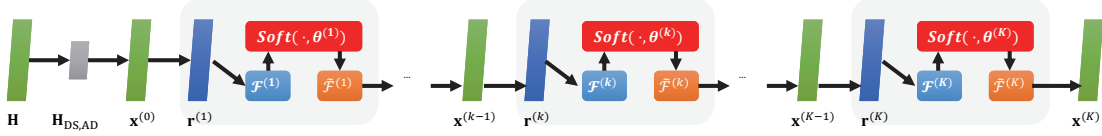


Figure 7.10: Network architecture of ISTA-Net.

the phases performing one iteration in the classic ISTA algorithm.

Figure 7.10 shows the deep learning network of the ISTA-Net. For each phase in ISTA-Net, it consists of two modules, namely the $\mathbf{r}^{(k)}$ **module** and the $\mathbf{x}^{(k)}$ **module**. The following items describe the operation in k -th phase as follows:

- **$\mathbf{r}^{(k)}$ Module:** This aims to produce intermediate result which is the same as ISTA algorithm. This step is to optimize the channel fidelity $\|\tilde{\mathbf{F}}_{\text{FD}}^T \mathbf{x}^{(k-1)} - \mathbf{H}_{\text{LR}}[i, :]^T\|_2^2$. To maintain the ISTA architecture while increasing the channel similarity, a trainable step size $\rho^{(k)}$ to vary across different phases is adopted so that the output of this module with input $\mathbf{x}^{(k-1)}$ for i -th antenna can be represented as:

$$\mathbf{r}^{(k)} = \mathbf{x}^{(k-1)} - \rho^{(k)} \tilde{\mathbf{F}}_{\text{FD}} (\tilde{\mathbf{F}}_{\text{FD}}^T \mathbf{x}^{(k-1)} - \mathbf{H}_{\text{LR}}[i, :]^T). \quad (7.7)$$

- **$\mathbf{x}^{(k)}$ Module:** It aims to compute $\mathbf{x}^{(k)}$ according to the intermediate result $\mathbf{r}^{(k)}$, which is given by

$$\mathbf{x}^{(k)} = \tilde{\mathcal{F}}^{(k)}(\text{soft}(\mathcal{F}^{(k)}(\mathbf{r}^{(k)}), \theta^{(k)})), \quad (7.8)$$

where a pair of functions $\mathcal{F}^{(k)}$ and $\tilde{\mathcal{F}}^{(k)}$ which are inverse of each other such that $\tilde{\mathcal{F}}^{(k)}(\mathcal{F}^{(k)}(\cdot)) = \mathcal{I}(\cdot)$ with $\mathcal{I}(\cdot)$ being an identity function. Such a constraint on $\mathcal{F}^{(k)}$ and $\tilde{\mathcal{F}}^{(k)}$ is called symmetry constraint.

7.4.3 SRISTA-Net Framework

The ISTA-Net can deal with non-uniform sampling but cannot exploit side information. Thus, in this subsection, we propose a new framework which combines ISTA-Net and

the proposed SRCsiNet to exploit the advantages of the two networks, which is termed as *SRISTA-Net*.

Figure 7.11 shows the deep learning network of the proposed network SRISTA-Net. We incorporate the SRCsiNet features into ISTA-Net by appending an additional block, Reciprocity Assisting (RA) Block, before the $\mathbf{r}^{(k)}$ module. This block aims to suppress the aliasing effects of the input $\mathbf{x}^{(k-1)}$ prior to solving the proximal mapping by applying the UL CSI assisted bandpass filter according to multipath reciprocity. We feed the magnitude of UL CSI $\mathbf{H}_{\text{UL,BD}}$ in the BD domain into two convolutional layers with ReLU and sigmoid functions, respectively, to obtain a bandpass filter Φ_{UL} .

Intuitively, for early phases, the model tends to heavily rely on UL CSI information and vice versa. Therefore, we design a weight matrix $\mathbf{W}^{(k)} \in \mathbb{C}^{N_b \times N_f}$ to adjust the dependency to the UL CSI at the k -th phase. We can rewrite the output of RA block as

$$\mathcal{R}^{(k)}(\mathbf{r}^{(k)}, \mathbf{H}_{\text{UL,BD}}) = \mathbf{W}^{(k)} \circ \Phi_{\text{UL}} \circ \mathbf{r}_{\text{BD}}^{(k)} + (1 - \mathbf{W}^{(k)}) \circ \mathbf{r}_{\text{BD}}^{(k)}, \quad (7.9)$$

where $\mathbf{r}_{\text{BD}}^{(k)}$ is the $\mathbf{r}^{(k)}$ after transformation to BD domain. We then feed the output into the $\mathbf{x}^{(k)}$ module in ISTA-Net to minimize the constraints of the L1 norm.

7.4.4 Loss Function Design

Given the training data pair $\{(\mathbf{H}_{\text{DS}}, \mathbf{H}_{\text{UL,BD}}, \mathbf{H})\}_{d=1}^D$, SRISTA-Net first transforms \mathbf{H}_{DS} into its AD version $\mathbf{H}_{\text{DS,AD}}$ as input and feeds into UL CSI information $\mathbf{H}_{\text{UL,BD}}$ in each phase to generate output $\mathbf{x}_d^{(K)}$. Note that \mathbf{H}_d , $\mathbf{x}_d^{(k)}$, and $\mathbf{r}_d^{(k)}$ are all in the AF domain. To reduce the discrepancy between \mathbf{H}_d and $\mathbf{x}^{(K)d}$ while maintaining the symmetry constraint $\tilde{\mathcal{F}}^{(k)}(\mathcal{F}^{(k)}(\cdot)) = \mathcal{I}(\cdot), \forall k = 1, \dots, K$, we design the following loss function:

$$\mathcal{L}_{\text{all}}(\Theta) = \mathcal{L}_{\text{discrepancy}} + \gamma \mathcal{L}_{\text{symmetry}}, \quad (7.10)$$

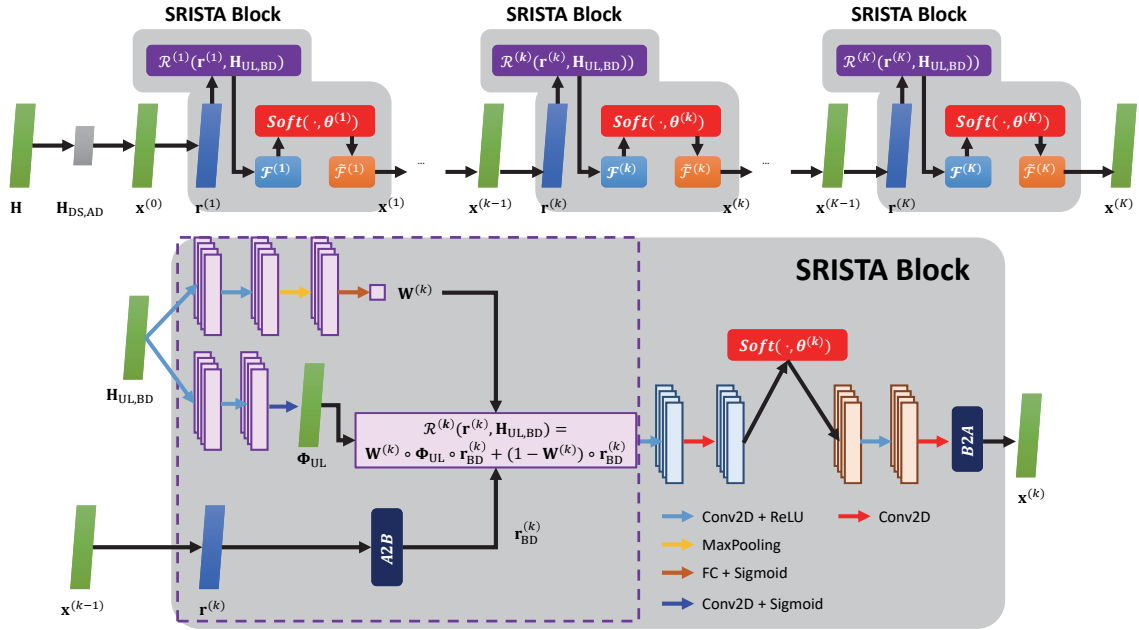


Figure 7.11: Network architecture of SRISTA-Net. For the construction of $\mathbf{W}^{(k)}$, we employ a pair of 2D convolutional layers followed by max pooling operations. This approach is designed to refine the output, focusing it more acutely on specific segments of the side information. Subsequently, integrating a sigmoid layer as the terminal activation mechanism compels $\mathbf{W}^{(k)}$ to execute a binary fusion of the processed and unprocessed outcomes, specifically between $\Phi_{UL} \circ \mathbf{r}_{BD}^{(k)}$ and $\mathbf{r}_{BD}^{(k)}$. As for the generation of Φ_{UL} , we apply BFD block in Eq. 7.4 mentioned in the previous section.

$$\mathcal{L}_{\text{discrepancy}} = \sum_{d=1}^D \left\| \mathbf{x}_d^{(K)} - \mathbf{H}_d \right\|_2^2, \quad (7.11)$$

$$\mathcal{L}_{\text{symmetry}} = \sum_{d=1}^D \sum_{k=1}^K \left\| \tilde{\mathcal{F}}^{(k)}(\mathcal{F}^{(k)}(\mathbf{q}_d^{(k)})) - \mathbf{q}_d^{(k)} \right\|_2^2, \quad (7.12)$$

where $\mathbf{q}_d^{(k)} = \mathcal{R}^{(k)}(\mathbf{r}^{(k)}, \mathbf{H}_{\text{UL,BD}})$ is the output of the RA block at the k -th phase. D , K and γ are the total number of training data size, the total number of SRISTA-Net phases, and the regularization parameter, respectively. In this chapter, we follow the original manuscript of ISTA-Net for the value of $\gamma = 0.01$.

7.4.5 Initialization

Like traditional iterative compressive sensing reconstruction, the proposed approach requires an initialization denoted by $\mathbf{x}^{(0)}$ as illustrated in Figure 7.11. From Eq.(7.6), we know $\mathbf{H}_{\text{LR}}[i, :]^T = \tilde{\mathbf{F}}_{\text{FD}}^T \mathbf{H}_{\text{AD}}[i, :]^T, \forall i = 1, \dots, N_b$. We take the LS solution to this problem for initialization such that

$$\mathbf{x}^{(0)} = \tilde{\mathbf{F}}_{\text{FD}}^* (\tilde{\mathbf{F}}_{\text{FD}}^T \tilde{\mathbf{F}}_{\text{FD}}^*)^{-1} \mathbf{H}_{\text{LR}}^T \quad (7.13)$$

To clarify the complex operations of SRISTA-Net, Alg. 7.1 shows the pseudo code of SRISTA-Net Framework.

7.5 Experimental Evaluations

7.5.1 Experiment Setup

Tests were focused on outdoor channels using widely used channel model software, QuaDriGa. The simulator considers a gNB with an 8×4 UPA and 32-element ULA serving single-antenna UEs, respectively, with half-wavelength uniform spacing. 2000 UEs uniformly distribute in the cell coverage which is rectangular region with size

Algorithm 7.1 SRISTA-Net Framework

Require: $\mathbf{H}_{\text{LR}}, \mathbf{H}_{\text{UL,BD}}, K, \gamma$ **Ensure:** Recovered DL CSI $\mathbf{x}^{(K)}$ in AD domain

- 1: **Initialize:** $\mathbf{x}^{(0)} = \tilde{\mathbf{F}}_{\text{FD}}^* (\tilde{\mathbf{F}}_{\text{FD}}^T \tilde{\mathbf{F}}_{\text{FD}}^*)^{-1} \mathbf{H}_{\text{LR}}^T$
 - 2: **for** $k = 1$ to K **do**
 - 3: **ISTA-Net r Module:**
 - 4: $\mathbf{r}^{(k)} = \mathbf{x}^{(k-1)} - \rho^{(k)} \tilde{\mathbf{F}}_{\text{FD}} (\tilde{\mathbf{F}}_{\text{FD}}^T \mathbf{x}^{(k-1)} - \mathbf{H}_{\text{LR}}^T)$
 - 5: **RA Block:**
 - 6: $\Phi_{\text{UL}} = f_{\text{BFD}}(\mathbf{H}_{\text{UL,BD}})$
 - 7: $\mathbf{W}^{(k)} = \text{Sigmoid}(\mathbf{W}_1 * \text{MaxPool}(\text{RuLU}(\text{Conv2D}(\text{RuLU}(\text{Conv2D}(\mathbf{H}_{\text{UL,BD}}))))))$
 - 8: $\mathbf{r}_{\text{BD}}^{(k)} = \mathbf{F}_{\text{BA}} \mathbf{x}^{(k-1)}$
 - 9: $\mathbf{r}^{(k)} = \mathbf{F}_{\text{BA}}^H (\mathbf{W}^{(k)} \circ \Phi_{\text{UL}} \circ \mathbf{r}_{\text{BD}}^{(k)} + (1 - \mathbf{W}^{(k)}) \circ \mathbf{r}_{\text{BD}}^{(k)})$
 - 10: **ISTA-Net x Module:**
 - 11: $\mathbf{x}^{(k)} = \tilde{\mathcal{F}}^{(k)}(\text{soft}(\mathcal{F}^{(k)}(\mathbf{r}^{(k)}), \theta^{(k)}))$
 - 12: **end for**
 - 13: **Loss Function:**
 - 14: $\mathcal{L}_{\text{all}}(\Theta) = \mathcal{L}_{\text{discrepancy}} + \gamma \mathcal{L}_{\text{symmetry}}$
 - 15: $\mathcal{L}_{\text{discrepancy}} = \sum_{d=1}^D \|\mathbf{x}_d^{(K)} - \mathbf{H}_d\|_2^2$
 - 16: $\mathcal{L}_{\text{symmetry}} = \sum_{d=1}^D \sum_{k=1}^K \|\tilde{\mathcal{F}}^{(k)}(\mathcal{F}^{(k)}(\mathbf{q}_d^{(k)})) - \mathbf{q}_d^{(k)}\|_2^2$
-

of $250(\text{m}) \times 300(\text{m})$. The scenario features given in 3GPP TR 38.901 UMa were followed, using $N_f = 667$ subcarriers with $15K$ -Hz spacing and $M_f = 55$ pilots with a downsampling ratio of $D_{\text{RS}} = 12$ as a common setting if not specified and assuming precise CSI estimates at the UEs. The NMSE metric was used to assess performance.

For DL-based models, we conducted training with a batch size of 32 for 1500 epochs, starting with a learning rate of 0.001 and setting an early stop criterion that validation loss does not improve for 100 epochs. We generate outdoor data sets using QuaDRiGa channel simulators. We consider 16 TTIs for each out of 2000 UEs. In total, the dataset consists of 32,000 channels. We used one-tenth of the channels for testing and validation, respectively. The remaining four-fifths channels are for training.

For ease in evaluating the degree of aliasing, it is common to use delay spread as a performance metric. A channel with a larger delay spread tends to suffer aliasing effects more severely since it contains more high-delay multipaths. We cluster all the 3200 test CSI data into 3 clusters according to their RMS delay spread: low (smaller

than 500 ns), medium (inbetween 500 ns and 1000 ns), high-delay spread (larger than 1000 ns). The low, medium and high delay spread clusters have 883, 1221 and 1095 test cases and are denoted as CL1, CL2 and CL3, respectively.

7.5.2 UL Assisted Bandpass Filter Design for Anti-aliasing

Figure 7.12 displays the NMSE performance of the UL masking method at various R levels compared to traditional interpolation across different CSI-RS placement densities. At a high CSI-RS density ($D_{RS} = 3$), the performance disparity between these approaches is minimal, notable mainly in the complete test dataset and CL1. However, a typical D_{RS} value, being either 12 or 24, introduces a more significant aliasing effect. For $D_{RS} = 12$, the performance divergence becomes more pronounced, as the NMSE metrics show effective mitigation of aliasing effects, particularly in the high-delay-spread cluster, CL3.

In Figure 7.13, the NMSE performance of the UL masking approach at varying R levels for $D_{RS} = 3, 6, 12$ is depicted. This figure reveals the sensitivity of the proposed method to the choice of the UL masking parameter R . In cases of CSI with intense aliasing effects, a higher R is necessary to effectively suppress the aliasing copies. Conversely, a large R might be excessively aggressive for channels with a low delay spread, potentially compromising the integrity of the actual delay peaks.

7.5.3 SRCsiNet

In addition to the two upsampling approaches mentioned in the previous subsection, we compare them with the proposed learning-based SRCsiNet and SR network, SRCNN [60] and a deep unfolding framework, ISTA-Net[48]. Figure 7.14 shows the NMSE performance of these alternatives for complete dataset and the three clusters. We can discover that ISTA-Net performs better than UL masking approach in CL1 due to the advantage of unfolding compressive sensing approach but performs poorly in

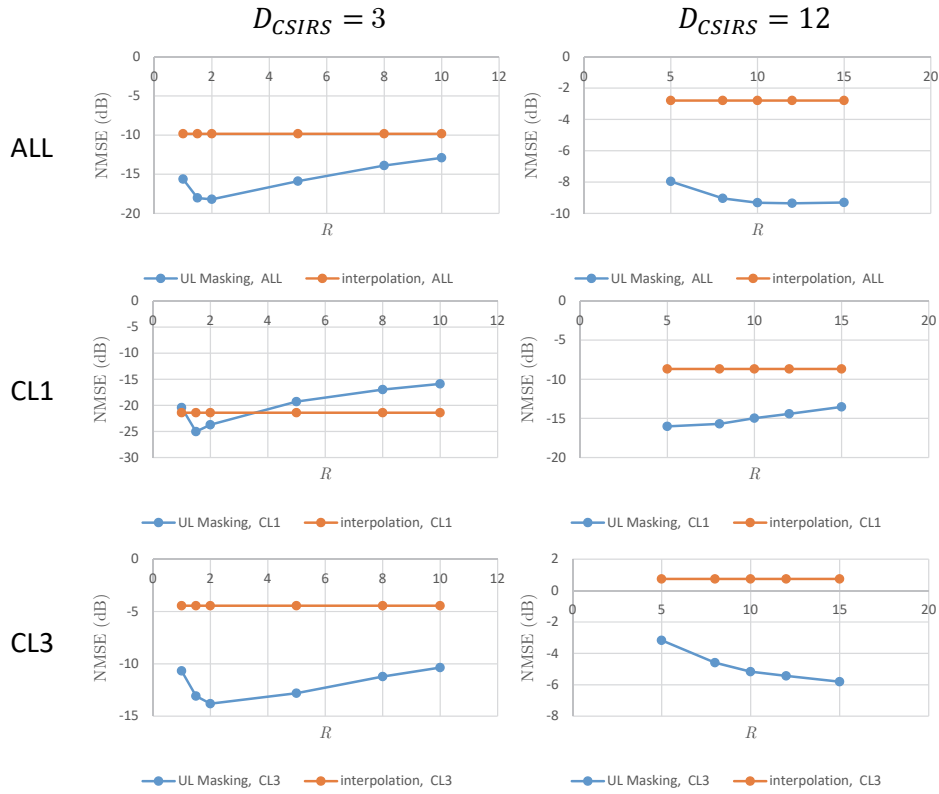


Figure 7.12: NMSE performance of the proposed UL-assisted anti-aliasing and traditional linear interpolation for different CSI-RS placement densities ($D_{CSI-RS} = 3, 12$).

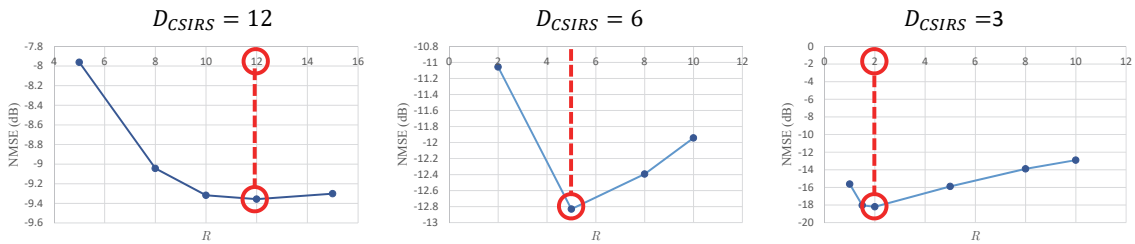


Figure 7.13: NMSE performance of the proposed UL-assisted anti-aliasing for different CSI-RS placement densities ($D_{CSI-RS} = 12, 6, 3$). We can clearly know that the optimal selection of the threshold level R varies with the aliasing effects. For the channels with strong aliasing effects, we require a larger R to suppress aliasing copies.

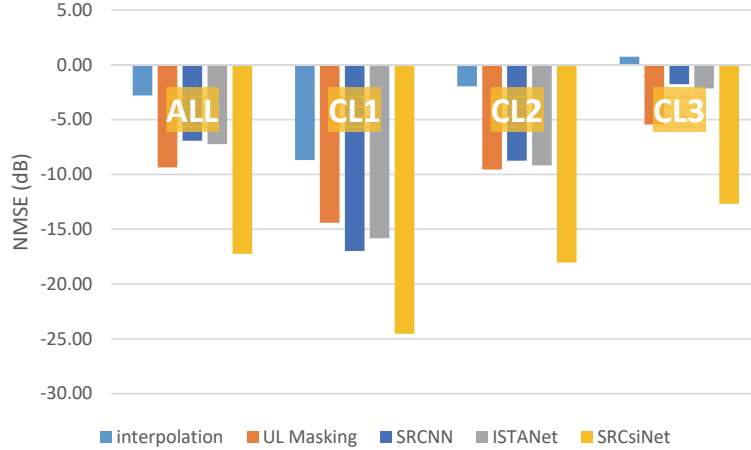


Figure 7.14: NMSE performance of the learning based UL-assisted framework, SRCsiNet, and the alternatives in comparison for different clusters (i.e., all all samples and the samples with low, medium, large delay spread).

CL3. That is because ISTA-Net does not introduce side information for dealing the aliasing effect. Clearly, by introducing UL CSI and providing flexibility in designing the bandpass filter, the overall performance can be improved by approximately 8 dB, which is significant. Figure 7.15 shows the visualization of SRCsiNet. We can find that the bandpass filter design can effectively suppress the aliasing peaks and retain the delicate detail of the true peaks at the same time.

7.5.4 End-to-end CSI Recovery

In this subsection, we would like to demonstrate the importance of optimizing the upsampling discrepancy to improve overall performance. Table 7.1 shows the NMSE performance from the end-to-end, feedback, and upsampling operation for SRISTA-Net, Interpolation, and ISTA-Net. End-to-end NMSE performance would be bounded by either feedback loss or upsampling discrepancy. However, we can first discover that end-to-end performance is generally bounded by upsampling loss in the considered UMa channels. This means that upsampling loss plays an critical role for improving the overall performance. Lastly, we can also find that the end-to-end NMSE performance

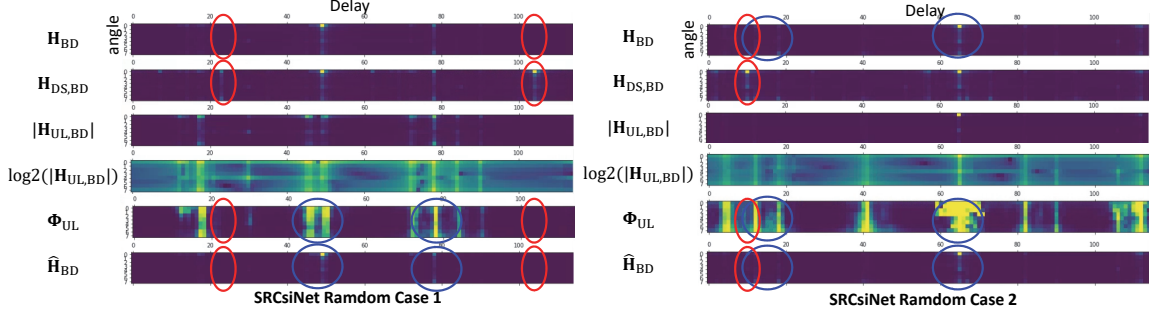


Figure 7.15: Visual illustration of the results of the SRCsiNet. For the limited space, these examples only show the first 128 delay taps (we have 660 delay taps in the experiment). Since $D_{CSIRS} = 12$, $\mathbf{H}_{DS,BD}$ is periodic in every 55 delay taps. We can know from the examples that the bandpass filter works very well since it can capture very delicate details which are belong to true peaks (denoted by **blue** color circles) and suppress the aliasing peaks effectively (highlighted by **red** circles).

Table 7.1: The end-to-end NMSE performance of SRISTA-Net, Interpolation and ISTA-Net for different numbers of virtual pilots under compression ratio is 4.

P = 0					P = 64					P = 128				
DualNet-MP + SRISTA-Net					DualNet-MP + SRISTA-Net					DualNet-MP + SRISTA-Net				
	ALL	CL1	CL2	CL3		ALL	CL1	CL2	CL3		ALL	CL1	CL2	CL3
$Loss$	-12.8	-16.2	-12.5	-9.8	$Loss$	-15.5	-19.4	-15.8	-11.9	$Loss$	-17.5	-21.6	-17.9	-13.7
$Loss_{FB}$	-14.5	-16.7	-13.6	-12.6	$Loss_{FB}$	-19.6	-23.2	-19.7	-16.3	$Loss_{FB}$	-22.2	-26.2	-22.7	-18.6
$Loss_{\uparrow}$	-17.2	-24.5	-18.0	-12.6	$Loss_{\uparrow}$	-17.6	-22.3	-18.7	-13.4	$Loss_{\uparrow}$	-19.4	-24.0	-20.1	-15.3
DualNet-MP + Interpolation					DualNet-MP + Interpolation					DualNet-MP + Interpolation				
	ALL	CL1	CL2	CL3		ALL	CL1	CL2	CL3		ALL	CL1	CL2	CL3
$Loss$	-2.7	-8.3	-1.9	0.7	$Loss$	-3.2	-8.9	-2.3	0.2	$Loss$	-3.6	-9.4	-2.8	-0.1
$Loss_{FB}$	-14.5	-16.7	-13.6	-12.6	$Loss_{FB}$	-19.6	-23.2	-19.7	-16.3	$Loss_{FB}$	-22.2	-26.2	-22.7	-18.6
$Loss_{\uparrow}$	-2.7	-8.6	-1.9	0.7	$Loss_{\uparrow}$	-3.2	-9.0	-2.3	0.3	$Loss_{\uparrow}$	-3.6	-9.5	-2.8	-0.1
DualNet-MP + ISTA-Net					DualNet-MP + ISTA-Net					DualNet-MP + ISTA-Net				
	ALL	CL1	CL2	CL3		ALL	CL1	CL2	CL3		ALL	CL1	CL2	CL3
$Loss$	-6.7	-13.5	-8.1	-1.9	$Loss$	-13.3	-18.3	-14.2	-9.0	$Loss$	-14.3	-19.5	-15.4	-10.0
$Loss_{FB}$	-14.5	-16.7	-13.6	-12.6	$Loss_{FB}$	-19.6	-23.2	-19.7	-16.36	$Loss_{FB}$	-22.2	-26.2	-22.7	-18.6
$Loss_{\uparrow}$	-7.2	-15.8	-9.1	-2.1	$Loss_{\uparrow}$	-14.5	-20.5	-15.9	-9.9	$Loss_{\uparrow}$	-15.3	-20.8	-16.5	-10.8

improvement is about 6-10 dB compared to other upsampling approaches without introducing UL CSI information.

7.5.5 Solving Overfitting problem

The SRISTA-Net architecture, necessitating 0.2 million parameters, faces a significant challenge due to its size relative to the training data, often leading to overfitting issues. This subsection highlights the effectiveness of Data Augmentation (DA) in our approach. Table 7.2 presents the NMSE performance for varying numbers of virtual pilots, comparing scenarios before and after implementing DA. A major hurdle in deploying learning-based models at gNB is the acquisition of real CSI data. In our

Table 7.2: NMSE performance of the SRISTA-Net with and without data augmentation (DA).

P	Method	ALL	CL1	CL2	CL3
0	SRISTA-Net	-14.62	-21.34	-15.41	-10.12
	SRISTA-Net + DA	-16.88	-23.15	-17.73	-12.43
256	SRISTA-Net	-17.20	-22.48	-18.34	-12.83
	SRISTA-Net + DA	-20.55	-23.89	-20.81	-17.18

experiments, the training of the deep learning model utilized less than 30,000 data points. We observed that overfitting becomes a significant issue when relying solely on the original training dataset. To counter this issue, we implemented circular shifting, as suggested by [63], on the original training data in the angle domain, effectively doubling the training dataset size. This augmentation was found to markedly enhance NMSE performance, demonstrating the benefits of increased training data.

7.5.6 Temporal Sensitivity of SRISTA-Net

SRISTA-Net significantly surpasses other alternatives in NMSE performance. However, it is important to note that previous experiments were conducted under the assumption that both CSI-RS and virtual pilots are present within the same time slot². Table 7.3 details the NMSE performance of SRISTA-Net, accounting for varying time gaps between CSI-RS and virtual pilots, alongside different counts of virtual pilots. Given the 10 ms periodicity of PBCH, PSS, and SSS, the maximum theoretical time difference between CSI-RS and virtual pilots is limited to under 5 ms. Our findings reveal that SRISTA-Net’s performance is highly susceptible to even minimal time differences, such as 5 ms. Interestingly, the NMSE performance in scenarios with a 5-ms gap is observed to be inferior compared to cases without any virtual pilots. In conclusion, when CSI-RS and virtual pilots coexist in the same time slot, leveraging the additional information is beneficial. Otherwise, it is preferable to upscale the DL CSI without incorporating data from virtual pilots.

²It’s assumed here that the CSI remains constant within the same time slot

Table 7.3: NMSE performance of SRISTA-Net for different time differences between CSI-RS and virtual pilots.

P = 64				
Time Difference	ALL	CL1	CL2	CL3
0ms	-17.6	-22.3	-18.7	-13.4
5ms	-13.6	-15.0	-14.1	-11.2
10ms	-9.2	-10.1	-9.5	-7.4
One-shot P=0	-17.2	-24.5	-18.0	-12.6
P = 128				
Time Difference	ALL	CL1	CL2	CL3
0ms	-19.40	-24.0	-20.1	-15.3
5ms	-11.7	-12.5	-11.9	-10.3
10ms	-6.2	-6.6	-6.3	-5.5
One-shot P=0	-17.2	-24.5	-18.0	-12.6

7.5.7 Complexity and Storage Requirements

Table 7.4 outlines the complexity and storage requirements of all previously mentioned approaches. It is observed that while SRISTA-Net and ISTA-Net have similar model sizes and required similar complexities, SRISTA-Net significantly surpasses ISTA-Net in terms of performance. However, this comparison also highlights a drawback of deep unfolding methods. Due to the recursive application of convolutional operations on full-size data, these models exhibit higher complexity relative to others. Fortunately, the upsampling module in these models is implemented at the gNB. Considering the demands of future AI-enhanced cellular systems, a gNB equipped with multiple GPUs is envisioned, enabling real-time operation of such complex models. Nonetheless, there is an ongoing need to reduce the complexity of deep unfolding approaches, potentially through techniques like pruning [64, 65] or other methods of model size reduction.

Finally but not least, to facilitate readers' understanding of our contributions, we provide Table 7.5 to highlight the key features of the proposed approaches compared

Table 7.4: Storage (PARA: model parameters) and complexity (FLOPs) comparison.

	PARA	FLOPs
Interpolation	0	109K
UL Masking	0	206K
SRCNN	63K	55.5M
ISTA-Net	196K	2G
SRCsiNet	7K	3.1M
SRISTA-Net	215K	2.01G

to previous rule-based and learning-based upsamplers.

Table 7.5: Comparison between CSI upsampling approaches in terms of different key features. High-DS recovery denotes the ability to recover high-DS CSIs (4 is the best). Note that ISTANet may perform better than UL Masking if introducing virtual pilots.

	High-DS Recovery	FDD Reciprocity Utilization	Virtual Pilot Utilization	Model Size	Complexity	Support Non-uniform Upsampling
Interpolation	1	X	X	N/A	Low	X
UL Masking	2	O	X	N/A	Low	X
ISTANet	2*	X	O	High	High	O
SRCsiNet (Ours)	3	O	X	Low	Medium	X
SRISTANet (Ours)	4	O	O	High	High	O

7.6 Conclusions

The chapter addresses a key challenge in massive MIMO FDD systems: the acquisition of DL CSI at the gNB, which is crucial for optimal performance. It identifies a significant issue in current systems, where the undersampling of CSI due to low-density pilot placement leads to aliasing effects, impairing CSI recovery. To deal with this issue, the chapter proposes a novel CSI upsampling framework for gNB, designed as a post-processing tool to fill the gaps caused by undersampling. This framework utilizes the principles of the DFT shifting theorem and multipath reciprocity, employing UL CSI to reduce aliasing effects. Additionally, the chapter presents a learning-based approach that combines the proposed algorithm with the ISTA-Net architecture, aiming to improve non-uniform sampling recovery. The chapter reports that both the rule-based and the deep learning methods demonstrate superior performance over traditional interpolation methods and current advanced techniques.

Appendix

To clarify the symbols to understand the problem formulation and equation better, we provide a summary of notations in Table 7.6.

Table 7.6: Notation Summary

Notation	Description
General Notations	
N_a	Number of antennas at gNB
N_f	Number of subcarriers in each subband
Δf	Subcarrier spacing
D_{RS}	Pilot (CSI-RS) spacing in subcarriers
S_F	Pilot sample rate in frequency
M_f	Number of pilots (CSI-RS) in a BWP
N_f	Number of subcarriers in a BWP
$(\cdot)^H$	Conjugate transpose
$\ \cdot\ _F$	Frobenius norm
\circ	Element-wise product operation
\mathbf{e}_i	i -th column vector of an identity matrix of size N_f
$\mathbf{Q}_{D_{RS}}$	Downsampling matrix with pilot rate D_{RS}
Ψ_{RS}	Downsampling index set
Δt_{\max}	Maximum delay tap
D	Number of random tests
\mathbf{N}	Channel estimation noise
Φ	Binary map for non-zero bin locations of \mathbf{H}_{BD}
Channel State Information (CSI) Matrices	
\mathbf{h}_i	RS CSI of the i -th antenna at gNB
\mathbf{H}	Full DL CSI matrix
\mathbf{H}_{RS}	RS CSI matrix
$\hat{\mathbf{H}}_{RS}$	Estimated RS CSI
\mathbf{H}_{DS}	DS CSI matrix
\mathbf{H}_{BD}	Full DL CSI in BD domain
$\mathbf{H}_{DS,BD}$	DS CSI in BD domain
$\hat{\mathbf{H}}$	Estimated DL CSI after upsampling
\mathbf{H}_{UL}	Full UL CSI matrix
$\mathbf{H}_{UL,BD}$	Full UL CSI matrix in BD domain
Transformations and Functions	
\mathbf{F}_{AB}	DFT matrix for antenna to beam domain
\mathbf{F}_{FD}	IDFT matrix for frequency to delay domain
$f_{en}(\cdot)$	Encoder function
$f_{de}(\cdot)$	Decoder function
$f_{\uparrow}(\cdot)$	Upsampling operation
$f_{TPR}(\cdot)$	True Peak Recovery function
$f_{BFD}(\cdot)$	Bandpass Filter Design function
$f_{AR}(\cdot)$	CSI Attention and Refinement function
Feedback and Loss Functions	
\mathbf{q}	Codeword for UL feedback
CR	Compression ratio
$Loss_{FB}$	Feedback loss
$Loss_{\uparrow}$	Upsampling loss
$\mathcal{L}_{\text{discrepancy}}$	Discrepancy loss
$\mathcal{L}_{\text{symmetry}}$	Symmetry loss
\mathcal{L}_{all}	Total loss
Compressive Sensing and ISTA-Net	
$\bar{\mathbf{F}}_{FD}$	Trimmed DFT transformation matrix
$\rho^{(k)}$	Trainable step size in ISTA-Net
$\mathcal{F}^{(k)}$	Function in ISTA-Net
$\bar{\mathcal{F}}^{(k)}$	Inverse function in ISTA-Net
$\mathbf{r}^{(k)}$	Intermediate result in ISTA-Net
$\mathbf{x}^{(k)}$	Result in ISTA-Net
$\mathcal{R}^{(k)}$	Reciprocity assisting function in SRISTA-Net
$\mathbf{W}^{(k)}$	Weight matrix in SRISTA-Net
Φ_{UL}	Bandpass filter generated using UL CSI
\mathbf{H}_{LR}	Low-resolution DL CSI
$\hat{\mathbf{H}}_{LR}$	Low-resolution CSI in AD domain
\mathbf{I}	Identity matrix
Φ_P	Pilot index set

Chapter 8

Plug-in UL-CSI-Assisted Precoder Upsampling Approach in Cellular FDD Systems

Acquiring downlink channel state information (CSI) is crucial for optimizing performance in massive Multiple Input Multiple Output (MIMO) systems operating under Frequency-Division Duplexing (FDD). Most cellular wireless communication systems employ codebook-based precoder designs, which offer advantages such as simpler, more efficient feedback mechanisms and reduced feedback overhead. Common codebook-based approaches include Type II and eType II precoding methods defined in the 3GPP standards. Feedback in these systems is typically standardized per subband (SB), allowing user equipment (UE) to select the optimal precoder from the codebook for each SB, thereby reducing feedback overhead. However, this subband-level feedback resolution may not suffice for frequency-selective channels. This chapter addresses this issue by introducing an uplink CSI-assisted precoder upsampling module deployed at the gNodeB. This module upsamples SB-level precoders to resource block (RB)-level precoders, acting as a plug-in compatible with existing gNodeB or base stations.

In this chapter, in Section 8.1, we first described the precoder upsampling problem in practical FDD system and introduced the precoder upsampling approach, SRPNet, along with the modified Type II/eTypeII precoder. Then, in Section 8.2, to tackle complexity issue, we proposed rule-based and learning-based UL-CSI/SSB assisted switches to avoid unnecessary processing SRPNet when applying to channels with low DS. In Section 8.3, test results demonstrate superior gain improvement after applying SRPNet and the complexity reduction by utilizing the PDP-based switch. Finally, we give conclusion in Section 8.4.

8.1 Type II/eTypeII based Precoder Upsampling

8.1.1 General Architecture

We propose a lightweight network deployed at the gNB that acts as a plug-in module, providing precoder upsampling from SB-level to RB-level. This architecture is compatible with existing modern cellular systems, such as 5G-NR. Figure 8.2 provides a high-level illustration of the proposed architecture, SRPNet. This network can effectively recover undersampled channels by exploiting the DFT shifting invariance property. Due to the UL/DL path reciprocity, the network can significantly suppress the aliasing effects caused by sub-Nyquist sampling. The details can be found in [66].

8.1.2 Modified Type II/eType II precoding

We discovered that the selected precoders according to Eq. 2.5 may lose multipath delay information. To maintain the signal structure of DL CSIs in Type II/eType II precoder design, we modify the precoder design criterion as follows:

$$\mathbf{w}_f = \operatorname{argmin}_{\mathbf{w}} \left\| \frac{\mathbf{h}_f}{\|\mathbf{h}_f\|_2} - \mathbf{w} \right\|_2 \quad (8.1)$$

This criterion ensures that the Type II/eType II precoder is close to the normalized DL CSI and preserves the phase information. Note that this criterion does not apply to the Type I precoder, as it lacks the degrees of freedom in choosing beams and combining coefficients.

For the eType II precoder, there is a critical problem to be solved. In the eType II precoder, to reduce the feedback overhead, as illustrated in the left part of Figure 8.1, truncation is performed, leaving the remaining delay components zero except for the first M_v delay taps. This truncation in the delay domain seems reasonable for most low-delay spread (DS) channels but performs poorly in capturing the high-delay components for precoders of high DS channels. Once the truncation is done on the UE side, it is impossible to recover on the gNB side. In the proposed approach, we replace the delay-domain truncation with frequency-domain downsampling. As illustrated in the right part of Figure 8.1, we uniformly sample M_v precoders in the RB domain and then transform them into the delay domain. We find that this modification preserves all the multipath information but might mistake low-delay components for high-delay ones due to sub-Nyquist sampling in the frequency domain, leading to aliasing effects. However, these aliasing effects can be alleviated by the following precoder upsampling module, SRPNet.

8.1.3 Precoder Upsampler, SRPNet

With the aid of delay domain sparsity, we introduce a lightweight neural network (NN), called the super-resolution precoder network (SRPNet), to suppress the aliasing effect due to sub-Nyquist sampling.

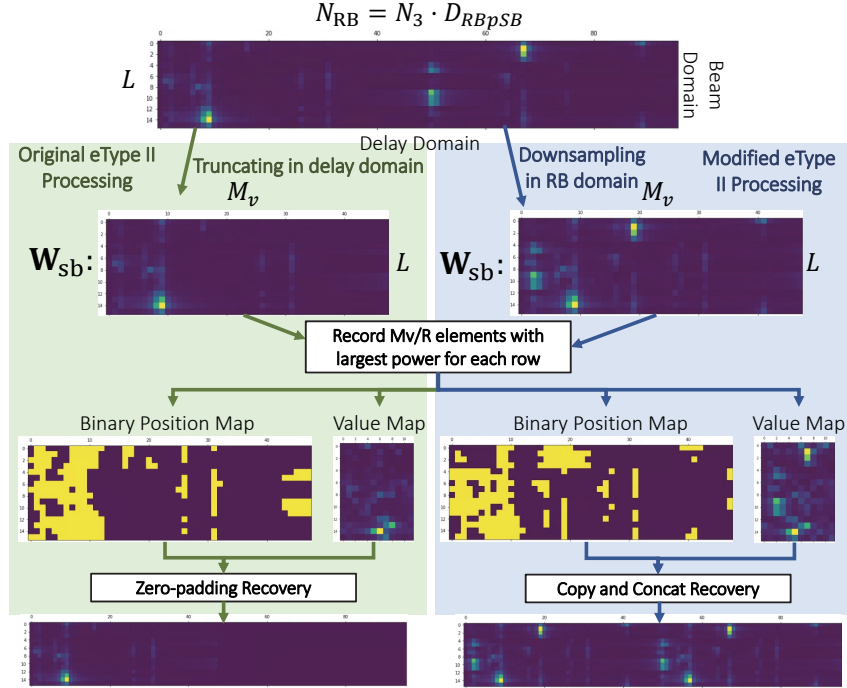


Figure 8.1: Illustration of the original and the modified eType II encoding processing for precoders of a high-DS DL CSI.

Architecture

This network consists of three modules: 1) bandpass filter (BPF) Design Module, 2) Initial Precoder Upsampling Module, and 3) Precoder Refinement Module, as illustrated in Figure 8.2.

- **BPF Design Module:** To obtain the delay profile of precoders, leveraging UL/DL multipath reciprocity, we feed the delay profile of the UL CSI into a convolutional network to infer a BPF that suppresses aliasing delay taps and preserves the true delay peaks of initial RB-level precoders in the delay domain.
- **Initial Precoder Upsampling Module:** We feed the modified Type II/eType II SB-level precoders to generate RB-level aliased precoders as initial precoders. These precoders may suffer severe aliasing effects but preserve all the true delay peaks at the same time.

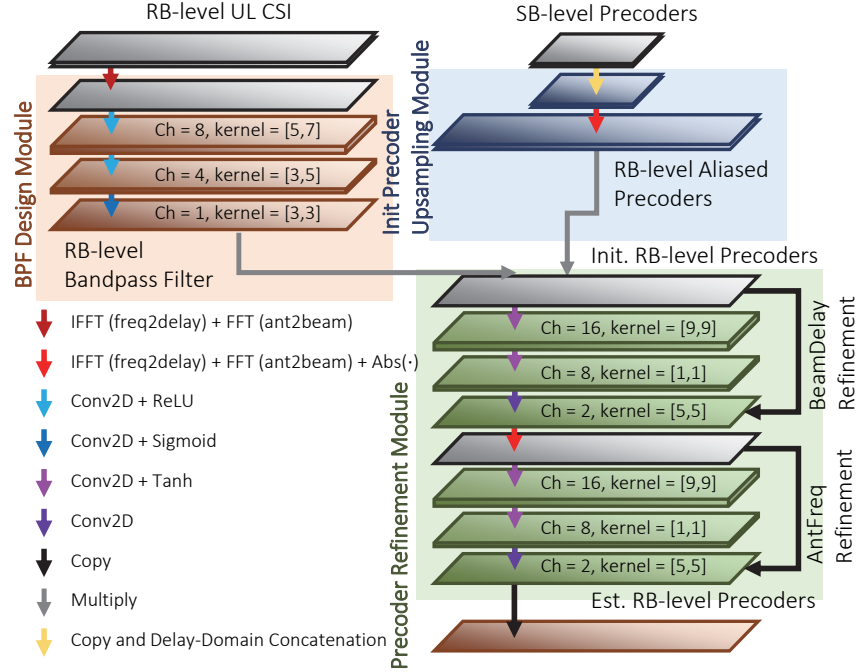


Figure 8.2: The model design of the SRPNet. It consists of three modules: 1) BPF Design Module, 2) Initial Precoder Upsampling Module, and 3) Precoder Refinement Module.

- Precoder Refinement Module:** First, we perform element-wise multiplication of the BPF and the RB-level aliased precoders. This guides the BPF Design Module to design a BPF instead of a confusing matrix¹. This opens up part of the black box of the NN. Then, we perform BD-domain and AF-domain refinement using convolutional NNs and shortcuts to generate the estimated RB-level precoders.

Lastly, we design the network with fully convolutional layers for scalability to any input size (i.e., different array sizes and bandwidths). For practical deployment considerations, SRPNet may not always be necessary for different scenarios such as low-DS DL CSIs. In the next section, we provide a PDP-based switch to determine when to utilize SRPNet.

¹If the BPF Design Module does not give a BPF, the aliasing peaks that are not suppressed will confuse the rest of the network.

8.2 UL-CSI/SSB Assisted Switch for Low-Complexity Precoder Upsampling

This section aims to provide a bridge for the proposed approach to practical cellular systems. We describe a PDP-based switch to help the system decide when to use simple linear interpolation or SRPNet to upsample SB-level precoders to RB-level.

Even with the design of an extremely lightweight and flexible network for precoder upsampling, SRPNet still leads to much higher computational complexity compared to linear interpolation. Given that for most channels with low delay spread (DS), linear interpolation can provide good enough precoder upsampling. Therefore, in this section, we propose several options to switch between SRPNet and linear interpolation.

8.2.1 PDP-based Switch

The key to determining whether to use SRPNet lies in evaluating how frequently the CSI varies with different RBs, which can be inferred from its delay profile. Although the gNB does not have the power delay profile (PDP) of DL CSI, it can still be inferred from the PDP correlation between UL and DL CSIs.

Threshold-based Switch

Assume we have RB-level PDP from UL CSI or SSB $\mathbf{PDP} \in \mathbb{C}^{N_{\text{RB}}}$. We make a decision s (that is, $s = 1$ means SRPNet utilization and vice versa) by applying the following measures, m , to trigger the switch on or off:

$$s = \begin{cases} 1, & m \geq \text{thres}, \\ 0, & \text{otherwise.} \end{cases}$$

- Maximum Excess Delay: the delay difference between significant multipath components.
- Mean Excess Delay: the mean delay weighted by its PDP.
- Root-Mean-Square (RMS) DS

Learning-based Switch

We also train a learning-based switch with a single-layer NN, which can be represented as:

$$s = f_{\text{switch}}(\mathbf{PDP}) = \text{Sigmoid}(\mathbf{f}^T \mathbf{PDP} + b),$$

by maximizing the gain-minus-cost metric $G - \lambda \cdot C$, where $G = s \cdot NG(\mathbf{W}_{\text{SRPNet}}, \mathbf{H}) + (1 - s) \cdot NG(\mathbf{W}_{\text{ITP}}, \mathbf{H})$ and $C = s \cdot C_{\text{SRPNet}} + (1 - s) \cdot C_{\text{ITP}}$. Here, $NG(\cdot)$ is the function to evaluate the average normalized gain between the precoders and DL CSIs. C_{SRPNet} and C_{ITP} are the computational complexities of SRPNet and linear interpolation, respectively. The ratio $\frac{C_{\text{SRPNet}}}{C_{\text{ITP}}}$ is roughly 1000. λ is a hyperparameter that determines the weight between computational cost and performance. In the test stage, we round s to determine the final outcome.

8.3 Experimental Evaluations

8.3.1 Experiment Setup

Tests were focused on outdoor channels using the widely used channel model software, QuaDriGa. The simulator considers a gNB with a 128-element uniform linear array (ULA) serving single-antenna UEs, with half-wavelength uniform spacing. 2000 UEs

are uniformly distributed in the cell coverage, which is a rectangular region of size $250 \text{ m} \times 300 \text{ m}$. The scenario features given in 3GPP TR 38.901 UMa were followed, using $N_{\text{RB}} = 96$ resource blocks (RBs) with a 20 MHz bandwidth part. The normalized gain

$$g = \sum_{f=1}^{N_{\text{RB}}} \frac{|\mathbf{h}_f^H \mathbf{w}|}{|\mathbf{h}_f| \cdot |\mathbf{w}|}$$

was used to assess performance.

For DL-based models, we conducted training with a batch size of 32 for 1500 epochs, starting with a learning rate of 0.001 and setting an early stop criterion if the validation loss did not improve for 100 epochs. We generated the outdoor datasets using the QuaDRiGa channel simulators. We considered 16 transmission time intervals (TTIs) for each of the 2000 UEs. In total, the dataset consists of 32,000 channels. We used one-tenth of the channels for testing and validation, respectively. The remaining four-fifths of the channels were used for training.

To evaluate the degree of aliasing, it is common to use DS as a performance metric. A channel with a larger DS tends to suffer from aliasing effects more severely since it contains more high-delay multipaths. We clustered all the 3200 test CSI data into 3 clusters according to their RMS DS: low (smaller than 500 ns), medium (between 500 ns and 1000 ns), and high DS (larger than 1000 ns). The low, medium, and high DS clusters have 883, 1221, and 1095 test cases, respectively.

8.3.2 Applying SRPNet to SB-level Type II Precoder

Figure 8.3 shows the capacity improvement ratio of precoders after applying SRPNet at different SNRs for low and high DS CSIs. Apparently, SRPNet improves capacity significantly especially for low SNRs. In addition, we can find that the benefit of the SRPNet becomes more obvious in high DS CSIs. That is because aliasing effect occurs

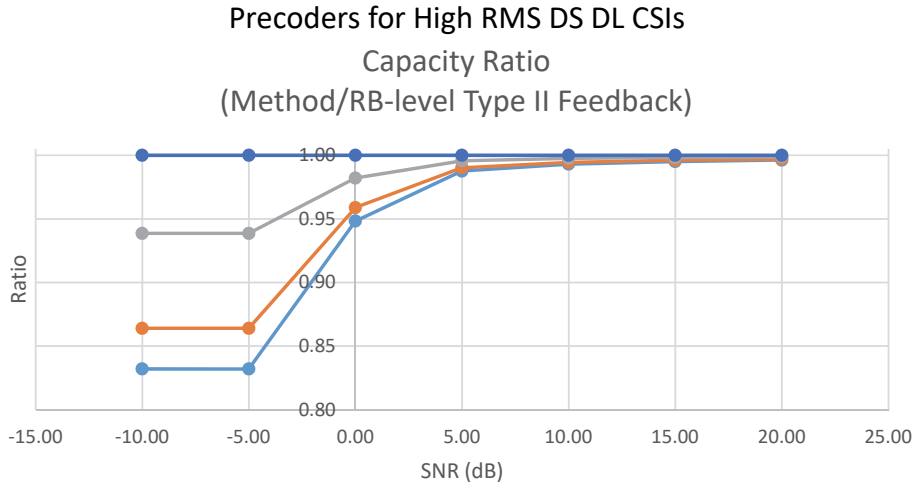
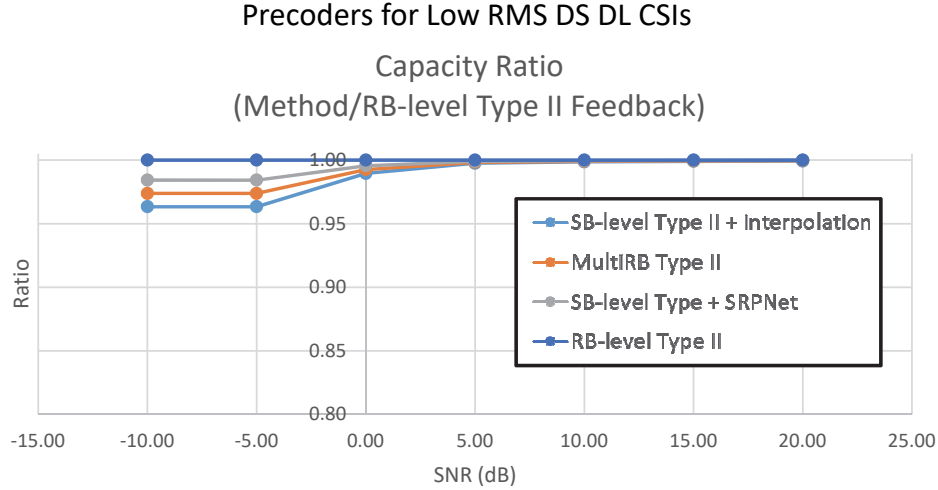


Figure 8.3: Capacity ratio of the SRPNet and other codebook-based approaches. Upper one is the results for the low RMS DS cluster. Bottom one is for the large DS cluster)

for high DS CSIs with higher probability and SRPNet can effectively upsample SB-level precoders to RB-level ones even if aliasing effects exist.

8.3.3 Applying SRPNet to SB-level eType II Precoder

Figure 8.4 demonstrates the normalized gain of Type II and eType II precoders before and after being applied SRPNet with different settings. Different points in a curve represent different configurations. For Type II precoders, we consider four different numbers of SBs ($N_3 = 3, 6, 12, 24$) representing different frequency downsampling rates

(from 96 RBs). For each curve of the eType II precoder, the anchor points from right to left represent $R = 2, 4, 8, 16$, and so on.

We observe a significant performance gap between the eType II precoder ($M_v = 24$) with and without SRPNet upsampling for high DS scenarios. Additionally, SRPNet-based eType II precoders outperform Type II and SRPNet-based Type II precoders, especially for low UL feedback overhead. This demonstrates the higher efficiency of the eType II precoder compared to Type II precoders after applying SRPNet. However, we also find that eType II precoders do not perform better even with $R = 1$. This is because eType II precoders find a common set of L beams for all SBs, which may not be optimal for each SB, leading to a performance bound.

8.3.4 Applying PDP-based Switch for Complexity Reduction

Figure 8.5 shows the normalized gain after applying the proposed PDP-based switches and a random switch to reduce computational complexity of precoder upsampling. We compare our proposed switches (Threshold-based switches and Learning-based switches) with a baseline random switch, which randomly chooses to utilize SRPNet or interpolation. The curve of the random switch is generated by setting different probabilities p to choose SRPNet ($p = 1$ for the rightmost point). It forms a straight line since both the normalized gain and the complexity are linear combinations of the outcomes of SRPNet and interpolation. We find that all the proposed switches perform better than the random switch, indicating that the PDP of UL CSI or SSB is beneficial for making the binary decision.

Among these rule-based switches, the one relying on maximum excess delay performs the best, since maximum excess delay is more direct and can better reflect when aliasing occurs (i.e., when the largest delay of significant paths exceeds the Nyquist measurable delay). The curve of the learning-based switch is built by training the model with different $\lambda = 1 \times 10^{-5}, 5 \times 10^{-5}, \dots, 1 \times 10^{-3}$. The learning-based approach

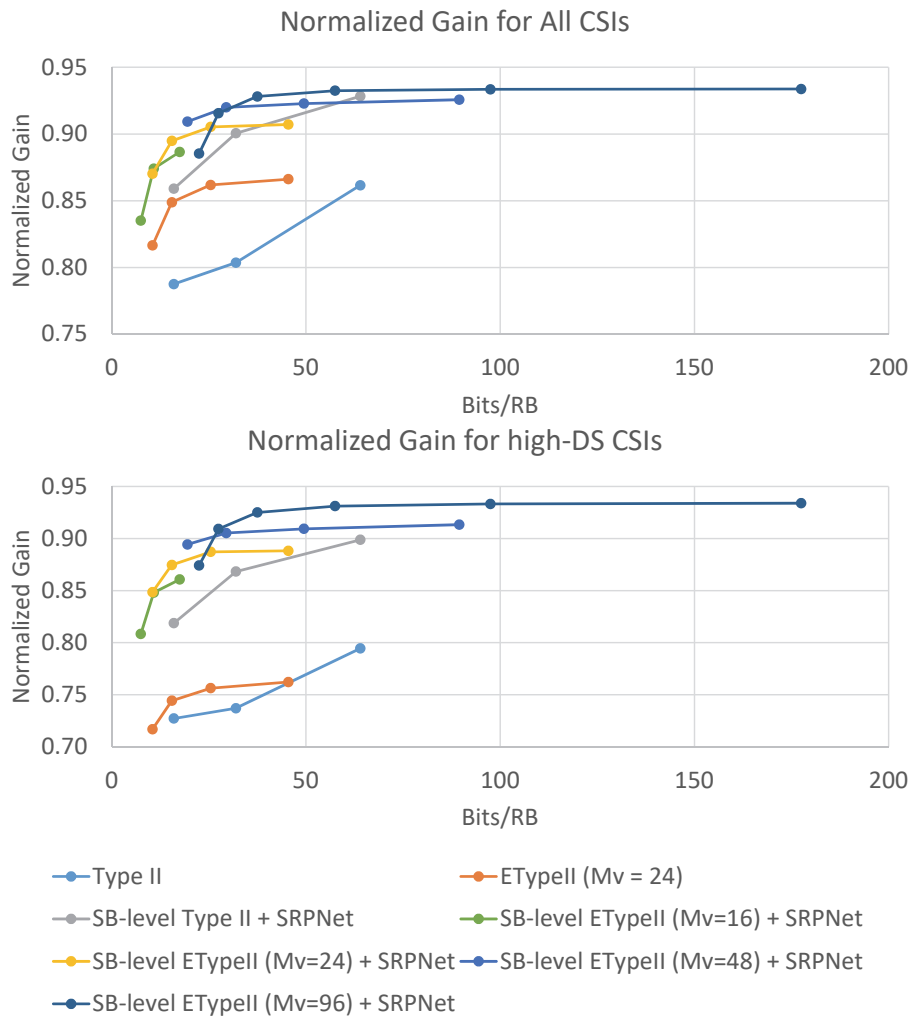


Figure 8.4: Normalized gain of Type II and eType II precoders before and after being applied SRPNet for all DL CSIs and high DS ones.

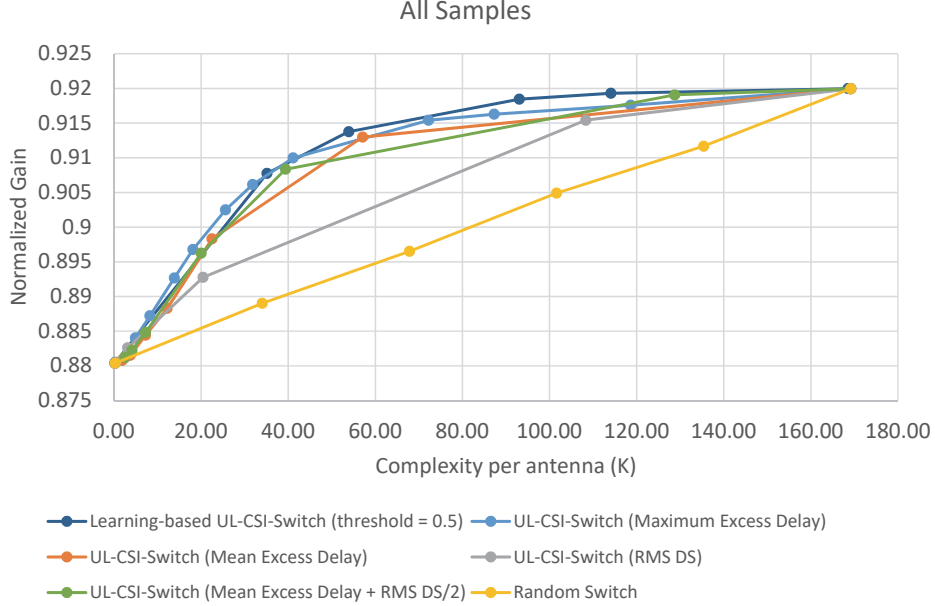


Figure 8.5: Normalized gain after applying the proposed PDP-based switches and random switch for computational complexity reduction of precoder upsampling.

performs the best among all the proposed switches and also has the lowest complexity. However, we still face the challenge of finding a mapping from choosing λ to achieving the desired complexity.

8.4 Conclusions

Acquiring accurate DL CSI is crucial for optimizing the performance of massive MIMO systems in FDD. Existing cellular systems use codebook-based precoder designs, such as Type II and eType II, which simplify feedback mechanisms and reduce overhead. However, standardized feedback per SB often falls short for frequency-selective channels. To address this issue, we introduced SRPNet, an uplink CSI-assisted precoder upsampling module deployed at the gNodeB. SRPNet improves SB-level precoders to RB-level precoders and is compatible with existing base stations. Our results demonstrated SRPNet’s effectiveness in improving normalized gain, particularly in high-DS scenarios. Additionally, we proposed a PDP-based switch to intelligently choose be-

tween SRPNet and linear interpolation, reducing computational complexity. Our findings showed that the proposed switches, especially the learning-based switch, outperformed random switches and achieved better performance with lower complexity. In summary, SRPNet and the PDP-based switch offer a robust solution for enhancing downlink CSI acquisition in massive MIMO systems. These advances significantly improve the efficiency and performance of modern cellular networks, particularly in scenarios with high frequency selectivity.

Chapter 9

Future Works and Conclusions

In this article, we focus on efficient explicit and implicit CSI feedback methods in FDD cellular systems, considering various practical concerns and achieving significant performance improvements. This area has recently garnered attention from both academia and industry. As pioneers in this field, we suggest two future directions:

9.1 Channel/Precoder Prediction against Channel Aging

In FDD cellular systems, UE estimates DL CSI and feeds it back to the gNB for precoder design for the next scheduled downlink transmission. For channels with high Doppler shifts, the precoder may become outdated by the time it is applied to the scheduled transmission due to the time difference (Δt) between DL training and data transmission. Predicting future CSI or precoder is crucial in practical scenarios to proactively address the effects of channel aging.

Figure 9.1 demonstrates two possible solutions in practical codebook-based precoder feedback: 1) gNB-side Precoder Prediction and 2) UE-side Channel Prediction. These can be implemented either at the gNB or UE side as modules for precoder prediction or

channel prediction, respectively. For gNB-side precoder prediction, the gNB, knowing Δt , predicts future precoders based on past ones. For UE-side channel prediction, the gNB first acknowledges Δt , then the UE predicts the future channel and selects the appropriate precoder for feedback.

For both approaches, they are very similar to the video prediction task, since we are trying to estimate the future CSI or precoder map based on historical ones, analogous to past reference video frames. However, common neural networks for the video prediction task, such as the recurrent neural network (RNN), long-short-term memory (LSTM) [67], ConvLSTM [68, 69], or Video Transformer (ViT) [70], require excessive computational power and time. This poses a severe computational burden for UE-side CSI prediction due to their limited available computing resources. Even for gNB-side precoder prediction, computation time matters significantly. A long computation time shortens the effective prediction horizon. Therefore, designing such a module requires ensuring low computational time while maintaining good prediction accuracy for long prediction horizons.

9.1.1 Problem Formulation: Future Precoder Prediction

In this subsection, we focus on the precoder prediction task. We first describe the problem formulation, propose heuristic approaches, provide some preliminary results, and give possible future directions.

Assume the gNB has a precoder matrix \mathbf{F} with dimensions $N_a \times N_{\text{RB}}$, consisting of precoders for each Physical Resource Block (PRB) in a Bandwidth Part (BWP). The gNB stores previous precoder matrices from the previous T time slots to predict the precoder matrix for the next time slot. Thus, the predicted precoder matrix can be expressed as

$$\hat{\mathbf{F}}_{T+1} = f_{\text{pred}}(\mathbf{F}_1, \mathbf{F}_2, \dots, \mathbf{F}_T), \quad (9.1)$$

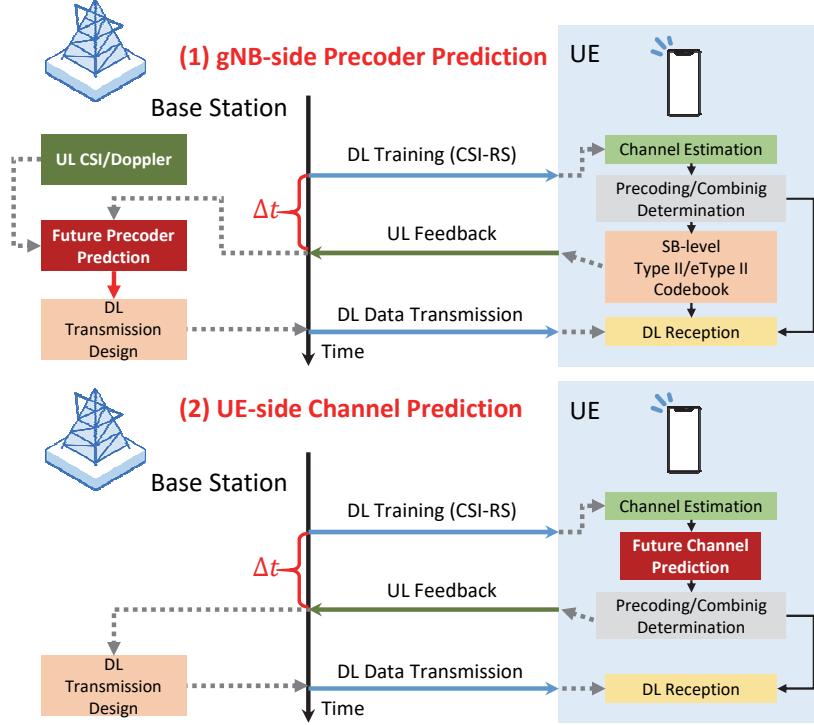


Figure 9.1: Illustration of (1) gNB-side precoder prediction and (2) UE-side channel prediction.

where \mathbf{F}_t is the precoder matrix for time slot t , and $f_{\text{pred}}(\cdot)$ is the prediction operation, which can be either a rule-based extrapolator or a learning-based method. To estimate the precoder matrix for $t \geq T + 2$, as illustrated in Figure 9.2, we append the last estimated precoder to the previous input frames and remove the first frame of the appended input as DL new input for predicting the next precoder matrix.

9.1.2 Proposed Approaches and Preliminary Results

We aim to find a non-linear mapping function f_{pred} which minimizes the MSE between the true and predicted precoder matrix. We design three different heuristic 3D-CNN based extrapolators, illustrated in Figure 9.3, and compare them with a well-known video predictor, SimVP [71]. The evaluation metric is based on the normalized gain (i.e., cosine similarity) between the precoder and DL CSI. We compare our approaches with the **sample-and-hold** method, which applies the last precoder that the gNB

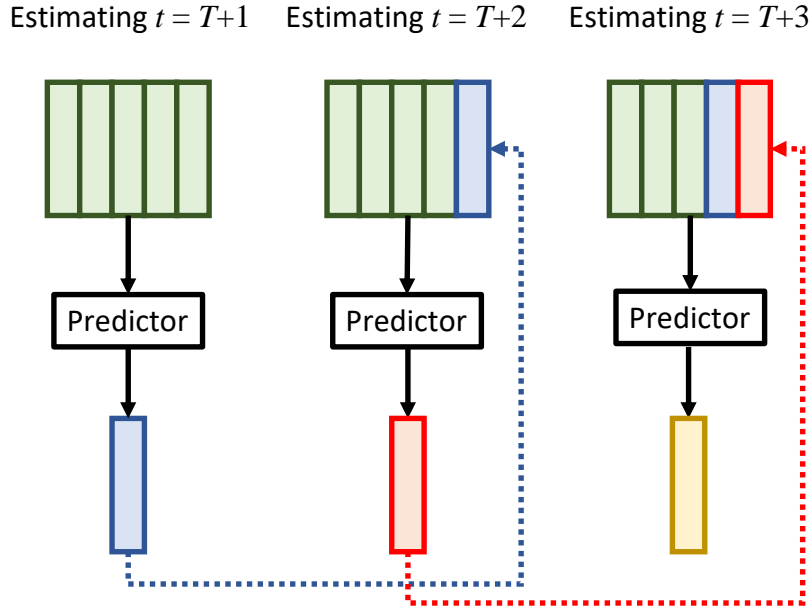


Figure 9.2: Illustration of the CSI/precoder precoder for the prediction for $t \geq T + 2$.

received to the future DL CSIs.

Figure 9.4 (a) demonstrates the preliminary results of the proposed approaches with different designs compared to traditional prediction algorithms. We find that shortcut designs significantly improve the initial performance for short prediction horizons. The process in the BD domain helps the model capture variations between time slots. Lastly, utilizing cosine similarity loss instead of MSE loss improves prediction for the first step but significantly degrades performance for subsequent steps.

Figure 9.4 (b) shows that the more complex SimVP method, used in video prediction tasks, outperforms our proposed approaches. However, for all the learning-based approaches, none outperform the simplest and most effortless sample-and-hold approach for predicting the precoders of DL CSIs after four time slots. This indicates significant room for improvement. We suggest introducing Doppler information into the model for better inference of channel variations. Additionally, a conditional generative model could better understand text information to control the content of generated videos. Similarly, we could exploit this idea to let the model understand how Doppler influences

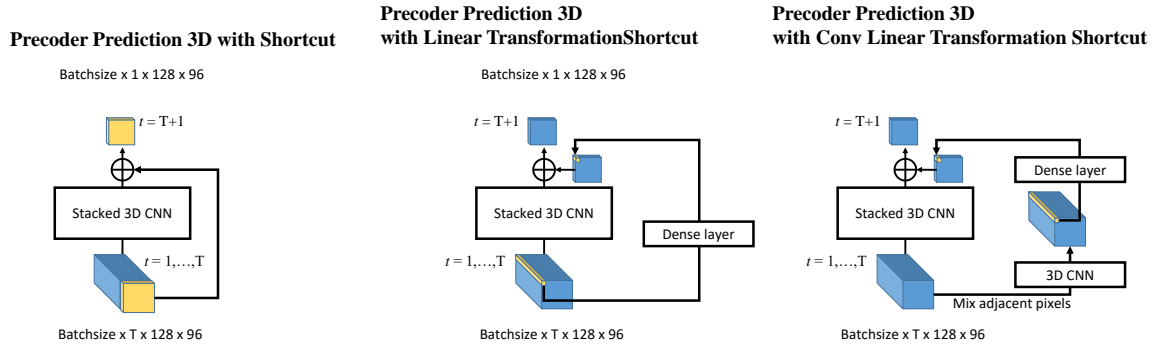


Figure 9.3: Design Model Architecture of (left) Precoder Prediction 3D with Shortcut (named PrecoderPrediction3D-sc in simulation plot): the core network consists of stacked 3D CNNs with a shortcut, (middle) Precoder Prediction 3D with Linear Transformation Shortcut (named PrecoderPrediction3D-sc-dense): same as the previous one with a linear transform shortcut, (right) Precoder Prediction 3D with Conv Linear Transformation Shortcut (named PrecoderPrediction3D-sc-dense-conv): same as the previous one, with an additional 3D CNN layer to merge adjacent pixels before linear transformation.

channel variations in subsequent time slots. Finally, ensemble different approaches to avoid performance degradation for long prediction horizons.

9.2 User Clustering for MU-MIMO

In previous works, we considered explicit and implicit CSI feedback for precoder designs. The goal is to provide the serving base station with the optimal precoder for each UE. In a Multi-User MIMO (MU-MIMO) scenario, as illustrated in Figure 9.5, the same time-frequency resources can be assigned to multiple UEs to boost throughput through resource reuse. However, the transmission signals to other UEs may act as strong interference for a specific UE if the precoders serving other UEs are correlated with its DL CSI. In this case, the gNB should perform **user scheduling** to assign UEs with low-correlation channels to the same time-frequency resource. Many previous works [72, 73] have addressed this, but the assumption of full knowledge of DL CSIs at the gNB is impractical in real FDD systems. In practical scenarios, only precoders are fed back to the gNB.

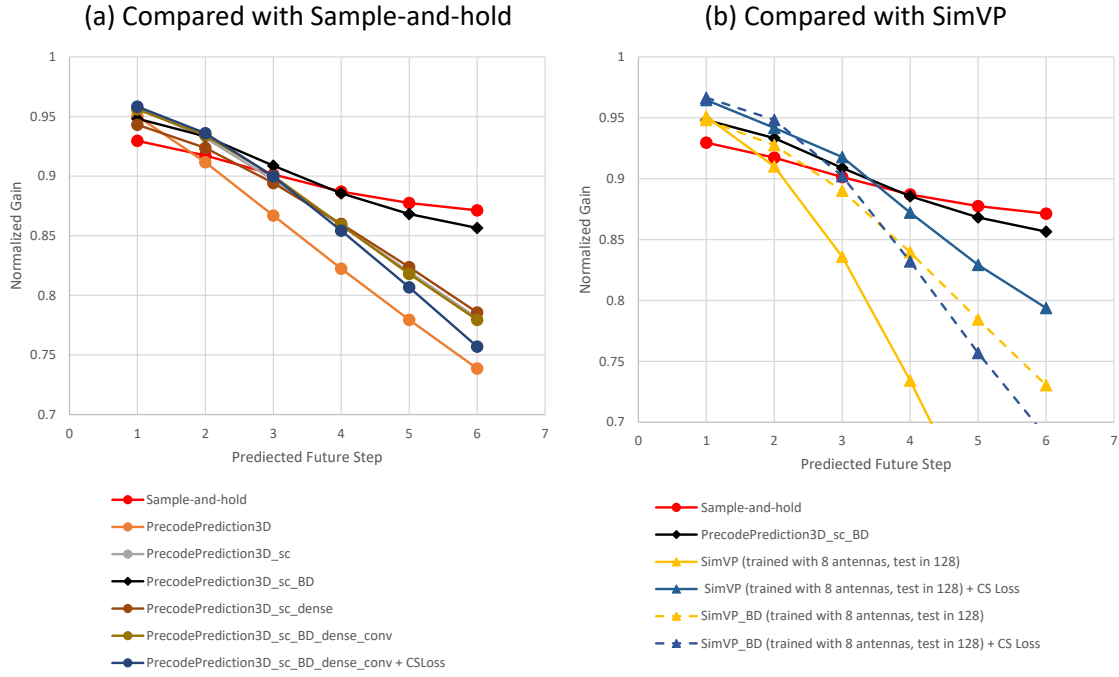


Figure 9.4: Performance comparison of the proposed approaches, SimVP, and sample-and-hold methods in terms of normalized gain performance.

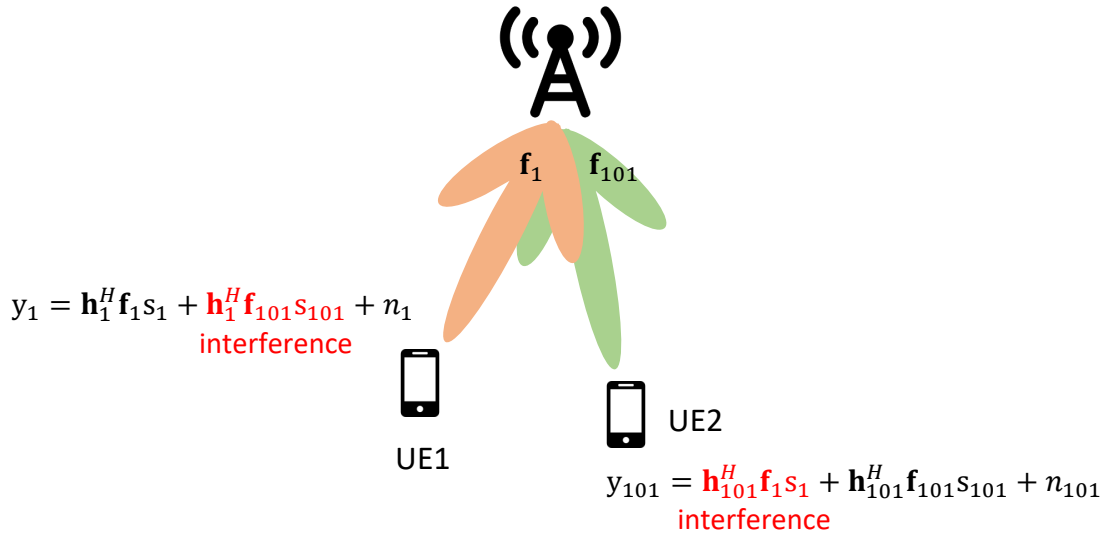


Figure 9.5: Illustration of mutual interference in MU-MIMO case. Note that we hope the mutual interference can be minimized. That is, the interference terms $\mathbf{h}_1^H \mathbf{f}_{101}$ and $\mathbf{h}_{101}^H \mathbf{f}_1$ can be minimized.

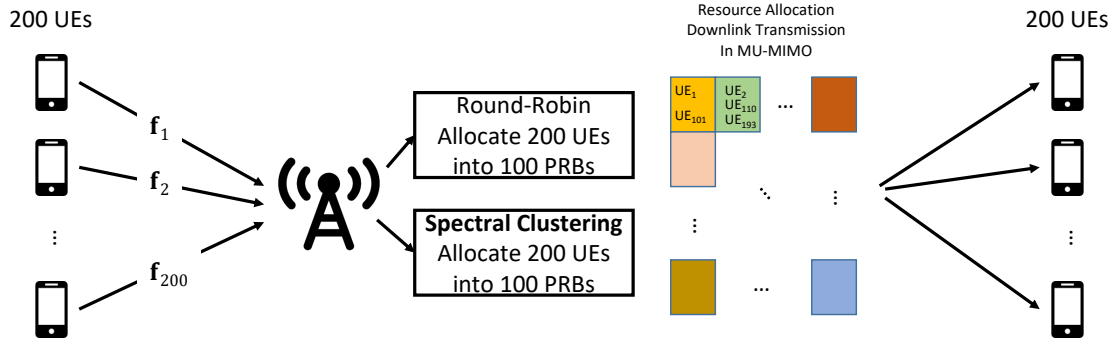


Figure 9.6: Illustration of the precoder-based user scheduling problem. Each color represents independent PRBs. The UEs in it denote the UEs that are assigned to be transmitted in the PRB.

We propose a precoder-based user scheduling approach based on a previous spectral clustering approach [73], illustrated in Figure 9.6. In the example of Figure 9.6, there are 200 UEs served by a gNB and 100 PRBs available. Assume each UE feeds its Type II precoder back to the base station. The gNB’s goal is to assign the 200 UEs to the 100 PRBs to maximize system capacity.

Figure 9.7 shows the capacity when the CSI-based [73] and precoder-based spectral clustering approaches are compared to the round-robin (RR)¹ user scheduling approach under different SNRs and numbers of antennas. We find that there is only a slight performance difference between the two approaches, and both significantly outperform RR. In summary, we suggest prospective research on the following: instead of passively assigning UEs based on their precoders, we should adjust the precoders to avoid mutual interference while maintaining the precoder gain of their own channel.

9.3 Conclusions

In this article, we have presented several new CSI feedback frameworks that address different challenges faced by existing learning-based techniques. In Chapter 3, we improved CSI feedback recovery performance by designing a UL-CSI-aided learning

¹Note that the RR approach evenly assigns all UEs to all PRBs randomly.

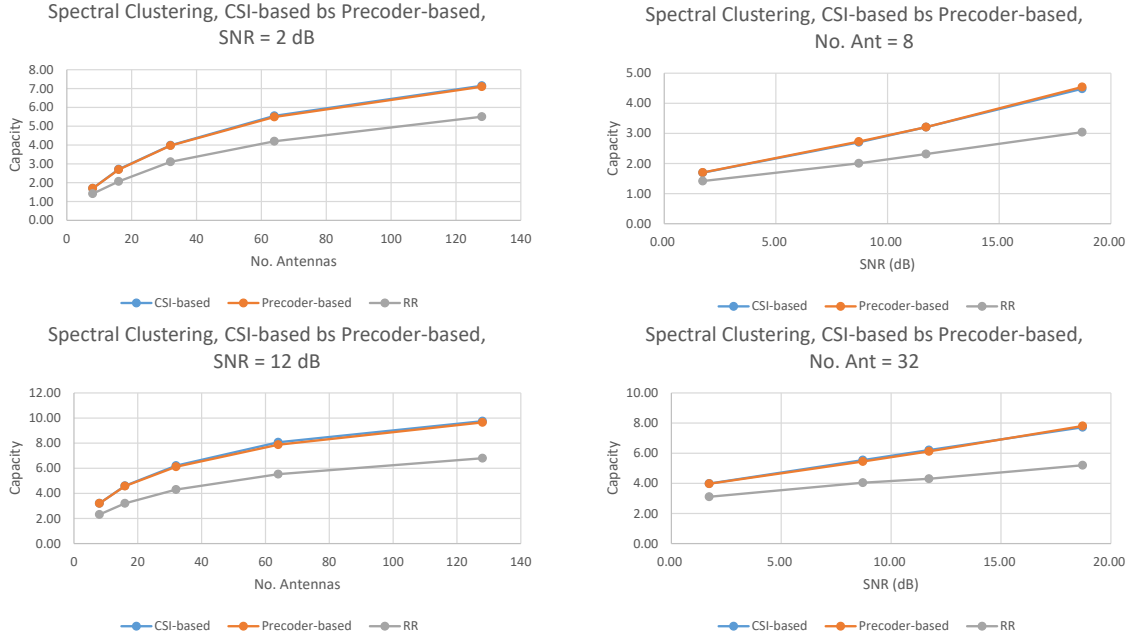


Figure 9.7: Illustration of (1) gNB-side precoder prediction and (2) UE-side channel prediction.

network that jointly optimizes magnitudes and phases. In Chapter 4, we aimed to save UL feedback overhead and reduce required resources for DL CSI training while maintaining recovery performance. We proposed a learning-based feedback framework that exploits FDD reciprocity to design CSI-RS precoding and CSI recovery schemes. In Chapter 5, we focused on significantly reducing model size by proposing a DCP-based CSI feedback along with tricks to further reduce storage and computational complexity. In Chapter 6, a JPEG-based CSI feedback framework was proposed. This framework is simple and does not require any prior training or knowledge about the channels, making it easy to implement in real-world wireless systems at low cost. In Chapter 7, we addressed the neglected undersampling issue prior to CSI feedback due to low-density pilot placement in current standards. We proposed a physics-inspired UL-CSI assisted CSI upsampling module to solve this problem, which can be applied to any similar CSI feedback frameworks. In Chapter 8, we focused on the undersampling issue in most cellular wireless FDD systems when performing codebook-based precoder feedback,

including Type II and eType II feedback. A plug-in precoder upsampling approach was proposed to upsample from SB-level precoders to RB-level ones to improve precoder gains.

In this chapter, we presented two future directions that we are focusing on. Firstly, we addressed the need for CSI and precoder prediction in practical FDD systems and presented some entry-level solutions to this task. We demonstrated the superior performance of the proposed approaches over the baseline sample-and-hold method and pointed out current limitations and future improvement directions that may be useful for prospective researchers. Lastly, we focused on the need for user scheduling in MU-MIMO FDD systems. We highlighted that some previous approaches perform well in user scheduling and provide decent performance but rely on an impractical assumption that the gNB has knowledge of DL CSIs. We modified a previous CSI-based spectral-clustering approach into a precoder-based one and achieved nearly the same performance as the original. Finally, we suggested further investigation into a precoder adjustment algorithm followed by user scheduling.

Appendix A

Appendix

Proof of Eq. (4.9):

For an $L \times N_b$ merging matrix \mathbf{T} with $L < N_b$, we have an underdetermined linear problem $\mathbf{y} = \mathbf{T}\mathbf{x}$. The minimum norm solution is simply

$$\mathbf{x}_{\text{mn}} = \mathbf{T}^H(\mathbf{T}\mathbf{T}^H)^{-1}\mathbf{T}\mathbf{x}, \quad (\text{A.1})$$

Based on singular value decomposition of \mathbf{T} by

$$\mathbf{T} = \mathbf{U} \begin{bmatrix} \boldsymbol{\Sigma} & \mathbf{0} \end{bmatrix} \mathbf{V}^H, \quad (\text{A.2})$$

where \mathbf{U} and \mathbf{V} respectively are left and right singular matrices corresponding to the $L \times L$ diagonal $\boldsymbol{\Sigma}$ of nonzero singular values. Let $\mathbf{V} = [\mathbf{v}_1 \ \mathbf{v}_2 \ \cdots \ \mathbf{v}_{N_b}]$ denote the corresponding right singular vectors. It is clear that

$$\mathbf{T}^H(\mathbf{T}\mathbf{T}^H)^{-1}\mathbf{T} = \mathbf{V} \begin{bmatrix} \mathbf{I}_{L \times L} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{V}^H = \sum_{i=1}^L \mathbf{v}_i \mathbf{v}_i^H \quad (\text{A.3})$$

Define a matrix $\tilde{\mathbf{I}} = \sum_{i=1}^L \mathbf{v}_i \mathbf{v}_i^H$. The minimum-norm solution is simply

$$\mathbf{x}_{\text{mn}} = \sum_{i=1}^L \mathbf{v}_i \mathbf{v}_i^H \mathbf{x} = \tilde{\mathbf{I}} \cdot \mathbf{x}. \quad (\text{A.4})$$

Since the singular vectors $\{\mathbf{v}_i\}$ are orthonormal, i.e., $\mathbf{v}_i^H \mathbf{v}_i = 1$, it is clear that

$$\begin{aligned} \text{Trace}\{\tilde{\mathbf{I}}\} &= \sum_{i=1}^L \text{Trace}\{\mathbf{v}_i \mathbf{v}_i^H\} \\ &= \sum_{i=1}^L \text{Trace}\{\mathbf{v}_i^H \mathbf{v}_i\} \end{aligned} \quad (\text{A.5})$$

$$= \sum_{i=1}^L 1 = L \quad (\text{A.6})$$

in which the equality of Eq. (A.5) holds because $\text{Trace}\{\mathbf{A}\mathbf{B}\} = \text{Trace}\{\mathbf{B}\mathbf{A}\}$.

Bibliography

- [1] I. Daubechies, M. Defrise, and C. Mol, “An Iterative Thresholding Algorithm for Linear Inverse Problems with a Sparsity Constrains,” *Commun. Pure Applied Math.*, vol. 57, 11 2004.
- [2] D. Donoho, A. Maleki, and A. Montanari, “Message Passing Algorithms for Compressed Sensing: I. Motivation and Construction,” in *IEEE Inf. Theory Workshop Inf. Theory*, 2010, pp. 1–5.
- [3] W. Yin, H. Jiang, and Y. Zhang, “An Efficient Augmented Lagrangian Method with Applications to Total Variation Minimization,” *Comput. Optimization and Appl.*, vol. 56, pp. 507–530, 12 2013.
- [4] C. A. Metzler, A. Maleki, and R. G. Baraniuk, “From Denoising to Compressed Sensing,” *IEEE Trans. Inf. Theory*, vol. 62, no. 9, pp. 5117–5144, 2016.
- [5] C. Wen, W. Shih, and S. Jin, “Deep learning for massive mimo csi feedback,” *IEEE Wireless Communications Letters*, vol. 7, no. 5, pp. 748–751, 2018.
- [6] X. Lin, “An overview of 5G Advanced evolution in 3GPP release 18,” *arXiv preprint arXiv:2201.01358*, 2022.
- [7] J. G. *et al.*, “Convolutional neural network-based multiple-rate compressive sensing for massive mimo csi feedback: Design, simulation, and analysis,” *IEEE Trans. Wirel. Commun.*, vol. 19, no. 4, pp. 2827–2840, 2020.

- [8] T. Wang, C.-K. Wen, S. Jin, and G. Y. Li, “Deep learning-based csi feedback approach for time-varying massive mimo channels,” *IEEE Wireless Communications Letters*, vol. 8, no. 2, pp. 416–419, 2019.
- [9] Z. Lu, J. Wang, and J. Song, “Multi-resolution CSI Feedback with Deep Learning in Massive MIMO System,” in *IEEE Intern. Conf. Communications (ICC)*, 2020, pp. 1–6.
- [10] Q. Yang, M. B. Mashhadi, and D. Gündüz, “Deep Convolutional Compression For Massive MIMO CSI Feedback,” in *IEEE Intern. Workshop Mach. Learning for Signal Process. (MLSP)*, 2019, pp. 1–6.
- [11] S. Ji and M. Li, “CLNet: Complex Input Lightweight Neural Network Designed for Massive MIMO CSI Feedback,” *IEEE Wirel. Commun. Lett.*, vol. 10, no. 10, pp. 2318–2322, 2021.
- [12] Z. Liu, L. Zhang, and Z. Ding, “An Efficient Deep Learning Framework for Low Rate Massive MIMO CSI Reporting,” *IEEE Trans. Commun.*, vol. 68, no. 8, pp. 4761–4772, 2020.
- [13] —, “Exploiting Bi-Directional Channel Reciprocity in Deep Learning for Low Rate Massive MIMO CSI Feedback,” *IEEE Wirel. Commun. Lett.*, vol. 8, no. 3, pp. 889–892, 2019.
- [14] Y.-C. Lin, Z. Liu, T.-S. Lee, and Z. Ding, “Deep Learning Phase Compression for MIMO CSI Feedback by Exploiting FDD Channel Reciprocity,” *IEEE Wireless Commun. Lett.*, vol. 10, no. 10, pp. 2200–2204, 2021.
- [15] Y. Ding and B. D. Rao, “Dictionary Learning-based Sparse Channel Representation and Estimation for FDD Massive MIMO Systems,” *IEEE Trans. Wirel. Commun.*, vol. 17, no. 8, pp. 5437–5451, 2018.

- [16] X. Zhang, L. Zhong, and A. Sabharwal, “Directional Training for FDD Massive MIMO,” *IEEE Trans. Wirel. Commun.*, vol. 17, no. 8, pp. 5183–5197, 2018.
- [17] Z. Liu, M. Rosario, and Z. Ding, “A Markovian Model-Driven Deep Learning Framework for Massive MIMO CSI Feedback,” *arXiv preprint arXiv:2009.09468*, 2020.
- [18] J. Guo *et al.*, “DL-based CSI Feedback and Cooperative Recovery in Massive MIMO,” *arXiv preprint arXiv:2003.03303*, 2020.
- [19] J. Guo, C.-K. Wen, and S. Jin, “CANet: Uplink-Aided Downlink Channel Acquisition in FDD Massive MIMO Using Deep Learning,” *IEEE Trans. Commun.*, vol. 70, no. 1, pp. 199–214, 2022.
- [20] Y. Sun, W. Xu, L. Liang, N. Wang, G. Y. Li, and X. You, “A Lightweight Deep Network for Efficient CSI Feedback in Massive MIMO Systems,” *IEEE Wirel. Commun. Lett.*, vol. 10, no. 8, pp. 1840–1844, 2021.
- [21] Y. Cui, J. Guo, C.-W. Wen, S. Jin, and S. Han, “Unsupervised Online Learning in Deep Learning-Based Massive MIMO CSI Feedback,” *IEEE Commun. Lett.*, vol. 26, no. 9, pp. 2086–2090, 2022.
- [22] J. Jiang *et al.*, “Federated Learning-Based Codebook Design for Massive MIMO Communication System,” in *Advances in Natural Computation, Fuzzy Systems and Knowledge Discovery*. Springer International Publishing, 2022, p. 1198–1205.
- [23] 3GPP, “NR; Physical Channels and Modulation,” 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 38.211, June 2020, version 16.6.0.
- [24] Y.-C. Lin, T.-S. Lee, and Z. Ding, “Deep Learning for Partial MIMO CSI Feedback by Exploiting Channel Temporal Correlation,” *arXiv preprint arXiv:xxxxxx*, 2021.

- [25] J. Guo, C.-K. Wen, S. Jin, and G. Y. Li, “Convolutional Neural Network-Based Multiple-Rate Compressive Sensing for Massive MIMO CSI Feedback: Design, Simulation, and Analysis,” *IEEE Trans. Wirel. Commun.*, vol. 19, no. 4, pp. 2827–2840, 2020.
- [26] H. Tang, J. Guo, M. Matthaiou, C.-K. Wen, and S. Jin, “Knowledge-distillation-aided Lightweight Neural Network for Massive MIMO CSI Feedback,” in *IEEE Veh. Technol. Conf. (VTC2021-Fall)*, 2021, pp. 1–5.
- [27] J. Guo, J. Wang, C.-K. Wen, S. Jin, and G. Y. Li, “Compression and Acceleration of Neural Networks for Communications,” *IEEE Wirel. Commun.*, vol. 27, no. 4, pp. 110–117, 2020.
- [28] X. Li, J. Guo, C.-K. Wen, S. Jin, and S. Han, “Multi-task Learning-based CSI Feedback Design in Multiple Scenarios ,” *arXiv preprint arXiv:2204.12698*, 2022.
- [29] Y. Wang *et al.*, “Multi-Rate Compression for Downlink CSI Based on Transfer Learning in FDD Massive MIMO Systems,” in *IEEE Veh. Technol. Conf. (VTC2021-Fall)*, 2021, pp. 1–5.
- [30] S. Jo and J. So, “Adaptive Lightweight CNN-Based CSI Feedback for Massive MIMO Systems,” *IEEE Wirel. Commun. Lett.*, vol. 10, no. 12, pp. 2776–2780, 2021.
- [31] Y. L. J. Guo, C.-K. Wen, “Overview of Deep Learning-based CSI Feedback in Massive MIMO Systems,” *arXiv preprint arXiv:2206.14383*, 2022.
- [32] Y. Yang, F. Gao, Z. Zhong, B. Ai, and A. Alkhateeb, “Deep Transfer Learning-Based Downlink Channel Prediction for FDD Massive MIMO Systems,” *IEEE Trans. Commun.*, vol. 68, no. 12, pp. 7485–7497, 2020.

- [33] J. Zeng *et al.*, “Downlink CSI Feedback Algorithm With Deep Transfer Learning for FDD Massive MIMO Systems,” *IEEE Trans. Cogn. Commun.*, vol. 7, no. 4, pp. 1253–1265, 2021.
- [34] Intel, “On NR Type I Codebook,” TSG RAN WG1 No 88 R1-1702205, 2021.
- [35] Samsung, “Type II CSI Reporting,” TSG RAN WG1 No 89 R1-1707962, 2021.
- [36] 3GPP, “Physical layer procedures for data ,” 3GPP TS 38.214 version 16.2.0 Release 16, 2020.
- [37] Y. Lin *et al.*, “Learning-Based Phase Compression and Quantization for Massive MIMO CSI Feedback with Magnitude-Aided Information,” *arXiv preprint arXiv:2103.00432*, 2021.
- [38] G. Parascandolo, H. Huttunen, and T. Virtanen, “Taming the Waves: Sine as Activation Function in Deep Neural Networks,” in *ICLR 2017*, 2017.
- [39] L. Liu *et al.*, “The COST 2100 MIMO Channel Model,” *IEEE Wirel. Commun.*, vol. 19, no. 6, pp. 92–99, 2012.
- [40] S. Jaeckel *et al.*, “QuaDRiGa: A 3-D Multi-Cell Channel Model with Time Evolution for Enabling Virtual Field Trials,” *IEEE Trans. Antennas and Propag.*, vol. 62, no. 6, pp. 3242–3256, 2014.
- [41] G. Morozov, A. Davydov, and V. Sergeev, “Enhanced CSI Feedback for FD-MIMO with Beamformed CSI-RS in LTE-A Pro Systems,” in *VTC-Fall*, 2016, pp. 1–5.
- [42] R. L. Haupt, “Array Beamforming,” *Timed Arrays: Wideband and Time Varying Antenna Arrays*, pp. 78–94, 2015.
- [43] C.-H. Lin, W.-C. Kao, S.-Q. Zhan, and T.-S. Lee, “BsNet: A Deep Learning-Based Beam Selection Method for mmWave Communications,” in *VTC-Fall*, 2019, pp. 1–6.

- [44] 3GPP, “Beam management,” 3GPP, Technical Report (TR) 38.802, Sep. 2017, version 16.6.0.
- [45] W. Debaenst, A. Feys, I. Cuiñas, M. G. Sánchez, and J. Verhaevert, “RMS Delay Spread vs. Coherence Bandwidth from 5G Indoor Radio Channel Measurements at 3.5 GHz Band,” *Sensors*, vol. 20, no. 3, 2020.
- [46] D. J. Ketchen and C. L. Shook, “The Application of Cluster Analysis in Strategic Management Research: An Analysis and Critique,” *Strategic Management Journal*, vol. 17, no. 6, pp. 441–458, 1996.
- [47] A. Beck and N. Teboulle, “A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems,” *Society for Industrial and Applied Mathematics*, vol. 2, no. 1, p. 183–202, Mar. 2009.
- [48] J. Zhang and B. Ghanem, “ISTA-Net: Interpretable Optimization-Inspired Deep Network for Image Compressive Sensing,” in *IEEE CVPR*, 06 2018, pp. 1828–1837.
- [49] M. Yang and N. Bourbakis, “An Overview of Lossless Digital Image Compression Techniques,” in *Midwest Symp. Circuits Syst.*, vol. 2, 2005, pp. 1099–1102.
- [50] Y.-C. Lin, T.-S. Lee, and Z. Ding, “A Scalable Deep Learning Framework for Dynamic CSI Feedback with Variable Antenna Port Numbers,” *IEEE Trans. Wirel. Commun.*, vol. 23, no. 4, pp. 3102–3116, 2024.
- [51] R. Chataut and R. Akl, “Massive MIMO Systems for 5G and beyond Networks—Overview, Recent Trends, Challenges, and Future Research Direction,” *Sensors*, vol. 20, no. 10, 2020.
- [52] M. Soltani, V. Pourahmadi, A. Mirzaei, and H. Sheikhzadeh, “Deep Learning-Based Channel Estimation,” *IEEE Commun. Lett.*, vol. 23, no. 4, pp. 652–655, 2019.

- [53] E. Balevi, A. Doshi, and J. Andrews, “Massive MIMO Channel Estimation With an Untrained Deep Neural Network,” *IEEE Trans. Wirel. Commun.*, vol. 19, no. 3, pp. 2079–2090, 2020.
- [54] L. Ramos *et al.*, “Mobile Channel Multipath Measurements and Statistical Characterization in Sub-6 GHz Bands,” in *IMOC 2023*, 11 2023.
- [55] K. Haneda *et al.*, “5g 3gpp-like channel models for outdoor urban microcellular and macrocellular environments,” in *2016 IEEE 83rd Vehicular Technology Conference (VTC Spring)*, 2016, pp. 1–7.
- [56] “Study on channel model for frequencies from 0.5 to 100 GHz,” *3GPP TR 38.901 version 14.3.0 Release 14*, 2018.
- [57] L. Tan and J. Jiang, *Digital Signal Processing: Fundamentals and Applications*, 2nd ed. USA: Academic Press, Inc., 2013.
- [58] Z. Zhong, L. Fan, and S. Ge, “FDD Massive MIMO Uplink and Downlink Channel Reciprocity Properties: Full or Partial Reciprocity?” in *GLOBECOM 2020 - 2020 IEEE Global Communications Conference*, 2020, pp. 1–5.
- [59] F. Rottenberg, R. Wang, J. Zhang, and A. F. Molisch, “Channel Extrapolation in FDD Massive MIMO: Theoretical Analysis and Numerical Validation,” in *2019 IEEE Global Communications Conference (GLOBECOM)*, 2019, pp. 1–7.
- [60] C. Dong, C. Chen, K. He, and X. Tang, “Image Super-Resolution Using Deep Convolutional Networks,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, 2016.
- [61] S. Gao, X. Liu, B. Zeng, S. Xu, Y. Li, X. Luo, J. Liu, X. Zhen, and B. Zhang, “Implicit Diffusion Models for Continuous Super-Resolution,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, June 2023, pp. 10 021–10 030.

- [62] J. Fang, H. Lin, X. Chen, and K. Zeng, “A Hybrid Network of CNN and Transformer for Lightweight Image Super-Resolution,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, June 2022, pp. 1103–1112.
- [63] Z. Liu, L. Wang, L. Xu, and Z. Ding, “Deep Learning for Efficient CSI Feedback in Massive MIMO: Adapting to New Environments and Small Datasets,” *arXiv preprint arXiv:2211.14785*, 2023.
- [64] G. T. Liang and J. L. Wang, S. Shi, and X. Zhang, “Pruning and quantization for deep neural network acceleration: A survey,” *Neurocomputing*, vol. 461, pp. 370–403, 2021.
- [65] S. Han, H. Mao, and W. J. Dally, “Deep Compression: Compressing Deep Neural Network with Pruning, Trained Quantization and Huffman Coding,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2015.
- [66] Y.-C. Lin, Y. Xin, T.-S. Lee, C. Zhang, and Z. Ding, “Physics-Inspired Deep Learning Anti-Aliasing Framework in Efficient Channel State Feedback,” *arXiv preprint arXiv:2403.08133*, 2024.
- [67] D. Madhubabu and A. Thakre, “Long-short term memory based channel prediction for siso system,” in *2019 International Conference on Communication and Electronics Systems (ICCES)*, 2019, pp. 1–5.
- [68] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-C. Woo, “Convolutional LSTM Network: a machine learning approach for precipitation nowcasting,” ser. NIPS’15. Cambridge, MA, USA: MIT Press, 2015, p. 802–810.
- [69] Y.-C. Lin, M.-X. Gu, C.-H. Lin, and T.-S. Lee, “Deep-Learning Based Decentralized Frame-to-Frame Trajectory Prediction Over Binary Range-Angle Maps for Automotive Radars,” *IEEE Trans. Veh. Techn.*, vol. 70, no. 7, pp. 6385–6398, 2021.

- [70] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lucic, and C. Schmid, “ViViT: A Video Vision Transformer,” *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6816–6826, 2021.
- [71] Z. Gao, C. Tan, L. Wu, and S. Z. Li, “SimVP: Simpler yet Better Video Prediction,” *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3160–3170, 2022.
- [72] C. Feres and Z. Ding, “An Unsupervised Learning Paradigm for User Scheduling in Large Scale Multi-Antenna Systems,” *IEEE Trans. Wirel. Commun.*, vol. 22, no. 5, pp. 2932–2945, 2023.
- [73] C.-H. Hsu, C. Feres, and Z. Ding, “Spectral Clustering Aided User Grouping and Scheduling in Wideband MU-MIMO Systems,” in *ICC 2023 - IEEE Int. Conf. Commun.*, 2023, pp. 4292–4297.