CLNet: Complex Input Lightweight Neural Network Designed for Massive MIMO CSI Feedback

Sijie Ji and Mo Li, *Fellow, IEEE*

***Abstract*—The Massive Multiple Input Multiple Output (MIMO) system is a core technology of the next generation communication. With the growing complexity of CSI, CSI feedback in massive MIMO system has become a bottleneck problem. Recently, numerous deep learning-based CSI feedback approaches demonstrate their efficiency and potential. However, most existing methods improve accuracy at the cost of computational complexity by adding more advanced deep learning blocks. This letter presents a novel neural network CLNet tailored for CSI feedback problem based on the intrinsic properties of CSI. CLNet proposes a forged complex-valued input layer to process signals and utilizes spatial-attention to enhance the performance of the network. The experiment result shows that CLNet outperforms the state-of-the-art method by average accuracy improvement of 5.41% in both outdoor and indoor scenarios with average 24.1% less computational overhead. Codes are available at GitHub.[1]**

***Index Terms*—Massive MIMO, FDD, CSI feedback, deep learning, complex neural network, attention mechanism, lightweight model.**

I. Introduction

THE MASSIVE multiple-input multiple-output (MIMO) technology is considered one of the core technologies of the next generation communication system, e.g., 5G. By equipping a large number of antennas, the base station (BS) can sufficiently utilize spatial diversity to improve the channel capacity. Especially, by enabling beamforming, a 5G BS can concentrate signal energy to a specific user equipment (UE) to achieve higher signal-to-noise ratio (SNR), less interference leakage and hence, higher channel capacity. However, beamforming is possibly conducted by the BS only when it has the channel state information (CSI) of the downlink at hand [1].

In frequency division duplexing (FDD) mode that most contemporary cellular systems operate in, the channel reciprocity does not exist. Therefore, the UE would have to explicitly feed back the downlink CSI to the BS, and the pilot-aided training overhead grows quadratically with the number of transmitting antennas, which might overturn the benefit of Massive MIMO itself [2]. Thus, CSI compression is needed before the feedback to reduce the overhead.

Manuscript received July 3, 2021; accepted July 24, 2021. Date of publication July 27, 2021; date of current version October 7, 2021. This work was supported by the Singapore MOE Tier 2 under Grant T2EP20220-0011. The associate editor coordinating the review of this article and approving it for publication was S. Sugiura. *(Corresponding author: Sijie Ji.)*

The authors are with the School of Computer Science and Engineering, Nanyang Technological University, Singapore (e-mail: sijie001@e.ntu.edu.sg; limo@ntu.edu.sg).

Digital Object Identifier 10.1109/LWC.2021.3100493

[1]https://github.com/SIJIEJI/CLNet

Traditional compressive sensing (CS) based methods rely heavily on channel sparsity and are limited by their efficiency in iteratively reconstructing the signals. Their performance is highly dependent on the wireless channel [3], and thus is not a desirable approach considering the diversified use cases of 5G networks.

The recent rapid development of deep learning (DL) technologies provides another possible solution for efficient CSI feedback in FDD massive MIMO system. Instead of relying on sparsity, the DL approaches utilize the auto-encoder framework [4]. The encoder learns a map to the low-dimensional compressed space and the decoder reconstruct the original data by a single run without the labeled data. It naturally overcomes the limits of CS-based approaches in channel sparsity and operation efficiency.

The first DL-based method, CsiNet [5], explored and demonstrated the efficiency of deep learning in CSI feedback. CsiNet significantly outperforms the traditional CS-based methods (LASSO, BM3D-AMP and TVAL3) under various compression rates.

Based on CsiNet, most of the subsequent DL-based methods utilize more powerful DL building blocks to achieve better performance with the sacrifice of computational overhead. CsiNet-LSTM [6] and Attention-CSI [7] introduced LSTM that significantly increases the computational overhead. CsiNet+ [8] comprehensively surveyed recent DL-based methods and proposed a parallel multiple-rate compression framework. The computational overhead of CsiNet+ is approximately x7 higher than the original CsiNet [9]. Recently, some methods start to reduce the complexity, for example, JCNet [10] and BcsiNet [11], however, their performance has also been reduced. So far, only CRNet [12] has outperformed CsiNet without increasing the computational complexity.

However, CSI or signals are represented in complex envelopes, which have their own physical meaning that is overlooked by previous works, only [13] considered this problem by adopting complex-valued three dimensional convolutional neural network [14]. However, as the complex kernel is hard to optimize through back-propagation, the network is hard to train and the computational complexity is inevitably greatly increased. Considering the limited computational resource and limited storage at UE side, this letter proposes a tailored DL network that can cope with **c**omplex number yet maintain **l**ightweight, CLNet, for CSI feedback problem. Eventually, CLNet outperforms CRNet with 5.41% higher accuracy and 24.1% less complexity on average. The main contributions are summarized as follows.

- CLNet proposes a simple yet effective way to organic integrate the real and imaginary parts into the real-valued neural network models.

- CLNet adopts spatial attention mechanism to let the DL model focus on the more informative clustered signal parts.

II. System Model and Preliminary

Consider a single cell FDD system using massive MIMO with N_t antennas at BS, where $N_t \gg 1$ and N_r antennas at UE side (N_r equals to 1 for simplicity). The received signal $y \in \mathbb{C}^{N_c \times 1}$ can be expressed as

$$y = \boldsymbol{A}x + z \tag{1}$$

where N_c indicates the number of subcarriers, $x \in \mathbb{C}^{N_c \times 1}$ indicates the transmitted symbols, and $z \in \mathbb{C}^{N_c \times 1}$ is the complex additive Gaussian noise. $\mathbf{A}$ can be expressed as $\mathrm{diag}(h_1^H p_1, \ldots, h_{N_c}^H p_{N_c})$, where $h_i \in \mathbb{C}^{N_t \times 1}$ and $p_i \in \mathbb{C}^{N_t \times 1}, i \in \{1, \ldots, N_c\}$ represent the downlink channel coefficients and beamforming precoding vector for subcarrier i, respectively.

In order to derive the beamforming precoding vector p_i, the BS needs the knowledge of corresponding channel coefficient h_i, which is fed back by the UE. Suppose that the downlink channel matrix is $\boldsymbol{H} = [h_1 \cdots h_{N_c}]^H$ which contains $N_c N_t$ elements. The number of parameters that need to be fed back is $2N_c N_t$, including the real and imaginary parts of the CSI, which is proportional to the number of antennas.

Because the channel matrix $\boldsymbol{H}$ is often sparse in the angular-delay domain. By 2D discrete Fourier transform (DFT), the original form of spatial-frequency domain CSI can be converted into the angular-delay domain, such that

$$\boldsymbol{H'} = \boldsymbol{F}_c \boldsymbol{H} \boldsymbol{F}_t^H \tag{2}$$

where $\boldsymbol{F}_c$ and $\boldsymbol{F}_t$ are the DFT matrices with dimension $N_c \times N_c$ and $N_t \times N_t$, respectively. For the angular-delay domain channel matrix $\boldsymbol{H'}$, every element corresponds to a certain path delay with a certain angle of arrival (AoA). In $\boldsymbol{H'}$, only the first N_a rows contain useful information, while the rest rows represent the paths with larger propagation delays are made up of near-zero values, can be omitted without much information loss. Let $\boldsymbol{H}_a$ denote the informative rows of $\boldsymbol{H'}$.

$\boldsymbol{H}_a$ is input into UE's encoder to produce the codeword $\boldsymbol{v}$ according to a given compression ratio η such that

$$\boldsymbol{v} = f_{\mathcal{E}}(\boldsymbol{H}_a, \Theta_{\mathcal{E}}) \tag{3}$$

where $f_{\mathcal{E}}$ denotes the encoding process and $\Theta_{\mathcal{E}}$ represents a set of parameters of the encoder.

Once the BS receives the codeword $\boldsymbol{v}$, the decoder is used to reconstruct the channel by

$$\hat{\boldsymbol{H}}_a = f_{\mathcal{D}}(\boldsymbol{v}, \Theta_{\mathcal{D}}) \tag{4}$$

where $f_{\mathcal{D}}$ denotes the decoding process and $\Theta_{\mathcal{D}}$ represents a set of parameters of the decoder. Therefore, the entire feedback process can be expressed as

$$\hat{\boldsymbol{H}}_a = f_{\mathcal{D}}(f_{\mathcal{E}}(\boldsymbol{H}_a, \Theta_{\mathcal{E}}), \Theta_{\mathcal{D}}) \tag{5}$$

The goal of CLNet is to minimize the difference between the original $\boldsymbol{H}_a$ and the reconstructed $\hat{\boldsymbol{H}}_a$, which can be expressed formally as finding the parameter sets of encoder and decoder satisfying

$$\left(\hat{\Theta}_{\mathcal{E}}, \hat{\Theta}_{\mathcal{D}}\right) = \underset{\Theta_{\mathcal{E}}, \Theta_{\mathcal{D}}}{\arg\min} \|\boldsymbol{H}_a - f_{\mathcal{D}}(f_{\mathcal{E}}(\boldsymbol{H}_a, \Theta_{\mathcal{E}}), \Theta_{\mathcal{D}})\|_2^2 \tag{6}$$

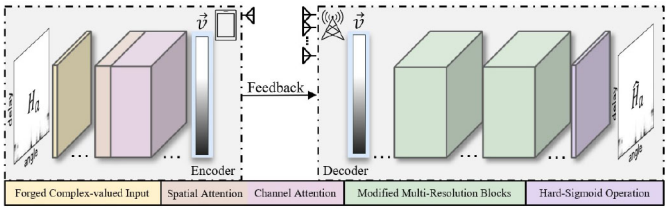

Fig. 1. The encoder and decoder architecture of CLNet.

III. CLNet Design

This section presents the design of the CLNet and its key components. Figure 1 depicts the overall architecture of CLNet, in which traditional convolution blocks are omitted for simplicity. Overall, CLNet is an encoder-decoder framework with four main building blocks that are tailored for the CSI feedback problem.

The performance of the CSI feedback scheme highly depends on the compression part, the encoder. The less information loss of the compression, the higher the decompression accuracy can be obtained. Due to the limited computing power and storage space of UE, deepening the encoder network design is not practical. Therefore, CLNet leverages the physical characteristics of CSI to achieve a lightweight yet informative encoder by two tailored blocks.

First, CSI is the channel frequency response with complex values that depict channel coefficients of different signal paths. The previous DL-based CSI feedback methods treat the real and imaginary parts of the CSI separately. Instead, the input CSI in CLNet first goes through the forged complex-valued input layer that embeds the real and imaginary parts together to preserve the physical information of the CSI (Section III-A). Second, different signal paths have different resolutions of cluster effect in the angular-delay domain, which corresponding to different angles of arrival and different path delays. Thus, we introduce the CBAM block [15] that serves as spatial-wise attention to force the neural network focus on those clusters and suppress the unnecessary parts (Section III-B).

Since the encoder becomes more powerful, the decoder can be correspondingly more lightweight, thus CLNet modifies the CRBlocks [12] in decoder by reducing the filter size from 1×9 to 1×3. To further reduce the computational cost, CLNet adopts the hard-Sigmoid activation fuction which is more hardware friendly than the conventional Sigmoid activation function (Section III-C).

A. Forged Complex-Valued Input

CSI is complex-valued channel coefficients such that:

$$\boldsymbol{H}(t) = \sum_{k=1}^{N} a_k(t) e^{-j\theta_k(t)} \tag{7}$$

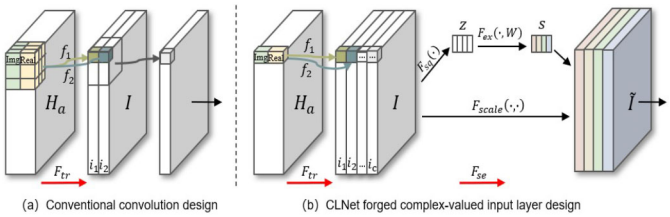

Fig. 2. Diagrammatic comparison of the conventional convolution and the CLNet forged complex-valued input layer.

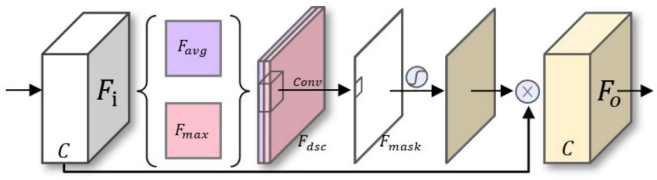

Fig. 3. Operation illustration of spatial-wise attention of CLNet.

where N is the number of signal paths. $a_k(t)$ and $\theta_k(t)$ indicate the signal attenuation and propagation phase rotation of the k-th path at time t, respectively. The BS relies on the physical meaning of CSI, the norm of real and imaginary parts describes the channel's attenuation to signal and the ratio of the real and the imaginary part describes the channel's phase rotation to the signal, to conduct the beamforming.

Since a typical deep learning neural network is designed based on real-valued inputs, operations, and representations. Existing DL-based CSI feedback methods simply separate the real and imaginary parts of the complex values as two independent channels of an image as the neural network input, which may destroy the original physical property of each complex-valued channel coefficient. Specifically, as Figure 2 (a) depicts, a conventional 3×3 kernel size entangles the real and imaginary parts of neighboring elements in $\boldsymbol{H}_a$, and as a result, the 9 complex CSI are interpolated as one synthesized value.

Mathematically, $\boldsymbol{F}_{tr} : \boldsymbol{H}_a \rightarrow \mathcal{I}$ is a convolutional transformation. Here, $\boldsymbol{H}_a \in \mathbb{R}^{N_a \times N_a \times 2}$ is a 3D tensor, extended from its 2D version by including an additional dimension to separately express the real and imaginary parts, and $\mathcal{I} \in \mathbb{R}^{N_a \times N_a \times C}$, where C indicates the number of convolutional filters applied to learn different weighted representations. The output of $\boldsymbol{F}_{tr}$ is $\mathcal{I} = [\boldsymbol{i}_1, \boldsymbol{i}_2, \ldots, \boldsymbol{i}_C]$, $\boldsymbol{i}_c \in \mathbb{R}^{N_a \times N_a}$. Let $a_n + b_n i$ denotes a CSI and w_n is the learnable weight of a convolutional filter f. The 3x3 convolution operation essentially is the sum of two multiplication such that:

$$i_1(1,1) = [a_1, \ldots, a_9] \cdot [w_1, \ldots, w_9] + [b_1, \ldots, b_9] \cdot [w_1, \ldots, w_9] \tag{8}$$

In such way, the real and imaginary parts of the same complex-valued signal are decoupled and different CSI metrics are mixed, thus losing the original physical information carried by the channel matrix.

The insight of CLNet is that by utilizing a 1×1 point-wise convolution, the real and imaginary parts of a complex-valued coefficient can be explicitly embedded such that:

$$i_1(1,1) = [a_1] \cdot [w_1] + [b_1] \cdot [w_1] \tag{9}$$

where the ratio between a and b is preserved, thus maintain the phase information and the amplitude of the signal be scaled by w. Since CNN shares the weight w, so the entire CSI matrix's amplitude is essentially scaled by the same w, the relative amplitude across subchannels is also preserved.

The output i_c, essentially, is a weighted representation of the original $\boldsymbol{H}_a$ and different filters learn different weighted representations, among which, some may be more important than others. Based on this, CLNet further adopts the SE block [16], which serves as the channel-wise attention in the forged complex-valued input layer. It assists the neural network to model the relationship of the weights so as to focus on the important features and suppress the unnecessary ones. A diagram of the SE block is shown in Figure 2 (b) with annotation $\boldsymbol{F}_{se}$.

The output $\mathcal{I}$ first goes through $\boldsymbol{F}_{sq}$ transformation by global average pooling to obtain the channel-wise statistics descriptor $\boldsymbol{z} \in \mathbb{R}^C$. Here, $\boldsymbol{F}_{sq}$ expands the neural network receptive field to the whole angular-delay domain to obtain the global statistical information, compensating the shortcoming of the insufficient local receptive field of 1×1 convolution used in the first step of the forged complex-valued input layer.

After that, the channel descriptor $\boldsymbol{z}$ goes through $\boldsymbol{F}_{ex}$ transformation, i.e., a gated layer with sigmoid activation to learn the nonlinear interaction as well as the non-mutually-exclusive relationship between channels, such that

$$\mathbf{s} = \boldsymbol{F}_{ex}(\boldsymbol{z}, \boldsymbol{W}) = \sigma(g(\boldsymbol{z}, \boldsymbol{W})) = \sigma(\boldsymbol{W}_2 \delta(\boldsymbol{W}_1 \boldsymbol{z})), \tag{10}$$

where δ is the ReLU function, $\boldsymbol{W}_1 \in \mathbb{R}^{\frac{C}{2} \times C}$ and $\boldsymbol{W}_2 \in \mathbb{R}^{C \times \frac{C}{2}}$. $\boldsymbol{F}_{ex}$ further explicitly models the inter-channel dependencies based on $\boldsymbol{z}$ and obtains the calibrated $\mathbf{s}$, which is the attention vector that summarizes all the characteristics of channel C, including intra-channel and inter-channel dependencies. Before being fed into the next layer, each channel of $\mathcal{I}$ is scaled by the corresponding attention value, such that

$$\tilde{\mathcal{I}}_{:,:,i} = \boldsymbol{F}_{scale}(\mathbf{s}, \mathcal{I}) = \mathbf{s}_i \mathcal{I}_{:,:,i}, \ \text{s.t.} \ i \in \{1, 2, \ldots, C\} \tag{11}$$

$\tilde{\mathcal{I}} \in \mathbb{R}^{N_a \times N_a \times C}$ is the final output of the forged complex-valued input layer, which preserves the CSI physical information while capturing dynamics by the channel-wise attention mechanism.

B. Attention Mechanism for Informative Encoder

On the other hand, in angular-delay domain, the channel coefficients exhibit the effect of clusters with different resolutions that correspond to the distinguishable paths that arrive with specific delays and AoAs. In order to pay more attention to such clusters, CLNet employs a CBAM block [15] to serve as spatial-wise attention to distinguish them with weights in the spatial domain as Figure 3 illustrates.

Based on the cluster effect in the angular-delay domain, spatial-wise attention uses the generated spatial statistical descriptors as the basis for assigning weights, forcing the network to focus more on the distinguishable propagation paths.

First, two pooling operations, i.e., average-pooling and max-pooling, are adopted across the input $\boldsymbol{F}_i$'s channel C to generate two 2D feature maps, $\boldsymbol{F}_{\mathbf{avg}} \in \mathbb{R}^{N_a \times N_a \times 1}$ and

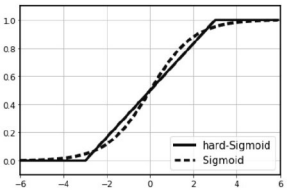

Fig. 4. Comparison between Sigmoid and hard-Sigmoid functions.

$\boldsymbol{F}_{\mathbf{max}} \in \mathbb{R}^{N_a \times N_a \times 1}$, respectively. CLNet concatenates the two feature maps to generate a compressed spatial feature descriptor $\boldsymbol{F}_{\mathbf{dsc}} \in \mathbb{R}^{N_a \times N_a \times 2}$, and convolves it with a standard convolution layer to produce a 2D spatial attention mask $\boldsymbol{F}_{\mathbf{mask}} \in \mathbb{R}^{N_a \times N_a \times 1}$. The mask is activated by Sigmoid and then multiplied with the original feature maps $\boldsymbol{F}_i$ to obtain $\boldsymbol{F}_o$ with spatial-wise attention.

$$\begin{aligned} \boldsymbol{F}_o &= \mathbf{CBAM}(\boldsymbol{F}_i) \\ &= \boldsymbol{F}_i(\sigma(\boldsymbol{f}_c([\mathrm{AvgPool}(\boldsymbol{F}_i); \mathrm{Max\,Pool}(\boldsymbol{F}_i)]))) \\ &= \boldsymbol{F}_i\big(\sigma\big(\boldsymbol{f}_c\big(\big[\boldsymbol{F}_{\mathbf{avg}}; \boldsymbol{F}_{\mathbf{max}}\big]\big)\big)\big) \end{aligned} \tag{12}$$

With spatial-wise attention, CLNet focuses the neural network on the more informative signal propagation paths in the angular-delay domain.

C. Reduction of Computational Cost

The often-used Sigmoid activation function contains exponential operation:

$$\sigma(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{e^x+1}. \tag{13}$$

In order to reduce time cost in the computation, CLNet uses the hard version of Sigmoid, its piece-wise linear analogy function, denoted as $h\sigma$ to replace the Sigmoid function [17],

$$h\sigma(x) = \frac{\min(\max(x+3, 0), 6)}{6} \tag{14}$$

Figure 4 compares the excitation curves of the hard-Sigmoid and Sigmoid functions. The hard-Sigmoid induces no discernible degradation in accuracy but benefits from its computational advantage of entailing no exponential calculations. In practice, the hard-Sigmoid can fit in most software and hardware frameworks and can mitigate the potential numerical quantization loss introduced by different hardware.

IV. Evaluation

This section presents the detailed experiment setting and the comparison with the state-of-the-art (SOTA) DL-based CSI feedback approach, in terms of accuracy and computational overhead.

1) Data Generation: To ensure a fair performance comparison, we use the same dataset as provided in the first work of DL-based Massive MIMO CSI feedback in [5], which is also used in later studies on this problem [6], [7], [8], [9], [12]. The generated CSI matrices are converted to angular-delay domain $\boldsymbol{H}_a \in \mathbb{R}^{32\times32\times2}$ by 2D-DFT. The total 150,000 independently generated CSI are split into three parts, i.e., 100,000 for training, 30,000 for validation, and 20,000 for testing, respectively.

2) Training Scheme and Evaluation Metric: The normalized mean square error (NMSE) between the original $\boldsymbol{H}_a$ and the reconstructed $\hat{\boldsymbol{H}}_a$ is used to evaluate the network accuracy:

$$NMSE = E\Big\{\|\boldsymbol{H}_a - \hat{\boldsymbol{H}}_a\|_2^2 / \|\boldsymbol{H}_a\|_2^2\Big\} \tag{15}$$

The complexity is measured by the flops (floating-point operations per second). The model was trained with the batch size of 200 and 8 workers on a single NVIDIA 2080Ti GPU. The epoch is set to 1000, as recommended in previous work [8], [12]. To further ensure the fairness, we fix the random seed of the computer.

3) CLNet Overall Performance: Table I shows the overall performance comparison among the proposed CLNet and related CSI feedback networks.

As for the complexity, generally, the LSTM-based networks (CSINet+ and Attn-CSI) require approximate five to seven-flods higher computational resources than the CNN-based networks (CSINet, CRNet[2] and CLNet). Furthermore, because LSTM's operation relies on the previous output as the input of the hidden layer and does not share parameters for parallel computation, it is difficult to reduce the complexity even if the compression rate increases. As shown in Table I, the CLNet is the lightest among all networks. Compared with the SOTA CRNet, the CLNet significantly reduces the computational complexity by 24.1% less flops on average. The flops of CLNet is 18.00%, 22.35%, 25.20%, 26.50%, 28.36% less than CRNet at the compression ratio η of 1/64, 1/32, 1/16, 1/8, 1/4, respectively. As the compression rate increases, the computational complexity degrades more.

In Table I, turn to the accuracy part, the best results in the lightweight network are shown in bold, and the best results in all networks are shown in italics. The result shows that CLNet consistently outperforms other lightweight networks at all compression ratios in both indoor and outdoor scenarios with 5.41% overall average improvement compared with the SOTA CRNet.[3] In indoor scenarios, CLNet obtains an average performance increase of 6.61%, with the most increase of 21.00% at the compression ratio of $\eta = 1/4$. In outdoor scenarios, the average improvement on NMSE is 4.21%, with the most increase of 10.44% at the compression ratio of $\eta = 1/32$. Compared to heavyweight networks, CLNet still achieves the best results at the compression ratio of 1/4, outperforming the second place CSINet+ by 6.54% and 3.87% in indoor and outdoor scenario, respectively. CLNet also achieves the best result in indoor scenario at the compression ratio equals to 1/64.

4) Ablation Study: Considering the limited interpretability of deep neural network, we further conduct the ablation study to better quantify the gain of the proposed forged complex-valued input layer and spatial-attention mechanism. The epochs of ablation studies are set to 500 in indoor scenarios, the rest settings remain the same as discussed in Section IV(1-2). Baseline is the CRNet with conventional convolution.

[2]Note that the CRNet paper reported flops is corrected by [13].

[3]We reproduce CRNet follow the open source code: https://github.com/Kylin9511/CRNet the higher performance they reported in the paper are from training with 2500 epoch.

TABLE I
NMSE(DB)[a] AND COMPLEXITY COMPARISON BETWEEN SERIES OF CSI FEEDBACK NETWORK AND THE PROPOSED CLNET

η	1/4			1/8			1/16			1/32			1/64		
Methods	FLOPS	NMSE indoor	NMSE outdoor	FLOPS	NMSE indoor	NMSE outdoor	FLOPS	NMSE indoor	NMSE outdoor	FLOPS	NMSE indoor	NMSE outdoor	FLOPS	NMSE indoor	NMSE outdoor
CLNet	***4.05M***	***-29.16***	***-12.88***	***3.01M***	**-15.60**	**-8.29**	***2.48M***	**-11.15**	**-5.56**	***2.22M***	**-8.95**	**-3.49**	***2.09M***	***-6.34***	**-2.19**
CRNet	5.12M	-24.10	-12.57	4.07M	-15.04	-7.94	3.55M	-10.52	-5.36	3.29M	-8.90	-3.16	3.16M	-6.23	**-2.19**
CSINet[5]	5.41M	-17.36	-8.75	4.37M	-12.70	-7.61	3.84M	-8.65	-4.51	3.58M	-6.24	-2.81	3.45M	-5.84	-1.93
CSINet+[8]	24.57M	-27.37	-12.40	23.52M	*-18.29*	*-8.72*	23.00M	*-14.14*	-5.73	22.74M	*-10.43*	-3.40	22.61M	/	/
Attn-CSI[7]	24.72M	-20.29	-10.43	22.62M	/	/	21.58M	-10.16	*-6.11*	21.05M	-8.58	*-4.57*	20.79M	-6.32	*-3.27*

[a] / means the performance is not reported.

TABLE II
NMSE (DB) COMPARISON OF ABLATION STUDY

η	**Baseline**	**1x1 Conv**	**1x1 Conv + SE**	**1x1 Conv + CBAM**	**1x1 Conv + SE + CBAM**
1/4	-21.702	-27.694	-27.903	-28.142	-28.984
1/8	-13.037	-15.171	-15.167	-15.321	-15.487
1/16	-10.212	-11.013	-11.231	-10.684	-11.217
1/32	-8.443	-8.525	-8.732	-8.613	-8.885
1/64	-6.023	-6.145	-6.201	-6.086	-6.297

TABLE III
DETAILED COMPLEXITY OF CRNET AND CLNET

η	**Method**	**Encoder at UE** flops(M)	**Encoder at UE** #params	**Decoder at BS** flops(M)	**Decoder at BS** #params
1/4	CLNet	1.34	1.049M	2.71	1.052M
	CRNet	1.20	1.049M	3.92	1.053M
1/64	CLNet	0.36	65.954K	1.73	69.210K
	CRNet	0.22	65.720K	2.94	70.386K

As Table II shown, by simply modifying the first layer from a conventional convolution layer to an 1x1 convolution as the forged complex input layer,its accuracy surpasses the baseline at all compression ratios with an average improvement of 10.964%, which demonstrates the efficacy of appropriately preserving the complex notation. After adding the SE block, the accuracy is slightly improved although there is no improvement at $\eta = 1/8$. The last two columns show that the spatial-attention slightly improves the accuracy at low compression rates, however, when combined with the SE block, its accuracy is further improved by 3.058% on average.

5) Encoder Complexity: Table III reveals that the CLNet encoder is actually slightly heavier than that of CRNet. However, the BS may need to execute several different models at the same time so a relatively light decoder would also be beneficial. In terms of storage space, CLNet and CRNet are roughly the same.

V. CONCLUSION

This letter studies the CSI feedback problem for massive MIMO under FDD mode, which is the key technology of 5G communication systems. Based on the understanding of the physical properties of the CSI data, a novel customized deep learning framework, CLNet, is proposed. The forged complex-valued input layer preserves the amplitude and phase information of the signal and enhances with spatial-attention mechanisms. The hard-Sigmoid function is adopted to eliminate the exponential calculations. The overall performance of CLNet has 5.41% higher accuracy than the state-of-the-art CRNet with 24.10% less computation overhead.

REFERENCES

[1] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3590–3600, Nov. 2010.

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

[3] P. Kyritsi, D. C. Cox, R. A. Valenzuela, and P. W. Wolniansky, "Correlation analysis based on MIMO channel measurements in an indoor environment," *IEEE J. Sel. Areas Commun.*, vol. 21, no. 5, pp. 713–720, Jun. 2003.

[4] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[5] 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.

[6] T. Wang, C.-K. Wen, S. Jin, and G. Y. Li, "Deep learning-based CSI feedback approach for time-varying massive MIMO channels," *IEEE Wireless Commun. Lett.*, vol. 8, no. 2, pp. 416–419, Apr. 2019.

[7] 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)*, Marrakesh, Morocco, 2019, pp. 1–5.

[8] 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.

[9] 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 system," 2021. [Online]. Available: arXiv:2105.00354.

[10] 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.

[11] 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.

[12] Z. Lu, J. Wang, and J. Song, "Multi-resolution CSI feedback with deep learning in massive MIMO system," in *Proc. Int. Conf. Commun. (ICC)*, 2020, pp. 1–6.

[13] Y. Zhang *et al.*, "CV-3DCNN: Complex-valued deep learning for CSI prediction in FDD massive MIMO systems," *IEEE Wireless Commun. Lett.*, vol. 10, no. 2, pp. 266–270, Feb. 2021.

[14] C. Trabelsi *et al.*, "Deep complex networks," 2017. [Online]. Available: arXiv:1705.09792.

[15] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.

[16] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, 2018, pp. 7132–7141.

[17] A. Howard *et al.*, "Searching for MobileNetV3," in *Proc. IEEE Int. Conf. Comput. Vis.*, Seoul, South Korea, 2019, pp. 1314–1324.