Quantization Adaptor for Bit-Level Deep Learning-Based Massive MIMO CSI Feedback

Xudong Zhang, Zhilin Lu, Rui Zeng, and Jintao Wang, Senior Member, IEEE

Abstract—In massive multiple-input multiple-output (MIMO) systems, the user equipment (UE) needs to feed the channel state information (CSI) back to the base station (BS) for the following beamforming. But the large scale of antennas in massive MIMO systems causes huge feedback overhead. Deep learning (DL) based methods can compress the CSI at the UE and recover it at the BS, which reduces the feedback cost significantly. But the compressed CSI must be quantized into bit streams for transmission. In this paper, we propose an adaptor-assisted quantization strategy for bit-level DL-based CSI feedback. First, we design a network-aided adaptor and an advanced training scheme to adaptively improve the quantization and reconstruction accuracy. Moreover, for easy practical employment, we introduce the expert knowledge of data distribution and propose a pluggable and cost-free adaptor scheme. Experiments show that compared with the state-of-the-art feedback quantization methods, this adaptor-aided quantization strategy can achieve better quantization accuracy and reconstruction performance with less or no additional cost.

Index Terms—Massive multiple-input multiple-output (MIMO), channel state information (CSI) feedback, deep learning, quantization adaptor, bit-level feedback.

I. INTRODUCTION

MASSIVE multiple-input multiple-output (MIMO) has been regarded as a promising technology to improve the spectrum and energy efficiency for 5G and beyond systems [1]. In massive MIMO systems, the base station (BS) needs to obtain the downlink channel state information (CSI) for beamforming to exploit the potential benefits of MIMO. However, in frequency division duplexing (FDD) systems, the downlink CSI has to be estimated at the user equipment (UE) and then fed back to the BS due to the lack of channel reciprocity [2]. The feedback overhead is unacceptable in massive MIMO systems due to the large scale of antennas. Thus, CSI compression before feedback is especially necessary.

The traditional methods based on compressed sensing (CS) [3] require the CSI matrices to be sparse enough. However, the practical system may not fully satisfy this requirement especially when the compression ratio is large. With the development of artificial intelligence (AI) and deep learning (DL), researchers have been inspired to use neural networks to solve this problem. CsiNet [4] is the first to introduce the DL-based auto-encoder [5] architecture into CSI feedback and shows its overwhelming superiority against traditional CS methods.

The DL-based auto-encoder mainly consists of an encoder and a decoder network. The CSI matrices estimated at the UE are first pre-processed for feedback. Then the encoder network extracts the features and compresses the CSI into a low-dimensional codeword. The codeword is transmitted back to the BS through the digital communication channel. At the BS, the received codeword is recovered to the CSI matrix by the decoder network. After some post-processing such as 2D-DFT and zero filling, the original CSI matrix is reconstructed.

In digital communication systems, signals must be transformed into bit streams for feedback. After compression by the encoder, the codeword consists of full precise float 32 (FP32) numbers. Direct transmission of the float numbers will cause unbearable overhead to the feedback channel. Thus, the codeword should be quantized first into low bit width before feedback. However, the quantization module will introduce quantization errors to the codeword and influence the reconstruction accuracy of the decoder. Specific optimization for quantization should be introduced to improve the system performance.

Therefore, to reduce the impact of quantization, we propose an adaptor-aided quantization strategy in this paper. The adaptor adjusts values in the codeword and aims at reducing the information loss brought by quantization. With more precise codewords, the reconstruction process at the decoder can be improved. We design two types of adaptors, namely network-based adaptor and expert knowledge-based adaptor. Notably, [6] proposes the first network-based adaptor, using an offset network before the decoder to improve the quantization accuracy. However, the network adaptor in [6] is too heavy for the resource-sensitive feedback task. And its training strategy can be further optimized for better performance.

More specifically, we first design a lightweight network-aided adaptor strategy which achieves better performance with less computation cost. Furthermore, we propose a pluggable and cost-free adaptor based on expert knowledge for more practical deployment. The main contributions of this paper are listed below.

- We design two lightweight adaptor network structures which can achieve excellent performance under indoor and outdoor scenarios. The networks cause less or no additional
We propose a cost-free expert knowledge-based adaptor considering the practical deployment. We analyze the characteristics of data distribution and use them as expert knowledge to design an adaptor to adjust the data distribution in the codeword. The adaptor brings no extra cost and modification to the original network and shows great performance and robustness under different channels.

Both types of adaptors also exhibit good generality across different compression network structures.

A novel advanced training scheme is designed to improve the learning ability of the adaptor networks. The training stages and schedulers are meticulously adjusted to fit the network to the task. With the advanced training strategy, the quantization adapting is conducted more sufficiently and the whole feedback performance is improved.

The rest of this paper is arranged as follows. Section II lists the related works and Section III introduces the system model for bit-level DL-based CSI feedback. Section IV describes the details of the adaptor design, including the network-aided adaptor structure and the pluggable adaptor design. Section V presents the advanced training scheme for adaptor networks. The numerical results and analysis are given in Section VI and the conclusion is drawn in Section VII.

II. RELATED WORK

After CsiNet in [4] introduces deep learning into the CSI feedback task, later researchers make much effort to improve the performance of this scheme with advanced network structures and training mechanisms. CRNet [7] proposes a cosine learning rate scheduler which is able to effectively boost training. The multi-resolution mechanism is explored in many works [7], [8], [9], [10] with parallel feature extraction paths. Based on the convolutional network, CsiNet+ proposed in [11] optimizes the convolutional kernels and increases its receptive field. CLNet [12] designs a forged complex-valued input layer and increases the feedback performance. Besides, [13] introduces the deep recurrent network to exploit the channel correlations which further benefits the CSI compression. References [14], [15] utilize the famous transformer architecture [16] and obtain performance improvement. Moreover, as a widely used method in deep learning, the attention mechanism is also introduced to increase the ability of feature extraction in [17], [18]. In this way, different parts of CSI matrices can be paid different attention to, and features of higher importance tend to be retained and transmitted.

In addition to performance improvement, the practical employment of the DL-based feedback scheme is promoted by many researchers. Deployment cost is an important factor in practical applications and many works try to reduce it. The elastic feedback scheme is proposed in ACRNet [19] to flexibly adjust the network scale according to different resource limitations. ShuffleCsiNet [20] replaces the vanilla convolutional layers with shuffle layers in [21] and reduces the parameters. Network quantization is also explored in references [19], [22], [23] to reduce the parameter storage cost.

In the meanwhile, codeword quantization is also critical to deploying the DL-based system with acceptable transmission overhead. Uniform quantization is utilized in [24], [25], [26] to quantize the float numbers in codewords into bit streams. [27] introduces the A-law quantizer and tests the influence of different values of hyper-parameter A on the quantization performance. In [6], researchers study the codeword data distribution and prove the effectiveness of the $\mu$-law quantizer. It also designs an offset network to compensate for the quantization errors. [28] equips the $\mu$-law quantizer into its system and improves the feedback performance through training. Moreover, the CQNet proposed in [29] designs an end-to-end training framework, in which a trainable forward quantization weight vector is learned to customize the quantization intervals for each element in the compressed codeword. It further includes the $\mu$-law compander and formulates an enhanced type of quantization module, named $\mu$-CQNet. Focusing on the non-differentiable problem brought by the quantization module, [30] designs a differentiable function to approximate the gradients.

There are many other quantization methods introduced to improve the bit-level system performance. [31] combines the quantization and channel coding into the design and proposes an entropy coding-based quantization scheme. Furthermore, a fully-convolutional CSI compression architecture with quantization and entropy coding blocks is proposed in [32]. The probability distributions of data are predicted by the network and a good rate-distortion performance is achieved. A clustering algorithm-based quantization method is proposed in [33]. It utilizes the k-means method with vector quantization to divide codewords into different quantization intervals. [34] proposes a nested dropout layer to define an importance ordering and distribute different bit widths to different codewords according to their importance. However, the quantization strategies based on vector quantization need to store the codebook at both the UE and BS sides, which brings extra cost.

In summary, codeword quantization is a critical problem considered in promoting the DL-based feedback system into practical deployment. Current quantization strategies based on quantizer design or the offset network still need to be further optimized to reduce the additional cost and improve performance. Therefore, this paper proposes an adaptor-aided quantization strategy that achieves better performance with less or even no additional cost.

III. SYSTEM MODEL

A. Massive MIMO-OFDM System

In this paper, we consider a single-cell massive MIMO-OFDM system with $N_t >> 1$ antennas at the BS and a single antenna at the UE. Orthogonal frequency division multiplexing (OFDM) is employed in this system with $N_c$ sub-carriers. The received signal at the $n^{th}$ sub-carrier $y_n \in \mathbb{C}$ can be derived as follows:

$$y_n = h_n^H s_n x_n + z_n,$$

where $h_n \in \mathbb{C}^{N_t \times 1}$, $s_n \in \mathbb{C}^{N_c \times 1}$, $x_n \in \mathbb{C}$, and $z_n \in \mathbb{C}$ denote the channel vector, preceding vector, data-bearing symbol, and additive noise at the $n^{th}$ sub-carrier, respectively. $(\cdot)^H$ represents
conjugate transpose. Let \( \hat{H} = [h_1, \ldots, h_{\tilde{N}_c}]^H \in \mathbb{C}^{\tilde{N}_c \times N_t} \) be the CSI matrix, which is essential for the BS to design the precoding vectors \( \{s_1, \ldots, s_{N_t}\} \). In the FDD system, the UE has to return the downlink CSI \( \hat{H} \) to the BS for lack of reciprocity. However, the \( 2\tilde{N}_cN_t \) scale of real elements in \( \hat{H} \) is unacceptably large for bit-level digital transmission and demonstrates the necessity of CSI compression.

With the channel sparsity in the angular-delay domain, the feedback overhead can be first reduced as follows:

\[
H = F_d H F_a^H, \tag{2}
\]

where \( F_d \) and \( F_a \) are the \( \tilde{N}_c \times \tilde{N}_c \) and \( N_t \times N_t \) DFT matrices, respectively. For the angular-delay domain channel matrix \( H \), only the first \( N_c \) rows contain large values. The values in the rest of the rows are near zero and can be left out without much loss of information. We use \( H_t \) to denote the first \( N_t \) rows of \( H \), which only contains \( 2N_cN_t \) elements but is still large for feedback. Therefore, DL-based CSI feedback networks are designed to further compress the CSI matrix.

### B. DL-Based CSI Feedback

In this paper, we consider the bit-level DL-based CSI feedback network structure as shown in Fig. 1. A widely referenced open-source network named CRNet [7] is utilized for the encoder and decoder. The CSI matrix \( \hat{H} \) is first transferred into the angular-delay domain through DFT as (2) and then reduced to \( H_t \) to encode. \( H_t \) is then feature-extracted and compressed into an \( M \times 1 \) codeword, where each element will be quantized into \( B \) bits to be transmitted back to the BS. It can be described as (3).

\[
v_q = Q(v) = Q(f_{en}(H_c, \Theta_{en})), \tag{3}
\]

where \( f_{en}(\cdot) \) and \( Q(\cdot) \) denote the encoding and quantizing process, and \( H_c, v, v_q, \) and \( \Theta_{en} \) are the CSI matrix, codeword, quantized codeword, and parameters of the encoder network, respectively. The compression ratio (CR) can be calculated as follows:

\[
CR = \frac{2N_cN_t}{M}. \tag{4}
\]

Once the BS receives the quantized codeword \( v_q \), the dequantization module and decoder will reconstruct the channel matrix \( \hat{H} \). The recovered channel matrix can be described as follows:

\[
\hat{H}_c = f_{de}(D(v_q), \Theta_{de}), \tag{5}
\]

where \( f_{de}(\cdot) \) and \( D(\cdot) \) represent the decoding and dequantization process. \( \Theta_{de} \) represents the parameters of the decoder network.

### C. Quantization Module

In this paper, we utilize the \( \mu \)-law quantizer which proved to perform well in CSI feedback tasks [6]. The \( \mu \)-law quantizer is a kind of non-uniform quantization. Through a non-decreasing companding function \( \Phi: (-1, 1) \rightarrow (-1, 1) \), it provides narrower quantization intervals near zero and wider intervals near one, thus performing well in operating zero-centralized data. Experiments show that the compressed codewords in CSI feedback tasks are concentrated near zero. The companding function in \( \mu \)-law can be described as follows:

\[
\Phi(x) = \text{sign}(x) \frac{\ln(1 + \mu |x|)}{\ln(1 + \mu)}, \tag{6}
\]

where \( x \) is the normalized input and \( \mu \) is a parameter to change the degree of companding and should be adjusted according to the distribution of given data.

When operating the \( \mu \)-law quantizer, the normalized data is first put into the companding function \( \Phi \) and quantized by a uniform quantizer, then expanded through function \( \Phi^{-1} \) to get the final quantized value. In practical deployment, the companding function \( \Phi \) can be approximated by the segmented polyline function, which greatly reduces the computing complexity. In this work, we use this \( \mu \)-law quantizer as a basic quantization method and further propose an adaptor-aided quantization strategy to adjust the codewords for better quantization and reconstruction performance.

## IV. CSI Feedback With Quantization Adaptor

### A. Bottleneck FC Adaptor Design

In this section, we propose a bottleneck FC adaptor network to improve the quantization performance. The adaptive learning ability of neural networks is mainly exploited. The adaptor can learn adaptively through training to adjust the distribution of the data and complement the quantization error. As to the structure design, fully-connected layers are used as the main feature-extracting module. A residual connection structure in [35] is used to make the adaptor learn and bridge the gap between original codewords and quantized codewords. In the meanwhile, the residual structure can also avoid overfitting.

Considering the practical employment, the FC layers occupy most of the parameter overhead. Thus we utilize the bottleneck
structure, in which the middle channels are fewer than the input and output ones. This structure can alleviate the heavy overhead brought by FC layers. Specifically, a 1/8 bottleneck is used for the bottleneck adaptor, and the parameter cost is reduced from $3M^2$ to $3M^2/4$ compared with the OffsetNet benchmark in [6]. The detailed structure is shown in Fig. 2.

In the feedback process, the received quantized codeword is first dequantized and then post-processed by the adaptor network. The characteristics of quantization noise can be learned by the adaptor through sufficient training so that the errors in codewords are compensated. There is less information loss in the codewords passed into the decoder and the reconstruction performance is then improved.

B. Parallel Bottleneck FC Adaptor Design

In outdoor scenarios, the channel features have more diversity of scales and the matrices can be denser. Thus, the parallel branch structure is more suitable to capture the features. Besides, inspired by [7], [8] which introduce multi-resolutions skills into CSI feedback with different convolutional kernel sizes, we design a parallel bottleneck FC adaptor. In this adaptor, we add a parallel branch with a different bottleneck scale to the original structure. Different sizes of bottlenecks can capture features of different scales and further make the recovery more accurate. This detailed structure is shown in Fig. 3.

Despite the additional branch, the total parameter cost can still be saved a lot compared with the OffsetNet benchmark [6] owing to the bottleneck structure. Specifically, with two bottlenecks of 1/8 and 1/16 sizes, the total parameters are reduced from $3M^2$ to $3M^2/8$, which reveals the validity of bottleneck structures in reducing parameters.

C. L1 Adaptor Design

Although the two bottleneck adaptor structures above can greatly reduce the deployment overhead, extra costs still exist which is not friendly to the resource-sensitive environment. Therefore, we further propose a totally cost-free adaptor design called L1 adaptor. This idea mainly comes from the knowledge that quantization performance depends on whether the characteristics of the quantizer match the data distribution. For the $\mu$-law quantizer, the quantization performance is better when the quantized data distribute close to zero. Therefore, we design the L1 adaptor to gather the compressed CSI codewords close to zero. As shown in Fig. 4, we add an L1 norm to the loss function. The gathering and sparsing abilities of the L1 regularization are exploited to narrow the values near zero and make their distribution match the $\mu$-law quantizer better. In this way, the distribution of codewords is better aligned with the demand of the $\mu$-law quantizer and better quantization performance is achieved. The whole loss function can be described as follows:

$$L(\Theta_{en}, \Theta_{de}) = \frac{1}{N} \sum_{i=1}^{N} ||\hat{H}_{ci} - H_{ci}||_2^2 + ||\vec{v}^i||_1,$$  \hspace{1cm} (7)
L1 adaptor only changes the training process and brings no adjustments to the network structure. Thus, it is plug-and-play and can be easily applied to a wide range of feedback networks. It does not bring any additional parameters to the encoder or decoder and is cost-free for inference. Compared with methods of OffsetNet [6] or adaptor networks, it has more advantages under resource-limited conditions.

V. TRAINING STRATEGY FOR NETWORK ADAPTOR

A. Two-Stage Training Pipeline Design

In this section, we introduce a two-stage training pipeline to improve the reconstruction performance and the quantization accuracy.

The first stage is to train the whole network parameters with the mean square error (MSE) loss as follows:

$$L_{MSE}(\Theta) = \frac{1}{N} \sum_{i=1}^{N} ||\hat{H}_{ci} - H_{ci}||^2_2,$$  \hspace{1cm} (8)

where $\Theta = \{\theta_{en}, \Theta_{NA}, \theta_{de}\}$ and $N$ represent the parameters of the whole network and the number of training samples. $\hat{H}_{ci}$ and $H_{ci}$ represent the recovered CSI matrix and original CSI matrix.

With this stage, the whole network is first trained to capture and compress the channel features. Through sufficient training, the network will converge to a stage with enough compression and reconstruction ability. Further performance optimization can be conducted on such a basis.

The second stage is mainly about training the adaptor network to adjust the data distribution for quantization error compensation. The encoder parameters are fixed and only the adapter and decoder parameters are updated in this stage. In this way, the learned information from the first stage can be retained as much as possible, avoiding degradation of reconstruction performance. The quantization error loss is added to the MSE loss as a regularization term against the quantization errors. In this stage, the whole loss function can be described as follows:

$$L(\Theta_{NA}, \Theta_{de}) = \frac{1}{N} \sum_{i=1}^{N} ||\hat{H}_{ci} - H_{ci}||^2_2 + \alpha ||v_i^c - v_i^{\tilde{q}}||^2_2$$  \hspace{1cm} (9)

where $\Theta_{NA}$ and $\Theta_{de}$ represent the parameters of the adapter and decoder. $N$ is the number of training samples. $\hat{H}_{ci}, H_{ci}, v_i^c,$ and $v_i^{\tilde{q}}$ represent the recovered CSI matrix, original CSI matrix, compressed codeword, and quantized compressed codeword of the $i^{th}$ sample, respectively. The parameter $\alpha$ is the regularization weight to adjust the influence of quantization improvement. Its scheduler design is described in the following sub-section.

After the second stage of training, the network can achieve better feedback performance through the quantization error compensation learned by the adaptor network. Experiments show that without the first training stage, the network can not converge to a well-trained state with sufficient compression ability.

B. Quantization Adapting Factor Scheduler Design

The quantization adapting factor $\alpha$ is a hyper-parameter to adjust the impact of quantization loss. An overlarge factor will destroy the reconstruction performance, while an excessively small factor can not fully improve the quantization performance. In this section, we design a scheduler for $\alpha$, making the quantization adapting factor vary through the training process and work sufficiently in the middle of the training.

As shown in Fig. 5, $\alpha$ is first linearly increased from 0.01 to 0.05, and then linearly decreased to zero in the end. The factor starts from a relatively small value to keep the performance of the original network. As the factor is increasing, the quantization error regularization term becomes larger and makes the network converge in the direction of reducing quantization errors. After the peak, $\alpha$ starts to decrease and a chance is provided for the network to fine-tune and get better performance on the basis of the previous adjustment about quantization. In this way, the quantization adapting can fully influence the training process and bring benefits to CSI compression and reconstruction. Ablation studies about the scheduler design are presented in the next section.

VI. SIMULATION RESULTS AND ANALYSIS

A. Simulation Settings

We follow the channel model with default settings in COST2100 [36], which is a widely used dataset in CSI feedback. We consider both the indoor scenario at 5.3 GHz and the outdoor scenario at 300 MHz. $N_t = 32$ transmit antennas in uniform linear array (ULA) at the BS are considered. We take $N_c = 1024$ sub-carriers for OFDM and only the first $N_c = 32$ rows of CSI matrix are reserved. 150,000 independently generated channel samples are split into three parts. The training, validation, and test dataset contain 100,000, 30,000, and 20,000 matrices, respectively. The batch size is set to 200.

Experiments are implemented using the Pytorch library. We adopt the same encoder and decoder structure and cosine annealing scheduler for the learning rate proposed in CRNet [7]. For the network-aided adaptor scheme, the network is trained for 1000 epochs for the first stage and 400 epochs for the second stage.
stage. For the L1 adaptor scheme, the network is only trained
for 1000 epochs.

We choose the OffsetNet in [6] and the \( \mu \)-CQNet in [29] as our
baselines. OffsetNet is a three-layer FC module with a residual
connection while \( \mu \)-CQNet is a \( \mu \)-law enhanced weighting quantization block. For a fair comparison, we deploy these methods
on the same autoencoder architecture CRNet [7]. We consider
the CR of 4, 8, and 16 and the quantization bit number \( B \) of 4 and 6. The parameter \( \mu \) in the companding function of the
\( \mu \)-law quantizer is set to 50 according to [6], in which researchers
select 50 as the best value through adjustment based on CSI
distribution maps.

For clearer representation, we provide some notations for
different methods here. We note the pure \( \mu \)-law quantizer, offset
network method [6], and \( \mu \)-CQNet [29] as \( \mu \)-law, OffsetNet,
and \( \mu \)-CQNet, respectively. Our proposed bottleneck FC adaptor
scheme, parallel bottleneck FC adaptor scheme, and L1 adaptor
scheme are denoted as BottleFC, ParaBottleFC, and L1Adaptor,
respectively.

### B. Quantization Performance

We propose the quantization performance of different methods
in this part. For evaluation, we compute the quantization signal-to-noise ratio (QSNR) as the criterion,

\[
QSNR = \mathbb{E}\left[ \frac{||v||_2^2}{||v - v_q||_2^2} \right],
\]

(10)

where \( v \) and \( v_q \) denote the compressed CSI codeword before
and after quantization. \( || \cdot ||_2 \) denotes the Euclidean norm.

Table I lists the QSNR for different methods under different scenarios. We can conclude that the adaptor scheme
achieves better performance than baselines in most experimental
scenarios. Through data adapting, the quantization errors can
be compensated, leading to higher QSNR. Moreover, the L1 adaptor
achieves better quantization performance than network adaptors in most scenarios. The main reason is that the L1
adaptor strictly constrains the data distribution ahead of the
quantizer. Such early-stage data restriction is more beneficial
to the QSNR performance compared with the network-based
error compensation after the vanilla quantization process.

For further clarification, we present the cumulative distribution functions (CDFs) of compressed CSI codewords with or
without the L1 adaptor. The quantization bit width is set to 6

![Fig. 6. Rate-distortion curves of different quantization strategies under CR=4.](image)

and functions under different scenarios are provided as shown
in Fig. 7. The x-axis denotes the compressed CSI value and the y-axis denotes the cumulative distribution probability of the corre-
sponding value. The CDFs mainly increase at zero which proves
the zero-centralized distribution of CSI codewords. Compared
with outdoor scenarios, CDFs in indoor scenarios are steeper
near zero which means CSI codewords in indoor scenarios
distribute more concentratedly. Drawn from the figure, we can
conclude that the codewords distribute more close to zero with
the L1 adaptor. This further makes the data distribution and the
\( \mu \)-law quantizer more compatible, thereby bringing quantization
performance gain.

### C. Reconstruction Performance

The whole system is evaluated with the reconstruction per-
formance of CSI feedback. We consider the distance between
the original CSI matrix \( \mathbf{H}_c \) and the reconstructed \( \hat{\mathbf{H}}_c \) with
normalized mean square error (NMSE) defined as follows:

\[
NMSE = \mathbb{E}\left[ \frac{||\mathbf{H}_c - \hat{\mathbf{H}}_c||_2^2}{||\mathbf{H}_c||_2^2} \right].
\]

(11)

The main NMSE performance comparison between different
methods is given in Table II. We add the Lloyd-Max quantization
method as another baseline, in which the Lloyd-Max quantizer
is performed on the codeword compressed by a well-trained net-
work. Note that NQ stands for the feedback with no quantization
module. For a clearer comparison, we provide the rate-distortion
curves of different methods in Fig. 6.
According to the results, the adaptor scheme achieves better performance than baselines in all scenarios. The better quantization accuracy brought by adaptors promotes the decoding and recovery process. Moreover, according to Table II, the effect of reconstruction is more sensitive to the quantization bit width and strategy in indoor scenarios than in outdoor scenarios. This is because the complex features of outdoor channels perform as the dominant challenge of feature extracting and compressing in feedback.

Besides, the network-aided adaptor scheme outperforms the L1 adaptor although the L1 adaptor achieves higher QSNR performance. Compared with hard constraints brought by the L1 adaptor, the learnable network adaptor can balance its attention paid to the quantization improvement and recovery performance. Thanks to the two-stage training pipeline and $\alpha$ scheduler, the network adaptor is able to optimize the feedback NMSE better under the pressure of quantization regularization. As we can see, the proposed network adaptors achieve the best performance under both indoor and outdoor scenarios.

What’s more, it can be seen that the performance gap between different quantization bits is larger for Lloyd-Max than for other methods. The reason is that Lloyd-Max can only perform on the quantization period. Though it can achieve great performance with multiple quantitative bits (e.g. 6 bits), it will suffer an inevitable performance degradation when the bits are decreasing. By contrast, other learning-based methods can adaptively adjust the data to compensate for the quantization loss in the reconstruction stage. This further demonstrates the effectiveness of our proposed methods.

It is interesting that the performances of our methods can approach or even exceed the NQ performance under certain conditions. This is because the L1 adaptor can resist overfitting and reduce the inadequacy of the original training. With network adaptors, the parameter space is slightly but finely enlarged to enable a better solution. This explains the outperformance of our methods.

Finally, we investigate into the compression ratios of the bottleneck structure in network adaptors and present the results in Table III. It is shown that our proposed compression ratios of the bottleneck structure achieve the best performance under corresponding scenarios. It can also be concluded that the parallel structure outperforms the single bottleneck path in outdoor scenarios, which reflects the complex features of the outdoor channel. The results further illustrate the importance of capturing features in different scales in the CSI compression and quantization task under outdoor scenarios.
Fig. 8. Validation NMSE(dB) performance of the $\mu$-law quantizer and our proposed adaptor-based methods.

### Table IV

**Complexity Comparison of Different Quantization Strategies**

| Complexity | CR | OffsetNet [6] | Parabottle/FC | Bottle/FC | $\mu$-CQN [29] | $\mu$-law & L1 Adaptor |
|------------|----|--------------|--------------|------------|----------------|---------------------|
| FLOPs      | 4  | 6.113M       | 3.485M       | 3.435M     | 5.346M         | 5.346M             |
|            | 8  | 4.534M       | 3.582M       | 3.435M     | 4.335M         | 4.335M             |
|            | 16 | 3.863M       | 3.820M       | 3.917M     | 3.815M         | 3.815M             |
| Parameter  | 4  | 2.891M       | 2.505M       | 2.169M     | 2.103M         | 2.103M             |
|            | 8  | 2.253M       | 2.079M       | 1.973M     | 1.860M         | 1.860M             |
|            | 16 | 1.579K       | 1.569K       | 1.553K     | 1.469K         | 1.469K             |

### D. Convergence and Complexity

The convergence performance is shown in Fig. 8. Despite the slightly heavier fluctuations of the network adaptor, both adaptor methods achieve stable convergence with enhanced NMSE performance at later training epochs. Moreover, at the beginning of the second stage for the network adaptor, the network suffers from a sudden performance degradation but quickly converges to an optimal state that achieves a lower NMSE. This reveals the training robustness of the proposed adaptor-based approaches.

The complexity comparison is listed in Table IV. Except for the proposed network adaptors, the L1 adaptor outperforms all the other methods with the lowest complexity, making it a practical cost-free upgrade for the quantization-aware feedback system. In the meanwhile, the two network-based adaptors achieve the best reconstruction performance with a slight complexity increase over the original system but much less than OffsetNet [6]. This proves the superiority and high complexity efficiency of our network adaptor design.

Considering the low complexity, our methods are friendly to realistic implementation. Besides, the L1 adaptor brings no changes to the original compression network structure and can be easily utilized on different structures. Furthermore, multiple complexity reduction algorithms such as parameter quantization and weight pruning can be easily utilized in our methods in ways like [23] and [22], thus improving the practicality of our proposed methods.

### E. Effectiveness of Quantization Adapting Factor Scheduler

In this part, some ablation studies are presented to prove the effectiveness of the quantization adapting factor scheduler. The scheduler controls the variation of the adapting factor and further adjusts the weights of quantization error regularization. Ablation studies show that the proposed scheduler which increases at first and then decreases works better because such a quantization adaptor mainly limits the network convergence in the middle of the training.

With thorough experiments, we discover that the feedback performance tends to be better when $\alpha$ lies in the range from 0 to 0.05. Based on this observation, we further design three types of schedulers.

We consider three types of schedulers. First, a constant adapting factor of 0.03 is introduced as a straightforward benchmark scheme. Then, a monotonically decreasing scheduler which ranges from 0.05 to 0 is tested to mainly restrict the network convergence at the early training stage. Finally, a piecewise linear scheduler as Fig. 5 shows is proposed with peak $\alpha = 0.05$ reached in the middle of the training. Compared with the previous linearly decreasing scheduler, we wish to see if the network can further benefit from freer early-stage training.

Results from Fig. 9 show that the piecewise linear scheduler provides the best quantization and reconstruction performance. This demonstrates the effectiveness of applying adapting restrictions mainly in the middle stage of training. It also suggests that too strong constraints in the early stage of training may lead to undermining the abilities learned from the previous stage, thereby reducing the performance of training.

### F. Generality on Different Network Structures

All the above-mentioned results are experimented on CR-Net [7] structure. However, the generality to different network structures is important for demonstrating the effectiveness of our methods. In this section, both types of proposed adaptors are tested on three different CSI compression networks including CsiNet [4], ACRNet [19], and CLNet [12]. The performances are listed in Table V.

Results show that the proposed quantization adaptor achieves stable performance improvement under different network structures. This is because the adaptor approach is based on the channel distribution instead of network structures. Thus, it can be applied to different network architectures operating on the CSI feedback task under massive MIMO channels. This experiment illustrates the robustness and generality of our methods.

### G. Robustness of L1 Adaptor on Different Channels

The L1 adaptor design is derived from the data distribution and may be affected by the communication environment. So
Fig. 9. Performance comparison between different quantization adapting schedulers. (a) NMSE performance. (b) QSNR performance.

in this part, we evaluate the performance of the L1 adaptor on another dataset to prove its robustness and effectiveness. Wireless-Intelligence [37] is a public channel dataset library, from which the nrCDLChannel model following 3GPP TR 38.901 [38] standard is used for the CSI generation. Simulation results show that the L1 adaptor can still obtain solid performance gain when processing the clustered delay line (CDL) channel models.

1) Dataset Parameters: The main parameter settings for CDL channel model simulation are listed in Table VI.

Fig. 10 presents two CSI samples in the angular-delay domain of the CDL channel model. As shown in the samples, the channel also has sparsity in the angular-delay domain and only the first 32 non-zero rows are meaningful for the feedback task.

2) Performance: In order to examine the robustness of the pluggable L1 adaptor, we test it on CDL channel models. Simulation results in Table VII show that this adaptor can improve the QSNR and NMSE performance at the same time under different scenarios. This proves that the data distribution refinement driven by the proposed L1 adaptor takes effect in a robust way. In other words, our design is proved to be generally helpful under different channel models and various scenarios.

VII. Conclusion

In this paper, we proposed an adaptor-aided quantization strategy for DL-based CSI feedback in massive MIMO systems. A network adaptor scheme was presented with two lightweight
network structures. To further reduce the computation cost, we introduced expert knowledge of data distribution and designed a robust L1 adaptor which was totally cost-free and improved the quantization performance through distribution adapting. Besides, an advanced training pipeline was specially designed to boost the performance. A quantization adapting factor scheduler was introduced to further optimize the quantization process. Experiments show that both adaptor schemes achieved better quantization and reconstruction performance with lower cost and overhead. The proposed methods also showed great generality over different compression networks and could be easily applied to multiple structures.

REFERENCES

[1] 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.
[2] M. S. Sim, J. Park, C.-B. Chae, and R. W. Heath, “Compressed channel feedback for correlated massive MIMO systems,” J. Commun. Netw., vol. 18, no. 1, pp. 95–104, Feb. 2016.
[3] 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., 2012, pp. 492–497.
[4] 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.
[5] P. Vincent, H. Larochelle, Y. Bengio, and P. Manzagol, “Extracting and composing robust features with denoising autoencoders,” in Proc. 25th Int. Conf. Mach. Learn., 2008, pp. 1096–1103.
[6] 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. IEEE Glob. Conf. Signal Inf. Process., 2019, pp. 1–5.
[7] Z. Lu, J. Wang, and J. Song, “Multi-resolution CSI feedback with deep learning in massive MIMO system,” in Proc. IEEE Int. Conf. Commun., 2020, pp. 1–6.
[8] Z. Lu, J. Guo, G. Liu, H. Zheng, and J. Xue, “MRFNet: A deep learning-based CSI feedback approach of massive MIMO systems,” IEEE Commun. Lett., vol. 25, no. 10, pp. 3310–3314, Oct. 2021.
[9] Y. Xu, M. Zhao, S. Zhang, and H. Jin, “DFFCNet: Exploiting diverse channel features for massive MIMO CSI feedback,” in Proc. 13th Int. Conf. Wireless Commun. Signal Process., 2021, pp. 1–5.
[10] B. Cheng, J. Zhao, and Y. Hu, “Multi-scale and multi-channel networks for CSI feedback in massive MIMO system,” J. Commun. Comput., vol. 9, pp. 132–141, 2021.
[11] J. Guo, C.-K. Wen, S. Jin, and G. Y. Li, “Convolutional neural network-based multiple-rate compressive sensing for massive MIMO CSI feedback: Design, simulation, and analysis,” IEEE Trans. Wireless Commun., vol. 19, no. 4, pp. 2827–2840, Apr. 2020.
[12] S. Ji and M. Li, “CLNet: Complex input lightweight neural network designed for massive MIMO CSI feedback,” IEEE Wireless Commun. Lett., vol. 10, no. 10, pp. 2318–2322, Oct. 2021.
[13] 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.
[14] Y. Cui, A. Guo, and C. Song, “TransNet: Full attention network for CSI feedback,” in Proc. IEEE Trans. Wireless Commun. Lett., vol. 11, no. 5, pp. 903–907, May 2022.
[15] Y. Xu, M. Yuan, and M.-O. Pun, “Transformer empowered CSI feedback for massive MIMO systems,” in Proc. 30th Wireless Opt. Commun. Conf., 2021, pp. 157–161.
[16] A. Vasswani et al., “Attention is all you need,” in Proc. 31st Int. Conf. Neural Inf. Process. Syst., 2017, pp. 6000–6010.
[17] Q. Cai, C. Dong, and K. Niu, “Attention model for massive MIMO CSI compression feedback and recovery,” in Proc. IEEE Wireless Commun. Netw. Conf. (WCNC), 2019, pp. 1–5.
[18] X. Song et al., “SALDR: Joint self-attention learning and dense refine for massive MIMO CSI feedback with multiple compression ratio,” IEEE Wireless Commun. Lett., vol. 10, no. 9, pp. 1899–1903, Sep. 2021.
[19] Z. Lu, X. Zhang, H. He, J. Wang, and J. Song, “Binarized aggregated network with quantization: Flexible deep learning deployment for CSI feedback in massive MIMO systems,” IEEE Trans. Wireless Commun., vol. 21, no. 7, pp. 5514–5525, Jul. 2022.
[20] Z. Cao, W.-T. Shih, J. Guo, C.-K. Wen, and S. Jin, “Lightweight convolutional neural networks for CSI feedback in massive MIMO,” IEEE Commun. Lett., vol. 25, no. 8, pp. 2624–2628, Aug. 2021.
[21] Z. Lu, Z. Zhao, X. Zhou, M. Lin, and J. Sun, “ShuffleNet: An extremely efficient convolutional neural network for mobile devices,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2018, pp. 6848–6856.
[22] J. Guo, J. Wang, C.-K. Wen, S. Jin, and G. Y. Li, “Compression and acceleration of neural networks for communications,” IEEE Wireless Commun., vol. 27, no. 4, pp. 110–117, Aug. 2020.
[23] Z. Lu, J. Wang, and J. Song, “Binary neural network aided CSI feedback in massive MIMO system,” IEEE Wireless Commun. Lett., vol. 10, no. 6, pp. 1305–1308, Jun. 2021.
[24] C. Lu, W. Xu, S. Jin, and K. Wang, “Bit-level optimized neural network for multi-antenna channel quantization,” IEEE Wireless Commun. Lett., vol. 9, no. 1, pp. 87–90, Jan. 2020.
[25] 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.
[26] H. Chang, X. Liang, H. Li, J. Shen, X. Gu, and L. Zhang, “Deep learning-based bistream error correction for CSI feedback,” IEEE Wireless Commun. Lett., vol. 10, no. 12, pp. 2828–2832, Dec. 2021.
[27] G. Fan, J. Sun, G. Gui, H. Gacanin, B. Adebisi, and T. Ohtsuki, “Fully convolutional neural network-based CSI limited feedback for FDD massive MIMO systems,” IEEE Trans. Cogn. Commun. Netw., vol. 8, no. 2, pp. 672–682, Jun. 2022.
[28] Z. Wang, X. Zhang, and Y. Liu, “Deep learning based CSI compression and quantization with high compression ratios in FDD massive MIMO systems,” IEEE Wireless Commun. Lett., vol. 10, no. 10, pp. 2101–2105, Oct. 2021.
[29] Z. Liu, L. Zhang, and Z. Ding, “An efficient deep learning framework for low rate massive MIMO CSI reporting,” IEEE Trans. Commun., vol. 68, no. 8, pp. 4761–4772, Aug. 2020.
[30] X. Liang, H. Chang, H. Li, X. Gu, and L. Zhang, “Changeable rate and novel quantization for CSI feedback based on deep learning,” IEEE Trans. Wireless Commun., vol. 21, no. 12, pp. 10100–10114, Dec. 2022.
[31] S. Ravula and S. Jain, “Deep autoencoder-based massive MIMO CSI feedback with quantization and entropy coding,” in Proc. IEEE Glob. Commun. Conf., 2021, pp. 1–6.
[32] M. B. Mashhadi, Q. Yang, and D. Gündüz, “Distributed deep convolutional compression for massive MIMO CSI feedback,” IEEE Trans. Wireless Commun., vol. 20, no. 4, pp. 2621–2633, Apr. 2021.
[33] J. Shen, X. Liang, X. Gu, and L. Zhang, “Clustering algorithm-based quantization method for massive MIMO CSI feedback,” IEEE Trans. Wireless Commun., vol. 11, no. 10, pp. 2155–2159, Oct. 2022.
[34] V. Rizzello, M. Nerini, M. Joham, B. Clerckx, and W. Utschick, “Learning representations for CSI adaptive quantization and feedback,” 2022, arXiv:2207.06924.
[35] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 770–778.
[36] L. Liu et al., “The COST 2100 MIMO channel model,” IEEE Wireless Commun., vol. 19, no. 6, pp. 92–99, Dec. 2012.
[37] OPPO, “Wireless-intelligence,” 2022. Accessed: Oct. 21, 2022 [Online]. Available: https://wireless-intelligence.com/.
[38] 3GPP, “Study on channel model for frequencies from 0.5 to 100 GHz,” 3rd Gener. Partnership Project (3GPP), Sophia Antipolis, France, Tech. Rep. 38.901, Dec. 2019.

Xudong Zhang received the B.S. degree in 2022 from the Department of Electronic Engineering, Tsinghua University, Beijing, China, where he is currently working toward the Ph.D. degree with the Department of Electronic Engineering. His research interests include deep learning for wireless communications, massive MIMO, and integrated sensing and communications.
Zhilin Lu received the B.S. and Ph.D. degrees in electrical engineering from Tsinghua University, Beijing, China, in 2018 and 2023, respectively. He is currently working with Lab for High Technology with Tsinghua University, Beijing. His research interests include deep learning for wireless communications and spectrum sensing.

Rui Zeng received the B.S. degree in 2021 from the Department of Electronic Engineering, Tsinghua University, Beijing, China, where he is currently working toward the Ph.D. degree. His research interests include deep learning for channel equalization, data detection, and modulation recognition.

Jintao Wang (Senior Member, IEEE) received the B.Eng. and Ph.D. degrees in electrical engineering from Tsinghua University, Beijing, China, in 2001 and 2006, respectively. From 2006 to 2009, he was an Assistant Professor with the Department of Electronic Engineering, Tsinghua University. Since 2009, he has been an Associate Professor and Ph.D. Supervisor. He has authored or coauthored more than 100 journal and conference papers and holds more than 40 national invention patents. His research interests include space-time coding, MIMO, and OFDM systems. He is the Standard Committee Member of the Chinese national digital terrestrial television broadcasting standard.