TransNet: Full Attention Network for CSI Feedback in FDD Massive MIMO System

Yaodong Cui, Aihuang Guo, and Chunlin Song

***Abstract*—Channel state information (CSI) is a key aspect of massive multi-input multi-output (MIMO) system. It depicts important properties of transmission channels such as scattering, fading, the attenuation of power with distance, etc. The quality and cost of CSI feedback between user equipment (UE) and base station (BS) play vital roles in the quality of the whole communication system. In this letter, a new deep learning (DL) method based on Google's famous Transformer architecture is presented for CSI feedback in frequency division duplex (FDD) massive MIMO system. Simulation results show that the presented inception network named TransNet outperforms other DL methods on the quality of CSI feedback.**

***Index Terms*—Massive MIMO, CSI feedback, deep learning, transformer architecture, inception network.**

I. Introduction

MASSIVE multi-input multi-output (MIMO) is one of the core technologies of the current 5G and the following 6G communication system. It is a method to increase sector throughput and capacity density by using large numbers of antennas at base station (BS). Channel state information (CSI) depicts crucial characters for signal transmission in a specific channel. In order to improve the communication quality of the whole wireless system, it is important for both transmitters and receivers to attain complete and accurate CSI. In time-division duplex (TDD) system, the downlink and uplink channels are divided by different time slots and have little difference in fading characteristics. It is useful to use channel reciprocity to conduct a back channel estimation. However, the same approach is not suitable for frequency division duplex (FDD) system, where the uplink and downlink work at different frequencies. As a result, user equipment (UE) of FDD system always needs to send CSI of downlink channels as complete and accurate as possible to BS by feedback link.

One of the key problems of this feedback process in massive MIMO system is that the cost is huge. With the number of antennas increasing in BS, the burden of downlink CSI's feedback increases in a linear way. To save the feedback cost, a practical and meaningful plan is to efficiently compress the downlink CSI at BS, send it to the feedback link and recover it as completely and accurately as possible at BS. This plan is of great concern in both academia and industry. In 2012, P. Kuo and his team first try to use compressed sensing (CS) methods to do this job [1]. Along with CS methods are some effective traditional algorithms such as LASSO [2], BM3D-AMP [3] and TVAL3 [4], etc. The CS methods rely heavily on the sparsity of CSI matrix, specifically, they require CSI matrix to be as large as enough and with most part as zeros or nearly zeros. Practically, even in massive MIMO system, the CSI matrices are not able to perfectly meet the mathematical sparsity requirement of the CS methods [5]. The CSI matrices of massive MIMO system are only approximately sparse in practice, giving obstacles on modeling for the CS methods.

In 2018, deep learning (DL) first shows its powerful ability on CSI feedback for massive MIMO system [6], the presented network architecture named CsiNet has overwhelming superiority against traditional CS methods. After that, some works extend the original scene of CsiNet in massive MIMO systems [7]–[9]. In 2019, the CsiNet+ presented in [10] improves the performance of CsiNet by updating the convolution kernel of CsiNet, while the floating point operations per second (FLOPS) of CsiNet+ is 7 times higher than the original CsiNet. Meanwhile, researchers try to introduce the attention mechanism in DL to the network by long short-term memory (LSTM). The presented CsiNet-LSTM [7], Attention-CSI [11], like CsiNet+, both attain performance improvement with increasing some computational overhead. Some latter works focus on how to reduce the computation burden [12], [13]. Among all these works, CRNet [14] first gets better results than the original CsiNet with lower FLOPS. CLNet [15] improves the performance of CRNet by giving different process to CSI's real part and imaginary part. In 2021, [16] first tries to use Google's Transformer [17] architecture to build the network for CSI feedback. While this letter merely gets slightly better results than the original CsiNet, the performance is not competitive with CRNet and CLNet. In this letter, Transformer's power on this problem is further excavated and a novel neural network named TransNet is presented. Compared with the work in [16], TransNet adopts a two-layer rather than single-layer Transformer architecture in the core structure of the network, and CSI is entered directly into the attention layer of encoder without extra processing, getting state of the art (SOTA) in recovery accuracy for compressed CSI. The open source codes are available at https://github.com/Treedy2020/TransNet.

The main contributions are summarized as follows.

- TransNet introduces attention mechanism by a two-layer Transformer architecture, enabling CSI to learn the connections between its parts in the process of feedback.

Manuscript received January 5, 2022; revised January 30, 2022; accepted February 2, 2022. Date of publication February 7, 2022; date of current version May 10, 2022. This work was supported by the Future Network Innovation Research and Application Project (2021). The associate editor coordinating the review of this article and approving it for publication was C.-K. Wen. *(Corresponding author: Yaodong Cui.)*

The authors are with the Department of Information and Communication Engineering, Tongji University, Shanghai 201804, China (e-mail: 2032974@tongji.edu.cn; tjgah@tongji.edu.cn; songchunlin@tongji.edu.cn).

Digital Object Identifier 10.1109/LWC.2022.3149416

- Comparative experiments with other DL methods are done, it is proved that TransNet has higher accuracy for CSI feedback and better training convergence property.

The rest of this letter is organized as follows. Section II describes the system model and the scenario of CSI feedback. In Section III, we explain detailed design about TransNet. We show the training scheme, evaluation metric, numerical results and analysis of TransNet in Section IV. The conclusion is placed in Section V.

II. System Model

We consider a FDD massive MIMO system where a single-cell has N_t transmitting antennas at the BS and N_r receiving antennas at UE, with $N_t \gg N_r$. In this scenario, N_r is set to 1 for simplicity. The whole FDD system has N_c sub-carriers. The CSI of UE's downlink channels can be described as

$$\widetilde{\mathbf{H}} = \begin{bmatrix} \widetilde{\mathrm{h}}_{0,0} & \widetilde{\mathrm{h}}_{0,1} & \cdots & \widetilde{\mathrm{h}}_{0,N_t-1} \\ \widetilde{\mathrm{h}}_{1,0} & \widetilde{\mathrm{h}}_{1,1} & \cdots & \widetilde{\mathrm{h}}_{1,N_t-1} \\ \vdots & \vdots & \ddots & \vdots \\ \widetilde{\mathrm{h}}_{N_c-1,0} & \widetilde{\mathrm{h}}_{N_c-1,1} & \cdots & \widetilde{\mathrm{h}}_{N_c-1,N_t-1} \end{bmatrix} \in \mathbb{C}^{N_c \times N_t}, \tag{1}$$

$\widetilde{\mathbf{H}}$ can be used at BS to improve the communication quality by beamforming. There is a necessity for BS to get $\widetilde{\mathbf{H}}$ of UE as complete as possible. Note that the matrix's dimension is $N_c \times N_t$, for the simplicity of subsequent expression, the elements of $\widetilde{\mathbf{H}}$ start with 0 for rows and columns. As the number of sub-carriers or antennas of BS increases, the feedback cost of $\widetilde{\mathbf{H}}$ increases in an approximately linear order.

In massive MIMO system case, $\widetilde{\mathbf{H}}$ is sparse in angular-delay domain [10], an effective way to reduce the feedback cost is first to do a two-dimensional discrete Fourier transform (DFT) for $\widetilde{\mathbf{H}}$ as

$$\mathrm{h}_{k,l} = \frac{1}{N_c \times N_t} \sum_{m=0}^{N_c-1} \sum_{n=0}^{N_t-1} \widetilde{\mathrm{h}}_{m,n} e^{-j2\pi\left(\frac{k}{N_c}m + \frac{l}{N_t}n\right)}, \tag{2}$$

where $k, l = 0, 1, 2, \ldots, N_c - 1; 0, 1, 2, \ldots, N_t - 1$. We define

$$\mathbf{H} = \begin{bmatrix} \mathrm{h}_{0,0} & \mathrm{h}_{0,1} & \cdots & \mathrm{h}_{0,N_t-1} \\ \mathrm{h}_{1,0} & \mathrm{h}_{1,1} & \cdots & \mathrm{h}_{1,N_t-1} \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{h}_{N_c-1,0} & \mathrm{h}_{N_c-1,1} & \cdots & \mathrm{h}_{N_c-1,N_t-1} \end{bmatrix} \in \mathbb{C}^{N_c \times N_t}. \tag{3}$$

$\mathbf{H}$ is in the angular-delay domain. In $\mathbf{H}$, nearly all the distinct non-zero values concentrate in the first N_a rows [10], it is practical to reduce the feedback cost by intercepting the first N_a rows of $\mathbf{H}$ to represent itself.

As it is shown in Fig. 1, let $\mathbf{H}_a$ represent the truncated matrices constructed by the first N_a rows of $\mathbf{H}$, the row vectors are as the same order as they are in $\mathbf{H}$. Although $\mathbf{H}_a$ has been reduced in dimension compared with $\widetilde{\mathbf{H}}$, there is still a need to do a further compression for $\mathbf{H}_a$ following (4).

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

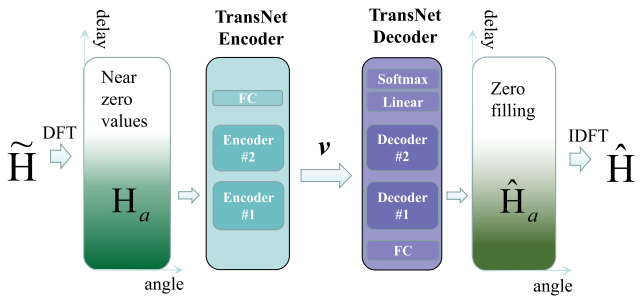

Fig. 1. Schematic diagram of TransNet aided downlink CSI feedback workflow. Note that v propagates in the feedback link, and there is no positional encoding process that exists in normal Transformer architecture.

where $f_{\mathcal{C}}(\cdot)$ denotes the compression process, and $\Theta_{\mathcal{C}}$ represents parameters of the designed compression module, and $\boldsymbol{v}$ is the compressed information. $\boldsymbol{v}$ transmits over the feedback link channels to the BS. At BS, the recovery process of $\mathbf{H}_a$ can be expressed as (5).

$$\hat{\mathbf{H}}_a = f_{\mathcal{R}}(\boldsymbol{v}, \Theta_{\mathcal{R}}) \tag{5}$$

where $f_{\mathcal{R}}(\cdot)$ denotes the recovery process and $\Theta_{\mathcal{R}}$ represents parameters of the designed recovery module. The original $\widetilde{\mathbf{H}}$ can be restored as $\hat{\mathbf{H}}$ by zero filling and inverse discrete Fourier transform (IDFT) for $\hat{\mathbf{H}}_a$.

By combining (4) and (5), we formulate the whole compression and recovery process as an optimization form by mean-squared error (MSE) distortion metric as

$$(\hat{\Theta}_{\mathcal{C}}, \hat{\Theta}_{\mathcal{R}}) = \underset{\Theta_{\mathcal{C}}, \Theta_{\mathcal{R}}}{\arg\min} \|\mathbf{H}_a - f_{\mathcal{R}}(f_{\mathcal{C}}(\mathbf{H}_a, \Theta_{\mathcal{C}}), \Theta_{\mathcal{R}})\|_2^2. \tag{6}$$

The compression module and recovery module of CSI constitute an encoder-decoder network, the compression module is the encoder of TransNet and the recovery module is the decoder of TransNet. Our purpose is to design and train $\Theta_{\mathcal{C}}$ and $\Theta_{\mathcal{R}}$ to minimize the distance between $\mathbf{H}_a$ and $\hat{\mathbf{H}}_a$.

III. Design of TransNet

In Fig. 1, we show that TransNet has 2 encoder layers at the TransNet encoder and 2 decoder layers at the TransNet decoder. Encoder#1 has the same structure with encoder#2 and decoder#1 has the same structure with decoder#2. In a complete compression process, the TransNet encoder's input $\mathbf{H}_a$ is first entered into encoder#1 and gets an output, then the output is entered into encoder#2 and gets another output. The output of encoder#2 is compressed as $\boldsymbol{v}$ at a fixed scale η by a full connected (FC) layer. $\boldsymbol{v}$ transmits through the feedback link and is received at the TransNet decoder of BS. Since this letter focuses on the feedback scheme for CSI, the uplink feedback of $\boldsymbol{v}$ is assumed as ideal, that is to say, the received $\boldsymbol{v}$ of the TransNet decoder at BS is actually the output of the TransNet encoder at UE.

We show details of single encoder and decoder layer in Fig. 2, they are based on the original Transformer encoder layer and decoder layer architectures in [17]. Transformer is a complicated neural network architecture, for simplicity, our following part of this section will introduce the core attention mechanism of the multihead attention layer and leave out

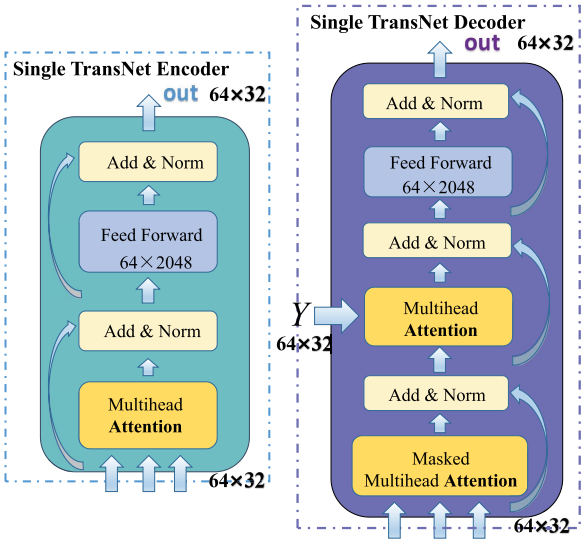

Fig. 2. Single encoder and decoder layer design of the proposed TransNet. Note that $\boldsymbol{Y}$ is gotten by $\boldsymbol{v}$, and the input and output tensors' shapes for each layer are 64 × 32.

detailed analysis for other layers of Fig. 2. Note that the shape of $\mathbf{H}_a$ is $N_a \times N_t$, and $\mathbf{H}_a$ is a complex matrix. In TransNet, we separate the real and imaginary part of $\mathbf{H}_a$ and reform them into a new $2N_a \times N_t$ real value matrix. Suppose the input of the TransNet encoder is a $2N_a \times N_t$ real value matrix $\boldsymbol{X}$, $\boldsymbol{X}$ is first sent into the multihead attention layer of the encoder#1, and gets multiplied by $\boldsymbol{W}_n^Q$, $\boldsymbol{W}_n^K$ and $\boldsymbol{W}_n^V$ as

$$\begin{cases} \boldsymbol{X}\,\boldsymbol{W}_n^Q = \boldsymbol{Q}_n \\ \boldsymbol{X}\,\boldsymbol{W}_n^K = \boldsymbol{K}_n, \\ \boldsymbol{X}\,\boldsymbol{W}_n^V = \boldsymbol{V}_n \end{cases} \quad \boldsymbol{W}_n^Q, \boldsymbol{W}_n^K, \boldsymbol{W}_n^V \in \mathbb{R}^{N_t \times (d^{\text{model}}/2)}, \tag{7}$$

where $n = 1, 2$, which means that (7) yields two sets of corresponding $\boldsymbol{Q}_n, \boldsymbol{K}_n, \boldsymbol{V}_n$ matrices. In the technical parlance of the Transformer, this can also be interpreted as the number of attention heads for TransNet is 2, and d^{model} is the dimension of TransNet. The multihead attention layer uses $\boldsymbol{Q}_n, \boldsymbol{K}_n$ and softmax function to get attention score matrices following

$$\boldsymbol{A}_n = \frac{\boldsymbol{Q}_n \boldsymbol{K}_n^T}{\sqrt{d^{\text{model}}/2}} = \left(a_{i,j|n}\right) \in \mathbb{R}^{2N_a \times 2N_a}, \tag{8}$$

where $a_{i,j|n}$ denotes the element of matrix $\boldsymbol{A}_n$ in i th row and j th column. We define the softmax normalized matrix as

$$Atten(\boldsymbol{A}_n) = \left(b_{i,j|n}\right) = \left(\frac{e^{a_{i,j|n}}}{\sum_{m=1}^{2N_a} e^{a_{i,m|n}}}\right) \in \mathbb{R}^{2N_a \times 2N_a}. \tag{9}$$

where $b_{i,j|n}$ denotes the element of matrix $Atten(\boldsymbol{A}_n)$ in i th row and j th column. Then $Atten(\boldsymbol{A}_n)$ and $\boldsymbol{V}_n$ are used to compute $\boldsymbol{Z}_n$ following (10).

$$\boldsymbol{Z}_n = Atten(\boldsymbol{A}_n)\,\boldsymbol{V}_n \in \mathbb{R}^{2N_a \times (d^{\text{model}}/2)} \tag{10}$$

Since $n = 1, 2$, we have $\boldsymbol{Z}_1, \boldsymbol{Z}_2 \in \mathbb{R}^{2N_a \times (d^{\text{model}}/2)}$. $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ are combined into a new block matrix, and the block matrix gets multiplied by a weight matrix $\boldsymbol{W}^O \in \mathbb{R}^{d^{\text{model}} \times d^{\text{model}}}$ following

$$\boldsymbol{Z} = \left[\boldsymbol{Z}_1 \vdots \boldsymbol{Z}_2\right] \boldsymbol{W}^O \in \mathbb{R}^{2N_a \times d^{\text{model}}}. \tag{11}$$

Intuitively, the whole process above means that for every single row $\boldsymbol{x}_i (i = 1, 2, \ldots, 2N_a)$ of the input $\boldsymbol{X}$, the multihead attention layer gives $\boldsymbol{x}_i$ normalized fractions for attention it need to pay to every single row $\boldsymbol{x}_j (j = 1, 2, \ldots, 2N_a)$ of $\boldsymbol{X}$. The input $\boldsymbol{X}$ and output $\boldsymbol{Z}$ of the multihead attention layer need to be combined by the so called skip connections mechanism [17] in the add & norm layer shown in Fig. 2. The add & norm layer exists in both encoder and decoder layers because it makes neural network have better convergence property, and getting the gradient of the network parameters during training becomes easier. After that, the output of the first add & norm layer is sent into the feed forward layer. The input and output of the feed forward layer, again, get combined by skip connections in the second add & norm layer.

The decoder layers are similar in structure with encoder layers, and the masked multihead attention layers of decoder layers work basically as same as the multihead attention layers. Specifically, the masked multihead attention layers prevent the leak of tags by the mask mechanism [17]. The compressed information $\boldsymbol{v}$ is converted into a 64 × 32 real matrix $\boldsymbol{Y}$ by sending into the FC layer of the TransNet decoder in Fig. 1 and reshaping. As shown in Fig. 2, $\boldsymbol{Y}$ is one of the input of single decoder layer, and is used to construct the $\boldsymbol{K}_n, \boldsymbol{V}_n (n = 1, 2)$ of the multihead attention layers of decoder#1 and decoder#2. In a complete recover process, $\boldsymbol{Y}$ is first sent into the masked multihead attention layer of decoder#1 and helps to construct the $\boldsymbol{Q}_n, \boldsymbol{K}_n, \boldsymbol{V}_n (n = 1, 2)$ of this layer. $\boldsymbol{Y}$ and the output $\boldsymbol{Z}$ of the masked multihead attention layer get combined by the add & norm layer. The output of the add & norm layer is sent into the multihead attention layer, and is used to construct $\boldsymbol{Q}_n (n = 1, 2)$ of this layer. The output of the multihead attention layer of decoder#1 goes through the same process in subsequent layers as we mentioned in encoder layer's part. The output of decoder#1 is the input of the masked multihead attention layer of decoder#2, and the output of decoder#2 is sent into the following linear layer and softmax layer of TransNet decoder in Fig. 1 and we get the restored $\hat{\mathbf{H}}_a$.

IV. Simulation Results and Analysis

1) *Dataset, Training Scheme and Evaluation Metric:* Following [6], [7], [10]–[16], we use the dataset generated by COST2100 [18] and compare the presented TransNet with CRNet and the SOTA CLNet. Two types of scenarios are considered: the indoor scenario at 5.3GHz and outdoor scenario at 300MHz. Same as CRNet and CLNet, the number of antennas at BS is set as $N_t = 32$, the number of subcarrier is $N_c = 1024$ for FDD system and $N_a = 32$ in angular domain. Dimension of TransNet, which means the second dimension of $\boldsymbol{Z}$ matrices in (masked) multihead attention layers of encoder and decoder layers, is set as $d^{\text{model}} = 32$. We set $d^{\text{model}} = 32$ with a consideration that it is actually the dimension of the row vectors of CSI matrix. Intuitively, it means that TransNet learns the features of the row vectors of the CSI matrices. The

TABLE I
NMSE(dB) AND FLOPS COMPARISON BETWEEN TRANSNET AND OTHER METHODS

η	1/4			1/8			1/16			1/32			1/64		
Methods	FLOPS	NMSE		FLOPS	NMSE		FLOPS	NMSE		FLOPS	NMSE		FLOPS	NMSE	
		indoor	outdoor		indoor	outdoor		indoor	outdoor		indoor	outdoor		indoor	outdoor
TransNet-1000ep	35.72M	**-32.38**	**-14.86**	34.70M	**-22.91**	**-9.99**	34.14M	**-15.00**	**-7.82**	33.88M	**-10.49**	**-4.13**	33.75M	**-6.08**	**-2.62**
TransNet-400ep		**-29.22**	**-13.99**		**-21.62**	**-9.57**		**-14.98**	**-6.90**		**-9.83**	**-3.77**		**-5.77**	**-2.20**
CsiNet+	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	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*
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
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
CsiNet	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

[1] / means the performance is not reported in original paper [10], [11].
[2] * in the upper right of the number indicates that this is the best result for this column other than TransNet.
[3] We use the results of other methods reported in [15], they reproduce CRNet following the open source code: https://github.com/Kylin9511/CRNet. In our following figures, we reproduce CLNet by the open source code: https://github.com/SIJIEJI/CLNet.

training, validation and test dataset contain 100,000, 30,000, and 20,000 matrices, respectively. We evaluate the accuracy of the network through the expectation of normalized expression of formula (6).

$$\mathrm{NMSE} = \mathrm{E}\left\{ \frac{\|\mathbf{H}_a - \hat{\mathbf{H}}_a(\Theta_{\mathcal{C}}, \Theta_{\mathcal{R}})\|_2^2}{\|\mathbf{H}_a\|_2^2} \right\} \tag{12}$$

The computing overhead of different DL methods is measured by FLOPS. For a fair comparison with CRNet and CLNet, as recommended in [10], [14], [15], we train TransNet with the batch size of 200 on a single NVIDIA 2060 GPU and update the parameters with a constant learning rate at 1×10^{-4}, and the largest training epoch is set as 1000.

2) *TransNet Overall Performance:* The main performance comparison between different methods and the proposed TransNet is shown in Table I. Large and small DL methods measured by FLOPS are separated by dotted line, the methods above the dotted line have FLOPS greater than 20M and the methods below the dotted line have FLOPS less than 6M. In order to demonstrate the convergent property and feature learning ability, we show performance of TransNet from training with 400 epochs and 1000 epochs under different η, where η is the compress scale for $\mathbf{H}_a$ in the TransNet encoder. We use TransNet-400ep and TransNet-1000ep to represent the performance from training with 400 epochs and 1000 epochs, respectively.

TransNet demonstrates its powerful feature learning capability, even it is the largest model measured by FLOPS for CSI feedback based on DL, with average FLOPS at 34.44M. The extra computing overhead is bearable for that TransNet is still fairly simple compared with the traditional computer vision models. The famous ResNet50, with the FLOPS of 3.9G, is over 100 times larger compared with TransNet. The core issue of CSI feedback is still the quality of the restored CSI, the FLOPS of the method is far from being a hindrance to the actual engineering.

As the results show in Table I, after 400 training epochs, TransNet starts to go beyond other DL methods in both indoor and outdoor scenarios with $\eta = 1/4, 1/8, 1/16$, respectively. In indoor scenarios, TransNet-400ep slightly outperforms the best results of CLNet and CsiNet+ with $\eta = 1/4, 1/16$, respectively. Specifically, with the compress scale $\eta = 1/8$, TransNet gets 3.33dB descend in NMSE than the SOTA CsiNet+, which means it gets a 53.55% gain of accuracy. In outdoor scenarios, TransNet-400ep gets 1.11dB, 0.85dB, 0.79dB gain in NMSE than other methods' best results with $\eta = 1/4, 1/8, 1/16$.

Our experiment results show that TransNet can perform well enough under different compression scales after 1000 training epochs. Note that here in Table I the best results of CLNet and CRNet are from training with 1000 epochs, higher performance is available for every method from training with more epochs. As we can see from Table I, TransNet-1000ep achieves the best NMSE at almost every compression scale in both indoor and outdoor scenarios. Compared with the best accuracy results of other methods, TransNet works at its best in indoor and large η scenarios. In the indoor scene with $\eta = 1/4, 1/8$, TransNet obtains NMSE gains of 3.22 dB and 4.62 dB, respectively. In the outdoor scene with $\eta = 1/4, 1/8$, TransNet obtains NMSE gains of 1.98dB and 1.27dB, respectively. It means TransNet improves accuracy by 52.36% and 65.49% in indoor scenarios, and 36.61% and 25.36% in outdoor scenarios with $\eta = 1/4, 1/8$, respectively. TransNet also performs well at $\eta = 1/16, 1/32$, achieving SOTA in NMSE except for outdoor scenarios with $\eta = 1/32$, where it is slightly less impressive than Attn-CSI. Specifically, TransNet gets a 32.55% gain in accuracy than other methods' best results in outdoor scenario with $\eta = 1/16$. Overall, the accuracy of TransNet is not obviously superior to other methods at $\eta = 1/64$, with slightly loss in feedback accuracy compared to the best results of other methods.

In Fig. 3, we show the test NMSE trends of TransNet and CRNet and CLNet during 1000 training epochs. For CRNet and CLNet, it is better to use the cosine warm up scheme as proposed in [14]. For TransNet, as the model is relatively large compared with CRNet and CLNet in FLOPS, we train it with a relatively smaller constant learning rate at 1×10^{-4}. As we can see, under the same largest training epoch, the NMSE of TransNet is significantly lower than CRNet and CLNet after about 40% of the training process. CRNet and CLNet are both DL methods based on convolutional neural network (CNN). TransNet is, after all, a relatively complex model from the perspective of FLOPS compared with them, while we still need a further analysis to its performance. In [17], a detailed comparison between the convolutional layer and the self attention layer is made, they are the core layers of CNN

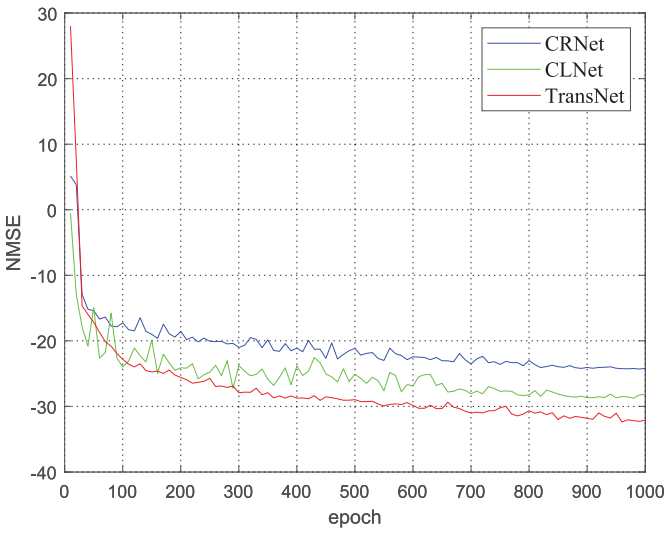

Fig. 3. Test NMSE (dB) trends of CRNet, CLNet and TransNet under different schemes from training with 1000 epochs, $\eta = 1/4$ and the scenario is indoor.

and Transformer architecture, respectively. It is concluded that compared with CNN, Transformer architecture has better capability to learn long-range dependencies. From the perspective of CSI matrix, this means TransNet, the relatively complex network with self attention layer as its core structure, has better capability to learn the relationship between CSI's row vectors x_i and x_j $(i, j = 1, 2 \cdots 2N_a)$ compared with traditional DL methods based on CNN when |*i*-*j*| is relatively large. Since in the (masked) multihead attention layer of TransNet, each row vector gives attention on other row vectors from CSI matrix, and every long-range dependence is better learned, this also means the relationships between local and global CSI are better captured during the compression and recovery process.

3) *Influence of Dimensions of Feature Expression:* Finally, we offer interesting observations about TransNet. In our main training results above, $d^{\text{model}} = 32$. We consider two other different dimensions of TransNet and show their performance under indoor scenarios from training with 400 epochs. The results are shown in Table II.

When $d^{\text{model}} = 16$, the results are not competitive in most of the η, while it shows slightly improvement compared with $d^{\text{model}} = 32$ at $\eta = 1/16$. When $d^{\text{model}} = 64$, it shows nearly the same results compared with $d^{\text{model}} = 32$ with $\eta = 1/8, 1/16$, respectively. While it has 0.89dB loss of NMSE at $\eta = 1/4$, it shows 1.17dB gain at $\eta = 1/32$. It is interesting that three different situations show nearly the same results at $\eta = 1/16$. Since in higher compression scale scenarios, especially when $\eta = 1/32$, $d^{\text{model}} = 64$ has better performance in NMSE, we recommend further study to use $d^{\text{model}} = 64$ or other high-dimensional expressions for features to advance feedback quality under high compressed scenarios for CSI.

V. Conclusion

In this letter, a novel neural network based on Transformer architecture named TransNet was proposed to handle with CSI feedback in FDD massive MIMO system. TransNet learns connections of CSI's various parts through a two-layer encoder-decoder network. Comparative experiments showed that with some bearable computing overhead, the proposed TransNet greatly outperformed other methods of DL in CSI feedback accuracy and got the SOTA.

TABLE II
NMSE(dB) Comparison of Different Dimensions of TransNet

η	$d^{\text{model}} = 32$		$d^{\text{model}} = 16$		$d^{\text{model}} = 64$	
	FLOPS	NMSE	FLOPS	NMSE	FLOPS	NMSE
1/4	35.72M	-29.22	35.68M	-26.62	35.78M	-28.33
1/8	34.70M	-21.62	34.64M	-16.94	34.73M	-22.51
1/16	34.14M	-14.98	34.11M	-15.14	34.21M	-15.17
1/32	33.88M	-9.83	33.85M	-8.59	33.95M	-11.00

References

[1] P. Kuo, H. T. Kung, and P. Ting, "Compressive sensing based channel feedback protocols for spatially-correlated massive antenna arrays," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, 2012, pp. 492–497.
[2] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Commun. Pure Appl. Math.*, vol. 57, no. 11, pp. 1413–1457, 2004.
[3] C. A. Metzler, A. Maleki, and R. G. Baraniuk, "From denoising to compressed sensing," *IEEE Trans. Inf. Theory*. vol. 62, no. 9, pp. 5117–5144, Sep. 2016.
[4] C. Li, W. Yin, and Y. Zhang, "User's guide for TVAL3: TV minimization by augmented lagrangian and alternating direction algorithms," Dept. CAAM, Rice Univ., Houston, TX, USA, Rep. 20, 2009.
[5] 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.
[6] 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.
[7] T. Wang, C. 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, Aug. 2019.
[8] Y. Yang, F. Gao, G. Y. Li, and M. Jian, "Deep learning-based downlink channel prediction for FDD massive MIMO system," *IEEE Commun. Lett.*, vol. 23, no. 11, pp. 1994–1998, Nov. 2019.
[9] Z. Liu, L. Zhang, and Z. Ding, "Exploiting bi-directional channel reciprocity in deep learning for low rate massive MIMO CSI feedback," *IEEE Wireless Commun. Lett.*, vol. 8, no. 3, pp. 889–892, Jun. 2019.
[10] 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.
[11] 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.
[12] 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.
[13] 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.
[14] Z. Lu, J. Wang, and J. Song, "Multi-resolution CSI feedback with deep learning in massive MIMO system," in *Proc. IEEE Int. Conf. Communications (ICC)*, 2020, pp. 1–6.
[15] 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.
[16] Y. Xu, M. Yuan, and M.-O. Pun, "Transformer empowered CSI feedback for massive MIMO systems," in *Proc. 30th Wireless Opt. Commun. Conf. (WOCC)*, 2021, pp. 157–161.
[17] A. Vaswani *et al.*, "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst. (ICONIP)*, vol. 30, 2017, pp. 5998–6008.
[18] L. Liu *et al.*, "The cost 2100 MIMO channel model," *IEEE Wireless Commun.*, vol. 19, no. 6, pp. 92–99, Dec. 2012.