Received: 24 May 2022 | Revised: 1 October 2022 | Accepted: 21 October 2022

DOI: 10.4218/etrij.2022-0209

ORIGINAL ARTICLE

ETRI Journal WILEY

Deep learning-based scalable and robust channel estimator for wireless cellular networks

Anseok Lee | **Yongjin Kwon** | **Hanjun Park** | **Heesoo Lee**

Intelligent Wireless Access Research Section, Mobile Communication Research Division, Telecommunications & Media Research Laboratory, Electronics and Telecommunications Research Institute, Daejeon, Republic of Korea

Correspondence
Anseok Lee, Intelligent Wireless Access Research Section, Mobile Communication Research Division, Telecommunications & Media Research Laboratory, Electronics and Telecommunications Research Institue, Daejeon, Republic of Korea.
Email: alee@etri.re.kr

Funding information
This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (2021-0-00972, Development of Intelligent Wireless Access Technologies)

Abstract
In this paper, we present a two-stage scalable channel estimator (TSCE), a deep learning (DL)-based scalable, and robust channel estimator for wireless cellular networks, which is made up of two DL networks to efficiently support different resource allocation sizes and reference signal configurations. Both networks use the transformer, one of cutting-edge neural network architecture, as a backbone for accurate estimation. For computation-efficient global feature extractions, we propose using window and window averaging-based self-attentions. Our results show that TSCE learns wireless propagation channels correctly and outperforms both traditional estimators and baseline DL-based estimators. Additionally, scalability and robustness evaluations are performed, revealing that TSCE is more robust in various environments than the baseline DL-based estimators.

KEYWORDS
channel estimation, deep learning, wireless cellular networks

1 | INTRODUCTION

5G wireless cellular networks are critical to mobile connectivity and many 5G-based applications, such as mobile communications, Internet of Things, and ultrareliable and low latency communication, according to Electronics and Telecommunications Research Institute (ETRI) [1]. Beyond 5G, 6G networks are expected to support a wide range of futuristic applications, including virtual and augmented reality services and holographic communication services. To meet the performance requirements of 6G, much effort will be expended; for example, higher frequencies will be extensively investigated to exploit much broader spectrums for extremely high throughput.

Artificial intelligence (AI), which is mainly represented as the machine learning (ML) and deep learning (DL) technologies, is one of the key ingredients for realizing 6G [2]. Breakthroughs in storage and computation technology, as well as the development of neural network models and algorithms, have largely contributed to recent advances in AI technologies. Many successful stories have emerged from AI research and applications in computer vision and natural language processing (NLP). Natural AI model capabilities such as feature discovery and complex representation played important roles in these areas. Wireless communications are also attempting to leverage advances in AI technologies, particularly in areas where the optimal solution does not exist or the complexity is too high to be implemented in practice, according to Shafin and others [3]. In 3GPP, a study of the benefits of augmenting the NR air interface with improved support of AI/ML-based algorithms is started [4]. Channel estimations are also being focused on using AI for improved estimation performances with reduced complexity.

Channel estimation is one of the fundamental problems in wireless communication because wireless signals are inevitably distorted, noised, and interfered with during wireless transmission medium experiences. By observing the distorted propagation channel, which is done by channel estimation, the transmitted signal can be correctly recovered and demodulated at the receiver. However, data rates are extremely high in 6G cellular networks, and accurate channel variations are required. Furthermore, increasing the number of transmission and reception antennas and data streams for high throughput makes channel estimation even more difficult, and communication bandwidths are becoming larger as higher frequency bands such as millimeter waves (mmWave) and terahertz (THz) bands are used.

In this paper, we present the two-stage scalable channel estimator (TSCE), a DL-based scalable and robust channel estimator. TSCE comprises two DL networks that can support different resource allocation sizes and reference signal (RS) patterns. The main contributions of this paper are as follows:

(1) We develop a novel DL architecture sequentially processing different resource units, which are a RS and a resource block (RB), for scalable and robust channel estimations.
(2) We propose a window averaging multihead self-attention (WA-MSA) operation for efficient global feature extractions of a large number of RSs. To our knowledge, the proposed DL-based channel estimator is the first channel estimator process sequence of RSs.
(3) We provide extensive simulation results, not only the channel estimation performances but also the generalization performances for scalability and robustness studies. To the best of our knowledge, this paper is the first extensive study considering both resource allocation sizes and RS configurations.

2 | RELATED WORKS

Recently, channel estimations leveraged by DL technologies have been widely studied for wireless cellular networks. Ye and others [5] presented a pioneer work of applying DL in channel estimation, which proposed a deep-neural network (DNN) for channel estimation and signal detections using the multilayer perceptron (MLP). By the ability of MLP to learn the characteristics of wireless channels, the proposed estimator outperforms traditional algorithms, such as least square (LS) and linear minimum mean squared error (LMMSE) algorithms. Furthermore, evaluations on nonideal environments are performed, and the results show that using the MLP results in less performance degradation than traditional methods. However, although MLP architecture is efficient in extracting features from received RSs via the high connection-density of nodes among hidden layers, MLP-based channel estimators are not scalable for various configurations, such as resource allocation sizes and the number of RSs, in general. Therefore, multiple DNNs might be defined and independently trained for each configuration because the number of inputs and outputs of the estimators are fixed.

Time-frequency responses of wireless propagation channels can be considered as 2D images, enabling the use of numerous AI technologies previously developed for image processing, which is a major application of recent AI. In Soltani and others [6], a pipelined image processing technique, which is super resolution (SR) and image restoration (IR), is proposed to obtain denoised full resources' channel responses from the interpolated resource grid using received RSs. Popular convolutional neural networks (CNNs)-based architectures, which are SRCNN [7] and DnCNN [8], are used as underlying networks for SR and IR of the estimator. In Li and others [9], deep residual network-based channel estimator [10] is introduced. The estimator in Li and others [9] proposes to perform postupsampling as a transpose convolution layer and all computation before the upsampling using the received pilot signal for low computations. Unlike MLP-based estimators, CNN-based estimators can be designed to be scalable to various input sizes; for example, fully convolutional networks can accept variable sizes of inputs. Although CNN-based estimators are scalable, it is still necessary to consider the robustness of scenarios and/or configurations on which the DL-based estimators are not trained.

Generative adversarial network (GAN) [11] is an important branch of DL and can generate realistic artificial images with the help of a discriminator network to determine whether the output of a generator network is real or not. In Radford and others [12], deep convolutional GAN (DCGAN) -based network [13] and an algorithm using the network for channel estimation were proposed. [14] also presented super resolution GAN (SRGAN)-based channel estimator [15], where the generator network is designed and trained to generate realistic channels using received pilot signals. The characteristics of scalability and robustness of each GAN-based channel estimator follow the design of the generator network, and the generator network of both Balevi and Andrews [12] and Ledig and others [15] is based on CNN.

Self-attention mechanism and transformer network [16] have great successes in various NLP applications, for example, Devlin and others [17], by its powerful feature extractions, and also for various image processing applications [18]. Several works [19-22] presented channel estimations based on transformer network or attention mechanisms. A transformer network can be designed to process variable lengths of sequences as inputs and outputs, which is beneficial for designing scalable channel estimators. However, several transformer-based estimators, including in Li and Peng [20] and Luan and Thompson [22], might not process variable size of resource grids through fully connected layers and/or upscaling modules, whose input and output dimensions should be predetermined to a specific configuration. Moreover, as shown in the later part of this paper, the baseline transformer-based estimator tends to learn a specific scenario in the training dataset rather than general relationships, like other baseline DL-based estimators. This study aims to design a channel estimator that has scalability while also having robustness by leveraging the feature extraction capabilities of the transformer network.

3 | SYSTEM MODEL

We consider orthogonal frequency division multiplexing (OFDM) wireless networks with T transmission antennas and R reception antennas. The received signal at the receiver's nth receive antenna on kth symbol, ith subcarrier can be represented as

$$y_{k,i,n} = h_{k,i,n} s_{k,i} + z_{k,i,n}, \tag{1}$$

where $s_{k,i}$ are transmission signals, $h_{k,i,n}$ are precoded channels, and $z_{k,i,n}$ are additive white gaussian noises. The precoded channels $h_{k,i,n}$ are

$$h_{k,i,n} = \sum_{m=1}^{N_s} \tilde{h}_{k,i,n,m} w_{k,i,m}, \tag{2}$$

where $\tilde{h}_{k,i,n,m}$ are channel responses between mth transmission antenna and nth reception antenna and $w_{k,i,m}$ are precoding coefficients of mth transmission antenna. For each reception antenna, the received signal of a nth reception antenna on a slot is

$$\mathbf{Y}_n = \mathbf{H}_n \otimes \mathbf{s} + \mathbf{z}_n, \tag{3}$$

where $\mathbf{H}_n$, $\mathbf{s}$, and $\mathbf{z}_n$ are channel response, transmitted signal, and noises on the resource grid (i.e., a grid of N_s symbols and N_f subcarriers), respectively, and the symbol $\otimes$ denotes the Hadamard product, and it is also known as the element-wise product. A portion of the resource grid is used to transmit known signals $\mathbf{s}$, also known as RSs, and the receiver exploits RSs to estimate entire channel responses $\mathbf{H}_n \in \mathbb{C}^{N_s \times N_f}$. The LS method is used for estimating the response on the positions of RSs as

$$\mathbf{h}^{\mathrm{LS}}_{p,n} = \frac{\mathbf{y}_{p,n}}{\mathbf{s}_p}, \tag{4}$$

where $\mathbf{y}_{p,n}$ are received signals on RSs positions and $\mathbf{s}_p$ is the transmitted RS. The simplest method for estimating the entire channel is to use two-dimensional (2D) interpolation of LS estimations.

The LMMSE is another well-known estimation method that outperforms the LS in terms of mean squared error (MSE). The LMMSE estimator can be expressed as follows:

$$\mathbf{H}^{\mathrm{LMMSE}}_n = \mathbf{R}_{Hh_p,n} \left(\mathbf{R}_{h_p h_p,n} + \frac{\sigma^2_{z,n}}{\sigma^2_s} \mathbf{I} \right)^{-1} \mathbf{h}^{\mathrm{LS}}_{p,n}, \tag{5}$$

where the $\mathbf{R}_{Hh_p,n}$ is the cross-correlation matrix between the channel of the entire resource grid and the RS positions and the $\mathbf{R}_{h_p h_p,n}$ is the autocorrelation matrix of the channel of the RS positions. $\sigma^2_{z,n}$ and σ^2_s are variances of noise and signal power at nth receive antenna, respectively. Note that to use LMMSE estimation, the correlation matrices must be known; however, accurate estimation of correlation values is problematic in many situations [23]. Correlation values can be modeled using the average delay spreads, and the maximum doppler frequency [24] and an additional reference signal, the tracking reference signals, is introduced to precisely measure delay and doppler spreads in the 5G NR system [25].

4 | TSCE

In this section, we present a DL-based channel estimator, TSCE, for scalable and robust channel estimations.

4.1 | Overall architecture

Figure 1 depicts an overview of the proposed estimator, TSCE, which is made up of two transformer networks: denoising network (DeNet) for denoising and upscaling network (UpNet) for upscaling. Note that the two-stage processing is similar to ChannelNet, where ChannelNet does upscaling first and then performs denoising. DeNet is designed to process sequences of RSs with positions and side information for robust denoising to various RS configurations. Since the length of RS sequences can be

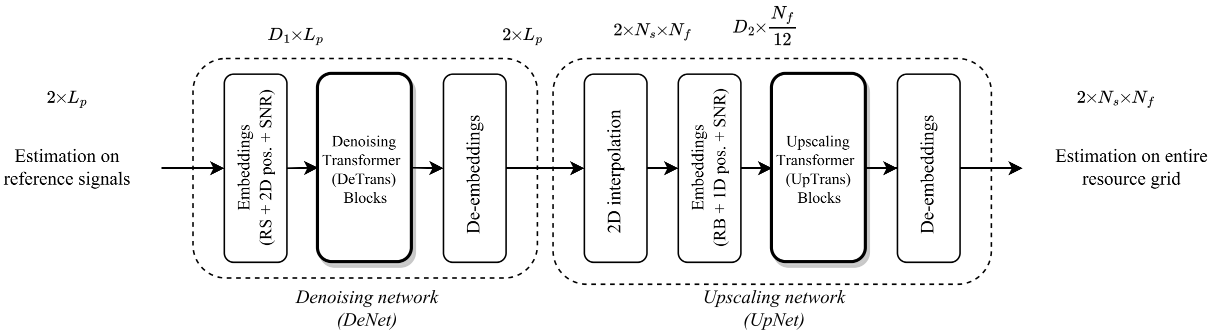

$2{\times}L_p$
Estimation on reference signals
$D_1{\times}L_p$
Embeddings (RS + 2D pos. + SNR)
Denoising Transformer (DeTrans) Blocks
$2{\times}L_p$
De-embeddings
Denoising network (DeNet)
$2{\times}N_s{\times}N_f$
2D interpolation
$D_2{\times}\frac{N_f}{12}$
Embeddings (RB + 1D pos. + SNR)
Upscaling Transformer (UpTrans) Blocks
De-embeddings
Upscaling network (UpNet)
$2{\times}N_s{\times}N_f$
Estimation on entire resource grid

FIGURE 1 The overview of TSCE network structure

very long for estimations of large bandwidths, computationally efficient feature extractions were required for DeNet. Unlike DeNet, UpNet processes a sequence of RBs for computationally efficient upscaling. Both DeNet and UpNet utilize a transformer network that uses a bi-LSTM layer for more efficient information flows among RS and RB sequences.

A resource grid $\mathbf{H}_p$ contains channel estimates $\mathbf{h}_p^{\mathrm{LS}} \in \mathbb{C}^{L_p}$ on known RSs, for example, demodulation RS (DMRS) in 5G NR systems. And the estimates $\mathbf{h}_p^{\mathrm{LS}}$ are given to TSCE. Because the estimates $\mathbf{h}_p^{\mathrm{LS}}$ are vector of complex numbers, the shape of input to the estimator is $2 \times L_p$, where L_p is number of RSs. The output of the estimator is the complex channel response of the entire resource grid $\hat{\mathbf{H}}$ and shaped $2 \times N_s \times N_f$. It should be noted that the resource grids of each receive antenna can be treated independently, so the receive antenna index n is omitted from this section.

4.2 | DeNet

A detailed structure of DeNet is in Figure 1. DeNet gets an input of estimated channel responses on RSs in real and imaginary parts, that is, $\mathbf{h}_p^{LS} \in \mathbb{R}^{2\times L_p}$. The first part of DeNet is an embedding where the responses on each RSs are converted into D_1 dimensions of RS embeddings using a learnable linear projection neural network layer. We also used a signal-to-noise (SNR) value as side information for the linear embedding layer. Since each RS has two-dimensional position information of subcarrier and symbol indices, the fixed 2D Sin-Cos position embeddings [26] are added into each RS embedding.

Then, denoising transformer (DeTrans) blocks receive the RS embeddings. The structure of DeTrans blocks is in Figure 2A. Each DeTrans block consists of three successive transformer encoders from Dosovitskiy and others [18]. First transformer encoders have window-based multihead self-attention (W-MSA) layer [27] instead of vanilla MSA and second transformer network replaces the MSA layer as window average-based MSA (WA-MSA) layer, which is proposed in this paper. We use the window size W as a number of RSs in an RB in both W-MSA and WA-MSA layers. For example, in 5G NR systems, the number of DMRS in an RB is six when the DMRS configuration type 1 and a DMRS symbol are used. The last transformer network of the DeTrans block replaces the MSA layer as a bi-LSTM layer for further enhanced estimation performances.

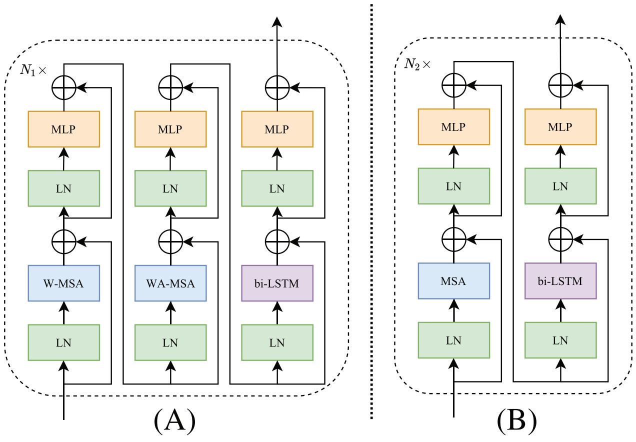


FIGURE 2 The structure of (A) DeTrans blocks and (B) UpTrans blocks

The final RS embeddings of DeTrans are recovered to the vector of denoised channel responses $\hat{\mathbf{h}}_p$ using a linear projection output layer of DeNet.

4.2.1 | WA-MSA

MSA in the original transformer [16] and vision transformer (ViT) [18] uses self-attentions among global embeddings, and it induces significant computations, especially with a large number of embeddings. The vanilla MSA is difficult to use due to its expensive computations when the bandwidths of the 6G communication system are expected to be vastly increased, and the number of RS in the entire resource grid can be extremely large.

To reduce computational overhead in extracting the context of RSs, we applied W-MSA in Liu and others [27], which splits the embeddings into W-sized windows and performs local self-attentions. W-MSA, in contrast, only extracts local information from adjacent RSs and is inefficient for obtaining global information. Therefore, we propose WA-MSA for self-attentions to exploit global channel variation with reduced computations. The operation of WA-MSA can be formulated as follows:

$$\mathbf{X}^{(i)} = \frac{1}{W} \sum_{k=iW}^{(i+1)W-1} \mathbf{x}^{(k)}, \text{ and} \tag{6}$$

$$\mathbf{Y}^{(i)} = \mathrm{MSA}(\mathbf{X}^{(i)}), \text{ for } i = 0, \ldots, \left(\frac{L_p}{W} - 1\right). \tag{7}$$

$$\mathbf{y}^{(k)} = \mathbf{x}^{(k)} + \left(\mathbf{Y}^{(\lfloor k/W \rfloor)} - \mathbf{X}^{(\lfloor k/W \rfloor)}\right), \text{ for } k = 0, \ldots, L_p - 1, \tag{8}$$

where $\mathbf{x}^{(k)}$ and $\mathbf{y}^{(k)}$ are input and output RS embeddings of WA-MSA and $\mathbf{X}^{(i)}$ and $\mathbf{Y}^{(i)}$ are window averaged input and output RS embeddings of MSA.

4.3 | UpNet

A detailed structure of UpNet is also in Figure 1. The denoised channel responses on RSs are inputted to UpNet, and a resource grid with denoised channel responses $\hat{\mathbf{H}}_p$ are reconstructed at the first stage of UpNet. Next, a linear 2D interpolation is performed to fill the response of resources other than RSs with responses of the nearest RS. The interpolated resource grid is divided into RB-sized (i.e., $N_s \times 12$) patches, and the patches are mapped into D_2 dimensions using a linear projection layer. The projection layer's outputs are referred to as RB embeddings. In addition to the responses in each RB, side information, such as SNR and positions of each RB, are also inputted to the linear embedding layer, same as in DeNet.

Upscaling transformer (UpTrans) blocks receive the RB embeddings. Figure 2B depicts the structure of UpTrans blocks. Each UpTrans block consists of a vanilla transformer encoder and a following modified transformer encoder, which replaces MSA as a bi-LSTM layer. Note that UpTrans blocks operate on RB embeddings, and the lengths of RB embeddings are much smaller than that of RS embeddings in general. Therefore, the computational complexity of UpTrans blocks is affordable with vanilla MSA. The output RB embeddings of UpTrans blocks are recovered into the entire channel responses $\hat{\mathbf{H}} \in \mathbb{R}^{2 \times N_s \times N_f}$ through the output layer in UpNet, which is a linear projection layer.

4.4 | Trainings

TSCE has greater training flexibilities than other DL-based estimators because it consists of two networks. For example, optimization algorithms and learning rates can be applied differently to each network. Furthermore, each network can be trained simultaneously or sequentially. Each network's loss functions can also be designed and applied differently during the training phase. For both networks, we used the normalized MSE (NMSE) as the loss function. The loss function for DeNet is

$$L_1 = \mathbb{E}\left[\frac{\|\mathbf{h}_p - \hat{\mathbf{h}}_p\|_2^2}{\|\mathbf{h}_p\|_2^2}\right] \tag{9}$$

and the loss function of UpNet is

$$L_2 = \mathbb{E}\left[\frac{\|\mathbf{H} - \hat{\mathbf{H}}\|_2^2}{\|\mathbf{H}\|_2^2}\right] \tag{10}$$

where the expectations are over the realization of channels $\mathbf{h}_p$ and $\mathbf{H}$.

We train our networks in a sequential manner, i.e., we train DeNet first, and then we use the trained DeNet to train UpNet. More specifically, inputs of UpNet are generated using the trained DeNet during the training of UpNet.

5 | EVALUATION RESULTS

In this section, we present evaluation results of channel estimation performances of TSCE as well as the baseline MLP-, CNN-, and ViT-based estimators. Furthermore, we also provide the generalization performances over various configurations of estimators.

5.1 | Training and evaluation setups

We consider a transmitter with a single transmission antenna and a receiver with a single receive antenna. The resource grid is as in 5G NR (New Radio) system [25] with 10 RBs and 20 RBs (12 subcarriers in each RB) in the subcarrier domain and a slot (14 symbols). The RS

TABLE 1 A summary of evaluation parameters

Parameter	Value
Subcarrier spacing	30 kHz
Propagation model	TDL-E/TDL-C
Delay spread	10 ns/100 ns
Maximum doppler frequency	100 Hz
Number of RBs	10/20
Number of symbols	14
Antenna configuration	1 × 1

configuration follows the DMRS configuration type 2 (four RS subcarriers in an RB) and on two symbols of a slot. We use the TDL (Tapped Delay Line) models defined by 3GPP [28]. The detailed evaluation environments are summarized in Table 1.

The training data consists of a pair of input and target data, where the input data is the channel response of RS positions and SNR value as SI, and the output data is the channel response of the entire resource grid. The training data set contains 50000 samples. We use 80% of the training data for training and the remaining 20% for validations. We train the networks for all DL-based estimators with 500 epochs and a batch size of 256.

For baseline MLP-based estimators, we used five layers of fully connected network layers, each followed by a rectified linear unit activation function as in Ye and others [5]. The number of hidden nodes in every layer is 256. To obtain the real and imaginary parts of each channel response, the number of input nodes is twice the number of RSs (i.e., $2 \times L_p$). The output layers have twice the number of nodes as the number of REs (i.e., $2 \times N_s \times N_f$). It should be noted that the number of input and output nodes in the MLP-based networks is determined by the estimation of resource size and the number of RSs. This is a drawback to applying the MLP-based estimator to real networks where the resource allocation and pilot resource can be arbitrarily configured.

For the baseline CNN-based channel estimator, we used ChannelNet [6] network model.[1]

We also applied ViT [18] as another baseline estimator. ViT was originally proposed for general image processing but can also be used for channel estimation applications. The detailed configuration of the transformer encoder of the baseline ViT-based estimator is in Table 3. The ViT-based estimator gets input of resource grids with estimations on the RS positions, splits the input resource grid into RB-size patches as in UpNet, and generates a final output of estimations of entire resource grids.

The detailed configurations of DeTrans and UpTrans blocks of TSCE are also in Table 2.

5.2 | Channel estimation performances

We perform a simulation to evaluate the channel estimation performances of traditional and DL-based estimators. The channel estimation performances for the TDL-E channel model with 10-ns delay spread and 100-Hz maximum doppler spread in 10 RBs are presented in Figure 3. The figure shows that TSCE achieves lower channel estimation error in terms of NMSE and traditional estimation technologies, such as LS and LMMSE estimators.[2] TSCE achieved about 21.2-dB and 8.3-dB lower NMSEs than LS and LMMSE estimators, respectively, at 0-dB SNR, as shown in the figure. More importantly, the MLP-based and ChannelNet estimators performed lower than TSCE. As shown in the figure, TSCE achieved approximately 2.6-dB and 4.3-dB lower NMSEs than the MLP-based estimator and ChannelNet at 0 dB SNR. The ViT-based estimator also outperformed other estimators, such as TSCE. Compared with the ViT-based estimator, performances in low SNRs were similar, but performance differences occurred as SNR increased. For example, at 15-dB SNR, TSCE showed about 0.94-dB lower NMSE than the ViT-based estimator.

Next, a similar simulation is repeated with another environment. The channel estimation performances for the TDL-C channel model with 100-ns delay spread in 20 RBs are shown in Figure 4, and TSCE also outperformed other estimators. To illustrate, TSCE can achieve NMSE of LMMSE obtained at 10-dB SNR (approximately −15 dB of NMSE) at about 5-dB SNR. TSCE can also achieve NMSE of MLP-based estimator and ChannelNet; both obtained at 10-dB SNR (approximately −17 dB of NMSE) at about 7-dB SNR. The ViT-based estimator also outperformed other estimators as well as TSCE, but its performances also tend to decrease in high SNRs compared to TSCE.

Although we evaluated the channel estimation performances by assuming system parameters of a 5G NR system, TSCE can also be applied to most of the other communication systems where the pilot signals are used to measure variations over the wireless channels, such as

[1] We used the official implementation of ChannelNet (https://github.com/MehranSoltani94/ChannelNet) with the following modifications: (1) We used real and imaginary parts as two channels of input data at the same time. (2) Instead of using Gaussian kernels, we used linear 2D interpolation.

[2] The LMMSE estimator uses estimated correlation matrix based on channel delay parameter [23] and doppler frequency estimations [29].

TABLE 2 Details of Transformer encoders of ViT-based estimator and TSCE

Model	Layers	Hidden size	MLP size	Heads	# Params
ViT-based	4	64	128	6	503 k
DeTrans	2	18	32	6	126 k
UpTrans	4	64	128	6	781 k

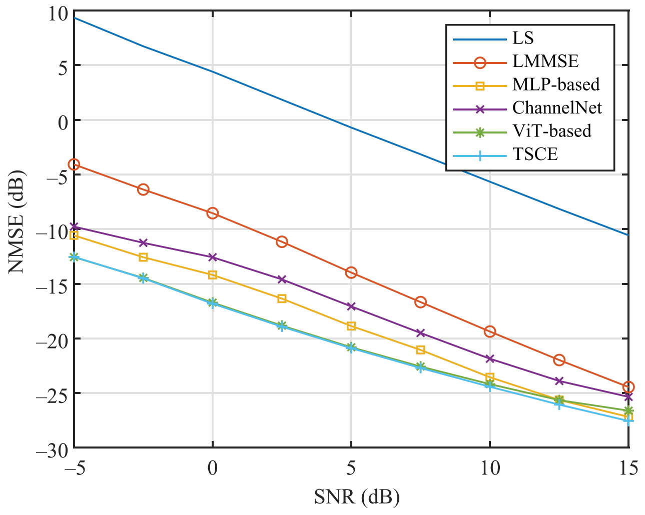


FIGURE 3 Channel estimation performances for TDL-E channel model with 10-ns delay spread in 10 RBs

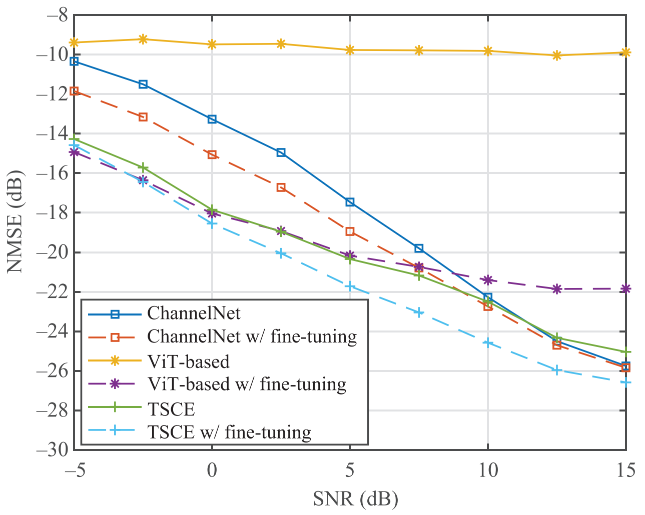


FIGURE 5 Channel estimation performances in 20 RBs using DL networks trained of 10 RBs

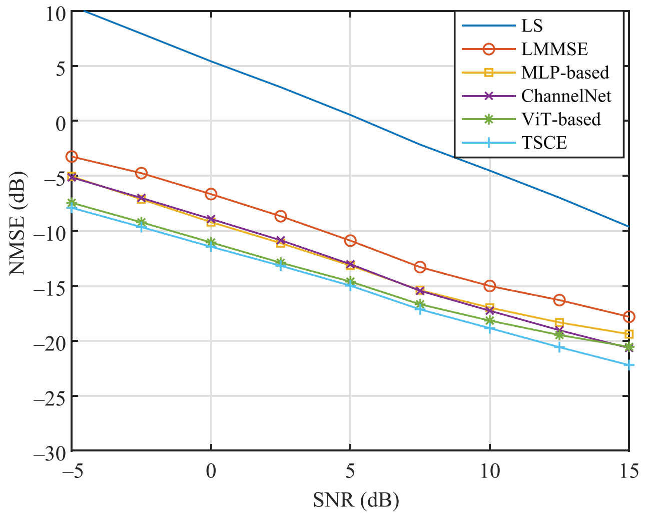


FIGURE 4 Channel estimation performances for TDL-C channel model with 100-ns delay spread in 20 RBs

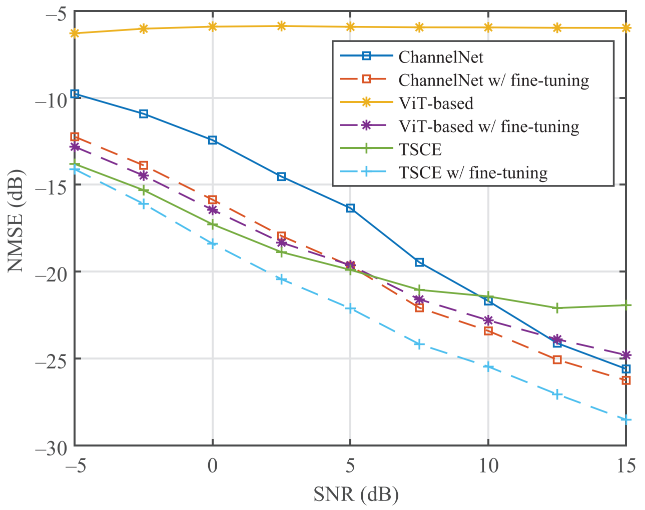


FIGURE 6 Channel estimation performances of DMRS configuration type 1 using DL networks trained of DMRS configuration type 2

Long-term Revolution (LTE) or Wireless Local Area Network (W-LAN) systems. Moreover, it is more affirmative to be applied in 6G cellular networks because TSCE does not require accurate channel covariances measurements, which can be acquired using additional and resource-consuming RSs in 5G NR, and this point will give great flexibility as well as reduced overheads in 6G cellular network designs.

TABLE 3 Computations and model sizes of DL-based estimators

Estimators	# FLOPs (10 RBs)	# Params (10 RBs)	# FLOPs (20 RBs)	# Params (20 RBs)
MLP-based	2.2 M	1.1 M	4 M	2 M
ChannelNet	2.3 G	682 k	4.6 G	682 k
ViT-based	10.96 M	503 k	21.02 M	503 k
TSCE	28.04 M	907 k	56 M	907 k

5.3 | Evaluations on scalability and robustness

As well as the estimation performance in environments where DL-based estimators trained on, scalable and robust estimations in other environments or configurations than the training data are important to be applied to real-world networks. To evaluate the scalability and robustness, we performed an experiment of estimating 20 RBs of channels using the DL-based estimators trained using 10 RBs of training data with the TDL-E channel model. It should be noted that we did not evaluate the MLP-based estimator since it does not support larger inputs than the predefined input size. Figure 5 shows that TSCE showed low estimation errors even when it was not further trained using the appropriate size (20 RBs) of training data, especially at low SNRs. Compared with ChannelNet, TSCE shows more than 4-dB lower NMSE at low SNRs without fine-tuning. We also performed an epoch of downstream fine-tuning with 20 RBs of training data. TSCE showed significantly enhanced performance in high SNRs with an epoch of fine-tuning, whereas ChannelNet did not show much enhancement with fine-tuning. The ViT-based estimator was shown as not properly working in a different environment than the training without fine-tuning. This is because the ViT-based estimator only learned estimations of a given size of the resource grid.

Next, we examined the impact of different RS configurations. We assess channel estimation performance using the DL-based estimators with the DMRS configuration type 1 (six RS subcarriers in an RB), whereas the DL-based estimators are trained using the dataset of DMRS configuration type 2. Figure 6 shows that TSCE gives more than 4 dB lowered NMSE than ChannelNet in low SNRs without any fine-tuning. We also perform an epoch of downstream fine-tuning with appropriate (DMRS configuration type 1) training data. TSCE also showed a significant gain in fine-tunings, especially in high SNRs. The ViT-based estimator also does not work properly in different RS configurations, and this is because the ViT-based estimator only processes the RB-size of patches and does not fully extract the information of RSs.

According to the evaluations in this subsection, it is indicated that TSCE is more robust in a different environment and configuration to the training, even without any further downstream fine-tunings. It is also indicated that the performances of TSCE in high SNRs can be significantly enhanced with an epoch of fine-tunings.

5.4 | Computations and model sizes

Table 3 shows the computation complexity in floating point operations per second (FLOPs) and the number of parameters of the DL-based estimators. Our model uses far fewer computations than ChannelNet while outperforming both channel estimation and generalization performances. The MLP-based channel estimator has the lowest computation complexity, but the number of parameters grows in proportion to the input and output sizes because the MLP-based estimators need to be defined according to different sizes of input and output.

6 | CONCLUSIONS

In this paper, we present TSCE, a DL-based scalable and robust channel estimator, for 6G wireless cellular networks. We propose a scalable DL architecture for channel estimation composed of two transformer networks, which are DeNet and UpNet, for capturing local and global contexts efficiently. For computationally efficient feature extractions of RSs, we also propose WA-MSA in DeNet. We show that TSCE achieves better channel estimation performances in terms of NMSE than other traditional and baseline DL-based channel estimators in various environments. We also focus on the scalability and robustness of estimators and demonstrate that TSCE is more robust in environments different from the training, even without further downstream fine-tunings.

CONFLICTS OF INTEREST

The authors declare that there are no conflicts of interest.

ORCID

Anseok Lee https://orcid.org/0000-0002-8196-6207

REFERENCES

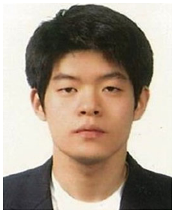
1. ETRI, *[5G insight white paper 2.0] 5G technologies and its way-forward*, 2017.
2. ETRI, *6G insight: Vision and technologies*, 2020.
3. R. Shafin, L. Liu, V. Chandrasekhar, H. Chen, J. Reed, and J. C. Zhang, *Artificial intelligence-enabled cellular networks: A critical path to beyond-5G and 6G*, IEEE Wirel. Commun. **27** (2020), no. 2, 212–217.
4. Qualcomm, *New SI: Study on artificial intelligence (AI)/Machine Learning (ML) for NR air interface*, RP-213599. 3GPP, 2021.
5. H. Ye, G. Y. Li, and B. H. Juang, *Power of deep learning for channel estimation and signal detection in OFDM systems*, IEEE Wirel. Commun. Lett. **7** (2018), no. 1, 114–117.
6. M. Soltani, V. Pourahmadi, A. Mirzaei, and H. Sheikhzadeh, *Deep learning-based channel estimation*, IEEE Commun. Lett. **23** (2019), no. 4, 652–655.
7. C. Dong, C. C. Loy, K. He, and X. Tang, *Image super-resolution using deep convolutional networks*, IEEE Trans Pattern Anal Machine Intell **38** (2016), no. 2, 295–307.
8. K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, *Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising*, IEEE Trans. Image Process. **26** (2017), no. 7, 3142–3155.
9. L. Li, H. Chen, H.-H. Chang, and L. Liu, *Deep residual learning meets OFDM channel estimation*, IEEE Wirel. Commun. Lett. **9** (2020), no. 5, 615–618.
10. K. He, X. Zhang, S. Ren, and J. Sun, *Deep residual learning for image recognition*, (Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition), 2016, pp. 770–778.
11. I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, *Generative Adversarial Networks*, arXive preprint, 2014. https://doi.org/10.48550/arXiv.1406.2661
12. E. Balevi and J. G. Andrews, *Wideband channel estimation with a generative adversarial network*, IEEE Trans Wirel Commun **20** (2021), no. 5, 3049–3060.
13. A. Radford, L. Metz, and S. Chintala, *Unsupervised representation learning with deep convolutional generative adversarial networks*, (International Conference on Learning Representations), 2016. https://doi.org/10.48550/arXiv.1511.06434
14. S. Zhao, Y. Fang, and L. Qiu, *Deep learning-based channel estimation with SRGAN in OFDM systems*, (IEEE Wireless Communications and Networking Conference, Nanjing, China), 2021. https://doi.org/10.1109/WCNC49053.2021.9417242
15. C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, *Photo-realistic single image super-resolution using a generative adversarial network*, (Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA), 2017, pp. 4681–4690.
16. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, *Attention is all you need*, Adv. Neural Inform. Process. Syst. **30** (2017). https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
17. J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, *BERT: Pre-training of deep bidirectional transformers for language understanding*, (Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies), 2019, pp. 4171–4186.
18. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, *An image is worth 16×16 words: Transformers for image recognition at scale*, arXive preprint ICLR, 2021. https://doi.org/10.48550/arXiv.2010.11929
19. A. Yang, P. Sun, T. Rakesh, B. Sun, and F. Qin, *Deep learning based OFDM channel estimation using frequency-time division and attention mechanism*, (IEEE GLOBECOM Workshops, Madrid, Spain), 2021. https://doi.org/10.1109/GCWkshps52748.2021.9682149
20. J. Li and Q. Peng, *Lightweight channel estimation networks for OFDM systems*, IEEE Wirel. Commun. Lett. **11** (2022), no. 10, 2066–2070.
21. Z. Chen, F. Gu, and R. Jiang, *Channel estimation method based on transformer in high dynamic environment*, (12th International Conference on Wireless Communications and Signal Processing, Nanjing, China), 2020, pp. 817–822.
22. D. Luan and J. Thompson, *Attention based neural networks for wireless channel estimation*, arXive preprint, 2022. https://doi.org/10.48550/arXiv.2204.13465
23. K. C. Hung and D. W. Lin, *Pilot-based LMMSE channel estimation for OFDM systems with powerdelay profile approximation*, IEEE Trans. Vehic. Technol. **59** (2010), no. 1, 150–159.
24. Y. S. Cho, J. Kim, W. Y. Yang, and C. G. Kang, *MIMO-OFDM wireless communications with MATLAB*®, Wiley-IEEE Press 2010.
25. 3GPP NR specification. http://www.3gpp.org/dynareport/38-series.htm
26. X. Chen, S. Xie, and K. He, *An empirical study of training self-supervised vision transformers*, (Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, Canada), 2021, pp. 9640–9649.
27. Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, *Swin transformer: Hierarchical vision transformer using shifted windows*, (Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, Canada), 2021, pp. 10012–10022.
28. 3rd generation partnership project; Technical Specification Group Radio Access Network; Study on channel model for frequencies from 0.5 to 100 GHz (Release 17), 2022.
29. J.-I. Kim, J.-H. Jang, and H.-J. Choi, *A low-complexity 2-D MMSE channel estimation for OFDM systems*, J. Korea Inform. Commun. Soc. **36** (2011), no. 5C, 317–325.

AUTHOR BIOGRAPHIES

Anseok Lee received his BS degree in Electrical Engineering from Kyungpook National University, Daegu, Korea, in 2006, and his MS degree in Electrical Engineering from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea, in 2008. Since February 2008, he has been with the Electronics and Telecommunications

Research Institute (ETRI), Daejeon, Korea, where he is currently a senior researcher. His research interests include wireless communication systems and artificial intelligence/machine learning.

Yongjin Kwon received his BS and MS degrees in Electrical Engineering from Korea Advanced Institute of Science and Technology, Daejeon, Republic of Korea, in 2009 and 2011, respectively. Since 2011, he has been with the Electronics and Telecommunications Research Institute, Daejeon, Republic of Korea, as a senior researcher. His research interests include future 6G wireless communication systems and intelligent communication systems.



Hanjun Park received his BS and MS degrees in Electrical and Electronic Engineering from Seoul National University. He joined ETRI in 2022 and is currently working as a senior researcher in Intelligent Wireless Access Research Section. Prior to joining ETRI, from Oct. 2018 to Jan. 2022, he worked as a senior engineer in the Air System Design Laboratory of Samsung Electronics. From Feb. 2011 to Sep. 2018, he worked as a senior engineer in the Advanced Standardization Laboratory of LG Electronics. His research interests include MIMO, LAA, UL Control, Power Control, DRX, and AI/ML-based air interface.

Heesoo Lee received his BS, MS, and PhD degrees in Industrial Engineering from the Korea Advanced Institute of Science and Technology (KAIST) in 1993, 1995, and 2001, respectively. In 2001, he joined the ETRI, where he is currently the Director of the Intelligent Wireless Access Research Section. He is working on core technologies for future wireless cellular communication, especially in the area of artificial intelligence, millimeter wave, OFDM, SC-FDMA, multiuser MIMO, interference management, relay, and so forth.

How to cite this article: A. Lee, Y. Kwon, H. Park, and H. Lee, *Deep learning-based scalable and robust channel estimator for wireless cellular networks*, ETRI Journal (2022), 1–10. https://doi.org/10.4218/etrij.2022-0209