Deep CSI Compression and Feedback for Massive MIMO: A Survey

Lu Zhaohua[1,2], Yi Chenyang[3], Wu Jie[3], Shao Bo[3], Xu Wei[3,4]

(1. State Key Laboratory of Mobile Network and Mobile Multimedia Technology, Shenzhen 518055, China;
2. ZTE Corporation, Shenzhen 518057, China;
3. National Mobile Communications Research Laboratory, Southeast University, Nanjing 211189, China;
4. Purple Mountain Laboratories, Nanjing 211111, China)

DOI: 10.12142/ZTECOM.202601003

https://kns.cnki.net/kcms/detail/34.1294.TN.20260304.1138.002.html, published online March 4, 2026

Manuscript received: 2024-09-20

Abstract: To achieve the potential performance gain of massive multiple-input multiple-output (MIMO) systems, base stations (BS) require downlink channel state information (CSI) fed back by users to execute beamforming design, especially in the frequency division duplex (FDD) systems. However, due to the enormous number of antennas in massive MIMO systems, the feedback overhead of downlink CSI acquisition is extremely large. To address this issue, deep learning (DL) techniques have been introduced to develop high-accuracy feedback strategies under limited backhaul constraints. In this paper, we provide an overview of DL-based CSI compression and feedback approaches in massive MIMO systems. Specifically, we introduce the conventional CSI compression and feedback schemes and the existing problems. Besides, we elaborate on various DL techniques employed in CSI compression from the perspective of network architecture and analyze the advantages of different techniques. We also enumerate the applications of DL-based methods for solving practical challenges in CSI compression and feedback. In addition, we brief the remaining issues in deep CSI compression and indicate potential directions in future wireless networks.

Keywords: deep learning; MIMO; CSI compression; limited feedback; FDD system

Citation (Format 1): Lu Z H, Yi C Y, Wu J, et al. Deep CSI compression and feedback for massive MIMO: a survey [J]. ZTE Communications, 2026, 24(1): 4 - 15. DOI: 10.12142/ZTECOM.202601003
Citation (Format 2): Z. H. Lu, C. Y. Yi, J. Wu, et al., "Deep CSI compression and feedback for massive MIMO: a survey," *ZTE Communications*, vol. 24, no. 1, pp. 4 - 15, Mar. 2026. doi: 10.12142/ZTECOM.202601003.

1 Introduction

With the increasing demand for data traffic and massive connectivity, advanced technologies such as multiple-input multiple-output (MIMO), non-orthogonal multiple access (NOMA), and ultra-dense networks (UDN) have been proposed to meet the high-throughput, high-reliability, and low-latency challenges in 5G and beyond 5G (B5G) wireless communication networks[1-2]. Massive MIMO is considered a key technology in B5G networks since it improves the spectral and energy efficiency of wireless communication networks by simultaneously serving a set of users with multiple antennas at the base station (BS)[3]. To achieve the potential multiplexing gain in massive MIMO systems, the BS requires downlink channel state information (CSI) for transmission design, such as beamforming and power allocation, to enhance the desired signal and eliminate multi-user interference.

Since the performance of massive MIMO systems relies directly on the accuracy of the CSI obtained at the BS, it is significant to develop practical CSI acquisition methods under various scenarios[4]. In time-division duplexing (TDD) systems, downlink CSI can be obtained at the BS from the uplink CSI by utilizing channel reciprocity. In frequency-division duplexing (FDD) systems, uplink and downlink operate in different frequency bands, and the channel reciprocity no longer holds. Consequently, downlink CSI in FDD systems needs to be estimated at the users and then fed back to the BS through a feedback link. However, the feedback overhead in massive MIMO systems is prohibitively large due to the fact that the dimension of CSI increases with the network scale. Thus, there is an urgent need for effective CSI compression and feedback methods to achieve acceptable accuracy under the constraint of limited backhaul.

In recent decades, deep learning (DL) has attracted growing attention in wireless communications due to its exceptional ability in feature extraction and function approximation, which

The corresponding author is Xu Wei.
This work was supported by ZTE Industry-University-Institute Cooperation Funds under Grant No. IA20240319003 and the NSFC under Grant No. 62571112..

makes it a potential methodology to address the intractable nonlinear challenges in signal processing[5]. The DL techniques are also introduced to CSI compression and feedback design in massive MIMO systems to overcome the drawbacks of conventional CSI feedback approaches such as substantial computational complexity and dependence on channel model assumptions. Given the superior performance of DL in image compression, DL-based CSI acquisition methods have the potential to learn the compression of the CSI matrix in a data-driven manner, thereby improving reconstruction accuracy and reducing feedback overhead. Furthermore, based on the domain knowledge of massive MIMO channels, model-driven DL methods can be applied to solve practical problems in CSI compression and feedback, such as network lightweighting and generalization enhancement.

In the rest of this paper, we first review the conventional CSI compression and feedback approaches in Section 2. The DL techniques deployed in CSI compression frameworks are summarized and elaborated in Section 3. Section 4 introduces the applications of DL techniques for solving practical challenges in CSI compression and feedback. The critical challenges and potential directions for CSI compression in future wireless networks are discussed in Section 5. Finally, Section 6 concludes this survey.

2 CSI Compression and Feedback for Massive MIMO

To reduce the tremendous feedback overhead in massive MIMO networks, researchers have proposed CSI compression methods utilizing channel correlations and environmental knowledge. A straightforward approach to address the challenges in feedback overhead is to feed back only statistical CSI[4]. However, this strategy achieves only satisfactory performance in limited scenarios such as slowly changing channels. To achieve the potential multiplexing gain in massive MIMO systems, it is necessary to investigate effective approaches for instantaneous CSI acquisition. In this section, we introduce the conventional CSI compression and feedback schemes based on two popular techniques, i.e., codebook-based methods and compressive sensing (CS).

2.1 Codebook-Based CSI Compression

An effective approach for instantaneous CSI acquisition with limited feedback relies on a pre-defined codebook for channel quantization. By employing a vector quantization codebook designed offline and known to both the BS and users, the users are only required to feed back the quantization index of the selected codeword in the codebook. In Ref. [6], noncoherent trellis-coded quantization (NTCQ) was proposed for channel quantization in massive MIMO. By leveraging the duality between source coding on the Grassmannian manifold and channel coding for noncoherent communication, the complexity of encoding grows linearly with the number of antennas. Moreover, codebook-based channel feedback techniques have been incorporated into wireless standards such as 3GPP LTE and IEEE 802.16m[4].

Codebook-based channel feedback also faces some technical challenges in practical implementation. Since the design of the codebook is closely related to the channel distribution, a specific design is difficult to adapt to different system scenarios. In addition, the codebook size increases exponentially with the number of antennas, and the computational expense for the look-up algorithm at the BS increases accordingly.

2.2 Compressive Sensing-Based CSI Compression

CS is a signal reconstruction framework for recovering sparse signals through sub-Nyquist sampling, which has been widely applied to signal processing in wireless communications[7]. Since the antenna arrays in massive MIMO have strong spatial correlations, the channel matrix is expected to exhibit sparsity in the spatial-frequency domain. Based on the channel sparsity assumption, CS was first applied to the design of the CSI compression and feedback scheme in Ref. [8]. Specifically, the channel matrix was estimated and compressed at the receiver via a predefined measurement matrix, which was randomly generated offline according to Gaussian distributions and known at both transmitter and receiver. Subsequently, the transmitter adopted the orthogonal matching pursuit (OMP) algorithm to recover the channel, leveraging the known measurement matrix and sparsifying bases. The two-dimensional discrete cosine transform and Karhunen-Loeve transform were employed as the sparsifying bases since they can offer a sparser representation of the signal. The dimensionality of the compressed channel was significantly reduced due to the channel sparsity, and the accuracy of recovery was acceptable with the sparsifying bases properly selected.

However, the CS-based channel compression and feedback still have some limitations in practical implementation. On the one hand, the CS-based CSI compression methods demand a channel sparsity assumption in a certain domain; this assumption may not strictly hold in practice, leading to inaccuracies in CSI recovery[9]. On the other hand, the signal reconstruction at the BS is generally solved by employing iterative algorithms such as OMP, linear programming (LP), and basis pursuit (BP)[8]. These iterative algorithms introduce substantial computational complexity and time delay, making the reconstruction process infeasible in practice. Consequently, numerous studies have turned to promising DL techniques to facilitate effective CSI acquisition under limited feedback, which will be introduced in the following sections.

3 Deep Learning Techniques for CSI Compression and Feedback

Due to its strong capability of data processing and function approximation, deep learning-based CSI acquisition is considered a promising approach to addressing the challenges of con-

ventional methods[10]. In this section, we focus on deep learning techniques for CSI compression and feedback. We first introduce the general framework of deep CSI acquisition. Then, we focus on deep CSI acquisition techniques for CSI matrices with specific correlations, i.e., the spatial correlation and temporal correlation. Finally, we analyze the computational complexity of various deep CSI acquisition methods.

3.1 General Framework of Deep CSI Compression

The autoencoder is a common framework adopted in deep CSI compression and feedback. Inspired by image processing, autoencoder-based deep CSI acquisition methods view the CSI matrix as an image. As depicted in Fig. 1, the autoencoder framework consists of an encoder and a decoder. The CSI matrix is first compressed into specific codewords by an encoder on the user side. Subsequently, the BS utilizes a decoder to reconstruct the CSI matrix from the received latent codewords, thereby facilitating efficient information feedback.

Various neural network (NN) architectures can be employed to design the autoencoder, such as convolutional neural networks (CNN)[11], long short-term memory (LSTM) networks[12], and the attention mechanism[13]. Different from images in computer vision, the CSI matrix contains inherent correlations due to the physical propagation environment. Since the performance of deep CSI compression and feedback is significantly affected by the NN architecture of autoencoders, appropriate NNs should be designed based on the specific characteristics of the CSI matrix, which will be discussed in the following two subsections.

The generative adversarial network (GAN)[14] is another effective framework of deep CSI compression and feedback, which learns the latent channel distribution to improve feedback performance. As shown in Fig. 2, GANs consist of two interlinked NNs, i.e., a generator and a discriminator, trained in tandem through an adversarial process. During the training phase, the generator produces samples mirroring the distribution of the training CSI data, whereas the discriminator aims to distinguish between authentic and synthesized samples. During the inference phase, only the generator is deployed as the decoder to reconstruct CSI matrices.

In Ref. [15], a deep convolutional generative adversarial network (DCGAN) framework was proposed to improve feedback accuracy in massive MIMO systems. Specifically, the generator of DCGAN learns to reconstruct high-quality CSI from the compressed vector, and the discriminator network evaluates the recovery quality. The GAN-based framework outperforms CS-based methods and achieves robust performance in outdoor channels. Moreover, a generative network termed PRVNet was proposed in Ref. [16] for CSI acquisition in MIMO-OFDM systems. By utilizing the generative framework of variational autoencoders, the PRVNet achieves robustness against various noise levels.

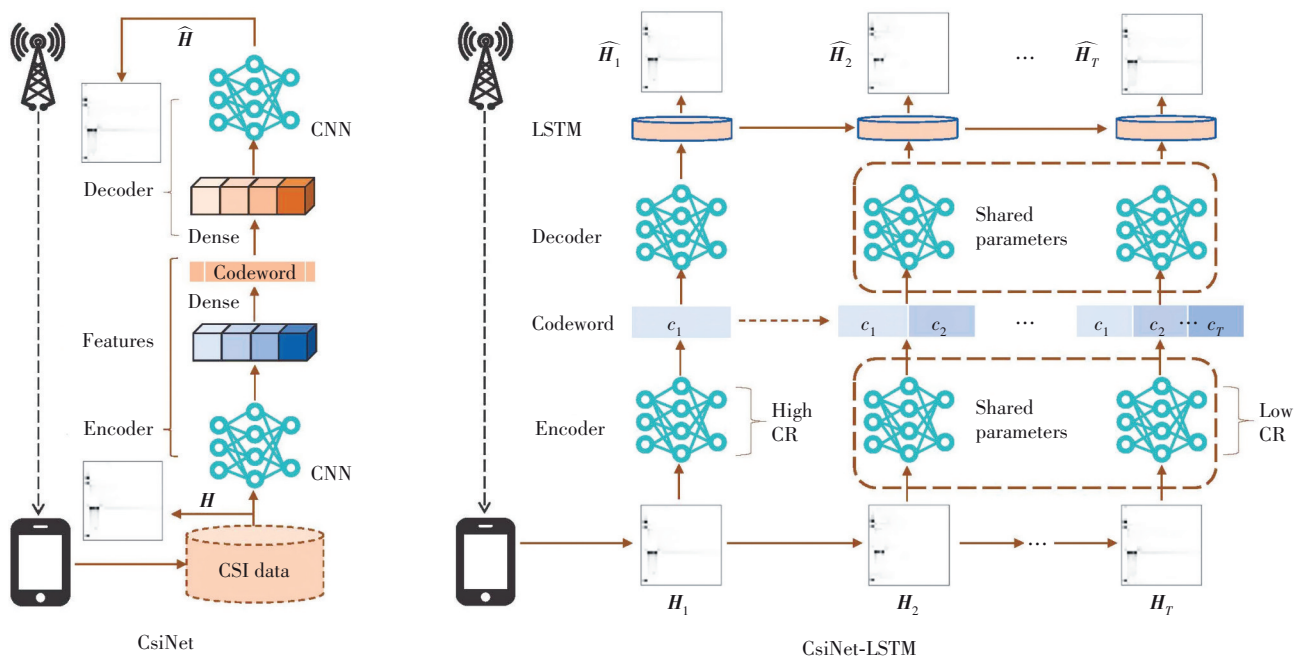

Figure 1. Illustration of autoencoder architecture

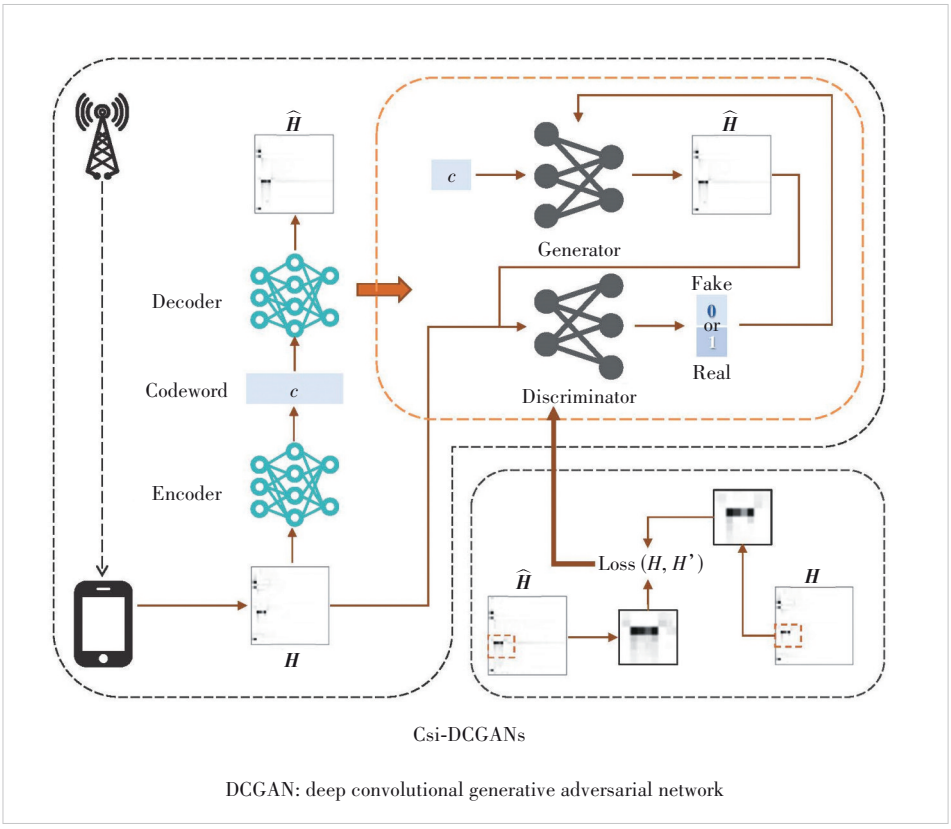

Figure 2. Illustration of generative adversarial network architecture

among transmit antennas. The encoder extracts CSI features with convolutional layers and compresses the CSI with fully connected layers, while the decoder adopts a symmetric structure and adjusts the number of layers and neurons. The reconstruction accuracy of CsiNet is significantly higher than that of the CS-based methods. CsiNet merely considers the CSI feedback in MIMO systems with a single user. However, in multiuser massive MIMO systems, the correlations among CSI matrices of nearby users can be exploited to improve the feedback performance. In Ref. [18], an autoencoder, termed DeepCMC, with fully convolutional layers, was proposed for CSI feedback in multiuser MIMO systems. In DeepCMC, the encoders are distributively deployed across the users, while the decoder at the BS jointly reconstructs the multiuser CSI. The decoder consists of separate decoder branches for users and combining kernels to fuse the side information of users. DeepCMC outperforms CsiNet by exploiting the correlations among users.

3.2 Spatially Correlated CSI Compression

The CSI matrix in the space-frequency domain contains inherent spatial correlations, such as the correlation across BS antennas and the correlation among users. CNN is a suitable NN architecture for CSI acquisition, as it learns the spatial correlation of CSI matrices through convolutional operations. Specifically, in each convolutional layer of CNN, convolutional kernels are used to perform element-wise multiplication and addition with the input data to capture local features. By sliding the convolutional kernels over the input data, an output feature map with detected features is generated. The mathematical representation of the convolutional layer is formulated as:

$$\boldsymbol{Y}_{c,i,j} = \sum_{m=1}^{C_{in}} \sum_{p=1}^{F} \sum_{q=1}^{F} \boldsymbol{X}_{m,i+p,j+q} \cdot \boldsymbol{W}_{c,m,p,q} + b_c \tag{1},$$

where $\boldsymbol{X} \in \mathbb{R}^{C_{in} \times H_{in} \times W_{in}}$ is the input data, $\boldsymbol{Y} \in \mathbb{R}^{C_{out} \times H_{out} \times W_{out}}$ is the output feature map, $\boldsymbol{W}_c \in \mathbb{R}^{F \times F}$ is the c-th convolutional kernel and $b_c \in \mathbb{R}$ is the corresponding bias. By stacking several convolutional layers, the CNN can adjust the receptive field to learn the spatial local correlation of the CSI matrix.

The CNN architecture was first applied to CSI compression and feedback in Ref. [17], where a CNN-based autoencoder, termed CsiNet, was proposed to learn the spatial correlations

3.3 Temporally Correlated CSI Compression

In time-varying channels, the temporal correlations should be considered in CSI acquisition. LSTM is capable of handling long-term dependencies in sequential data due to its gated mechanisms, making it suitable for extracting temporal correlations to improve CSI feedback. Specifically, LSTM processes sequential data by stacking a series of LSTM cells, which is formulated as:

$$\begin{aligned}
\boldsymbol{i}_t &= \sigma\left(\boldsymbol{W}_i \boldsymbol{x}_t + \boldsymbol{U}_i \boldsymbol{h}_{t-1} + \boldsymbol{b}_i\right) \\
\boldsymbol{f}_t &= \sigma\left(\boldsymbol{W}_f \boldsymbol{x}_t + \boldsymbol{U}_f \boldsymbol{h}_{t-1} + \boldsymbol{b}_f\right) \\
\boldsymbol{o}_t &= \sigma\left(\boldsymbol{W}_o \boldsymbol{x}_t + \boldsymbol{U}_o \boldsymbol{h}_{t-1} + \boldsymbol{b}_o\right) \\
\boldsymbol{C}_t &= \boldsymbol{f}_t \odot \boldsymbol{C}_{t-1} + \boldsymbol{i}_t \odot \tanh\left(\boldsymbol{W}_c \boldsymbol{x}_t + \boldsymbol{U}_c \boldsymbol{h}_{t-1} + \boldsymbol{b}_c\right) \\
\boldsymbol{h}_t &= \boldsymbol{o}_t \odot \tanh\left(\boldsymbol{C}_t\right)
\end{aligned} \tag{2},$$

where $\boldsymbol{i}_t$ is the input gate that decides what information should be added to the cell state, $\boldsymbol{f}_t$ is the forget gate that decides what information should be discarded from the cell state, $\boldsymbol{o}_t$ is the output gate that decides what information should be output from the cell state. Additionally, $\boldsymbol{C}_t$ represents the memory cell, $\boldsymbol{h}_t$ denotes the hidden vector, and t signifies the time

step. By maintaining and updating the cell state through these gates, LSTMs can effectively handle long-term dependencies of input data.

CsiNet-LSTM was proposed in Ref. [19] for CSI acquisition in time-varying massive MIMO systems. In CsiNet-LSTM, CNN-based encoders are deployed at the user side, while the decoders at the BS are composed of CNNs and LSTMs to capture the spatial and temporal features of CSI matrices. The reconstruction accuracy of CsiNet-LSTM is higher than that of CsiNet due to the temporal correlation extraction. Advancing from CsiNet-LSTM, ConvlstmCsiNet was proposed in Ref. [20] for further improvement in reconstruction quality. By adopting pseudo-3D blocks to maintain the independence of temporal and spatial features, ConvlstmCsiNet achieves remarkable feedback accuracy and robustness at low compression ratios.

Although LSTM-based deep CSI feedback methods effectively extract the temporal correlation features of CSI matrices, they ignore the weight assignment of CSI features. An attention mechanism can be deployed in CSI feedback networks to assign more weight to dominant features, thereby enhancing the representation of temporal features and improving the performance of CSI reconstruction. Specifically, the attention mechanism generates a set of weight factors to describe the importance of features, and allocates more weights to the feature maps with more information. The attention weights are calculated as:

$$\alpha_{ij} = \frac{\exp\left(f\left(\boldsymbol{s}_i, \boldsymbol{s}_j\right)\right)}{\sum_k \exp\left(f\left(\boldsymbol{s}_i, \boldsymbol{s}_k\right)\right)} \tag{3},$$

where $\boldsymbol{s}_i \in \mathbb{R}^d$ is the input feature, and f is the alignment model operated by dense layers. Therefore, the output of the attention layer is:

$$\mathbf{c}_i = \sum_j \alpha_{ij} \mathbf{s}_j \tag{4},$$

where $i, j = 1, \cdots, T$ denotes the length of the input sequence.

An LSTM-attention network was proposed in Ref. [21] to improve CSI feedback accuracy by leveraging the attention mechanism to fine-tune the temporal features of CSI data. The autoencoder first deploys LSTM units to exploit the temporal correlation of massive MIMO channels, and subsequently incorporates the attention mechanism to prioritize and weight feature importance. The CSI recovery accuracy is significantly improved by adopting the attention mechanism. Moreover, a CNN-LSTM-A network was proposed in Ref. [22] for CSI feedback in MIMO systems with high user mobility, where the attention mechanism is introduced to assign more weights to dominant features in each time step.

3.4 Computational Complexity Analysis

In this subsection, we evaluate the computational complexity of deep CSI compression and feedback methods in the training and inference stages. Since the complexity of the training stage is mainly influenced by the backpropagation process for updating trainable parameters, we assess the complexity of the training stage based on the parameter size of NN models, which is referred to as space complexity (SC). Meanwhile, the complexity of the inference stage is evaluated by the number of floating-point operations (FLOPs), referred to as time complexity (TC).

We consider the downlink of an FDD massive MIMO-OFDM system with N_t antennas at the BS and a single-antenna user. The system adopts OFDM transmission with N_c subcarriers. Since the complexity of deep CSI feedback models is primarily dominated by the convolutional layers and dense layers, we analyze the time and space complexity of these layers, respectively. According to Refs. [23 – 24], the time and space complexity of convolutional layers are

$$TC_C = O\left(\sum_{l=1}^{L} W_l H_l C_{l-1} C_l F_l^2\right) \tag{5},$$

$$SC_C = O\left(\sum_{l=1}^{L} C_{l-1} C_l F_l^2\right) \tag{6},$$

where C_l is the number of channels in layer l, and F_l is the convolution kernel size in layer l. In deep CSI feedback models, the width and height of input feature satisfies $W_l H_l = 2N_t N_c, \forall l$. Thus, the time and space complexity of convolutional layers in CSI feedback NN models are

$$\mathrm{TC}_C^{\mathrm{CSI}} = 4N_t N_c \sum_{l=1}^{L} C_{l-1} C_l F_l^2 \tag{7},$$

$$\mathrm{SC}_C^{\mathrm{CSI}} = \sum_{l=1}^{L} C_{l-1} C_l F_l^2 \tag{8},$$

where the time complexity consists of the number of multiplication and addition in convolution operations, each occupying $2N_t N_c \sum_{l=1}^{L} C_{l-1} C_l F_l^2$ FLOPs. The space complexity indicates the parameter size of kernels.

Similarly, the time and space complexity of dense layers are

$$\mathrm{TC}_D = O\left(\sum_{l=1}^{L} N_{l-1} N_l\right) \tag{9},$$

$$\mathrm{SC}_D = O\left(\sum_{l=1}^{L} N_{l-1} N_l + N_l\right) \tag{10},$$

where N_l denotes the feature dimension in layer l. In deep CSI feedback models, the input and output features are $N_0 = 2N_t N_c, N_L = 2N_t N_c \gamma$ at the encoder and $N_0 = 2N_t N_c \gamma, N_L = 2N_t N_c$ at the decoder, where γ is the compression ratio. Thus, the time and space complexity of dense layers in CSI feedback NN models are

$$\mathrm{TC}_D^{\mathrm{CSI}} = 4N_tN_c(1+\gamma)(N_1+N_{L-1}) + 4\sum_{l=2}^{L-1}N_{l-1}N_l \quad (11),$$

$$\mathrm{SC}_D^{\mathrm{CSI}} = 2N_tN_c(1+\gamma)(N_1+N_{L-1}+1) + 2\sum_{l=2}^{L-1}(N_{l-1}N_l+N_l) + N_1 + N_{L-1} \quad (12),$$

where the space complexity indicates the parameter size of weights and biases.

The time and space complexity of various feedback NNs are shown in Table 1. Since the feature dimension is related to the dimension of the CSI matrix and the feedback length, the computational complexity increases with the compression ratio. Moreover, the complexity of convolutional layers is generally lower than that of dense layers. Since LSTM contains more dense layers, the complexity of NN in Ref. [17] is lower than that in Refs. [20] and [21].

Table 1. Computational complexity of deep CSI feedback models

Complexity		1/4	1/8	1/16	1/32
Time complexity	Ref. [17]	21 659 648	5 668 864	3 571 712	2 523 136
	Ref. [20]	121 708 544	97 591 296	86 319 104	80 879 616
	Ref. [21]	-	-	-	-
Space complexity	Ref. [17]	2 103 904	1 055 072	530 656	268 448
	Ref. [20]	28 326 904	22 296 312	19 477 624	18 117 432
	Ref. [21]	10 247 148	-	7 484 688	7 024 272

4 Applications of Deep CSI Compression and Feedback

In the 5G R18 standard[25], various methods have been adopted to improve the CSI compression and feedback, including the historical CSI-based prediction methods that utilize historical CSI data to predict future CSI, non-AI/ML prediction methods such as filters and interpolation algorithms, and the advanced prediction models utilizing AI/ML technologies. Current 5G communication protocols focus on enhancing the accuracy and real-time performance of CSI feedback through DL techniques, optimizing the transmission efficiency of 5G networks. DL techniques have been widely applied in 5G communications. In this section, we discuss the applications of DL techniques for solving practical challenges in CSI compression and feedback designs. Specifically, we emphasize the representative research achievements in network lightweighting, reconstruction performance improvement, CSI generalization enhancement, and the joint design of CSI compression, feedback, and precoding. These approaches are significant to the practical implementation of CSI acquisition, as they reduce deployment costs and enhance compression efficiency.

4.1 Network Lightweighting

Network lightweighting is crucial to the practical deployment of DL-based CSI compression and feedback networks, aiming to reduce the network size deployed on both BS and users, thereby saving hardware costs. To date, numerous effective network lightweighting methods have been proposed and applied in the CSI compression and feedback of MIMO systems.

Due to more stringent computational and memory constraints on the users than on the BS, the primary objective of network lightweighting methods is to reduce the size of the encoder network at the user end. The most common approach involves designing innovative convolutional structures to decrease the number of parameters at the encoder[26]. Typical lightweight structures include multi-branch convolutions, dimensionality reduction sampling of CSI feature maps, etc.

On the other hand, the inherent characteristics of CSI are also leveraged for the implementation of network lightweighting. The observed similarity in the probability distributions of the real and imaginary parts of the CSI matrix enabled a method where only the real part of the CSI matrix is inputted into the network for training[27]. Subsequently, the trained network was reused for the compression and feedback of the imaginary part. This approach, without compromising performance, effectively reduces the network parameters by approximately half.

Furthermore,the real and imaginary parts of CSI matrix also carry inherent physical information. This characteristic has been utilized in studies focusing on network lightweighting. One research approach in Ref. [28] involved transforming a real-valued NN designed for lightweight purposes into a complex-valued NN, achieving equivalent network performance with fewer parameters required. Another research in Ref. [29] focused on the design of a pseudo-complex-valued input layer, while retaining the real-valued NN. This approach allows the input CSI matrix to undergo equivalent complex-valued operations, thereby reducing the computational overhead by 24%.

4.2 Performance Improvement

In the context of DL-based CSI compression and feedback, enhancing the efficiency of CSI acquisition represents a paramount research direction. Extracting the inherent features of the CSI matrix to improve the compression performance from a physical perspective is considered a highly promising research avenue.

Researchers initially focused on the sparsity characteristics of the CSI matrix, which vary with different channel scenarios and compression rates. According to existing DL-based theory, dense images are more aptly processed using convolu-

tional kernels of smaller size for feature extraction, while sparse images benefit from larger convolutional kernels. To enable CSI encoders and decoders to effectively extract features of CSI in various scenarios, a multi-path parallel convolutional structure has been proposed and applied to both[30 - 31]. By employing parallel convolutional layers with different kernel sizes to extract CSI features, this architecture significantly enhances the efficiency of CSI compression and feedback across diverse scenarios and compression rates.

In the realm of DL for image compression, a series of efficient techniques have been developed by investigating the structural features of images. In CSI compression and feedback, the CSI matrix is often viewed as an image, thereby enabling the utilization of image characteristics of the CSI matrix to enhance compression efficiency. The CSI image can be divided into many small blocks, where some blocks contain a high level of self-information and the image features within are referred to as shape features; conversely, blocks with less self-information are characterized by their texture information. By preserving blocks with high self-information shape features and discarding those with low self-information texture features during compression, researchers have achieved efficient CSI compression and feedback[32 - 33]. Distinct from black-box neural networks, this method incorporates CSI prior knowledge and significantly reduces the complexity of the encoder network.

The previous researchers designed NNs from the perspective of the image features of CSI. In contrast, other researchers aim to design CSI compression and feedback networks based on the extraction of physical CSI features. In Ref. [34], the authors discovered that the line-of-sight (LoS) propagation path characteristics and non-line-of-sight (NLoS) path features can be effectively extracted by different NNs. Consequently, the authors employed a dual-feature fusion NN that combines a CNN with an attention enhancement network structure to achieve improved compression performance. In Ref. [35], the authors considered the similarity of the CSI matrix across different polarization directions caused by dual-polarized antennas and introduced a decoupled representation learning method. This method reduces the redundant information shared across different polarization directions of CSI, thereby enhancing compression performance.

4.3 Generalization Enhancement

To achieve satisfactory CSI feedback performance, DL-based approaches require a substantial amount of CSI training data, which is exceedingly costly in real-world scenarios. Furthermore, when the channel environment changes, DL networks trained under different channel conditions cannot be applied to new channel environments. This necessitates CSI data re-collection and network re-training, thereby significantly increasing the deployment cost of NN applications. Hence, generalization enhancement is a valuable direction for research.

To address the issues of insufficient training samples and the maladaptation of NNs to new channel environments, direct transfer learning and meta-learning, two methods of deep transfer learning, have been introduced into the CSI compression and feedback domain for generalization enhancement[36 - 38]. Fig. 3 presents a schematic illustration of employing transfer learning methods for CSI compression and feedback. Both direct transfer learning and meta-learning are effective strategies for addressing model generalization and adaptability issues. They achieve this by utilizing the transfer of existing knowledge and adopting a "learning to learn" strategy to rapidly adapt to new tasks, respectively. In Refs. [36 - 38], deep transfer networks employing direct transfer and

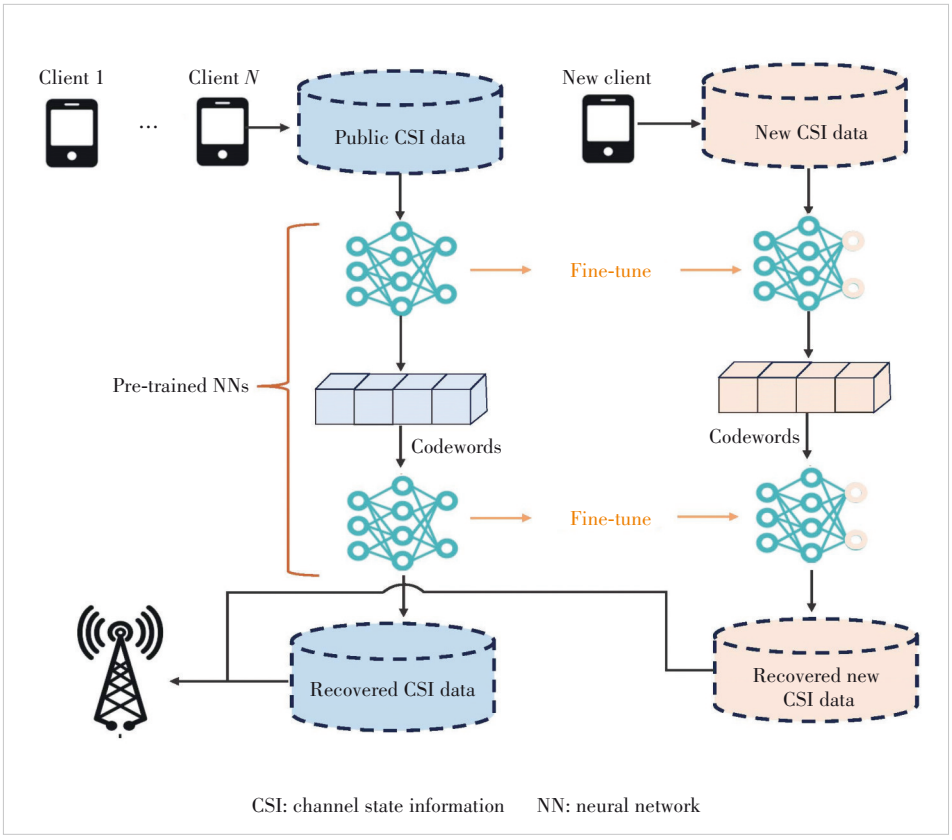

Figure 3. Transfer learning model for CSI feedback

meta-learning methods required only a minimal amount of training data to achieve satisfactory CSI compression and feedback performance. Moreover, through fine-tuning, these networks can quickly adapt to new channel environments, significantly reducing training overhead. Specifically, in Ref. [36], the downlink channel prediction was formulated as a deep transfer learning (DTL) problem, and a direct transfer algorithm based on a fully-connected NN architecture was proposed. The authors also designed a meta-learning algorithm that trains the network by alternately performing intra-task and inter-task updates, which is then adapted to new environments using a small amount of labeled data. Simulation results show that, compared with methods without transfer learning, the direct transfer algorithm and the meta-learning algorithm can improve downlink CSI prediction accuracy by up to 50%. In Refs. [37] and [38], the authors proposed a model-agnostic meta-learning (MAML) approach to address the issue of needing a large number of wireless channel environment samples for training deep neural networks (DNNs) as pre-trained models. By fine-tuning the pre-trained model with a relatively small number of samples, they achieved CSI feedback models for different wireless channel environments at a lower training cost. Simulation results show that the MAML method achieves 7 – 8 dB gains in CSI feedback NMSE compared with the deep transfer learning under different channel models.

Transfer learning methods can also be applied to the reconstruction of MIMO channels, enabling the acquisition of sufficient CSI training data without the need for extensive measurement and labeling of actual channels. Ref. [39] utilized untrained neural networks (UNN) to obtain a large amount of equivalent wireless channel data through minimal measurements (e.g., only a few time snapshots). In this transfer learning training, the UNN acquired prior knowledge about the propagation environment, thereby enabling the reconstruction of wireless channels.

Moreover, transfer learning techniques can also address the inflexibility issue in compression ratios within DL-based CSI compression and feedback methods[40]. In real-world communication scenarios, the channel conditions between the BS and users are constantly changing, necessitating adjustment to the compression ratio. Consequently, the users and BS need to store network parameters for various compression rates, which increases the hardware overhead. By employing transfer learning approaches, this overhead can be saved and the CSI compression performance can be enhanced.

Finally, Ref. [41] addressed the issue of data silos and online training adaptation caused by offline NN training, and proposed a DL approach based on interactive federated and transfer learning (IFTL). This method enables downlink CSI prediction and online update capabilities. The transfer learning approach takes into account various factors, including the asynchrony of different clients, and achieves good performance in the channel environments of different cells. These research findings collectively underscore the potential and promising prospects of transfer learning methods in CSI compression and feedback, making it a promising direction for future research.

4.4 Joint Design of CSI Compression, Feedback, and Precoding

In wireless communication, the purpose of CSI compression and feedback is to facilitate more efficient design of the precoding matrix, thereby enhancing the communication rate in millimeter-wave MIMO (mMIMO) systems. Hence, merely improving the accuracy of CSI compression and feedback without considering precoding matrix design does not guarantee optimal communication performance. To address this, researchers have integrated pilot estimation, CSI compression and feedback, and precoding matrix design into a unified process, optimizing communication performance from the perspective of communication rates. This approach is more practically meaningful than efforts focused solely on achieving higher CSI reconstruction accuracy, as it addresses the overall effectiveness issue in communication systems.

Ref. [42] presented a joint framework for pilot design, CSI compression and feedback, and precoding design in downlink multi-user massive MIMO systems based on DL. The authors observed that this joint design problem could be modeled as a distributed source coding issue. Specifically, the pilot design employed an unbiased single-layer fully connected network for equivalence, with the network's weights serving as the pilot sequences to be optimized. The pilot was input into the CSI compression and feedback network, which output quantized codewords. This network effectively accomplished three tasks: channel estimation, channel compression, and quantization of the compressed codewords. Finally, the quantized codewords were input into a precoding network, which output the precoding matrix optimized by the NN. Simulation results indicate that this approach closely matches the performance of traditional precoding schemes with perfect CSI while requiring significantly less pilot and codeword feedback overhead. Ref. [43] proposed a similar joint training framework while introducing a training strategy distinct from that in Ref. [42]. This approach can circumvent the need for retraining across different network scales intended for scalable designs, thereby reducing training overhead.

Furthermore, to address the need for flexible pilot adaptation due to the limited saturation levels of amplifiers in massive MIMO systems, Ref. [44] devised a DL-based closed-loop massive MIMO system joint optimization scheme. In this scheme, the authors represented the process of generating the beamforming (BF) matrix as a functional optimization problem. The functions of adaptive pilot length, CSI compression, and BF were all substituted by NNs, where each functional block learned the optimal strategy to maximize network utility during the training process.

In real-world communication scenarios, signaling overhead and CSI mismatches can arise due to transmission delays. To address this issue, Ref. [45] developed a dual-timescale DNN architecture composed of long-term and short-term DNNs. The analog precoder, designed by the long-term DNN based on CSI statistical information, was updated once over multiple time slots. In contrast, the digital precoder was optimized at each time slot by the short-term DNN based on a low-dimensional equivalent CSI matrix. Moreover, a corresponding dual-timescale training method was developed, which achieved lower bit error rates and pilot overhead reduction.

To summarize, the main contributions of the above papers are listed in Table 2.

5 Technical Challenges and Future Directions

Although various research has been proposed to solve the practical problems in massive MIMO CSI acquisition, there are still some open issues remaining to be investigated. In the following content, we enumerate the key challenges and future directions of DL-based CSI acquisition in future wireless networks.

5.1 Technical Challenges

The major challenge of DL-based CSI feedback lies in the enormous number of training samples required in the training process. Most existing works utilize the true CSI in massive MIMO downlinks as the training label and learn to minimize the error between true CSI and reconstructed CSI. However, CSI estimation in massive MIMO downlinks incurs overwhelming overhead, as the number of orthogonal pilots increases linearly with the number of antennas. This makes the training phase of autoencoders for CSI compression and reconstruction expensive and time-consuming. Furthermore, the testing dataset in practical implementation often exhibits domain discrepancies with training datasets due to time delay, which may lead to performance degradation[46]. Consequently, it is necessary and challenging to obtain the appropriate training dataset at an acceptable cost.

The cooperation between the BSs and users in DL-based CSI feedback design also brings up various challenges. Since most existing DL-based designs employ end-to-end learning of the encoder and decoder, the users and BSs are required to exchange compressed CSI and calculated gradients, respectively, leading to tremendous signaling overhead. Moreover, the network training process demands huge computational and storage resources, which are unaffordable for the users. Meanwhile, some studies investigate novel DL architectures to deploy end-to-end training on the BS[47]. In this case, the encoder is obtained at the BS and thus brings up extra problems such as the intellectual property among manufacturers.

Another critical issue is enhancing the generalization capability of DL-based CSI feedback models. Most existing designs focus on autoencoder architectures over a specific channel distribution or have limited scalability towards some specific factors, which may suffer from severe performance degradation in real-time implementations, such as high mobility scenarios[48]. In addition, it is infeasible to train diverse models for different channel scenarios due to the high training cost. Therefore, the generalization capability of DL-based CSI feedback models is significant in practical use and remains a crucial challenge for future investigation.

5.2 Future Directions

The artificial intelligence-native air interface (AI-AI) is a disruptive framework that incorporates conventional signal processing modules by deploying AI models to the air interface, which is considered a promising evolution of the network design in the B5G and 6G systems[49 – 50]. Different from the existing systems decoupling the source coding, channel coding, and data transmission into a block-to-block architecture, the goal of the AI-native air interface is to initially build an AI-based communication framework considering the impact of

Table 2. Summary of recent papers on deep CSI feedback

Advantages	Key Techniques	References
Network lightweighting	Design an innovative structure of multi-branch convolutions	Ref. [26]
	Exploit the similarity of real and imaginary parts of CSI	Refs. [27 – 29]
Performance improvement	Exploit the sparse characteristics of CSI	Refs. [30 – 31]
	Exploit the image characteristics of the CSI matrix	Refs. [32 – 33]
	Extract CSI features based on physical propagation environment	Refs. [34 – 35]
Generalization enhancement	Adopt model-agnostic meta-learning approaches	Refs. [36 – 38]
	Adopt deep transfer learning techniques	Refs. [39 – 40]
	Adopt interactive federated and transfer learning	Ref. [41]
End-to-end design	Joint framework of pilot design, CSI feedback, and precoding	Refs. [42 – 43]
	Consider adaptive pilot length for mm-wave MIMO systems	Ref. [44]
	Design a dual-timescale network to reduce signaling overhead	Ref. [45]

CSI: channel state information MIMO: multiple-input multiple-output

hardware imperfections and radio environment. According to 3GPP Release 18, CSI feedback is included as one of the three representative specific use cases in AI-native air interface[51]. The key challenge of designing deep CSI feedback models for the AI-native air interface is the model generalization capability over scenarios and configurations. Training dataset mixing and online learning are two potential approaches to this challenge[52]. By mixing the CSI samples generated with different channel models, a training dataset that covers various channel distributions can be formed, thereby improving the generalization ability of the model. Online learning is required when new channel distribution occurs. Moreover, advanced learning-based techniques such as transfer learning and meta-learning could be deployed to accelerate online learning and reduce training overhead.

Terahertz (THz)-band communication is regarded as a crucial technique to support the increasing demand for communication capacity and bandwidth in future wireless mobile communications. To alleviate the high propagation loss and power limitation in THz communications, densely packed nano-antenna arrays are employed to construct ultra-massive MIMO (UM-MIMO) systems[53]. However, CSI acquisition in UM-MIMO systems is more challenging than that in massive MIMO due to the expanding number of antennas, beam squint, and hybrid-field effects[54]. Moreover, training labels for DL-based CSI feedback design are more difficult to acquire. Model-driven deep learning is a promising approach to these challenges. For example, deep unfolding is a model-driven technique that unfolds iterative algorithms into a layer-wise neural network[55]. Since the CSI reconstruction at the BS is generally achieved via iterative algorithms, deep unfolding can be adopted for CSI feedback design in UM-MIMO systems. Furthermore, unfolding-based DL networks rely on the architecture of the underlying iterative algorithm and incorporate inherent domain knowledge. Thus, the number of trainable parameters in unfolding-based DL networks is considerably lower than that of black-box DNNs, thereby reducing training overhead.

Semantic communication is a novel framework that takes into account the meaning of transmission messages in signal processing designs[56–57]. With the increasing demand for content-based services in 5G and beyond, semantic communication is considered a promising technique to meet the tremendous requirements by exploiting the semantic aspects of communication not included in Shannon's information theory. Semantic communication-based deep CSI feedback is a potential approach to mitigating feedback overhead, while it faces critical challenges in semantic expression modeling. To address this issue, a task-oriented deep CSI feedback framework can be employed. For example, in a data hiding-based CSI feedback system where downlink CSI is hidden within the transmitted images, the autoencoder architecture used for image compression can be adopted to design the CSI feedback network[58]. Moreover, in a precoding-oriented CSI feedback system where the BS aims to design multiuser precoding vectors, an end-to-end DNN architecture can be employed, and a specific loss function can be designed to strike a trade-off between sum-rate performance and feedback overhead[59].

6 Conclusions

In this paper, we provide a comprehensive overview of DL-based CSI compression and feedback techniques in massive MIMO systems. We focus on the critical challenges in conventional CSI acquisition approaches, such as quantized codebooks and compressive sensing. Specifically, we analyze the advantages of various DL techniques applied to CSI compression, including CNN, LSTM, GAN, and attention mechanisms. The applications of DL-based methods for solving practical challenges in CSI compression and feedback such as network lightweighting and generalization enhancement are also discussed. Finally, we emphasize the existing critical challenges and promising future directions.

References

[1] Xu W, Yang Z H, Ng D W K, et al. Edge learning for B5G networks with distributed signal processing: semantic communication, edge computing, and wireless sensing [J]. IEEE journal of selected topics in signal processing, 2023, 17(1): 9–39. DOI: 10.1109/JSTSP.2023.3239189
[2] He M, Li X, Ni J. Physical layer security for mmWave communications: challenges and solutions [J]. ZTE communications, 2022, 20(4): 41–51. DOI: 10.12142/ZTECOM.202204006
[3] Lu L, Li G Y, Swindlehurst A L, et al. An overview of massive MIMO: benefits and challenges [J]. IEEE journal of selected topics in signal processing, 2014, 8(5): 742–758. DOI: 10.1109/JSTSP.2014.2317671
[4] Love D J, Heath R W, Lau V K N, et al. An overview of limited feedback in wireless communication systems [J]. IEEE journal on selected areas in communications, 2008, 26(8): 1341–1365. DOI: 10.1109/JSAC.2008.081002
[5] Zhao M K, Huang Y, Li X. Federated learning for 6G: a survey from perspective of integrated sensing, communication and computation [J]. ZTE communications, 2023, 21(2): 25–33. DOI: 10.12142/ZTECOM.202302005
[6] Choi J, Chance Z, Love D J, et al. Noncoherent trellis coded quantization: a practical limited feedback technique for massive MIMO systems [J]. IEEE transactions on communications, 2013, 61(12): 5016–5029. DOI: 10.1109/TCOMM.2013.111413.130379
[7] Gao Z, Dai L L, Han S F, et al. Compressive sensing techniques for next-generation wireless communications [J]. IEEE wireless communications, 2018, 25(3): 144–153. DOI: 10.1109/MWC.2017.1700147
[8] Kuo P H, Kung H T, Ting P G. Compressive sensing based channel feedback protocols for spatially-correlated massive antenna arrays [C]//IEEE Wireless Communications and Networking Conference (WCNC). IEEE, 2012: 492–497. DOI: 10.1109/WCNC.2012.6214417
[9] Liu Z Y, Zhang L, Ding Z. An efficient deep learning framework for low rate massive MIMO CSI reporting [J]. IEEE transactions on communications, 2020, 68(8): 4761–4772. DOI: 10.1109/TCOMM.2020.2993626
[10] Wang T Q, Wen C K, Wang H Q, et al. Deep learning for wireless physical

layer: opportunities and challenges [J]. China communications, 2017, 14 (11): 92 – 111. DOI: 10.1109 /CC.2017.8233654
[11] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition [J]. Proceedings of the IEEE, 1998, 86(11): 2278 – 2324. DOI: 10.1109/5.726791
[12] Hochreiter S, Schmidhuber J. Long short-term memory [J]. Neural computation, 1997, 9(8): 1735 – 1780. DOI: 10.1162/neco.1997.9.8.1735
[13] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [EB/OL]. [2024-08-12]. https://proceedings. neurips. cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
[14] Creswell A, White T, Dumoulin V, et al. Generative adversarial networks: an overview [J]. IEEE signal processing magazine, 2018, 35(1): 53 – 65. DOI: 10.1109/MSP.2017.2765202
[15] Tolba B, Elsabrouty M, Abdu-Aguye M G, et al. Massive MIMO CSI feedback based on generative adversarial network [J]. IEEE communications letters, 2020, 24(12): 2805 – 2808. DOI: 10.1109/LCOMM.2020.3017188
[16] Hussien M, Nguyen K K, Cheriet M. PRVNet: a novel partially-regularized variational autoencoders for massive MIMO CSI feedback [C]// Wireless Communications and Networking Conference (WCNC). IEEE, 2022: 2286 – 2291. DOI: 10.1109/WCNC51071.2022.9771642
[17] Wen C K, Shih W T, Jin S. Deep learning for massive MIMO CSI feedback [J]. IEEE wireless communications letters, 2018, 7(5): 748 – 751. DOI: 10.1109/LWC.2018.2818160
[18] Mashhadi M B, Yang Q Q, Gündüz D. Distributed deep convolutional compression for massive MIMO CSI feedback [J]. IEEE transactions on wireless communications, 2021, 20(4): 2621 – 2633. DOI: 10.1109/TWC.2020.3043502
[19] Wang T Q, Wen C K, Jin S, et al. Deep learning-based CSI feedback approach for time-varying massive MIMO channels [J]. IEEE wireless communications letters, 2019, 8(2): 416 – 419. DOI: 10.1109/LWC.2018.2874264
[20] Li X Y, Wu H M. Spatio-temporal representation with deep neural recurrent network in MIMO CSI feedback [J]. IEEE wireless communications letters, 2020, 9(5): 653 – 657. DOI: 10.1109/LWC.2020.2964550
[21] Li Q, Zhang A H, Liu P C, et al. A novel CSI feedback approach for massive MIMO using LSTM-attention CNN [J]. IEEE access, 2020, 8: 7295 – 7302. DOI: 10.1109/ACCESS.2020.2963896
[22] Zhang Z F, Zheng Y, Gan C Q, et al. Massive MIMO CSI reconstruction using CNN-LSTM and attention mechanism [J]. IET communications, 2020, 14(18): 3089 – 3094. DOI: 10.1049/iet-com.2019.1030
[23] Xia W C, Zheng G, Zhu Y X, et al. A deep learning framework for optimization of MISO downlink beamforming [J]. IEEE transactions on communications, 2020, 68(3): 1866 – 1880. DOI: 10.1109/TCOMM.2019.2960361
[24] He K M, Sun J. Convolutional neural networks at constrained time cost [C]// Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2015: 5353 – 5360. DOI: 10.1109/CVPR.2015.7299173
[25] 3GPP TR 38.843. Study on artificial intelligence (AI)/machine learning (ML) for NR air interface (Release 18) [S]. 2023
[26] Cao Z, Shih W T, Guo J J, et al. Lightweight convolutional neural networks for CSI feedback in massive MIMO [J]. IEEE communications letters, 2021, 25(8): 2624 – 2628. DOI: 10.1109/LCOMM.2021.3076504
[27] Sun Y Y, Xu W, Liang L, et al. A lightweight deep network for efficient CSI feedback in massive MIMO systems [J]. IEEE wireless communications letters, 2021, 10(8): 1840 – 1844. DOI: 10.1109/LWC.2021.3083331
[28] Li H Z, Zhang B Y, Chang H R, et al. CVLNet: a complex-valued lightweight network for CSI feedback [J]. IEEE wireless communications letters, 2022, 11(5): 1092 – 1096. DOI: 10.1109/LWC.2022.3157263
[29] Ji S J, Li M. CLNet: complex input lightweight neural network designed for massive MIMO CSI feedback [J]. IEEE wireless communications letters, 2021, 10(10): 2318 – 2322. DOI: 10.1109/LWC.2021.3100493
[30] Lu Z L, Wang J T, Song J. Multi-resolution CSI feedback with deep learning in massive MIMO system [C]//International Conference on Communications (ICC). IEEE, 2020: 1 – 6. DOI: 10.1109/icc40277.2020.9149229
[31] Lu Z L, Zhang X D, He H Y, et al. Binarized aggregated network with quantization: flexible deep learning deployment for CSI feedback in massive MIMO systems [J]. IEEE transactions on wireless communications, 2022, 21(7): 5514 – 5525. DOI: 10.1109/TWC.2022.3141653
[32] Yin Z Q, Xie R J, Xu W, et al. Self-information domain-based neural CSI compression with feature coupling [J]. IEEE transactions on vehicular technology, 2023, 72(10): 13661 – 13665. DOI: 10.1109/TVT.2023.3272560
[33] Yin Z Q, Xu W, Xie R J, et al. Deep CSI compression for massive MIMO: a self-information model-driven neural network [J]. IEEE transactions on wireless communications, 2022, 21(10): 8872 – 8886. DOI: 10.1109/TWC.2022.3170576
[34] Zhang S Q, Xu W, Jin S, et al. Dual-propagation-feature fusion enhanced neural CSI compression for massive MIMO [J]. IEEE transactions on communications, 2023, 71(9): 5182 – 5198. DOI: 10.1109/TCOMM.2023.3282227
[35] Fan S H, Xu W, Xie R J, et al. Deep CSI compression for dual-polarized massive MIMO channels with disentangled representation learning [J]. IEEE transactions on communications, 2024, 72(9): 5564 – 5580. DOI: 10.1109/TCOMM.2024.3384256
[36] Yang Y W, Gao F F, Zhong Z M, et al. Deep transfer learning-based downlink channel prediction for FDD massive MIMO systems [J]. IEEE transactions on communications, 2020, 68(12): 7485 – 7497. DOI: 10.1109/TCOMM.2020.3019077
[37] Zeng J, Sun J L, Gui G, et al. Downlink CSI feedback algorithm with deep transfer learning for FDD massive MIMO systems [J]. IEEE transactions on cognitive communications and networking, 2021, 7(4): 1253 – 1265. DOI: 10.1109/TCCN.2021.3084409
[38] Zeng J, He Z R, Sun J L, et al. Deep transfer learning for 5G massive MIMO downlink CSI feedback [C]//Proceedings of IEEE Wireless Communications and Networking Conference (WCNC). IEEE, 2021: 1 – 5. DOI: 10.1109/wcnc49053.2021.9417349
[39] Boas B V, Zirwas W, Haardt M. Transfer learning capabilities of untrained neural networks for MIMO CSI recreation [C]//International Conference on Communications. IEEE, 2022: 1288 – 1293. DOI: 10.1109/ICC45855.2022.9838738
[40] Wang Y T, Sun J L, Wang J, et al. Multi-rate compression for downlink CSI based on transfer learning in FDD massive MIMO systems [C]//The 94th Vehicular Technology Conference (VTC2021-Fall). IEEE, 2021: 1 – 5. DOI: 10.1109/VTC2021-Fall52928.2021.9625585
[41] Sun J L, Zhang Y B, Gui G, et al. Interacting federated and transfer learning-aided CSI prediction for intelligent cellular networks [J]. IEEE transactions on vehicular technology, 2023, 72(12): 15776 – 15787. DOI: 10.1109/TVT.2023.3290660
[42] Sohrabi F, Attiah K M, Yu W. Deep learning for distributed channel feedback and multiuser precoding in FDD massive MIMO [J]. IEEE transactions on wireless communications, 2021, 20(7): 4044 – 4057. DOI: 10.1109/TWC.2021.3055202
[43] Jang J, Lee H, Kim I M, et al. Deep learning for multi-user MIMO systems: joint design of pilot, limited feedback, and precoding [J]. IEEE transactions on communications, 2022, 70(11): 7279 – 7293. DOI: 10.1109/TCOMM.2022.3209887
[44] Jee J, Park H. Deep learning-based joint optimization of closed-loop FDD mmWave massive MIMO: pilot adaptation, CSI feedback, and beamforming [J]. IEEE transactions on vehicular technology, 2024, 73(3): 4019 – 4034. DOI: 10.1109/TVT.2023.3327276
[45] Hu Q Y, Cai Y L, Kang K, et al. Two-timescale end-to-end learning for channel acquisition and hybrid precoding [J]. IEEE journal on selected areas in communications, 2022, 40(1): 163 – 181. DOI: 10.1109/JSAC.2021.3126050
[46] Zhang H Y, Lu Z L, Zhang X D, et al. Data augmentation for bridging the delay gap in DL-based massive MIMO CSI feedback [J]. IEEE wireless communications letters, 2024, 13(5): 1315 – 1319. DOI: 10.1109/LWC.2024.3368558
[47] Cui Y M, Guo J J, Cao Z, et al. Lightweight neural network with knowledge distillation for CSI feedback [J]. IEEE transactions on communications,

2024, 72(8): 4917 – 4929. DOI: 10.1109/TCOMM.2024.3377724
[48] Zhou T, Liu X P, Xiang Z W, et al. Transformer network based channel prediction for CSI feedback enhancement in AI-native air interface [J]. IEEE transactions on wireless communications, 2024, 23(9): 11154 – 11167. DOI: 10.1109/TWC.2024.3379123
[49] Yan J T, Chen T, Xie B, et al. Hierarchical federated learning: architecture, challenges, and its implementation in vehicular networks [J]. ZTE communications, 2023, 21(1): 38 – 45. DOI: 10.12142/ZTECOM.202301005
[50] Xu W, Huang Y M, Wang W, et al. Toward ubiquitous and intelligent 6G networks: from architecture to technology [J]. Science China information sciences, 2023, 66(3): 130300. DOI: 10.1007/s11432-023-3704-8
[51] Lin X Q. An overview of the 3GPP study on artificial intelligence for 5G new radio [PP/OL]. arXiv (2023-08-10) [2024-08-20]. https://doi. org/10.48550/arXiv.2308.05315
[52] Guo J J, Wen C K, Jin S, et al. AI for CSI feedback enhancement in 5G-advanced [J]. IEEE wireless communications, 2024, 31(3): 169 – 176. DOI: 10.1109/MWC.010.2200304
[53] Sarieddeen H, Alouini M S, Al-Naffouri T Y. Terahertz-band ultra-massive spatial modulation MIMO [J]. IEEE journal on selected areas in communications, 2019, 37(9): 2040 – 2052. DOI: 10.1109/JSAC.2019.2929455
[54] Wang K Y, Gao Z, Chen S, et al. Knowledge and data dual-driven channel estimation and feedback for ultra-massive MIMO systems under hybrid field beam squint effect [J]. IEEE transactions on wireless communications, 2024, 23(9): 11240 – 11259. DOI: 10.1109/TWC.2024.3380638
[55] Balatsoukas-Stimming A, Studer C. Deep unfolding for communications systems: a survey and some new directions [C]//IEEE International Workshop on Signal Processing Systems (SiPS). IEEE, 2019: 266 – 271. DOI: 10.1109/sips47522.2019.9020494
[56] Deng L T, Z Y K. Deep learning-based semantic feature extraction: a literature review and future directions [J]. ZTE communications, 2023, 21 (2): 11 – 17. DOI: 10.12142/ZTECOM.202302003
[57] Shi G M, Xiao Y, Li Y Y, et al. From semantic communication to semantic-aware networking: model, architecture, and open problems [J]. IEEE communications magazine, 2021, 59(8): 44 – 50. DOI: 10.1109/MCOM.001.2001239
[58] Guo J J, Wen C K, Jin S. Deep data hiding-based CSI feedback overhead elimination: An initial investigation [C]//IEEE International Conference on Communications. IEEE, 2022: 5347 – 5352. DOI: 10.1109/ICC45855.2022.9839120
[59] Carpi F, Venkatesan S, Du J F, et al. Precoding-oriented massive MIMO CSI feedback design [C]//International Conference on Communications. IEEE, 2023: 4973 – 4978. DOI: 10.1109/ICC45041.2023.10278955

Biographies

Lu Zhaohua received his BS degree in electrical engineering and PhD degree in signal processing from Tianjin University, China in 2001 and 2006, respectively. Since 2006, he has been engaged in mobile communication physical layer technology at ZTE Corporation, including MIMO, interference control, artificial intelligence, etc. He has published more than 30 papers and held over 200 authorized patents.

Yi Chenyang received her BS degree in electrical engineering from Southeast University, China in 2020. She is currently working toward her PhD degree with the School of Information Science and Engineering, National Mobile Communications Research Laboratory, Southeast University. Her current research interests include massive MIMO, mmWave communications, and artificial intelligence for wireless communications.

Wu Jie received his BS degree in electrical engineering from Southeast University, China in 2022, where he is currently pursuing his MS degree in communication and information engineering. His recent research interests include deep learning for CSI compression and feedback in wireless communications.

Shao Bo received his BS degree in electrical engineering from Xidian University, China in 2023. He is currently pursuing his MS degree in communication and information engineering at Southeast University, China. His recent research interests include deep learning for CSI compression and feedback in wireless communications and massive MIMO systems.

Xu Wei (wxu@seu.edu.cn) received his BS degree in electrical engineering and his MS and PhD degrees in communication and information engineering from Southeast University, China in 2003, 2006, and 2009, respectively. Between 2009 and 2010, he was a post-doctoral research fellow with the Department of Electrical and Computer Engineering, University of Victoria, Canada. He is currently a professor at the National Mobile Communications Research Laboratory, Southeast University. He was an adjunct professor of the University of Victoria, Canada from 2017 to 2020, and a distinguished visiting fellow of the Royal Academy of Engineering, UK in 2019. He has co-authored over 100 refereed journal papers in addition to 36 domestic patents and four US patents granted. His research interests include information theory, signal processing and machine learning for wireless communications. He is currently an editor of *IEEE Transactions on Communications* and a senior editor of *IEEE Communications Letters*. He received the Best Paper Awards from a number of prestigious IEEE conferences including IEEE Globecom/ICCC, etc. He received the Science and Technology Award for Young Scholars of the Chinese Institute of Electronics in 2018. He is an IEEE Fellow and IET Fellow.

New Member of ZTE Communications Editorial Board

Wang Ling received his BSc, MSc, and PhD degrees in electronic engineering from Xidian University, China in 1999, 2002, and 2004, respectively. From 2004 to 2007, he worked at Siemens and Nokia Siemens Networks, Beijing. Since 2007, he has worked at Northwestern Polytechnical University (NPU), China, where he was promoted to Professor in 2012. Currently, he serves as the Vice President of NPU. His research interests include spectrum sensing, array processing, smart antennas, and cognitive radio. He has published over 100 research papers and holds more than 60 authorized invention patents.