Convolutional Neural Network-Based Multiple-Rate Compressive Sensing for Massive MIMO CSI Feedback: Design, Simulation, and Analysis

Jiajia Guo, Chao-Kai Wen, *Member, IEEE*, Shi Jin, *Senior Member, IEEE*, and Geoffrey Ye Li, *Fellow, IEEE*

***Abstract*— Massive multiple-input multiple-output (MIMO) is a promising technology to increase link capacity and energy efficiency. However, these benefits are based on available channel state information (CSI) at the base station (BS). Therefore, user equipment (UE) needs to keep on feeding CSI back to the BS, thereby consuming precious bandwidth resource. Large-scale antennas at the BS for massive MIMO seriously increase this overhead. In this paper, we propose a multiple-rate compressive sensing neural network framework to compress and quantize the CSI. This framework not only improves reconstruction accuracy but also decreases storage space at the UE, thus enhancing the system feasibility. Specifically, we establish two network design principles for CSI feedback, propose a new network architecture, CsiNet+, according to these principles, and develop a novel quantization framework and training strategy. Next, we further introduce two different variable-rate approaches, namely, SM-CsiNet+ and PM-CsiNet+, which decrease the parameter number at the UE by 38.0% and 46.7%, respectively. Experimental results show that CsiNet+ outperforms the state-of-the-art network by a margin but only slightly increases the parameter number. We also investigate the compression and reconstruction mechanism behind deep learning-based CSI feedback methods via parameter visualization, which provides a guideline for subsequent research.**

***Index Terms*— Massive MIMO, FDD, CSI feedback, deep learning, compressive sensing, quantization, multiple-rate.**

Manuscript received May 17, 2019; revised October 15, 2019; accepted January 16, 2020. Date of publication January 28, 2020; date of current version April 9, 2020. This work was supported in part by the National Key Research and Development Program under Grant 2018YFA0701602, in part by the National Natural Science Foundation of China (NSFC) for Distinguished Young Scholars of China under Grant 61625106, and in part by the NSFC under Grant 61531011. The work of Chao-Kai Wen was supported in part by the Ministry of Science and Technology of Taiwan under Grant MOST 108-2628-E-110-001-MY3 and in part by Qualcomm through a Taiwan University Research Collaboration Project. This article was presented in part at the 2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP), Ottawa, Canada, November 2019. The associate editor coordinating the review of this article and approving it for publication was X. Chen. *(Corresponding author: Shi Jin.)*

Jiajia Guo and Shi Jin are with the National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, China (e-mail: jiajiaguo@seu.edu.cn; jinshi@seu.edu.cn).

Chao-Kai Wen is with the Institute of Communications Engineering, National Sun Yat-sen University, Kaohsiung 80424, Taiwan (e-mail: chaokai.wen@mail.nsysu.edu.tw).

Geoffrey Ye Li is with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: liye@ece.gatech.edu).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TWC.2020.2968430

I. Introduction

MASSIVE multiple-input multiple-output (MIMO) is a critical technology for 5G and beyond systems [2]–[4]. In massive MIMO systems, base stations (BSs), equipped with a large number of antennas, can recover information received from user equipment (UE) at low signal-to-noise-ratio (SNR) and simultaneously serve multiple users [5]–[7]. However, BSs should obtain the instantaneous channel state information (CSI) to acquire these potential benefits and the accuracy of the obtained CSI directly affects the performance of the massive MIMO systems [5]. For the uplink, BSs can easily estimate CSI accurately through the pilots sent by the UE. However, the downlink CSI is difficult to achieve, especially in frequency-division duplexing (FDD) systems, which are employed by the most cellular systems nowadays. In time-division duplexing (TDD) systems, downlink CSI can be inferred from uplink CSI utilizing the reciprocity [8]. However, in FDD systems, weak reciprocity is present, thereby making it hard to infer downlink CSI by observing uplink CSI [9].

In traditional MIMO systems, downlink CSI in FDD systems is first estimated at the UE by the pilots and then fed back to the BS. However, this feedback strategy is infeasible in massive MIMO because the substantial antennas at the BS greatly increase the dimension of CSI matrix, thereby leading to a large overhead [8], [10]. To address this issue, the CSI matrix should be efficiently compressed [10], [11], which can be based on compressive sensing (CS) or deep learning (DL). The CS-based methods exploit the sparsity of massive MIMO CSI in certain domain [12]. In [13], CS has been first applied to CSI feedback in the spatial-frequency domain, which exploits the high spatial correlation of CSI resulting from the limited distance among antennas in massive MIMO. In [14], a hidden joint sparsity structure in the user channel matrices has been found and exploited due to the shared local scatterers. CS techniques simplify the encoding (compression) process; but, the decoding (decompression) process turns into solving an optimization problem and demands substantial computing sources and time [11], thereby making it difficult to implement in many practical communication systems.

DL recently made tremendous strides in several aspects, including computer vision, natural language processing, and wireless communications [15]–[17]. The DL-based

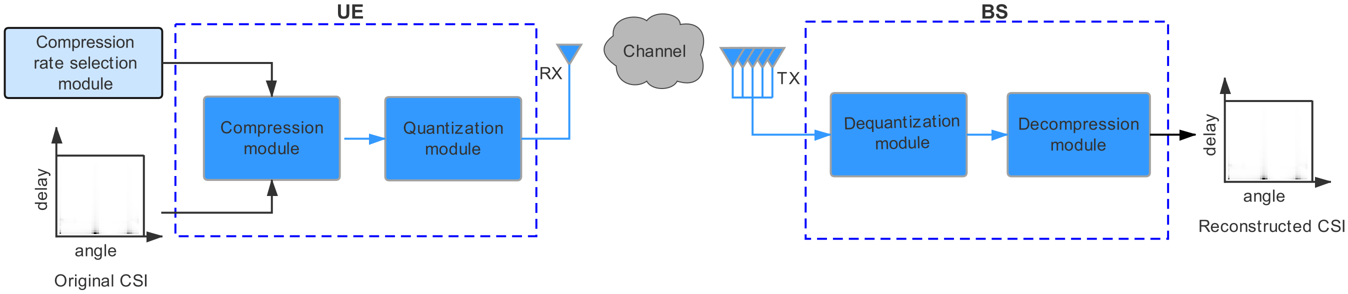


Fig. 1. Overview of the multiple-rate bit-level compressive sensing CSI feedback framework. UE compresses CSI matrix with a selected CR, quantizes the measurement vector, and then transmits it. Once BS receives the transmitted bitstream, it dequantizes bitstream and then decompresses measurement vector.

non-iterative methods have shown outstanding performance on image compression. Traditional algorithms use iterative methods to solve image reconstruction optimization problem. The ReconNet in [18] recovers images utilizing stacked convolutional layers without iteration, thereby reducing the reconstruction time by a margin. From [19], it is a good strategy to generate a preliminary reconstruction at the decoder via a linear mapping network and then use a residual network to refine estimates. In [20], the random Gaussian measurement matrix is replaced by a learned measurement matrix at the encoder.

The DL-based image compression technique is first introduced to massive MIMO CSI feedback in [21] based on the autoencoder architecture in [22]. In the CsiNet in [21], the encoder acts as the role of compression module instead of randomized Gaussian matrix at the UE and the decoder is regarded as the decompression module at the BS. Then, CsiNet-LSTM in [23] improves the reconstruction accuracy by considering the temporal correlations of CSI utilizing long-short time memory (LSTM) architecture [24]. From [25], this neural network architecture can be modified to significantly reduce the number of network parameters. Based on the reciprocity between bi-directional channels, uplink CSI information can help reconstruct the downlink CSI in [26]. The work in [27] reduces impact of the feedback transmission errors and delays. It has been shown in [28] that the performance of DL-based measurement matrix is better than that of randomized measurement matrices (e.g., Gaussian and Bernoulli distribution). In [29], a novel joint convolutional residual network architecture, JC_ResNet, is introduced to efficiently extract channel features.

In the above-mentioned DL-based work, DL-based models are regarded as a black box and have no interpretation of why excellent performance can be obtained. Meanwhile, the impact of the quantization process has been ignored, thereby leading to substantial errors in practical wireless communication systems. The CSI feedback in massive MIMO systems should be drastically compressed while the coherence time is short and vice versa. Therefore, the compression rate (CR) must be adjusted according to the environments. The iterative algorithms should be able to work for different CRs. However, the existing DL-based methods can only compress the CSI matrix with a fixed CR. The UE has to store several CS network architectures and corresponding parameter sets to realize multiple-rate CSI compression, which is infeasible due to the limited storage space at the UE.

In this work, we propose a multiple-rate compressive sensing framework as shown in Fig. 1, which will not only improve the reconstruction accuracy but also bridge the gap between DL-based methods and practical deployment. First, we introduce a new network architecture, namely, CsiNet+, modified from CsiNet, which exploits the sparsity characteristics of CSI in angular-delay domain and the refinement theory. Then, we develop a novel framework and training strategy for quantization, which does not need extra storage space for different quantization rates at the UE. Subsequently, two different network frameworks are developed for variable CR compression, thereby greatly saving storage space. Finally, we discuss the compression and reconstruction mechanism of DL-based methods via parameter visualization and evaluate the performance.

The major contributions of this work are summarized as follows:

- After investigating the characteristics of CSI from the aspect of sparsity and the key idea of refinement theory in DL, we propose two network design principles for CSI feedback, which provides a guideline for future network design. We propose a new network architecture, named CsiNet+, which improves the original CsiNet.
- We introduce a novel quantization framework and training strategy, which is especially suitable to CSI feedback in massive MIMO systems. This framework and training strategy require no architectural change or parameter update at the UE. Neural networks are used to offset the quantization distortion. Furthermore, different quantization rates can be realized without increasing parameter number or computational resource at the UE.
- We propose two different variable rate frameworks, namely, SM-CsiNet+ and PM-CsiNet+. They also reduce the parameter number by 38.0% and 46.7%, respectively, thereby greatly saving the storage space at the UE. This work is the first to address variable CR issue in DL-based CSI feedback.
- We investigate the compression and reconstruction mechanism of DL-based CSI feedback via parameter visualization and obtain insightful understanding of DL-based CSI feedback, which is the first to reveal the reason behind the excellent performance of DL-based methods, and

provides important guidelines for subsequent research in this area.

The rest of this work is organized as follows. In Section II, we introduce the system model, including channel model and CSI feedback process. Then, the novel network architecture CsiNet+ and quantization framework are presented in Section III. Section IV introduces two different variable rate frameworks. In Section V, we provide the experiment details and numerical results of the proposed networks and frameworks, and reveal the compression and reconstruction mechanism. Section VI finally concludes our work.

II. System Model

After introducing the massive MIMO-orthogonal frequency division multiplexing (OFDM) system in this section, we will describe the CSI feedback process.

A. Massive MIMO-OFDM System

We consider a single-cell FDD massive MIMO-OFDM system, where there are $N_t(\gg 1)$ transmit antennas at the BS and a single receiver antenna at the UE, OFDM is with N_c subcarriers. The received signal at the n-th subcarrier can be expressed as follows:

$$y_n = \tilde{\mathbf{h}}_n^H \mathbf{v}_n x_n + z_n, \tag{1}$$

where $\tilde{\mathbf{h}}_n$ and $\mathbf{v}_n \in \mathbb{C}^{N_t\times 1}$ are the channel frequency response vector and the precoding vector at the n-th subcarrier, respectively, x_n represents the transmitted data symbol, z_n is the additive noise or interference, and $(\cdot)^H$ represents conjugate transpose. The CSI matrix in the spatial-frequency domain can be expressed in matrix form as $\tilde{\mathbf{H}} = [\tilde{\mathbf{h}}_1, \tilde{\mathbf{h}}_2, \ldots, \tilde{\mathbf{h}}_{N_c}]^H \in \mathbb{C}^{N_t\times N_c}$.

In the FDD system, UE estimates the downlink channel and then feeds this information (CSI) to the BS. With the downlink CSI, the BS calculates precoding vector $\mathbf{v}_n \in \mathbb{C}^{N_t\times 1}$ via singular value decomposition. The number of feedback parameters is $2N_cN_t$, which is proportional to the number of antennas. Excessive feedback in massive MIMO system greatly occupies the precious bandwidth.

We consider reducing feedback overhead by exploiting the sparsity of CSI in the angular-delay domain. The CSI matrix in the spatial-frequency domain can be converted into the angular-delay domain by 2D discrete Fourier transform (DFT) as follows:

$$\mathbf{H} = \mathbf{F}_{\rm d}\tilde{\mathbf{H}}\mathbf{F}_{\rm a}, \tag{2}$$

where $\mathbf{F}_d$ is a $N_c \times N_c$ DFT matrix and $\mathbf{F}_a$ is a $N_t \times N_t$ matrix. Due to the sparsity of massive MIMO channel in the angular-delay domain, most elements in the delay domain are near zero and only the first N_c' ($<N_c$) rows exhibit distinct non-zero values because the time delay among multiple paths only lies in a particularly limited period. Therefore, we directly truncate the channel matrix rows to the first N_c' rows that are with distinct non-zero values. Meanwhile, the channel matrix is also sparse in a defined angle domain by performing DFT on spatial domain channel vectors if the number of the transmit antennas $N_t \to \infty$ [30].

In this paper, we regard the 2D channel matrix as an image and the normalized absolute values of CSI matrix are regarded as the gray-scale values to visualize the sparsity of the retained $N_c' \times N_t$ channel matrix $\mathbf{H}$ in the angular-delay domain, which has been demonstrated in the literature, such as [31], [32].

B. CSI Feedback Process

Once the channel matrix $\mathbf{H}$ in the angular-delay domain is estimated at the UE, compression, quantization, and entropy encoding[1] will be used in turn to reduce CSI feedback overhead. The compressed CSI matrix can be expressed as follows:

$$\mathbf{H}_c = \mathcal{Q}(f_{\rm com}(\mathbf{H}, \Theta_1)), \tag{3}$$

where $f_{\rm com}(\cdot)$ and $\mathcal{Q}(\cdot)$ denote the compression and quantization processes, respectively, and Θ_1 represents parameters of the compression module (encoder).

Once the BS receives the compressed CSI matrix, dequantization and decompression will be used to recover the channel matrix in the angular-delay domain,

$$\hat{\mathbf{H}} = f_{\rm decom}(\mathcal{D}(\mathbf{H}_c, \Theta_2)), \tag{4}$$

where $\mathcal{D}(\cdot)$ and $f_{\rm com}(\cdot)$ represent the dequantization and decompression functions, respectively, and Θ_2 denotes the parameters in the decompression module (decoder). Therefore, the optimization compression and recovery can be formulated by combining (3) and (4) together with the mean-squared error (MSE) distortion metric as the following:

$$(\hat{\Theta}_1,\hat{\Theta}_2) = \underset{\Theta_1,\Theta_2}{\arg\min}\ \|\mathbf{H} - f_{\rm decom}(\mathcal{D}(\mathcal{Q}(f_{\rm com}(\mathbf{H}, \Theta_1))), \Theta_2)\|_2^2. \tag{5}$$

III. CSI Compression Based on Convolutional Neural Networks

In this section, we describe the proposed framework, which mainly includes neural network architecture, and quantization and dequantization sub-modules.

A. Network Architecture for Channel Dimension Reduction

The CsiNet in [21], an encoder-decoder structure, has demonstrated promising performance in CSI compression and reconstruction. The encoder first extracts CSI features via a convolutional layer with two 3×3 filters, followed by an activation layer. Then, a fully connected (FC) layer with M neurons is adopted to compress the CSI features to a lower dimension. The CR of this encoder can be calculated by:

$$CR = \frac{2N_tN_c}{M}. \tag{6}$$

At the decoder, the first layer is also an FC layer with $2N_tN_c$ neurons, and the output vector is reshaped with the same shape of the original channel matrix. The above layers produce the initial estimate of channel matrix $\mathbf{H}$. Then, the output is fed into two RefineNet blocks [21], which are designed

[1] Since entropy encoding is lossless, we do not take it into consideration in the following parts.

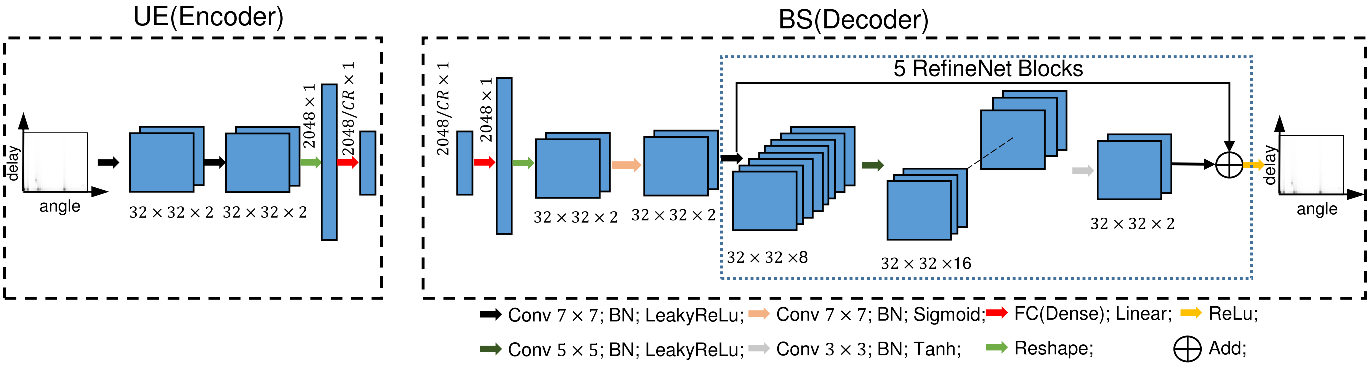
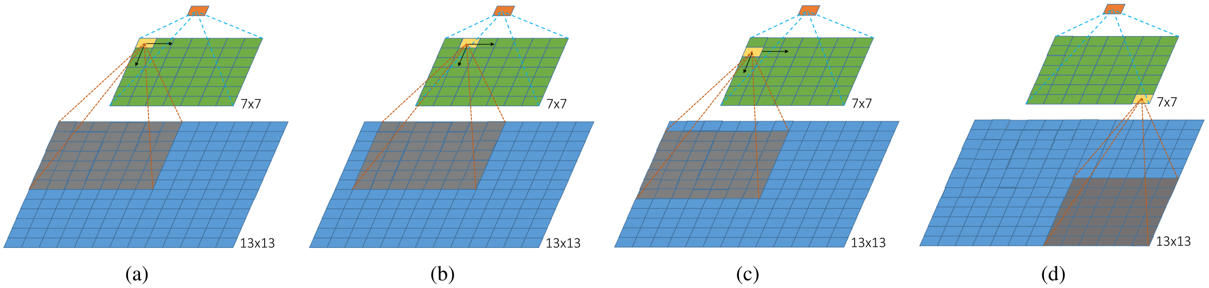


Fig. 2. Overview of CsiNet+ architecture. The left module is an encoder at the UE, compressing the CSI matrix. Meanwhile, the right module is a decoder at the BS, reconstructing CSI matrix from the received compressive measurements.



(a) (b) (c) (d)

Fig. 3. The illustration of the receptive field for two serial 7×7 convolution operations with stride 1×1. The upper 'pixel' is determined by the middle 7×7 area. Each 'pixel' in the middle 7×7 area is determined by the down 7×7 area, which is overlapped by each other.

to continuously refine the reconstruction and contain three convolutional layers and identity shortcut connections [33]. The last layer of CsiNet is a convolutional one with batch normalization (BN) [34] and Sigmoid activation layer, scaling the output to the $[0, 1]$ range.

The proposed architecture of neural network, CsiNet+, as shown in Fig. 2, is based on the CsiNet with two main modifications: convolutional kernel size and refinement process.

1) Modification 1: Wireless communication channels often exhibit a block-sparse structure, that is, a matrix that exhibits nonzero values occurring in clusters [35]. CsiNet and other CS-based feedback methods [23], [25] regard this sparsity of the channel matrix as a precondition. However, CsiNet first uses a convolutional layer with 3×3 filters to extract the features of block-sparse channel matrix, which is inappropriate in this scenario. In the image scenario, 3×3 filters can extract the edge information within a particularly small receptive field. In contrast, the visualization results of output of the first convolutional layer in CsiNet indicates that most output coefficients are near-zero, which is similar to the input channel matrix and contains less information. Specifically, if the receptive field is located in a large 'blank' area, the nine coefficients are still zero after being convoluted with 3×3 filters. Thus, these convolution operations can be regarded as futile. Meanwhile, the sparsity is unable to exhibit in the fully 'non-blank' area.

In [36], a new block activation unit with the block size of six has been proposed to handle the block-sparse vector in wideband wireless communication systems. [37] also finds that increasing the receptive field shows the improvement in the super-resolution accuracy. Inspired by them, we use two 7×7 convolutional layers with stride 1×1 in CsiNet+ to replace the first 3×3 convolutional layer of CsiNet at the encoder. As shown in Fig. 3, the two serial convolutional layers with 7×7 filters present a 13×13 receptive field, which is hardly located in the 'blank' area due to the large convolutional kernel size. Meanwhile, hardly no fully 'non-blank' area exists, so sparsity can effectively exhibit. For the same reason, we also replace the two serial 3×3 convolutional layers in RefineNet block with 7×7 and 5×5 ones, which is called as CsiNet-M1.

2) Modification 2: The key idea of the refinement is to improve the estimates from the initial ones via stacking convolutional layers and identity shortcut connections [18]. Each RefineNet block is optimized to ensure that its output is the same as the residual between its input and ground truth as much as possible, which can be expressed as,

$$\hat{\mathbf{H}}_{\text{res}} = \mathbf{H} - \hat{\mathbf{H}}_{\text{in}}, \tag{7}$$

where $\hat{\mathbf{H}}_{\text{in}}$ is the initial estimate and $\hat{\mathbf{H}}_{\text{res}}$ is the expected residual. Fundamentally, the output of the last RefineNet block should be the final estimate; otherwise, the refinement will be disturbed. In [19], [38], the RefineNet block is directly used as the last layer. However, in CsiNet, a convolutional layer follows the last RefineNet block, thereby disturbing the refinement. Therefore, we remove this convolutional layer in proposed CsiNet+.

Training the entire neural network DR^2-Net in [19] is conducted in two steps. First, the encoder and the first FC layer at the decoder are trained using a large learning rate to obtain a preliminary reconstructed image. Second, the encoder and decoder are trained jointly in an end-to-end manner using a smaller learning rate. Obviously, the two-step training

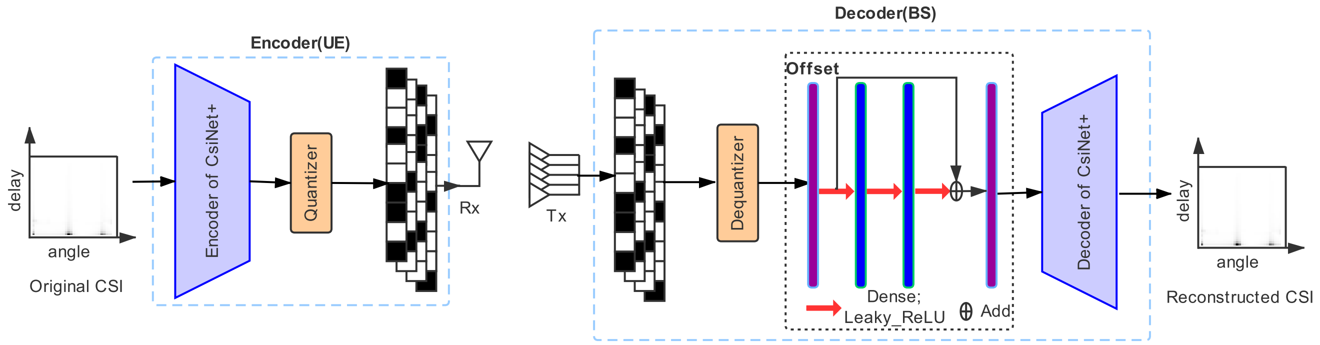


Fig. 4. Proposed bit-level CsiNet+ framework. The original CSI is first compressed at the encoder (UE), and then quantization is adopted to generate a bitstream. At the decoder (BS), the received measurement vectors are first dequantized and then fed into several neural networks.

strategy is inefficient and time-consuming. In contrast, CsiNet in [23] is trained via end-to-end learning, where a good initial estimate $\hat{\mathbf{H}}_{\text{in}}$ is difficult to obtain. Therefore, we add a 7×7 convolutional layer with a BN layer and Sigmoid activation between the FC layer and the first RefineNet block considering the disadvantages of the DR^2-Net and the CsiNet. The output of this added convolutional layer is regarded as the initial estimate $\hat{\mathbf{H}}_{\text{in}}$, whose quality has been improved by the extra layer. Similar to the CsiNet, the parameters of the encoder and decoder are updated during training via an end-to-end approach, which is called as CsiNet-M2.

The above are the two modifications in CsiNet+, as shown in Fig. 2. The left module is the encoder at the UE, which compresses the CSI matrix, and the right module is a decoder at the BS, which reconstructs the CSI matrix from compressed channel. The loss function of CsiNet+ is MSE.

B. Quantization and Dequantization

The output of the encoder at the UE in CsiNet+ needs to be converted into bitstream for transmission (feedback). Therefore, the output of the CsiNet+ encoder should be first quantized. Once the BS receives bitstream, dequantization is first used before feeding into the neural networks, as in Fig. 4.

1) Quantization Method: In [27], uniform quantization is used to discretize measurement vectors. However, it is not optimal for compressed CSI even if uniform quantization provides good quantization for strong signal. Therefore, in this work, we adopt a μ-law non-uniform quantizer, which is optimized by adapting a companding function $f(\cdot)$ as,

$$f(x) = \pm \frac{\ln(1+\mu|x|)}{1+\mu}, \tag{8}$$

where $x \in [-1, 1]$ is the weak signal and μ is a constant that determines companding amplitude.

2) Offset Module: After dequantizing the bitstream at the BS, an offset module [39] is first used to minimize the quantization distortion as follows,

$$\hat{f}_{\text{off}} = \arg\min_{f_{\text{off}}} (f_{\text{com}}(\mathbf{H}, \Theta_1) - f_{\text{off}}(\mathbf{H}_c)), \tag{9}$$

where f_{off} denotes the offset process. In order to model the function f_{off}, an offset neural network is designed to minimize the distortion, as shown in Fig.4. The offset network is based on residual learning and consisted of three FC layers in which there are N neurons.

3) Training Strategy: Since the quantization function is non-differentiable, the gradient of the entire bit-level network cannot be passed while using backpropagation learning algorithm, thereby making it impossible to train networks in the end-to-end way. The widely used solution is to set the quantization gradient (i.e., $round(\cdot)$ gradient) to a constant. Then, the entire neural networks including encoder and decoder are trained in an end-to-end way. This training strategy solves the gradient backpropagation problem due to the quantization function, but the network can only work with specific quantization bits. If requiring different quantization bit rates, different neural networks are required and substantial parameters need storing, which is inapplicable in CSI feedback due to the limited storage space in the UE. Occupying great storage space just for different quantization bit rates is not worthy (the number of encoder parameters sometimes is over 1 million).

In contrast with the existing quantization training strategy [39]–[41] that jointly trains the encoder and decoder with a fix quantization bit rate, we do not always train networks via end-to-end learning or train different encoders for different quantization bit rates. Specifically, we first train CsiNet+ without quantization via an end-to-end approach with a large learning rate. Next, we use non-uniform quantization to discrete the measurement vectors and the dequantizer at the BS conducts the inverse operation of quantization to recover the vectors, which generates the training set of the offset network. Then, the offset network is optimized by the Adam optimizer [42] with MSE loss function. Once the offset network is trained, we fix bit-level CsiNet+'s parameters of the encoder and fine-tune the offset and decoder networks with a small learning rate to further minimize the quantization distortion effect, which can be formulated as follows,

$$(\hat{\Theta}_2, \hat{\Theta}_3) = \arg\min_{\Theta_2, \Theta_3} F, \tag{10}$$

where

$$F = \|\mathbf{H} - f_{\text{off}}(f_{\text{decom}}(\mathcal{D}(\mathcal{Q}(f_{\text{com}}(\mathbf{H}, \hat{\Theta}_1)), \Theta_3)), \Theta_2)\|_2^2. \tag{11}$$

In (11), $\hat{\Theta}_1$ and Θ_3 denote the learned parameters at the encoder in CsiNet+ and the parameters in the offset network, respectively. Therefore, the UE only needs to store one parameter set regardless of quantization bit rates. The problem of different quantization bit rates is solved at the BS by training

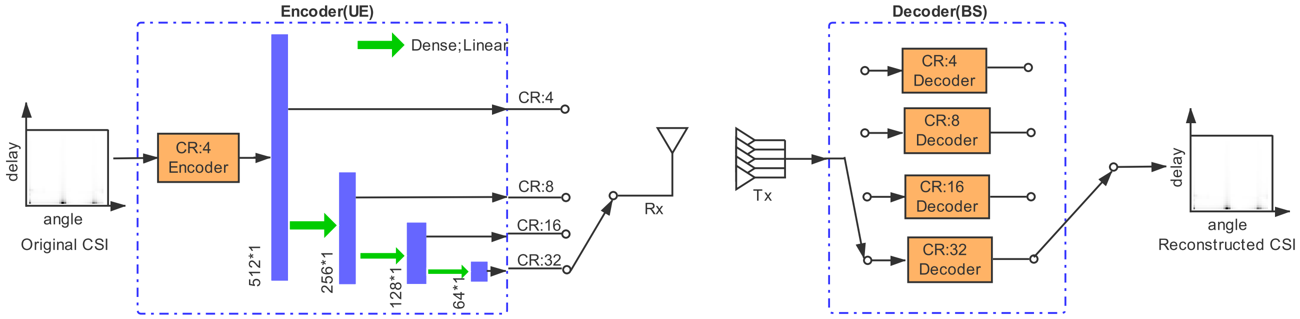


Fig. 5. Serial multiple-rate compression framework. The key idea of SM-CsiNet+ is that high compression measurement vectors can be generated from the low ones.

TABLE I
PARAMETER NUMBERS OF CSINET+ ENCODERS WITH DIFFERENT CRs

CR / Number	4	8	16	32
Total	1,049,500	524,956	262,684	131,548
FC layer	1,048,576	524,288	262,144	131,072
Proportion	99.91%	99.87%	99.79%	99.64%

distinct decoders for different quantization bit rates, which is feasible due to the great storage space at the BS.

IV. MULTIPLE-RATE CSI FEEDBACK

Although CSI compression can reduce feedback overhead, accuracy of reconstructed CSI at the BS is sacrificed, which may adversely affect MIMO communication network performance. Hence, communication systems sometime need adjust the CR according to the environments, as mentioned in Section I. In contrast with the traditional iterative algorithms that can work with different CRs, the existing DL-based methods can only compress CSI matrix with a fixed CR and have to train and store a different neural network for a different CR, thereby occupying large storage space at the UE. In this part, we focus on a multiple-rate framework, which can compress the CSI matrix at different CRs to save the storage space at the UE. As before, we neglect the decoder parameter number at the BS because the storage space of the BS is enough.

The CsiNet+ encoder is mainly composed of two convolutional layers, two BN layers, and one FC layer. N_{Conv}, N_{BN}, and N_{FC} represent the parameter number of the convolutional layer, BN layer, and FC layer, respectively, which are calculated by,

$$\begin{aligned} N_{\text{Conv}} &= C_{\text{in}}(K^2+1)\ C_{\text{out}}, \\ N_{\text{BN}} &= 4C_{\text{out}}, \\ N_{\text{FC}} &= N_{\text{out}}(N_{\text{in}}+1), \end{aligned} \tag{12}$$

where C_{in} and C_{out} are the numbers of the input and output features of convolutional layer, K denotes the convolutional kernel size, and N_{out} and N_{in} represent the numbers of the input and output neurons of the FC layer, respectively. As shown in Table I, the FC layer contains almost all model parameters that consume the most memory. The number of the parameters for the FC layer at the fourfold compression encoder module is 1,048,576 and occupies 99.9% of 1,049,500 overall parameters. Therefore, it is critical to use the multiple-rate compression framework to decrease the parameter number of the FC layers. We will reuse FC layers to decrease the encoder parameters. There are two kinds of multiple-rate compression: serial multiple-rate framework (SM-CsiNet+) and parallel multiple-rate framework (PM-CsiNet+). In the following, we will introduce SM-CsiNet+ and PM-CsiNet+ in detail.

A. Serial Multiple-Rate Compression Framework: SM-CsiNet+

In general, highly compressed measurement vectors can be generated from the low ones, as in Fig. 5. For instance, we can first compress the CSI matrix by fourfold and then continue to compress the compressed CSI matrix by twofold to obtain eightfold compression. This method decreases the FC layer parameter number from 2048×256 to 512×256 for eightfold compression compared with compressing from the original CSI matrix. Meanwhile, the first two convolutional layers, which are used to extract features, are also shared by different compression encoders, thereby further decreasing the number of encoder parameters. Similarly, if we want to compress CSI by 16-fold, then we can keep on compressing the aforementioned compressed vector by twofold.

We train this multiple-rate compression framework by an end-to-end approach. We concatenate the output of different decoders, generating a $32 \times 32 \times 8$ matrix as the output of the entire framework. The label of this framework is obtained by repeating the original CSI matrix by fourfold. We still use MSE as the loss function, which is calculated as follows,

$$\begin{aligned} &L_{\text{Total}}(\Theta) \\ &= c_4 L_4(\Theta_4) + c_8 L_8(\Theta_8) + c_{16} L_{16}(\Theta_{16}) + c_{32} L_{32}(\Theta_{32}), \end{aligned} \tag{13}$$

where L_{N}, c_{N}, and Θ_{N} are the MSE loss, weight, and the learnable parameters of the N-fold compression network. We can balance the magnitude of loss terms into similar scales by setting hyperparameter c_{N}. In the practical environments, the UE selects a suitable CR, and then the encoder compresses the CSI matrix to generate the corresponding measurement vectors. Once the BS receives these

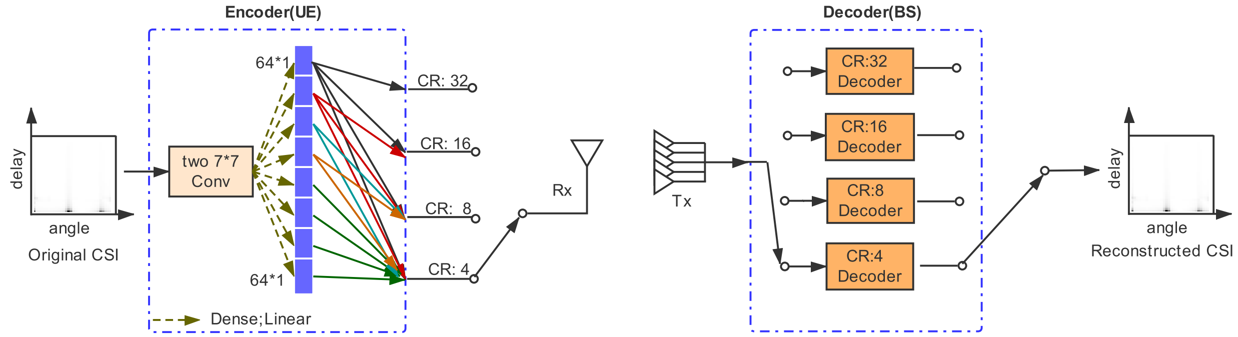


Fig. 6. Parallel multiple-rate compression framework.

measurement vectors, it decompresses them using the corresponding decoder network.

The parameter number of SM-CsiNet+ at the UE (ignoring the BN parameter number N_{BN}) can be calculated as:

$$N_{\mathrm{SM}} = N_{\mathrm{Conv}} + N_{\mathrm{FC}_{2048-512}} + N_{\mathrm{FC}_{512-256}} + N_{\mathrm{FC}_{256-128}} + N_{\mathrm{FC}_{128-64}}, \quad (14)$$

where N_{Conv} and $N_{\mathrm{FC}_{\mathrm{N_{in}-N_{out}}}}$ denote the parameter numbers of the first two convolutional layers and the FC layer with N_{in} input and N_{out} output neurons, respectively. The parameter number of the CsiNet+ at the UE using different encoders to realize different CRs can be calculated as:

$$N_{\mathrm{CsiNet+}} = 4 \times N_{\mathrm{Conv}} + N_{\mathrm{FC}_{2048-512}} + N_{\mathrm{FC}_{2048-256}} + N_{\mathrm{FC}_{2048-128}} + N_{\mathrm{FC}_{2048-64}}. \quad (15)$$

The calculation result shows that the parameter number of SM-CsiNet+ at the UE is 1,221,532, while that of the methods using different encoders to realize different CRs is 1,968,688. The proposed serial framework decreases the parameter number by approximately 38.0%, thereby greatly saving the storage space of the UE. Furthermore, the parameter number reduction at the UE will be larger if the more CRs need to be realized.

B. Parallel Multiple-Rate Compression Framework: PM-CsiNet+

Although the serial multiple-rate compression framework, SM-CsiNet+ in Fig. 5, greatly decreases the parameter number at the UE, it still occupies more storage space than the fourfold compression encoder. Here, we develop a parallel multiple-rate compression framework that is with the same parameter number as fourfold compression encoder.

The key idea of the proposed SM-CsiNet+ is to generate measurement vectors with a large CR from that with a small CR. By contrast, parallel compression framework first compresses the CSI matrix with a large CR and then generates measurement vectors with a small CR via connecting those with large CRs in turn, as shown in Fig. 6. This scheme is based on the fact that, the compression using a larger CR just reserves the essential information and drops the non-essential information compared with the compression using a small CR. Therefore, if the CR decreases, we can extract and feedback more information from the original CSI matrix to help the reconstruction at the BS. For instance, the size of 16-fold measure vectors is 128×1, and that of 32-fold measurement vectors is 64×1, which is half of 16-fold measurement vectors. Therefore, we can generate 16-fold measurement vectors via compressing CSI matrix by 32-fold twice and then connecting the measurement vectors together. Similarly, an eightfold measurement vectors can be generated from two 16-fold measurement vectors or four 32-fold measurement vectors.

From another perspective, parallel framework can be also regarded as generating measurement vectors with a large CR from that with a small CR. The encoder of this framework is the same as that of fourfold compression. Specially, we first compress CSI matrix by fourfold and then select the part of the measurement vectors to generate vectors with larger CRs. For example, when we need to compress CSI matrix by 32-fold, we just select the first 64 elements in the compressed vectors of fourfold compression. Similarly, when 16-fold compression is needed, only the first 128 elements in the compressed vectors are selected.

Different from the rate-adaptive compressive sensing neural networks in [43] using a three-stage training strategy, we still use an end-to-end approach to train parallel framework, similar to the serial framework. Meanwhile, the loss function here is also the same as that of the serial framework.

The parameter number of this framework at the UE is 1,049,500 and the same as that of fourfold encoder. This framework decreases parameter number by approximately 46.7%, thereby greatly saving the storage space at the UE.

V. Simulation Results and Discussions

In this section, we first describe the details of our experiments. Then, we evaluate the effects of the proposed two modifications on the reconstruction accuracy, compare with the existing state-of-the-art methods, and analyze the compression and reconstruction mechanism of CsiNet+. Next, we evaluate the accuracy of the proposed quantization framework. Finally, we analyze the two proposed multiple-rate compression frameworks.

A. Experimental Setting

1) Data Generation: We use the same dataset[2] generated in [21] to fairly compare CsiNet+ with CsiNet. Here, we test the proposed networks and frameworks under COST

[2] https://drive.google.com/drive/folders/1_lAMLk_5k1Z8zJQlTr5NRnSD6ACaNRtj?usp=sharing

TABLE II
NMSE (dB) PERFORMANCE OF RECONSTRUCTED CSI

	CR	CsiNet	CsiNet-M1	CsiNet-M2
Indoor	4	-17.36	-20.80	-24.80
	8	-12.70	-14.52	-15.23
	16	-8.65	-11.77	-12.21
	32	-6.24	-8.75	-8.65
Outdoor	4	-8.75	-10.14	-10.78
	8	-7.61	-8.11	-8.55
	16	-4.51	-4.99	-4.44
	32	-2.81	-1.87	-2.78

2100 MIMO channel model [31]. The dataset includes two representative types of CSI matrices, namely, the indoor and outdoor rural scenarios, which are at the carrier frequency of 5.3 GHz and 300 MHz, respectively. There are $N_c = 1024$ subcarriers and $N_t = 32$ uniform linear array (ULA) antennas at the BS, respectively. The complex CSI matrix in the angular-delay domain is first truncated to 32×32. The other parameters are the same as [31].

Here, we use no k-fold cross-validation because the dataset can be manually created without size limitation. The generated datasets are randomly divided into three parts, namely, training, validation, and testing sets, with 100,000, 30,000, and 20,000 samples, respectively. During the experiment, the training set is used to update the model parameters.

2) Hyperparameter Setting: Network models are initialized by a truncated normal initializer and optimized by the Adam optimizer [42]. CsiNet+ and two multiple-rate frameworks, SM-CsiNet+ and PM-CsiNet+, are all trained from the scratch. Meanwhile, bit-level CsiNet+ is fine-tuned from those without quantization. The batch sizes are all 200, while the epoch of the former three models and that of the latter model are 1000 and 200, respectively. Moreover, the initial learning rate of the former models is 0.001 and that of the latter is 0.0001. The learning rate will decay by half if the loss does not decrease in 20 epochs. Variables c_4, c_8, c_{16}, and c_{32} are 30, 6, 2, and 1, respectively.

3) Evaluation Metric: We utilize normalized MSE (NMSE) to measure CSI reconstruction accuracy, which is calculated as follows:

$$\mathrm{NMSE} = \mathrm{E}\{\|\mathbf{H} - \hat{\mathbf{H}}\|_2^2 / \|\mathbf{H}\|_2^2\}. \tag{16}$$

B. Performance of the Proposed CsiNet+

In this subsection, we first study the effects of two modifications on CSI reconstruction accuracy. Then, we compare the performance of CsiNet+ with the state-of-the-art model CsiNet. Finally, we visualize the parameters of the FC layer at the UE to explain the mechanism of the compression and reconstruction.

1) Effect of Modification 1: As mentioned in Section III-A.1, we replace small convolutional kernels with 7×7 filters, thereby making full use of CSI block sparsity in the angular-delay domain. Table II lists the performance of CsiNet-M1. Evidently, CsiNet-M1 outperforms CsiNet in both indoor and outdoor scenarios. However, the error reduction of outdoor scenario is much smaller than that of indoor scenario. Although CSI matrices in both indoor and outdoor are sparse in the angular-delay domain, the outdoor CSI matrix exhibits much smaller 'blank' area than the indoor CSI matrix. The large kernel mainly works in the large 'blank' area, thereby leading to the relatively smaller performance improvement in the outdoor scenario.

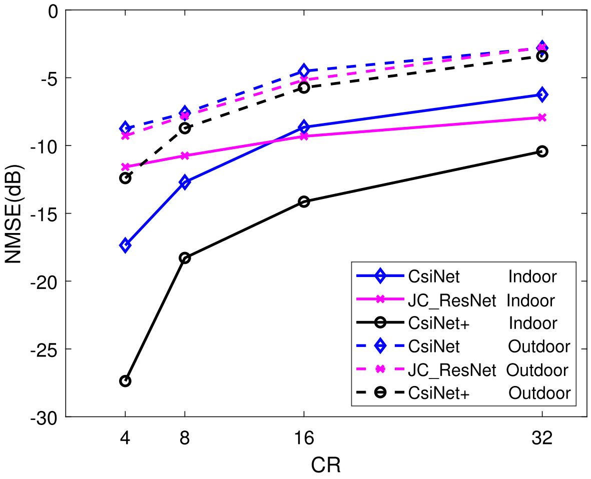


Fig. 7. $NMSE$ (dB) performance comparison between CsiNet+, CsiNet, and JC_ResNet. CsiNet+ shows noticeable accuracy advantages under all CRs.

TABLE III
PARAMETER NUMBER COMPARISON BETWEEN CSINET+, JC_RESNET AND CSINET

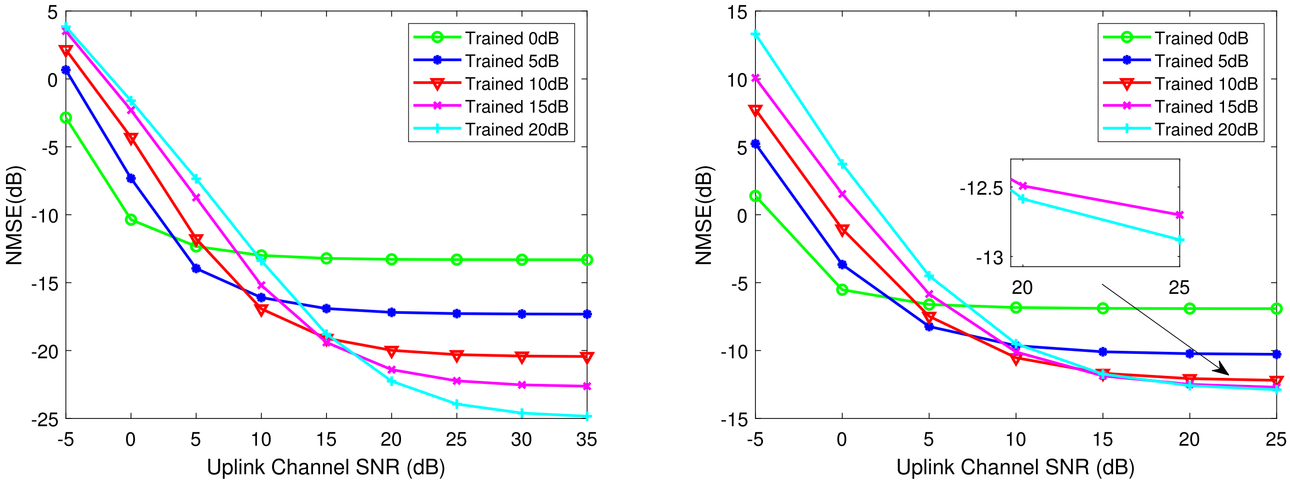
CR Number	4	8	16	32
CsiNet	2,103,904	1,055,072	530,656	268,448
JC_ResNet	2,113,696	1,064,864	540,448	278,240
CsiNet+	2,122,340	1,073,508	549,092	286,884

2) Effect of Modification 2: The motivation of Modification 2 is from the refinement theory. Table II shows the performance of CsiNet-M2. From the table, CsiNet-M2 outperforms not only CsiNet but also CsiNet-M1, which demonstrates that the proposed modification efficiently improves the refinement performance. Moreover, the performance improvement evidently decreases with the increase of CR since the information loss of high CR compression is unable to be offset even though the refinement has been strengthened.

3) Comprehensive Performance of CsiNet+: Here, we compare our proposed CsiNet+ with the state-of-the-art CsiNet and JC_ResNet in reconstruction accuracy and parameter number.

CsiNet+ shows noticeable accuracy advantages under all CRs, especially on small CRs, while the parameter number of the networks slightly increases, as shown in Fig. 7 and Table III.

We use fourfold compression as an example and compare the time complexity of CsiNet, JC_ResNet, and CsiNet+. The processing time of CsiNet+ is 0.12 ms when tested on the 1080Ti GPU, while CsiNet and JC_ResNet require 0.07 ms and 0.32 ms, respectively. Although the parameter numbers



(a) $\boldsymbol{CR}$ = 4 Indoor

(b) $\boldsymbol{CR}$ = 4 Outdoor

Fig. 8. $NMSE$ (dB) vs SNR for the two scenarios when $CR = 4$.

are similar to CsiNet, the processing time increases since the processing time is not only affected by parameter numbers but also dependent on the floating point operations (FLOPs) [44]. Due to the increase of convolutional kernel size in CsiNet+ from 3×3 in CsiNet to 7×7 and 5×5, the FLOPs of CsiNet+ are significantly increased. Although the processing time has greatly increased, it still meets the requirement of practical CSI reconstruction.

4) Network Robustness to Channel Noise: We here study the reconstruction robustness of the proposed CsiNet+ to the varying uplink channel conditions. Different from the above simulation that ignores the uplink channel noise, only an additive white Gaussian noise (AWGN) is added during the network training, which corrupts the feedback information of the measurement vectors. The NMSE performance for the fourfold compression versus the AWGN channel SNR is illustrated in Fig. 8 where each curve represents the performance of the CsiNet+ trained for a specific uplink channel SNR and deployed in different channel conditions. These simulation results provide an essential insight into the reconstruction performance of the DL-based CSI feedback methods if they are deplyed to the practical communication systems, where the channel conditions not only differ from the training channel but also are varying. We can observe that when the testing SNR are smaller than the training one, i.e., the testing channel condition is worse than the training one, the reconstruction performance will drastically fall.

On the other hand, when the testing SNR increases above the training one, the reconstruction performance is gradually improved intially and saturates if the testing SNR in above a certain value. It is worth noting that the accuracy saturation occurs beyond the training SNR, i.e., when the testing channel conditions are better than the training ones, the performance of the pre-trained CsiNet+ is better than that at the training period. Therefore, if we have known the statistical channel conditions, to achieve a better reconstruction performance, the training SNR should be slightly smaller than the statistical one.

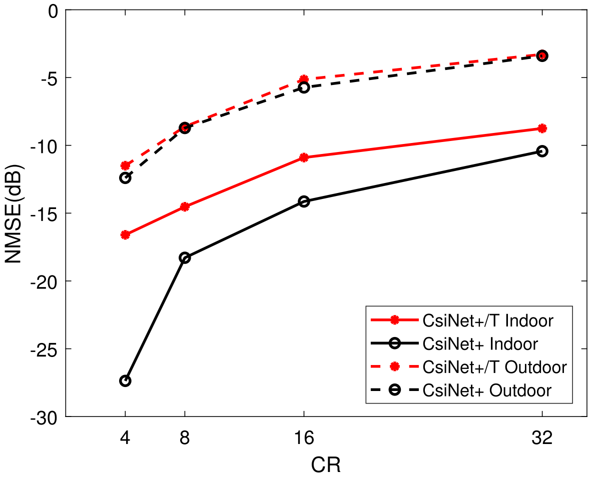
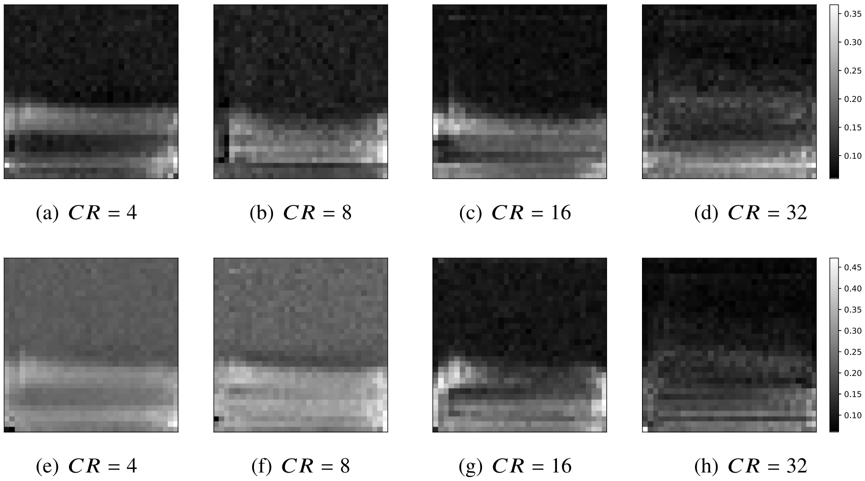


Fig. 9. $NMSE$ (dB) performance comparison between CsiNet+ and CsiNet+/T. CsiNet+/T denotes the method that uses a neural network to handle two scenarios simultaneously.

5) Neural Network Capacity: In practical applications, UE will not always stay at a scenario and it may move from indoor to outdoor. We investigate whether a neural network can handle two or more different scenarios simultaneously. To test CsiNet+'s reconstruction performance for two scenarios, we use both of indoor and outdoor CSI to train CsiNet+. From Fig. 9, the reconstruction accuracy of the indoor scenario CSI decreases a lot but there is little accuracy reduction of the outdoor scenario CSI reconstruction. During training, the loss of the outdoor CSI is much larger than that of the indoor CSI and the MSE loss function primarily smoothes the outdoor CSI reconstruction error, ignoring the indoor CSI [45].

Although the reconstruction accuracy is not as high as before, this method has proved the feasibility of DL-based CSI feedback operating at different scenarios. The proposed CsiNet+ is of high expressivity by widening convolutional layers and deepening the network [46], making it possible for a single model to reconstruct CSI of different scenarios. Similarly, [47] also finds a single denoising convolutional neural network model can yield excellent results for three



(a) $CR = 4$ (b) $CR = 8$ (c) $CR = 16$ (d) $CR = 32$

(e) $CR = 4$ (f) $CR = 8$ (g) $CR = 16$ (h) $CR = 32$

Fig. 10. FC layer parameter heatmaps of the indoor scenario for different CRs. The larger the values are, the more attention the FC layer gives to the corresponding area. Each FC layer exhibits two heatmaps, namely, the upper and bottom rows, because the output of the convolutional layers at the encoder is $32 \times 32 \times 2$.

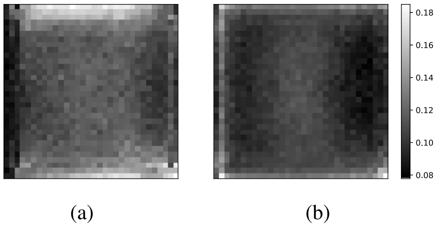


Fig. 11. FC layer parameter heatmaps of outdoor scenario for fourfold compression.

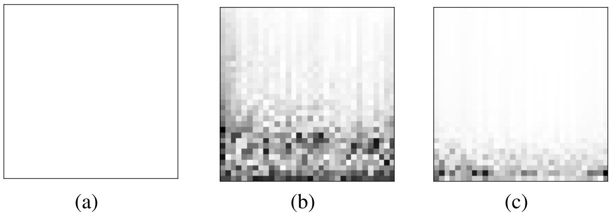


Fig. 12. CsiNet+'s ability to learn statistical CSI. a: all-zero input; b: output of CsiNet+; c: mean of training dataset. Although the output and mean CSI are not visually exactly the same, some remarkable similarities can be observed.

general image denoising tasks, i.e., super-resolution, blind Gaussian denoising, and JPEG deblocking, because of its high capacity.

6) Compression and Reconstruction Mechanism: We will map the FC layer parameters to the output of the former layer to observe the compression process. We use 32-fold compression as an example. First, we reshape the FC layer parameter $\Theta_{\mathrm{fc}} \in \mathbb{R}^{2048\times 64}$ to $2 \times 32 \times 32 \times 64$. Then, we calculate the average of the absolute value of Θ_{fc} over Axis 4, thereby obtaining a $2 \times 32 \times 32$ matrix Θ'_{fc}. The normalized absolute FC layer parameters Θ'_{fc} are regarded as the values of heatmaps. The larger the values are, the more attention the FC layer gives to the corresponding area.

From Fig. 10, the areas of great interest to the FC layers are the bottoms of the feature maps. Meanwhile, the remaining area is full of near-zero values and contains little information. Therefore, the DL-based CS feedback methods utilize FC layers to determine the non-zero areas and then generate measurement vectors by mainly exploiting the information of these areas. With the increase of CRs, the areas of interest become smaller, and the upper areas are given little attention to, as shown in Fig. 10. Specifically, high compression is at the expense of the loss of incidental information. Evidently, the reconstruction accuracy drops with the increase of CR.

In both CsiNet and CsiNet+, the reconstruction accuracy of the outdoor scenario is much lower than that of the indoor scenario. We can compare the heatmaps of indoor and outdoor scenarios in Fig. 10 and 11. From Fig. 11, CsiNet+ is unable to efficiently extract the information of the outdoor CSI since the outdoor CSI exhibits much little sparsity than the indoor CSI.

The CS-based algorithms only use signal sparsity as prior information and neglect other signal characteristics. DL-based methods can extract all useful features from data automatedly. The proposed CsiNet+ learns not only the sparse structure of CSI but also the statistical CSI. We feed a $32 \times 32 \times 2$ all-zero matrix, as in Fig. 12(a), into the trained CsiNet+. Fig. 12(b) and 12(c) are the corresponding output of CsiNet+ and the mean of CSI matrices at the indoor scenario, respectively. Although the output and mean CSI are not visually exactly the same, some remarkable similarities can be observed. The upper parts contain large 'blank' areas while the bottoms contain the most information. Therefore, the DL-based CSI feedback method CsiNet+ can automatically learn statistical CSI.

The DL-based methods make use of not only limited feedback but also the statistical CSI learned automatedly. Therefore, even when the feedback information is little or noisy, the DL-based methods can reconstruct high-quality CSI compared with traditional CS-based methods.

7) Computational Complexity: In this section, we provide a brief discussion of the computational complexity of the proposed CsiNet+. The most computationally costly parts in the neural networks are the convolutional and FC layers, both of which involve the multiplications and additions. For the

TABLE IV

NMSE (dB) PERFORMANCE OF THE PROPOSED QUANTIZATION METHODS FOR DIFFERENT SCENARIOS. KEY FOR METHOD ABBREVIATIONS: UQ: UNIFORM QUANTIZATION; NUQ: NON-UNIFORM QUANTIZATION; NUQ+O: NON-UNIFORM QUANTIZATION WITH THE OFFSET NETWORK

	Indoor															
CR	4				8				16				32			
B	3	4	5	6	3	4	5	6	3	4	5	6	3	4	5	6
CsiNet+	-27.37				-18.29				-14.14				-10.43			
UQ	-12.97	-16.78	-20.35	-23.53	-11.33	-14.39	-16.22	-17.79	-9.64	-12.11	-13.43	-13.93	-7.12	-8.93	-9.96	-10.29
NUQ	-14.82	-18.55	-22.00	-24.95	-12.45	-15.39	-17.20	-18.02	-10.35	-12.56	**-13.64**	**-14.04**	-7.82	-9.36	-10.09	**-10.35**
NUQ+O	**-15.27**	**-19.06**	**-22.13**	**-24.97**	**-12.81**	**-15.65**	**-17.23**	**-18.03**	**-10.48**	**-12.58**	**-13.64**	-14.02	**-7.91**	**-9.37**	**-10.10**	**-10.35**
	Outdoor															
CR	4				8				16				32			
B	3	4	5	6	3	4	5	6	3	4	5	6	3	4	5	6
CsiNet+	-12.40				-8.72				-5.73				-3.40			
UQ	-7.97	-10.37	-11.66	-12.18	-6.51	-7.87	-8.43	-8.65	-4.05	-4.93	-5.41	-5.61	**-2.54**	**-2.99**	**-3.35**	-3.34
NUQ	-9.45	-11.17	-11.96	-12.28	-6.53	-7.90	-8.44	-8.64	-4.65	-5.33	-5.59	-5.68	-2.45	-2.92	-3.27	-3.37
NUQ+O	**-9.96**	**-11.54**	**-12.42**	**-12.83**	**-6.67**	**-7.94**	**-8.51**	**-8.81**	**-4.94**	**-5.49**	**-5.70**	**-5.75**	-2.53	-2.93	-3.18	**-3.79**

TABLE V

NMSE(dB) PERFORMANCE WITH FIXED BITSTREAM LENGTH

Scenario	Indoor						Outdoor					
Total bits	1536		738		384		1536		738		384	
CR/B	4/3	8/6	8/3	16/6	16/3	32/6	4/3	8/6	8/3	16/6	16/3	32/6
NMSE(dB)	-15.27	-18.03	-12.81	-14.02	-10.48	-10.35	-9.96	-8.81	-6.67	-5.75	-4.94	-3.79

convolution operation, the computational cost is $K \times K \times D \times N \times W \times H$, where K, D, N, and $W \times H$ denote the filter size, the number of the input feature channels, the number of the convolutional filters, and the size of the input feature [44], [48]. The computational cost of the FC layer is $N_{\text{in}} \times N_{\text{out}}$, where N_{in} and N_{out} are the numbers of the input and output neurons, respectively. N_{in} and N_{out} are both determineded by the size of the input CSI. Therefore, the computational complexity of the proposed CsiNet+ for both of the compression and reconstruction is $\mathcal{O}(N'_c \times N_t)$, which indicates that the computational complexity of the proposed DL-based CSI feedback algorithm is linear in the size of the CSI matrix, as only the feature map size and FC layer neurons depend on the antenna and subcarrier numbers while all other factors are constant and independent.

For the CS-based traditional CSI feedback approach, the compressed measurements are achieved by a simple matrix-vector multiplication, whose computational complexity is also $\mathcal{O}(N'_c \times N_t)$. However, the computational complexity of the CS-based reconstruction algorithms is much higher. We here use fourfold compression as an example to compare the speed of the entire feedback. As mentioned in Section V-B.3, due to the acceleration of the high-performance GPU, the entire compression and reconstruction process of the CsiNet+ only needs 0.12 ms while the LASSO [49], BM3D-AMP [50], and TVAL3 [51] need about 0.32 s, 0.55 s, and 0.24 s, respectively. Obviously, the reconstruction speed of the DL-based algorithm is much higher than that of the CS-based traditional algorithm.

Although the inference speed of the proposed method is high, the training process is much more time-consuming compared with the traditional methods. Fortunately, the neural networks can be first off-line trained using powerful GPUs and then online fine-tuned, which can be accelerated by techniques, e.g., meta-learning [52].

C. Quantization Evaluation

Table IV shows the NMSE performance of our proposed bit-level CsiNet+. From the table, non-uniform quantization outperforms uniform quantization by a margain as we imagine because the measurement vectors are weak signals. Non-uniform quantization with the offset network achieves the best quantization performance. From the perspective of refinement theory, the stacked FC layers in the offset network refine the output of the dequantizer, thereby minimizing quantization distortion.

As we can imagine, the reconstruction is becoming more and more accurate with the increase of quantization bits. Moreover, CsiNet+Q6 even exhibits a similar performance as the original CsiNet+ without quantization. Second, the indoor scenario is more sensitive to the change of quantization bits compared with the outdoor scenario. Similarly, low compression is much more sensitive than high compression. This phenomenon may be attributed to that high-accuracy reconstruction is based on the full use of measurement vectors. Therefore, little distortion of measurement vectors will lead to a large decrease in reconstruction accuracy. Besides, CsiNet+Q6 at the outdoor scenario even outperforms that with quantization because the encoder cannot efficiently extract the information of outdoor CSI and quantization drops some redundant information.

In practical scenario, CR and quantization bits B together determine the overhead of CSI feedback. For example, if the feedback bitstream contains 1536 bits, we can have two choices, $CR = 4$, $B = 3$, or $CR = 8$, $B = 6$. The NMSE of the former at the indoor scenario is –15.27 dB, while that of the latter is –18.03 dB, as shown in Table V. From Table V,

TABLE VI

NMSE (dB) PERFORMANCE OF THE PROPOSED SM-CSINET+ AND PM-CSINET+

	CR	CsiNet	CsiNet+	SM-CsiNet+	PM-CsiNet+
Indoor	4	-17.36	-27.37	**-27.90**	-27.60
	8	-12.70	-18.29	**-18.49**	-17.70
	16	-8.65	**-14.14**	-13.45	-12.25
	32	-6.24	**-10.43**	-9.89	-8.24
Outdoor	4	-8.75	**-12.40**	-11.91	-12.02
	8	-7.61	**-8.72**	-8.25	-8.10
	16	-4.51	**-5.73**	-5.31	-5.07
	32	-2.81	**-3.40**	-3.22	-3.00

TABLE VII

NMSE (dB) PERFORMANCE OF PM-CSINET+ WITH DIFFERENT c_{16} AT OUTDOOR SCENARIO

c_{16} \ CR	4	8	16	32
2	-12.02	-8.10	-5.07	-3.00
20	-11.78	-7.90	-5.39	-2.98
200	-11.84	-7.51	-5.56	-2.73

at the indoor scenario, with a fixed bitstream length, compared with decreasing quantization bits, the increase of CR has less bad effect on reconstruction accuracy. In other words, the CSI reconstruction accuracy is more senstive to quantization distortion. At the outdoor scenario, to the opposite, the CSI reconstruction accuracy is more senstive to compression errors. This gives a guideline to practical deployment that, even if the length of feedback bitstream is fixed, suitable CR and quantization bits must be selected to achieve optimal performance at the practical scenario, which also shows the necessity of the proposed quantization strategy and multiple-rate frameworks.

D. Performance of Multiple-Rate Compression Framework

Table VI shows the NMSE (dB) performance of the multiple-rate compression frameworks. From the table, multiple-rate compression frameworks are not lossless compared with the direct compression network. The serial framework, SM-CsiNet+, has similar reconstruction accuracy to the direct one, as shown in Table VI. When the CR is 4 or 8 for the indoor scenario, the serial compression network performs better than the direct one. In other cases, its accuracy is slightly worse than the direct one by approximately 0.5 dB. The parallel framework, PM-CsiNet+, is approximately 2 dB worse than the direct one.

From Table VI, SM-CsiNet+ outperforms PM-CsiNet+ by a margin, because the parameter number of PM-CsiNet+ is approximately 85.9% of that of SM-CsiNet+ and the stacked FC layers increase the depth of SM-CsiNet+. In general, the DL-based methods exhibit enhanced fitting ability with the increase of parameter number and layer depth in neural networks [15].

Further more, SM-CsiNet+ even performs better than CsiNet+ at the indoor scenario when the CR is four or eight, which can be explained by the regularization theory. We use fourfold compression as an example. In the testing period, SM-CsiNet+ for fourfold compression is the same as CsiNet+ in not only the network architecture but also the parameter number. Therefore, the performance improvement of SM-CsiNet+ results from the training period rather than network architecture or parameter number. If only focusing on fourfold compression during training, we can regard the subsequent compression networks as an additional regularization term [53]. If the fourfold compression measurements are inefficient, other high-compression measurements generated from the fourfold compression measurements cannot contain enough useful information, thereby leading to the poor reconstruction accuracy of other high compression. Specifically, the subsequent compression networks force the former compression networks to extract useful information as much as possible, which is the reason behind the excellent performance of multiple-rate frameworks at low CRs.

As mentioned in (13) and Section V-A.2, c_4, c_8, c_{16}, and c_{32} are used to balance the magnitude of loss terms into similar scales and set as 30, 6, 2, and 1, respectively. During the experiments, the performance changes of different compression sub-networks differ if we change these parameters. For example, we change c_{16} into 20 and 200, respectively. Table VII shows that the reconstruction accuracy of 16-fold compression is improved, with the increase of c_{16}, because the increase of c_{16} forces the training process to focus more on the 16-fold compression sub-network at the expense of other sub-networks' performance.

This makes the proposed multiple-rate frameworks more suitable for practical applications. We can increase the weight of the preferred sub-network in (13) because the UE commonly possesses the preferred CR, which is decided by its communication environment.

VI. CONCLUSION

In this paper, we have proposed a multiple-rate bit-level compressive sensing DL-based framework for CSI feedback problem described in Section I. We focus on the CSI reconstruction accuracy and feasibility at the UE.

We first introduce a network architecture, CsiNet+, modified from CsiNet. In contrast with other DL-based methods, which regard neural networks as a black box and only care about the reconstruction accuracy, we explain the motivation of the modifications and the compression mechanism via parameter visualization of the FC layer and determine CsiNet+'s ability to learn statistical CSI automatedly.

We propose a novel framework and training strategy for quantization problem. Other quantization strategies need to train different networks for distinct quantization bits. By contrast, we do not need any parameter update at the UE, and just fine-tune the parameters in the decoder at the BS for different quantization bits. Besides, we introduce an offset network to minimize quantization distortion.

Lastly, we propose two frameworks SM-CsiNet+ and PM-CsiNet+ to overcome the problem that DL-based methods need to store different parameter sets for distinct CRs, thereby leading to great storage space waste. The proposed two frameworks reduce parameter number by 38.0% and 46.7% compared with the existing DL-based method that train and store a different neural network for a different CR. SM-CsiNet+

outperforms PM-CsiNet+ at the expense of increasing the parameter number. SM-CsiNet+ even performs better than CsiNet+ at the low CR.

Although the DL-based CSI feedback methods have shown promising results, some extensive challenges are worth exploring further in the future.

First, when the UE moves quickly, the channel between the BS and the UE will change dynamically. Though the LSTM architecture has been introduced to extract the CSI temporal correlations for the low-speed scenario [23], [25], how to adapt to the variable channel conditions for the high-speed scenario is still a problem worth studying. Then, although the DL-based multiple-rate CsiNet+ saves storage space, it still needs to store large numbers of network parameters and requires very high computation compared with the conventional algorithms. Therefore, the efficient and lightweight network architecture should be explored and then the neural networks should be compressed using pruning, quantization, etc [54]. Meanwhile, more channel model should be taken into consideration since the most DL-based CSI feedback methods are only tested on the COST 2100 channel model. Robustness of the machine learning-based approaches also need to be carefully addressed. Besides, all experiments are based on simulation and a public real-world dataset is greatly needed to evaluate the performance of the DL-based CSI feedback methods.

REFERENCES

[1] T. Chen, J. Guo, S. Jin, C.-K. Wen, and G. Y. Li, "A novel quantization method for deep learning-based massive MIMO CSI feedback," in *Proc. GlobalSIP*, Ottawa, Canada, 2019.

[2] F. Boccardi, R. W. Heath, Jr., A. Lozano, T. L. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 74–80, Feb. 2014.

[3] T. L. Marzetta, "Massive MIMO: An introduction," *Bell Labs Tech. J.*, vol. 20, pp. 11–22, Mar. 2015.

[4] V. W. Wong, R. Schober, D. W. K. Ng, and L.-C. Wang, *Key Technologies for 5G Wireless Systems*. Cambridge, U.K.: Cambridge Univ. Press, 2017.

[5] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 186–195, Feb. 2014.

[6] L. Lu, G. Y. Li, A. L. Swindlehurst, A. Ashikhmin, and R. Zhang, "An overview of massive MIMO: Benefits and challenges," *IEEE J. Sel. Top. Signal Process.*, vol. 8, no. 5, pp. 742–758, Oct. 2014.

[7] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta, "Energy and spectral efficiency of very large multiuser MIMO systems," *IEEE Trans. Commun.*, vol. 61, no. 4, pp. 1436–1449, Apr. 2013.

[8] H. Xie, F. Gao, and S. Jin, "An overview of low-rank channel estimation for massive MIMO systems," *IEEE Access*, vol. 4, pp. 7313–7321, 2016.

[9] M. S. Sim, J. Park, C.-B. Chae, and R. W. Heath, Jr., "Compressed channel feedback for correlated massive MIMO systems," *IEEE/KICS J. Commun. Netw.*, vol. 18, no. 1, pp. 95–104, Feb. 2016.

[10] D. J. Love, R. W. Heath, Jr., V. K. Lau, D. Gesbert, B. D. Rao, and M. Andrews, "An overview of limited feedback in wireless communication systems," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 8, pp. 1341–1365, Oct. 2008.

[11] Z. Qin, J. Fan, Y. Liu, Y. Gao, and G. Y. Li, "Sparse representation for wireless communications: A compressive sensing approach," *IEEE Signal Process. Mag.*, vol. 35, no. 3, pp. 40–58, May 2018.

[12] Z. Gao, L. Dai, S. Han, C.-L. I, Z. Wang, and L. Hanzo, "Compressive sensing techniques for next-generation wireless communications," *IEEE Wireless Commun.*, vol. 25, no. 3, pp. 144–153, Jun. 2018.

[13] P.-H. Kuo, H. T. Kung, and P.-A. Ting, "Compressive sensing based channel feedback protocols for spatially-correlated massive antenna arrays," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2012, pp. 492–497.

[14] X. Rao and V. K. N. Lau, "Distributed compressive CSIT estimation and feedback for FDD multi-user massive MIMO systems," *IEEE Trans. Signal Process.*, vol. 62, no. 12, pp. 3261–3271, Jun. 2014.

[15] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

[16] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing," *IEEE Comput. Intell. Mag.*, vol. 13, no. 3, pp. 55–75, Aug. 2018.

[17] Z. Qin, H. Ye, G. Y. Li, and B.-H.-F. Juang, "Deep learning in physical layer communications," *IEEE Wireless Commun.*, vol. 26, no. 2, pp. 93–99, Apr. 2019.

[18] K. Kulkarni, S. Lohit, P. Turaga, R. Kerviche, and A. Ashok, "ReconNet: Non-iterative reconstruction of images from compressively sensed measurements," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 449–458.

[19] H. Yao, F. Dai, S. Zhang, Y. Zhang, Q. Tian, and C. Xu, "DR2-Net: Deep residual reconstruction network for image compressive sensing," *Neurocomputing*, vol. 359, pp. 483–493, Sep. 2019.

[20] W. Cui, F. Jiang, X. Gao, W. Tao, and D. Zhao, "Deep neural network based sparse measurement matrix for image compressed sensing," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 3883–3887.

[21] C.-K. Wen, W.-T. Shih, and S. Jin, "Deep learning for massive MIMO CSI feedback," *IEEE Wireless Commun. Lett.*, vol. 7, no. 5, pp. 748–751, Oct. 2018.

[22] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proc. 25th Int. Conf. Mach. Learn. (ICML)*, Jul. 2008, pp. 1096–1103.

[23] 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 Commun. Lett.*, vol. 8, no. 2, pp. 416–419, Apr. 2019.

[24] K. Greff, R. K. Srivastava, J. Koutnik, B. R. Steunebrink, and J. Schmidhuber, "LSTM: A search space Odyssey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 10, pp. 2222–2232, Oct. 2017.

[25] C. Lu, W. Xu, H. Shen, J. Zhu, and K. Wang, "MIMO channel information feedback using deep recurrent network," *IEEE Commun. Lett.*, vol. 23, no. 1, pp. 188–191, Jan. 2019.

[26] Z. Liu, L. Zhang, and Z. Ding, "Exploiting bi-directional channel reciprocity in deep learning for low rate massive MIMO CSI feedback," *IEEE Wireless Commun. Lett.*, vol. 8, no. 3, pp. 889–892, Jun. 2019.

[27] Y. Jang, G. Kong, M. Jung, S. Choi, and I.-M. Kim, "Deep autoencoder based CSI feedback with feedback errors and feedback delay in FDD massive MIMO systems," *IEEE Wireless Commun. Lett.*, vol. 8, no. 3, pp. 833–836, Jun. 2019.

[28] P. Wu, Z. Liu, and J. Cheng, "Compressed CSI feedback with learned measurement matrix for mmWave massive MIMO," Mar. 2019, *arXiv:1903.02127*. [Online]. Available: https://arxiv.org/abs/1903.02127

[29] C. Lu, W. Xu, S. Jin, and K. Wang, "Bit-level optimized neural network for multi-antenna channel quantization," Sep. 2019, *arXiv:1909.10730*. [Online]. Available: https://arxiv.org/abs/1909.10730

[30] C.-K. Wen, S. Jin, K.-K. Wong, J.-C. Chen, and P. Ting, "Channel estimation for massive MIMO using Gaussian-mixture Bayesian learning," *IEEE Trans. Wireless Commun.*, vol. 14, no. 3, pp. 1356–1368, Mar. 2015.

[31] L. Liu *et al.*, "The COST 2100 MIMO channel model," *IEEE Wireless Commun.*, vol. 19, no. 6, pp. 92–99, Dec. 2012.

[32] B. Wang, F. Gao, S. Jin, H. Lin, and G. Y. Li, "Spatial- and frequency-wideband effects in millimeter-wave massive MIMO systems," *IEEE Trans. Signal Process.*, vol. 66, no. 13, pp. 3393–3406, Jul. 2018.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[34] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," Feb. 2015, *arXiv:1502.03167*. [Online]. Available: https://arxiv.org/abs/1502.03167

[35] Y. C. Eldar and H. Bolcskei, "Block-sparsity: Coherence and efficient recovery," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Apr. 2009, pp. 2885–2888.

[36] Y. Bai, B. Ai, and W. Chen, "Deep learning based fast multiuser detection for massive machine-type communication," Jul. 2018, *arXiv:1807.00967*. [Online]. Available: https://arxiv.org/abs/1807.00967

[37] Q. Wang, H. Fan, Y. Cong, and Y. Tang, "Large receptive field convolutional neural network for image super-resolution," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 958–962.

[38] X. Xie, C. Wang, J. Du, and G. Shi, "Full image recover for block-based compressive sensing," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2018, pp. 1–6.

[39] W. Cui, F. Jiang, X. Gao, S. Zhang, and D. Zhao, "An efficient deep quantized compressed sensing coding framework of natural images," in *Proc. ACM Multimedia Conf. Multimedia Conf. (MM)*, Oct. 2018, pp. 1777–1785.

[40] B. Sun, H. Feng, K. Chen, and X. Zhu, "A deep learning framework of quantized compressed sensing for wireless neural recording," *IEEE Access*, vol. 4, pp. 5169–5178, 2016.

[41] J. Cai and L. Zhang, "Deep image compression with iterative non-uniform quantization," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 451–455.

[42] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," Dec. 2014, *arXiv:1412.6980*. [Online]. Available: https://arxiv.org/abs/1412.6980

[43] S. Lohit, R. Singh, K. Kulkarni, and P. Turaga, "Rate-adaptive neural networks for spatial multiplexers," Sep. 2018, *arXiv:1809.02850*. [Online]. Available: https://arxiv.org/abs/1809.02850

[44] P. Molchanov, S. Tyree, T. Karras, T. Aila, and J. Kautz, "Pruning convolutional neural networks for resource efficient inference," in *Proc. ICLR*, Apr. 2017, pp. 1–17.

[45] J. Guo, H. Du, J. Zhu, T. Yan, and B. Qiu, "Relative location prediction in CT scan images using convolutional neural networks," *Comput. Methods Programs Biomed.*, vol. 160, pp. 43–49, Jul. 2018.

[46] M. Raghu, B. Poole, J. Kleinberg, S. Ganguli, and J. S. Dickstein, "On the expressive power of deep neural networks," in *Proc. ICML*, 2017, pp. 2847–2854.

[47] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian Denoiser: Residual learning of deep CNN for image denoising," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3142–3155, Jul. 2017.

[48] E. Bourtsoulatze, D. Burth Kurka, and D. Gunduz, "Deep joint source-channel coding for wireless image transmission," *IEEE Trans. Cogn. Commun. Netw.*, vol. 5, no. 3, pp. 567–579, Sep. 2019.

[49] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Commun. Pure Appl. Math.*, vol. 57, no. 11, pp. 1413–1457, Nov. 2004.

[50] 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, Sep. 2016.

[51] C. Li, W. Yin, H. Jiang, and Y. Zhang, "An efficient augmented Lagrangian method with applications to total variation minimization," *Comput. Optim. Appl.*, vol. 56, no. 3, pp. 507–530, Dec. 2013.

[52] J. Vanschoren, "Meta-learning," in *Automated Machine Learning: Methods, Systems, Challenges*, F. Hutter, L. Kotthoff, and J. Vanschoren, Eds. Cham, Switzerland: Springer, 2019, pp. 35–61, doi: 10.1007/978-3-030-05318-5_2.

[53] G. Yang *et al.*, "DAGAN: Deep de-aliasing generative adversarial networks for fast compressed sensing MRI reconstruction," *IEEE Trans. Med. Imag.*, vol. 37, no. 6, pp. 1310–1321, Jun. 2018.

[54] J. Guo, J. Wang, C.-K. Wen, S. Jin, and G. Y. Li, "Compression and acceleration of neural networks for communications," Jul. 2019, *arXiv:1907.13269*. [Online]. Available: https://arxiv.org/abs/1907.13269

Jiajia Guo received the B.S. degree from the Nanjing University of Science and Technology, Nanjing, China, in 2016, and the M.S. degree from the University of Science and Technology of China, Hefei, China, in 2019. He is currently pursuing the Ph.D. degree in information and communications engineering with Southeast University, China. His current research interests include deep learning for wireless communications and massive MIMO.

Chao-Kai Wen (Member, IEEE) received the Ph.D. degree from the Institute of Communications Engineering, National Tsing Hua University, Taiwan, in 2004. He was with the Industrial Technology Research Institute, Hsinchu, Taiwan, and MediaTek Inc., Hsinchu, from 2004 to 2009. Since 2009, he has been with National Sun Yat-sen University, Taiwan, where he is a Professor with the Institute of Communications Engineering. His research interest centers on the optimization in wireless multimedia networks.

Shi Jin (Senior Member, IEEE) received the B.S. degree in communications engineering from the Guilin University of Electronic Technology, Guilin, China, in 1996, the M.S. degree from the Nanjing University of Posts and Telecommunications, Nanjing, China, in 2003, and the Ph.D. degree in information and communications engineering from Southeast University, Nanjing, in 2007. From June 2007 to October 2009, he was a Research Fellow with the Adastral Park Research Campus, University College London, London, U.K. He is currently with the faculty of the National Mobile Communications Research Laboratory, Southeast University. His research interests include space time wireless communications, random matrix theory, and information theory. He and his coauthors have been awarded the 2011 IEEE Communications Society Stephen O. Rice Prize Paper Award in the field of communication theory and a 2010 Young Author Best Paper Award by the IEEE Signal Processing Society. He serves as an Associate Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, the IEEE COMMUNICATIONS LETTERS, and *IET Communications*.

Geoffrey Ye Li (Fellow, IEEE) is currently a Professor with the Georgia Institute of Technology, Atlanta, GA, USA. Before moving to Georgia Tech, he was with AT&T Labs—Research, Red Bank, NJ, USA, as a senior and then a Principal Technical Staff Member, from 1996 to 2000, and a Post-Doctoral Research Associate with The University of Maryland, College Park, Maryland, from 1994 to 1996. His general research interests include statistical signal processing and machine learning for wireless communications. In these areas, he has published over 500 journals and conference papers in addition to over 40 granted patents. His publications have been cited around 40,000 times, and he has been recognized as the World's Most Influential Scientific Mind, also known as a Highly Cited Researcher, by Thomson Reuters almost every year.

He was awarded as the IEEE Fellow for his contributions to signal processing for wireless communications in 2005. He received several prestigious awards from the IEEE Signal Processing Society (Donald G. Fink Overview Paper Award in 2017), the IEEE Vehicular Technology Society (James Evans Avant Garde Award in 2013 and Jack Neubauer Memorial Award in 2014), and the IEEE Communications Society (Stephen O. Rice Prize Paper Award in 2013), Award for Advances in Communication in 2017, and Edwin Howard Armstrong Achievement Award in 2019. He also received the 2015 Distinguished Faculty Achievement Award from the School of Electrical and Computer Engineering, Georgia Tech. He has organized and chaired many international conferences, including a Technical Program Vice-Chair of IEEE ICC'03, Technical Program Co-Chair of IEEE SPAWC'11, General Chair of IEEE GlobalSIP'14, Technical Program Co-Chair of IEEE VTC'16 (Spring), and a General Co-Chair of IEEE VTC'19 (Fall). He has been involved in editorial activities for over 20 technical journals, including founding the Editor-in-Chief of IEEE 5G Tech Focus.