CSI Reconstruction in UM-MIMO Systems Based on Improved ViT and TNT Networks

1st Yuchen Xia
School of Information Science and Engineering
Shandong University
Qingdao, China
yuchenxia@mail.sdu.edu.cn

2nd Bin Qi*
School of Information Science and Engineering
Shandong University
Qingdao, China
qib@mail.sdu.edu.cn

***Abstract*—Channel state information (CSI) plays a crucial role in sixth-generation communication systems, allowing them to adapt to changing channel conditions and optimize transmission quality. A compression restoration model, CSI-Net, based on ResNet, is proposed. However, it does not perform well under high compression ratios. Two novel networks are proposed to improve CSI reconstruction under high compression ratios in this paper. We notice that the Vision Transformer (ViT) model effectively extracts local information for overall features through attention mechanisms, and considering the potential correlation of CSI, we use the Transformer model to reconstruct CSI-Net. Furthermore, the Transformer in Transformer (TNT) model considers the local correlation of data, and we optimize CSI-Net accordingly based on the relevant structure. The improved ViT and TNT networks provide an excellent CSI compression and reconstruction solution. The reconstructed CSI is highly correlated with the original CSI, even under high compression ratios.**

***Keywords*-CSI, ViT, Attention mechanisms, TNT, compression & reconstruction.**

I. Introduction

In mobile communication networks, with the movements of devices, the channels between base stations and users often exhibit different characteristics due to factors such as space and time. In Frequency Division Duplexing (FDD) systems, to ensure high reliability and high data rates in communication, the receiver needs to provide feedback of the CSI to the transmitter after receiving data. In order to greatly increase the channel capacity, ultra massive multiple-input multiple-output (UM-MIMO), a major technology of the sixth-generation wireless communication [1], is widely researched. Although the cost of CSI feedback is low for single-input single-output (SISO) systems, for UM-MIMO systems, the cost of transmitting complete CSI feedback hinders the improvement of transmission efficiency significantly.

The improvements in transmission efficiency based on traditional encoding methods are limited, so an innovative scheme to use deep learning based on ResNet for CSI compression and reconstruction is proposed [2]. When the compression ratio is low, CSI-Net can encode and reconstruct the original CSI quite well. However, at higher compression ratios, it cannot reconstruct the CSI very well. CSI-Net is a network structure based on ResNet, which inherits the characteristics of classic deep neural networks and extracts the features through convolution and other methods. To address the issue of poor reconstruction performance at high compression ratios, many researchers proposed different optimization directions and structures for the CSI-Net network [3], [4]. However, these improvements either require the addition of separate memory networks after the CSI-Net model or their optimization effects are relatively not significant.

With the proposal and application of attention mechanisms, a new solution has emerged [5]. The Transformer model is built on the attention mechanism, and through this mechanism, we can better capture the features and dependencies of the input. In recent years, based on the feature extraction capability of the Transformer model, some researchers have proposed the ViT [6] to process image information. ViT adds a conversion step from images to sequences on top of the Transformer model, transforming image segmentation into sequences and image relationships into dependencies between sequences. Further, to obtain the relationships between superpixels within patches, some researchers proposed the TNT model [7]. This model can balance both local and global relationships, as well as fine-grained local information. With these network structures and by leveraging the characteristics of CSI matrices, we have achieved compressed reconstruction of CSI matrices under high compression ratios. The main contributions of this paper can be concluded as follows:

1) We introduce attention mechanism and Transformer to build the improved ViT and TNT networks for CSI-Net, which replace the original ResNet-based networks and effectively improve the efficiency of feature extraction.
2) Following the principles of ViT and TNT, the encoder's efficiency and effectiveness are improved by preprocessing the CSI, ultimately significantly improving the reconstruction performance.
3) To adapt the ViT and TNT networks to the function of compression and reconstruction, we adjust the CSI preprocessing by changing the target mapping space. The improved ViT and TNT Networks achieve high-quality CSI reconstruction at high compression ratios in UM-MIMO systems.

In this paper, the COST-2100-Channel dataset [8] is used.

The section II explains the methodology used, while the section III provides a detailed description of the structures of the improved ViT and TNT networks. The results are compared and analyzed in the section IV, leading to the conclusions.

II. METHODOLOGY

In this section, the mechanisms and principles required for the model will be explained so that readers can gain a clear understanding of the model solely through this paper. The core mechanisms will be divided into four parts: Attention Mechanism, Transformer, ViT & TNT, and Model Processing Procedure & Reconstruction Evaluation Metric.

A. Attention Mechanism

In traditional deep-learning networks, it is often necessary to increase the number of network layers to improve model accuracy. ResNet addresses the problems of network degradation and gradient vanishing/exploding that occur with large numbers of layers in traditional networks, allowing the network to continue to develop deeper. However, as the network depth increases, it causes a sharp expansion of parameters, which reduces the efficiency of the model in processing information and results in an explosive increase in training time.

The attention mechanism can be trained to allocate more computation to valuable parts of the input while reducing or ignoring attention to other parts, resulting in better processing efficiency with the same parameter size. Given an input $Source$, in the attention mechanism, we view the input as a collection of information and transform $Source$ into a sequence of $\langle Key, Value\rangle$ pairs. Key represents the feature vector of the information, while $Value$ represents the weight value. During training, given a target $Query$, attention can be expressed as

$$\begin{aligned}&\mathrm{Attention}(Query, Source, Value)\\&=\sum_{i=1}^{L_x}\text{Similarity}\left(Query, Key(i)\right)*Value(i).\end{aligned} \tag{1}$$

In the formula, L_x represents the number of $\langle Key, Value\rangle$ pairs obtained after transformation. By adjusting the weight values through training, we can extract the content according to its importance. Therefore, using the attention mechanism, we can more easily obtain higher-order features with the same model size, achieving higher performance.

B. Transformer

Based on the aforementioned attention mechanism, the Transformer model was proposed. In Transformer, self-attention is used, and its implementation can be represented

$$\begin{aligned}&\mathrm{Attention}(Query, Key, Value)\\&=\mathrm{softmax}\left(\frac{Query*Key^T}{\sqrt{d_k}}\right)*Value.\end{aligned} \tag{2}$$

In the formula, d_k represents the dimensionality of the key vectors. The purpose of introducing d_k is to ensure that the attention values are not too large or too small, and to make the gradient calculation more stable during backpropagation. The original Transformer is developed for natural language processing applications such as machine translation, with inputs and outputs in sequential format. To make it applicable in other fields, various variations have been derived, such as Sandwich Transformers, Universal Transformers, Transformer-XL, Reformer, and Vision Transformer [9]–[12].

C. ViT & TNT

To implement a Transformer-based model for image processing, ViT proposes to divide the image into patches and flatten them into a sequence to fit the input of the encoder. The model's objective is to recognize and classify images, and the Multilayer Perceptron (MLP) is simplified into fully connected layers to achieve efficient image recognition with a smaller parameter size.

Indeed, for images, the information is often not limited to the values of the sequence itself and the relationships between the sequences after segmentation but also exists in the relationships within the sequence. Therefore, the TNT model proposes to apply another Transformer to the segmented sequences to extract local detailed information of images further. This approach has achieved higher accuracy than the ViT model on corresponding datasets. Therefore, it provides a possibility for CSI-Net to achieve better results.

D. Model Processing Procedure & Reconstruction Evaluation Metric

The input data is assumed as M_x, a matrix with a size of $2 \times 32 \times 32$. In the traditional transmission process, a total of 2048 data units are needed to be transmitted to reconstruct the data. To improve the speed of communication information exchange, we attempt to transmit fewer data units to reconstruct the data within an acceptable level of accuracy. Therefore, the key task of the model is compression and reconstruction. We assume that the coded sequence after compression and encoding by the model is C_x, which is a matrix with a size of $d \times 1$. The compression ratio γ of the model can be easily obtained by

$$\gamma = \frac{d}{2048}. \tag{3}$$

In this scenario, the overall model implements encoding $G(x)$ and decoding $D(x)$ such that $G(x)$ and $D(x)$ satisfy the following relationships

$$G(M_x) = C_x, \tag{4}$$

$$D(C_x) = M_x'. \tag{5}$$

M_x' in the (5) serves as the output of the model. In order to reconstruct the CSI based on the compressed codeword C_x, the output M_x' needs to approach the actual value M_x. And for the inverse function of (4) $\mathrm{IF}(y) = G^{-1}(x)$, in order to reconstruct the original CSI, the following formula must hold

$$\mathrm{IF}(y) = D(x). \tag{6}$$

However, there is a problem with such an evaluation criterion: it is difficult to measure the similarity between $\mathrm{IF}(y)$ and $D(x)$.

Therefore, we measure the quality of reconstruction by comparing the output results with the input results. In statistics, normalized mean squared error (NMSE) and correlation coefficient (ρ) are commonly used to measure the correlation between data. NMSE and ρ satisfy

$$\mathrm{NMSE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|M_x(x_i) - M_x'(x_i)|^2}{|M_x(x_i)|^2} \tag{7}$$

$$\rho = \frac{\sum_{i=1}^{n}(M_x(x_i) - \overline{M_x})(M_x'(x_i) - \overline{M_x'})}{\sqrt{\sum_{i=1}^{n}(M_x(x_i) - \overline{M_x})^2}\sqrt{\sum_{i=1}^{n}(M_x'(x_i) - \overline{M_x'})^2}} \tag{8}$$

where n is the number of samples, M_x is the actual value of the input, M_x' is the output of the model, and $\overline{M_x}$ and $\overline{M_x'}$ are the mean values of M_x and M_x'. In practical implementation, to achieve modularity, (8) is often rewritten in the following form

$$\rho = r(M_x, M_x') = \frac{\mathrm{Cov}(M_x, M_x')}{\sqrt{\mathrm{Var}[M_x]\,\mathrm{Var}[M_x']}}. \tag{9}$$

A smaller NMSE or a larger ρ indicates a better reconstruction result. Among these two evaluation criteria, NMSE is more affected by the number of training epochs. When NMSE enters a range close to 0, there are often significant differences due to different training conditions. In comparison, ρ is more effective in measuring the relationship between the input and output.

III. CSI-ViT AND CSI-TNT

In the COST-2100-Channel dataset, the size of the CSI we obtained is $2\times32\times32$. This information covers the amplitude-frequency and phase-frequency response information between antennas. Departing from traditional processing methods, we treat the input as an image consisting of two layers with a size of 32×32.

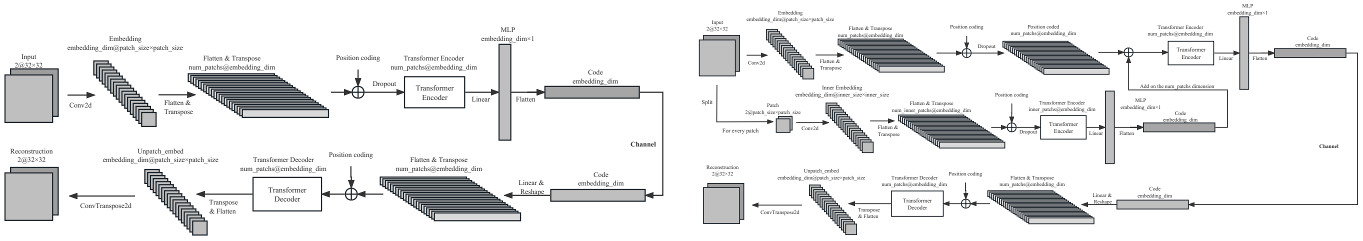
From the Fourier expansion of the signal, we know that different frequency points are related during the transmission process, often multiples of the fundamental frequency. However, these positions are often not contiguous in the CSI matrix, representing the channel response in both the time and frequency domains. Therefore, unlike in the traditional ViT model, where the image is directly sliced into patches and embedded into the Transformer Encoder, where the image is first convolved to map its features to the required dimensions. At the same time, the initial features of the image are extracted through this convolution process. When selecting patches, choosing too large patches can lead to a decrease in the effectiveness of feature extraction by the subsequent Transformer Encoder. On the other hand, choosing too small patches can increase the number of sequences input to the Transformer Encoder, reducing efficiency and making overfitting more likely. Therefore, based on an input size of 32×32, 4×4 patches are chosen to balance these considerations. At this step, we still have a multi-layered small image. In order to input it into the Transformer Encoder, we flatten it and exchange its dimensions. At this point, we obtain the transformed initial sequence for each small block. In the compression and reconstruction stage, we do not need to extract macroscopic object information from the image like in the ViT model. Therefore, we do not need to introduce an additional ”cls” token to record this information to save computational costs. However, in the data provided by the COST-2100-Channel dataset, the position information of the small blocks is important for the overall reconstruction process. Thus, we still need to introduce position encoding to capture the spatial information of the small blocks. After introducing position encoding, the sequence generated from an image is already equipped with positional information. This sequence is then fed into the Transformer Encoder. Within the Transformer Encoder, the sequence information is processed using the attention mechanism. After processing, the Transformer Encoder outputs a sequence of sequence groups with dimensions identical to the input of the Transformer Encoder. For this output sequence, each processed sequence of each segmented image block contains information from other blocks. Referencing all of this information would introduce redundancy. Therefore, an MLP is used to extract information from all the processed sequences, resulting in an encoding with dimensions identical to each small block-generated sequence. The output from the MLP is in the form of a column vector, which can be flattened into a row vector for convenience, although this step is optional. The resulting encoding can then be transmitted through a channel.

An output dimension needs to be provided to implement the seq2seq problem in the classic Transformer structure. Therefore, when the receiver receives the encoding from the channel, it needs to be expanded to the same dimension as the input of the Transformer Encoder to specify the output dimension for the Transformer Decoder. Additionally, to restore the original positional information, the expanded sequence group needs to be augmented with position encoding before being fed into the Transformer Decoder. It should be noted that the output of the Transformer Decoder is still the result of embedding the image. Deconvolution is required to reconstruct the image from the embedded result.

The above is the CSI-ViT model, based on ViT but reconstructed and improved, as shown in Fig. 1a.

Although the CSI-ViT model has better performance in CSI restoration compared to other models, and the time cost is relatively acceptable, in some fields where high precision is required, sacrificing some time for improved accuracy is acceptable. Therefore, based on CSI-ViT, modifications were made to the encoding end of the model to improve the accuracy of image restoration, resulting in the CSI-TNT model, as shown in Fig. 1b.

As mentioned in the previous section on CSI-ViT, the size of the initial image cropping can affect the final reconstruction result, and CSI-ViT attempts to find a balance point. In CSI-TNT, we aim to achieve better reconstruction results through improved partitioning and feature relationship extraction. Inspired by the TNT model, we further improved



(a) CSI-ViT.

(b) CSI-TNT.

Fig. 1. The specific implementation of the encoder-decoder architecture for CSI-ViT and CSI-TNT networks.

the already effective CSI-ViT. In CSI-ViT, the image is already cropped, allowing the Transformer Encoder to extract the relationship between the corresponding patches and the entire image. However, it cannot extract the relationships within each patch effectively. Therefore, we perform a similar operation on each small block of the original input image that was cropped, which is then sent to the Transformer Encoder after undergoing Embedding and Position coding. After passing through MLP, we obtain the feature information for each small patch. These pieces of information are added to the sequence group generated by Embedding and Position coding for the entire image. This allows the Transformer Encoder to process the details better when the entire sequence is sent through it.

IV. SIMULATIONS

We first generated datasets of two environments using the COST-2100-Channel model: 1) the indoor picocellular scenario at the 5.3 GHz band and 2) the outdoor rural scenario at the 300 MHz band. There are 100,000 samples in the training set, 30,000 samples in the validation set, and 20,000 samples in the testing set. The testing set is completely separate from the training and validation sets. We trained the model using the training set, saved the best-performing model on the validation set, and finally tested the model on the testing set.

Compared to the 1,000 rounds of training for CSI-Net and 3,000 rounds of training for CSA-Net [13], the model restructured with Transformer has better convergence. Therefore, we only trained it for 200 rounds and validated its effectiveness by comparing the results with those obtained through fewer training rounds. To improve accuracy, we set a dynamic learning rate with an initial value of 0.001. When the number of training rounds reaches 100 and 150, the learning rate is reduced to one-tenth of the previous stage to avoid oscillation near the optimal point caused by a relatively large learning rate. Additionally, we used the Adam optimizer to prevent the model from getting trapped in local optima. The batch size is set to 128.

The models included in the comparison were the original CSI-Net and CSA-Net, which only use attention mechanisms. The results of reproducing CSI-Net on the data were slightly worse than those reported in the original paper. Therefore, the data in the table directly quote the results from the original paper. Among the two branch models of the CSA-Net, we selected the best results obtained in any environment at any compression ratio. In Table I, we test the numerical values of two metrics for the four models at different compression ratios in different environments.

TABLE I
NMSE IN dB AND COSINE SIMILARITY ρ

γ	Method	Indoor		Outdoor	
		NMSE	ρ	NMSE	ρ
$\frac{1}{4}$	CSI-Net	-17.36	0.99	-8.75	0.91
	CSA-Net	-34.18	0.99	-14.73	0.95
	CSI-Vit	**-45.48**	**0.9693**	**-43.33**	**0.9711**
	CSI-TNT	**-67.34**	**0.9998**	**-62.41**	**0.9995**
$\frac{1}{16}$	CSI-Net	-8.65	0.93	-4.51	0.79
	CSA-Net	-15.54	0.96	-6.29	0.86
	CSI-Vit	**-55.84**	**0.9973**	**-54.69**	**0.9979**
	CSI-TNT	**-64.02**	**0.9995**	**-62.84**	**0.9997**
$\frac{1}{32}$	CSI-Net	-6.24	0.89	-2.81	0.67
	CSA-Net	-11.24	0.94	-4.45	0.79
	CSI-Vit	**-55.47**	**0.9972**	**-52.79**	**0.9971**
	CSI-TNT	**-63.49**	**0.9995**	**-55.91**	**0.9981**
$\frac{1}{64}$	CSI-Net	-5.84	0.87	-1.93	0.59
	CSA-Net	-6.56	0.86	-2.86	0.67
	CSI-Vit	**-48.35**	**0.9859**	**-41.34**	**0.9501**
	CSI-TNT	**-50.93**	**0.9926**	**-44.71**	**0.9732**
$\frac{1}{128}$	CSI-Net	-	-	-	-
	CSA-Net	-	-	-	-
	CSI-Vit	**-39.73**	**0.9086**	**-33.07**	**0.6380**
	CSI-TNT	**-40.39**	**0.9296**	**-33.58**	**0.6891**

The precision of the CSI-Net/CSA-Net models in the original paper is shown to two decimal places in the table. Due to unsatisfactory restoration results, testing was not conducted for these two models at a compression ratio of $\frac{1}{128}$, resulting in no data for these models at this compression ratio.

In the original CSI-Net model, the decrease in restoration performance in indoor environments is less apparent as the compression ratio increases, but the restoration performance is still unsatisfactory under high compression ratios. In outdoor environments, even a slight increase in compression ratio causes a sharp drop in restoration performance, and even at lower compression ratios, ideal results cannot be achieved. The CSI-ViT and CSI-TNT models can maintain a correlation coefficient of 0.95 or higher between the reconstructed CSI matrix and the original CSI matrix, even when the compression ratio reaches $\frac{1}{64}$, and can reconstruct the CSI matrix well. At a compression ratio of $\frac{1}{64}$, compared to CSI-Net, CSI-ViT, and CSI-TNT respectively improve by 11.59% and 12.26% in indoor environments, and by 36.01% and 38.32% in outdoor environments. Compared to CSA-Net, they respectively improve by 12.59% and 13.26% in indoor environments and by 28.01% and 30.32% in outdoor environments. This greatly improves the quality of the reconstructed CSI matrix under high compression ratios.

To demonstrate the quality of the reconstruction, we show in Fig. 2 the CSI input to the model as well as the output reconstructed by the CSI-Net, CSI-ViT, and CSI-TNT models under a compression ratio of 1/32 in an outdoor environment. To facilitate distinction, we normalized the images during the output process. Under this condition, the values of ρ for CSI-Net, CSI-ViT, and CSI-TNT are 0.67, 0.9971, and 0.9981, respectively. We can clearly see the superiority of CSI-ViT and CSI-TNT through Fig. 2.

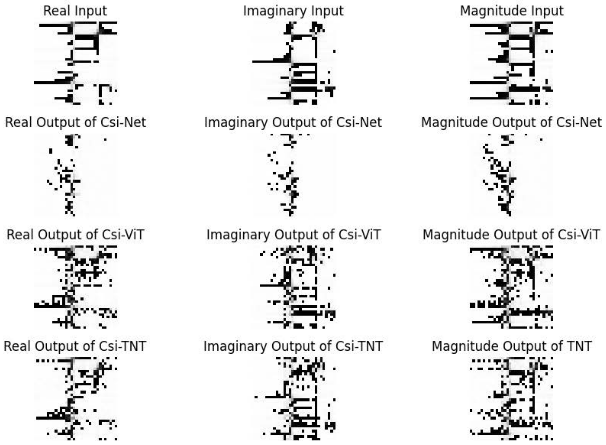


Fig. 2. The input CSI and the output reconstructed by the CSI-Net, CSI-ViT, and CSI-TNT models under a compression ratio of $\frac{1}{32}$ in an outdoor environment.

To achieve better reconstruction, the model generates an encoding for each local region, and the final encoding is the sum of all the local region encodings. Therefore, with a relatively low compression ratio, the model has a slightly larger number of parameters and computations. As the compression ratio increases, the model's reconstruction quality remains high, but the number of parameters and computations sharply decreases, resulting in both low computational complexity and high correlation at high compression ratios. To measure the complexity of the model, we introduce the concept of floating point operations (FLOPs), which represent the total number of floating point operations required by the model. In Fig. 3, we test and output the FLOPs of the model. We can clearly see that the model has achieved a decent CSI reconstruction.

The advantage of CSI-ViT and CSI-TNT lies in high compression environments. From Fig. 3, we can see that when the compression ratio is $\frac{1}{32}$ or higher, the increase in computational complexity is not significant, but it can greatly increase the correlation between the reconstructed CSI matrix and the original CSI matrix. The characteristic of CSI-ViT and CSI-TNT models is that their parameter count decreases as the compression ratio increases. This enables these models to achieve better performance than the original CSI-Net under high compression ratios, with relatively shorter time and computational costs. A clear example is CSI-ViT at a compression ratio of $\frac{1}{128}$, which not only achieves better reconstruction performance than CSI-Net and CSA-Net at a compression ratio of $\frac{1}{64}$ but also has a smaller computational cost than CSA-Net.

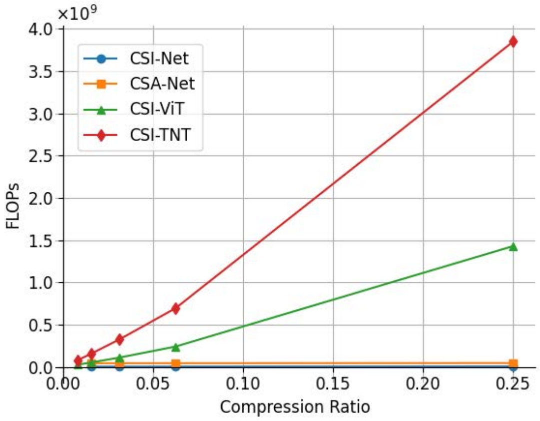


Fig. 3. The relationship between FLOPs and compression ratio for different models.

V. Conclusion

This paper introduced the ViT and TNT models into the traditional CSI-Net. By incorporating attention mechanisms and reasonable preprocessing of CSI, such as adjusting the mapping dimensions of information and performing preliminary feature extraction, and by improving the ViT and TNT models to efficiently complete the CSI compression and reconstruction process, the CSI-ViT and CSI-TNT networks were constructed. These two networks effectively addressed the weak CSI matrix reconstruction capability of CSI-Net under high compression ratios, thus promoting the further reduction in CSI feedback time and enabling communication networks to achieve even lower latency.

References

[1] C.-X. Wang et al., “On the road to 6G: Visions, requirements, key technologies and testbeds,” IEEE Communications Surveys & Tutorials, Early Access, Feb. 2023.

[2] C.-K. Wen, W.-T. Shih, and S. Jin, “Deep learning for massive MIMO CSI feedback,” IEEE Wireless Communications Letters, vol. 7, no. 5, pp. 748-751, Oct. 2018.

[3] S. Ji and M. Li, “CLNet: Complex input lightweight neural network designed for massive MIMO CSI feedback,” IEEE Wireless Communications Letters, vol. 10, no. 10, pp. 2318-2322, Oct. 2021.

[4] Y. Sun, W. Xu, L. Fan, G. Y. Li, and G. K. Karagiannidis, “AnciNet: An efficient deep learning approach for feedback compression of estimated CSI in massive MIMO systems,” IEEE Wireless Communications Letters, vol. 9, no. 12, pp. 2192-2196, Dec. 2020.

[5] A. Vaswani et al., “Attention is all you need,” Advances in neural information processing systems, vol. 30, Dec. 2017.

[6] A. Dosovitskiy et al., “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” ArXiv, vol. abs/2010.11929, Jun. 2021.

[7] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang, “Transformer in Transformer,” ArXiv, vol. abs/2103.00112, Dec. 2021.

[8] L. Liu et al., “The COST 2100 MIMO channel model,” IEEE Wireless Communications, vol. 19, no. 6, pp. 92-99, Dec. 2012.

[9] O. Press, N. A. Smith, and O. Levy, “Improving transformer models by reordering their sublayers,” arXiv preprint arXiv:1911.03864, Apr. 2020.

[10] M. Dehghani, S. Gouws, O. Vinyals, J. Uszkoreit, and L. Kaiser, “Universal Transformers,” ArXiv, vol. abs/1807.03819, Mar. 2019.

[11] Z. Dai, Z. Yang, Y. Yang, J. G. Carbonell, Q. V. Le, and R. Salakhutdinov, “Transformer-XL: Attentive Language Models beyond a Fixed-Length Context,” ArXiv, vol. abs/1901.02860, Jun. 2019.

[12] N. Kitaev, L. Kaiser, and A. Levskaya, “Reformer: The Efficient Transformer,” ArXiv, vol. abs/2001.04451, Feb. 2020.

[13] Q. Liu, J. Sun, S. Qiu, Y. Lv, and X. Du, “A Convolutional Self-Attention Network for CSI Reconstruction in MIMO System,” Wireless Communications and Mobile Computing, vol. 2023, pp. 1-10, Jan. 2023.