The following publication S. Singh, A. Trivedi and D. Saxena, "Channel Estimation for Intelligent Reflecting Surface Aided Communication via Graph Transformer," in IEEE Transactions on Green Communications and Networking, vol. 8, no. 2, pp. 756-766 is available at https://doi.org/10.1109/TGCN.2023.3339819.

Channel Estimation for Intelligent Reflecting Surface aided Communication via Graph Transformer

Shatakshi Singh, Aditya Trivedi, *Senior Member, IEEE* and Divya Saxena, *Member, IEEE*

Abstract

Intelligent reflecting surface (IRS) is a potential technology for enhancing communication systems' performance. Accurate cascaded channel estimation between the base station (BS), IRS, and the user is vital for optimal system performance. However, incorporating IRS increases channel estimation complexity due to additional dimensions from each element, leading to higher training overhead. To reduce training overhead, existing approaches assume the sparse cascaded channel which may not be valid in dense multipath propagation and non-line-of-sight settings. We propose a novel technique to address this issue by leveraging the spatial correlation among IRS elements' channels. By dividing the IRS surface into groups, we estimate the channel for some groups via the least square (LS) method. To estimate the channels for the remaining groups, a graph transformer-based IRS channel estimation (G-TIRC) model is proposed, which includes a graph neural network (GNN) and transformer model. The GNN finds the correlations among the different groups by embedding the channel information. Then, the attention mechanism within the transformer extracts useful correlations to accurately predict the channels for the unknown groups. The experiments demonstrate the effectiveness of the G-TIRC model in achieving accurate channel estimation with reduced pilot overhead compared to other state-of-the-art methods.

Shatakshi Singh is with the Department of Information Technology (IT), Atal Bihari Vajpayee-Indian Institute of Information Technology and Management (ABV-IIITM), Gwalior, Madhya Pradesh (MP), India, 474015.
E-mail: shatakshisingh704@gmail.com.

Aditya Trivedi is with the Department of IT, ABV-IIITM, Gwalior, MP, India, 474015. E-mail: atrivedi@iiitm.ac.in.

Divya Saxena is with the Department of Computing, The Hong Kong Polytechnic University, Kowloon, Hong Kong. E-mail: divsaxen@comp.polyu.edu.hk.

Index Terms

Cascaded channel estimation, intelligent reflecting surface (IRS), attention mechanism, graph transformer.

I. Introduction

Intelligent reflecting surface (IRS) is a new and promising technology that has emerged as a breakthrough concept for enhancing the performance of wireless communication systems [1], [2]. An IRS is an array of passive reflecting elements that can manipulate the electromagnetic waves incident on it, allowing them to be steered, amplified, or attenuated [3], [4]. These reflecting elements can be individually controlled and configured to achieve specific communication objectives such as enhancing signal strength, coverage, and capacity [5], [6]. IRS has several advantages over traditional wireless communication systems. It is power-efficient and cost-effective as it uses only passive elements and requires less hardware. To achieve better spectral efficiency (SE) and energy efficiency (EE) via the deployment of the IRS, the active beamforming at the base station (BS) and the passive beamforming at the IRS are designed in [7]–[10]. However, this research is conducted under the implicit assumption that channel state information (CSI) is completely known at the BS, which is not a practical assumption [11].

Independent estimation of channels from the BS to the IRS, and IRS to the user is practically challenging due to the IRS's limited signal processing capabilities. Instead, only the cascaded user-IRS-BS channels can be predicted at the BS or user [12]. Moreover, it is important to consider the channel training overhead incurred during the estimation of cascaded channels. This training overhead consumes valuable system resources, such as bandwidth and power. Consequently, it has the potential to impact the overall EE and SE of the system. Thus, it is crucial to strike a balance between the need for accurate channel estimation and the associated training overhead.

In [13], a binary reflection-controlled least-squares (LS) channel estimation scheme is introduced, which involves activating a single reflecting element while deactivating the rest within each time slot for channel estimation. In [14], a minimum variance unbiased channel estimator is proposed, which estimates the channel with all elements of the IRS switched on simultaneously. Further, a three-phase LS channel estimation scheme is introduced in [15]. However, the estimation errors in the two phases can propagate in the third phase of channel estimation, resulting in low estimation accuracy. A linear minimum mean-squared error (LMMSE) estimator leverages prior knowledge

of channel distributions is proposed in [16]. However, this method involves multidimensional integration which is intractable in practice. In [17], a systematic pattern of pilot signals is presumed for channel estimation, and this approach leverages parallel factor (PARAFAC) tensor modeling of the received signals. The authors in [18] introduced a convolutional deep residual network (CDRN) for estimating cascaded channels in a multi-user scenario. This approach treats channel estimation as a denoising task, improving the accuracy of cascaded channel estimates obtained from noisy LS estimation.

The aforementioned approaches estimate the cascaded channels for all the reflecting elements of an IRS simultaneously. As a consequence, they require longer pilot lengths that scale with the number of reflecting elements, leading to significant pilot overhead. In [19] and [20], active elements are placed on the IRS to perform channel estimation. The estimated CSI from these active elements is then utilized to predict the CSI from the remaining elements. In [19], the orthogonal matching pursuit (OMP) algorithm is used to recover full CSI. In [20], the LS estimate from active elements is treated like a low-resolution image, and a convolutional neural network (CNN) is trained to extract a high-resolution image. While these methods potentially reduce training overhead, they also lead to increased hardware complexity due to the inclusion of active elements within the IRS.

Several studies present the channel estimation problem as a sparse signal recovery task [21], [22]. This method makes use of compressive sensing (CS) methods like OMP [23] and structured OMP [24] to estimate cascaded channels with fewer pilot symbols. Further, in [25] SOMP and CNN are combined to improve the channel estimation accuracy. In [26], the authors tackled the channel estimation challenge in multi-input multi-output (MIMO) systems by leveraging intrinsic low-rank properties through a two-phase algorithm, combining sparse matrix factorization and matrix completion techniques. In [27] and [28], the variational approximate message passing (VAMP) algorithm is employed for estimating cascaded channels under the assumption of sparsity in the angular domain and with the consideration of specific probability distributions. If these assumptions do not align well with the true characteristics of the channel, the performance of the VAMP may be negatively affected, leading to less accurate channel estimation.

To address the high channel training overhead issue, the authors in [29] and [30] divide IRS elements into groups. Within each group, only the effective cascaded channel for all elements is estimated collectively. This approach significantly reduces the number of required pilot symbols, as the estimation is performed at the group level rather than for individual elements. The

partitioning of the IRS elements into groups presents a trade-off, as it can negatively impact the performance of passive beamforming. When we have access to only per-group effective channels, it becomes necessary to set the reflection coefficients identically for all elements within each group. This limitation curtails the design flexibility and restricts the available degrees of freedom for passive beamforming.

A. Motivation and contribution

Current approaches for channel estimation in IRS systems have the shortcoming of usually not taking into account the correlations among the channels of IRS elements, that arise due to their spatial proximity. By availing of the advantage of these correlations, we can potentially decrease the pilot overhead for channel estimation. Furthermore, it's worth highlighting that a reduction in channel estimation overhead can have a direct and significant impact on both EE and SE within the communication system. By decreasing the pilot overhead, valuable resources such as power can be conserved for the data transmission phase.

In this study, we initially partition the elements of the IRS into groups, drawing inspiration from grouping techniques described in the existing literature [29]. Pilot signals are transmitted to estimate cascaded channels for some of the groups, while for the rest of the groups, it is assumed that there exists some spatial correlation among the channels of the IRS elements as elements are aligned very close to each other. By exploiting this spatial proximity, the correlations among the channels of IRS groups are extracted to predict the channels for the next groups, thus saving pilots and reducing the complexity of the channel estimation process.

In the areas of image processing and natural language processing (NLP), transformer models have achieved amazing success [31], [32], and [33]. Their achievements in these domains have paved the way for exploring their application in the field of wireless communication, such as accurate time-varying channel estimation [34].

We are motivated by the capability of the transformer to capture correlations and dependencies in the data well. Thus, we exploit transformer models' capability to capture the intricate connections and relationships among the channels of the groups of IRS elements to perform channel estimation for unknown IRS groups. As a result, for the channels that have not previously been observed, the model is able to successfully learn about them by taking knowledge from known groups, producing precise predictions for IRS cascaded channels.

In the context of channel prediction for IRS, a transformer architecture is proposed for the

first time to the best of our knowledge for predicting the channel response of the unknown group of IRS elements based on the channel responses of the known groups. The attention mechanism learns to assign different weights to the correlations among groups based on their relevance to accurately predict the channels for the unknown groups. By attending to the most relevant correlations, the attention mechanism helps in effectively capturing the spatial correlation information and incorporating it into the channel estimation process.

We make the following contribution to estimate cascaded channels in the IRS-aided communication system:

- Our proposed technique for predicting the cascaded channels, involves dividing the elements of IRS into groups based on their position. Then, use the LS method to estimate the channels for some of these groups by transmitting pilot symbols. Leveraging the spatial correlation among the groups, we can estimate the channels for the remaining groups by utilizing the channel information from the previously estimated groups. By employing this approach, it becomes possible to achieve an accurate estimation of the channels of all IRS elements while simultaneously reducing the required number of pilot symbols.
- To extract and utilize the correlation information among the groups, we adopt a graph transformer-based approach. In this approach, the spatial relationship among groups is obtained by the embedding created by graph neural network (GNN) layers of the graph transformer. Furthermore, we integrate the position information of each IRS group into the graph embedding of each group.
- The embedding concatenated with position information is the input to the attention layer of the graph transformer. The attention layer identifies the most relevant correlations for predicting the channel responses of unknown groups. By leveraging spatial correlations and position relationships, the proposed method improves the accuracy of channel prediction by focusing on important groups. This enables efficient channel estimation in IRS systems while reducing the required pilot symbols.
- Through extensive experimental evaluation, it has been shown that the proposed graph transformer-based IRS cascaded channel estimation (G-TIRC) model, surpasses the existing methods in terms of channel estimation accuracy and significantly reduces the pilot symbol overhead by leveraging spatial correlations and position information. Further, we demonstrated the impact of reduced pilot overhead on the uplink SE and EE of the system.

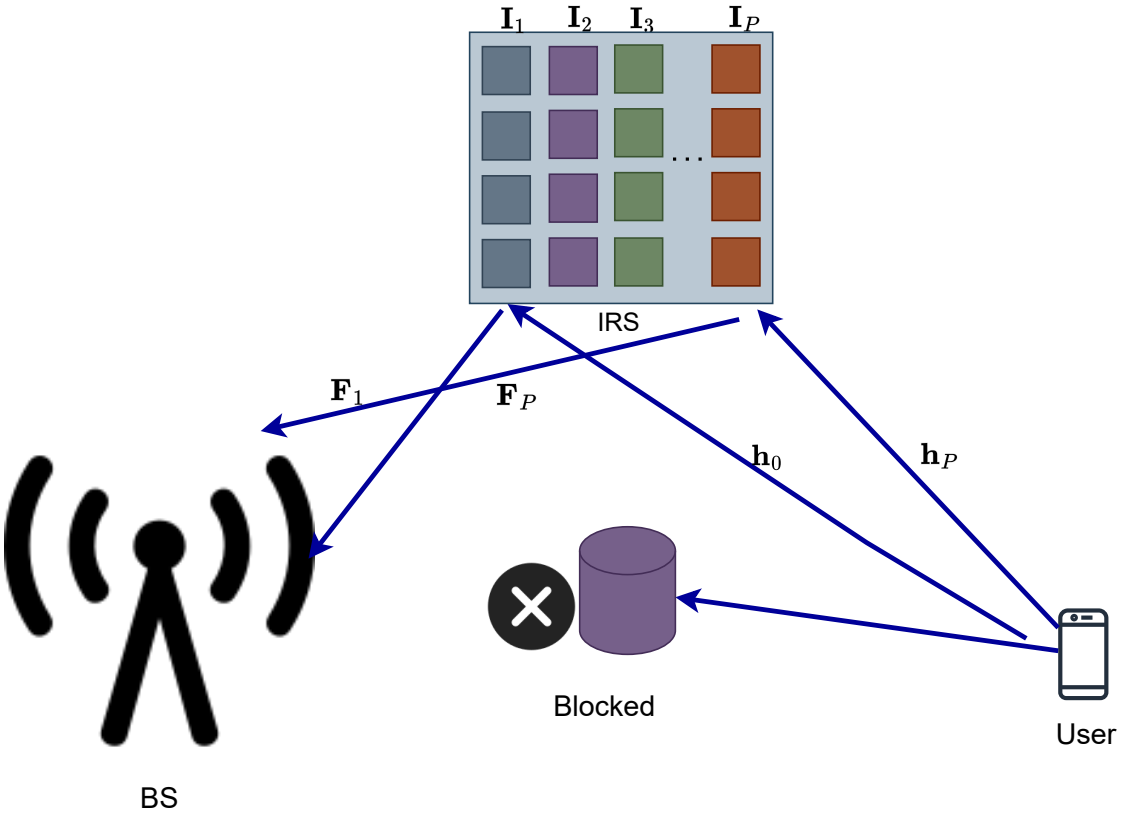

Fig. 1. IRS aided communication system

The subsequent sections of the paper are structured as follows: Section II provides a description of the system model and outlines the problem formulation. The proposed G-TIRC model is explained in Section III. Experimental settings and results are presented in Section IV. The conclusion of the paper can be found in Section V.

Notations: A transpose and inverse operations are denotes as $(\cdot)$ and $(\cdot)^{-1}$, respectively. $\|\cdot\|_F$ denotes the Frobenius norm of a matrix and the diagonal matrix is given as diag$(\cdot)$. The real numbers and complex numbers are denoted $\mathbb{R}$ and $\mathbb{C}$, respectively. $\mathbf{v}$ and $\mathbf{V}$ represent the vector and matrix, respectively.

II. System Model

We consider a communication system reinforced by IRS as exhibited in Fig. 1, where the direct communication link between the BS and the user is obstructed or unavailable. The IRS consists of a I number of passive reflecting elements, which can reflect and manipulate the

wireless signals transmitted between the BS and the user. The BS is endowed with a B number of antennas and the user has a single antenna. The channel from the user to IRS is denoted as $\mathbf{h} \in \mathbb{C}^{I\times 1}$. Further, the channel from IRS to the BS is given as $\mathbf{F} \in \mathbb{C}^{B\times I}$. The I elements of IRS are partitioned into P groups $(I_1, I_2, .., I_P)$, each column represents one group as shown in Fig. 1. Each group consists of $L = I/P$ elements. A time division duplex (TDD) protocol that leverages the channel reciprocity principle to acquire the downlink CSI by utilizing uplink channel estimation is considered.

For channel estimation of the p-th group, the user sends pilot symbols $\mathbf{x}_p \in \mathbb{C}^{1\times\tau}$ with $\mathbf{x}_p\mathbf{x}_p^T = 1$ to the BS, where $\tau \geq L$. At the BS, the received signal can be represented as:

$$\mathbf{Y}_p = \mathbf{F}_p\mathtt{diag}(\mathbf{s}_p)\mathbf{h}_p\mathbf{x}_p + \mathbf{N}_p, \tag{1}$$

where $\mathbf{F}_p \in \mathbb{C}^{B\times L}$ is the channel from the IRS to the BS and $\mathbf{h}_p \in \mathbb{C}^{L\times 1}$ is the channel from the user to the IRS for the p-th group of the IRS elements. $\mathbf{N}_p$ represents the additive white noise, whose entries are independent and identically distributed (i.i.d.) Gaussian has zero mean and unit variance. $\mathbf{s} = [e^{j\varphi_1}, .., e^{j\varphi_i}, .., e^{j\varphi_I}]^T$, where φ_i is phase shift of i-th element. The vector $\mathbf{s}_p \in \mathbb{C}^{L\times 1}$ represents the phase shift of p-th group of IRS elements. It is obvious that $\mathtt{diag}(\mathbf{h}_p)\mathbf{s}_p$ = $\mathtt{diag}(\mathbf{s}_p)\mathbf{h}_p$. The above equation can be rewritten as:

$$\mathbf{Y}_p = \mathbf{G}_p\mathbf{s}_p\mathbf{x}_p + \mathbf{N}_p, \tag{2}$$

where

$$\mathbf{G}_p = \mathbf{F}_p\mathtt{diag}(\mathbf{h}_p) = \begin{bmatrix} F_{11} & F_{12} & \cdots & F_{1L} \\ F_{21} & F_{22} & \cdots & F_{2L} \\ \vdots & \vdots & \ddots & \vdots \\ F_{B1} & F_{B2} & \cdots & F_{BL} \end{bmatrix} \begin{bmatrix} h_1 & 0 & \cdots & 0 \\ 0 & h_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & h_L \end{bmatrix} \in \mathbb{C}^{B\times L}, \tag{3}$$

is the cascaded channel from the user to the BS. The signal retrieved at the BS from the user is given as follows:

$$\widehat{\mathbf{Y}}_p = \mathbf{G}_p\mathbf{s}_p + \mathbf{W}_p, \tag{4}$$

where $\widehat{\mathbf{Y}}_p = \mathbf{Y}_p\mathbf{x}_p^T \in \mathbb{C}^{B\times 1}$ and $\mathbf{W}_p = \mathbf{N}_p\mathbf{x}_p^T \in \mathbb{C}^{B\times 1}$. The LS estimate of the cascaded channel can be written as:

$$\tilde{\mathbf{G}}_p^{\mathrm{LS}} = \widehat{\mathbf{Y}}_p\mathbf{s}_p^{-1}. \tag{5}$$

Similarly, the LS estimate for other groups can be estimated.

1) Problem Formulation: Considering an IRS with a large number of elements, we divide the elements into P groups. Then, we estimate the channels of some groups using (5). For simplicity, we divide the groups column-wise as shown in Fig. 1 (the same color represents the element corresponding to one group), and estimated the channels for starting $P - P'$ groups through LS as $[\tilde{\mathbf{G}}_1, \tilde{\mathbf{G}}_2, ..., \tilde{\mathbf{G}}_{P-P'}]$. Our objective is to estimate the channels for the rest of the groups $[\mathbf{G}_{P-P'+1}, ..., \mathbf{G}_{P-2}, \mathbf{G}_{P-1}, \mathbf{G}_P]$ with the channel knowledge of aforementioned known groups.

III. Graph Transformer-based IRS channel estimation

A. Embedding

We represent each group of IRS cascaded channels as a node of the graph. To extract the correlation information among groups, embeddings are created using GNN incorporated into the graph transformer. The GNN updates the embedding of each group by layers of computation. The output of the GNN is the final embedding, which provides us with the correlation information for each group with respect to the other groups.

The GNN works by propagating messages between the nodes and aggregating them to update the node embedding. Each layer of the GNN updates the node embedding using the information from the previous layer. As a result, the updated embeddings capture higher-level information about the nodes and their correlations.

For the purpose of applying GNN to obtain embedding groups of IRS channels, the initial step involves extracting the real and imaginary components of the LS estimate for each group $[\tilde{\mathbf{G}}_1^{LS}, \tilde{\mathbf{G}}_2^{LS}, ..., \tilde{\mathbf{G}}_{P-P'}^{LS}]$, then each group is of the form $\tilde{\mathbf{G}}_P^{LS} \in \mathbb{R}^{2BL\times 1}$. The GNN consists of an input layer, that takes features from the IRS channel groups as initial representation vector $\tilde{\mathbf{G}}_p^0 = \tilde{\mathbf{G}}_p^{LS}$, $p = 1, 2, 3, ..., P - P'$. We then train one layer of fully connected (FC) layer to obtain $\tilde{\mathbf{G}}_p^1$.

In the second layer, to obtain the embedding of the p-th group, we apply the aggregation and combination function to propagate the feature information from the corresponding and neighboring nodes. The aggregation function can be given as:

$$F_{agg}\left(\{\tilde{\mathbf{G}}_j^0\}_{j\in N(p)}\right). \tag{6}$$

Finally, to update the information of $\tilde{\mathbf{G}}_p$, we combine its information from the previous layer with the information of its neighboring groups as:

$$\tilde{\mathbf{G}}_p^1 = F_{combine}\left(\tilde{\mathbf{G}}_p^0, F_{agg}\left(\{\tilde{\mathbf{G}}_j^0\}_{j\in N(p)}\right)\right). \tag{7}$$

Here, the aggregation function is an element-wise mean pooling. The combination function is implemented by the FC layer.

Further, we concatenate the position information of groups of the IRS to the graph embedding. The final embedding $\mathbf{E} = [\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_n] \in \mathbb{R}^{(u^{'}+3=u)\times n}$, where $u^{'} + 3$ represents the feature dimension of embedding, where 3 represents the dimension of the location of the IRS group and n represents the number of groups to be embedded.

After GNN embeds the channel information and calculates the correlation values among groups, the attention mechanism comes into play. It assigns weights to the embeddings based on their relevance and significance in the channel estimation process. This allows the attention mechanism to focus on the most informative correlations and ignore irrelevant or less important ones.

B. Attention Mechanism

The attention mechanism in the G-TIRC model is crucial for capturing relevant information and dependencies among different groups of IRS elements' channels. It applies a linear transformation to the embeddings, allowing the model to identify the relationships between the groups.

By assigning different weights to the groups through the attention mechanism, the model highlights the importance of each group in predicting the channels for the unknown groups. This enables the model to focus on the most relevant correlations and extract useful information for accurate channel estimation.

In the attention mechanism, there are three types of vectors involved: query $\mathbf{q}_i \in \mathbb{R}^{d\times 1}$, key $\mathbf{k}_i \in \mathbb{R}^{d\times 1}$, and value $\mathbf{v}_i \in \mathbb{R}^{u\times 1}$. Here, d is the feature dimension of the key vector. The key, query, and value vectors are given as:

$$\mathbf{k}_i = \mathbf{\Theta}^k \mathbf{e}_i, i = 1, ..., n, \tag{8}$$

$$\mathbf{q}_i = \mathbf{\Theta}^q \mathbf{e}_i, i = 1, ..., n, \tag{9}$$

$$\mathbf{v}_i = \mathbf{\Theta}^v \mathbf{e}_i, i = 1, ..., n, \tag{10}$$

where $\mathbf{\Theta}^k \in \mathbb{R}^{d\times u}$, $\mathbf{\Theta}^q \in \mathbb{R}^{d\times u}$, and $\mathbf{\Theta}^v \in \mathbb{R}^{u\times u}$ are the trainable weight matrix for key, query, and value, respectively. The matrix can be written as:

$$\mathbf{K} = \mathbf{\Theta}^k \mathbf{E}, \tag{11}$$

$$\mathbf{Q} = \mathbf{\Theta}^q \mathbf{E}, \tag{12}$$

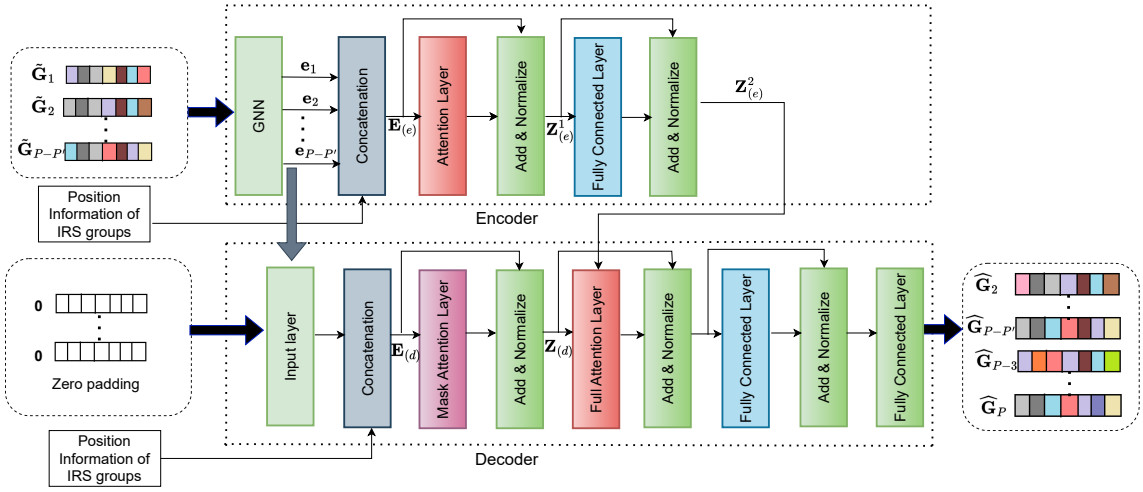

Fig. 2. Architecture of the G-TIRC model.

$$\mathbf{V} = \mathbf{\Theta}^{v}\mathbf{E}. \tag{13}$$

By incorporating an attention mechanism, the model gains the capability to concentrate on the most pertinent groups for predicting the channels of the unknown groups more accurately. The attention matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ is obtained by computing the product between the key matrix $\mathbf{K}^T$ and the query matrix $\mathbf{Q}$. The softmax function is then applied to these scores to obtain a probability distribution over the input. The attention matrix is denoted as:

$$\mathbf{A} = \text{Softmax}\left(\frac{\mathbf{K}^T\mathbf{Q}}{\sqrt{d}}\right). \tag{14}$$

Finally, the value matrix is weighted by this probability distribution and summed to obtain the output of the attention mechanism:

$$\mathbf{C} = \mathbf{V}\mathbf{A}, \tag{15}$$

where $\mathbf{C} \in \mathbb{R}^{u\times n}$.

C. Graph transformer-based IRS channel estimation (G-TIRC) model

Fig. 2 illustrates the architecture of the G-TIRC model, which comprises an encoder and a decoder component. The first two layers of the encoder are GNN layers that are integrated into the transformer to obtain the embedding of the IRS channel groups. The GNN layers are followed by the concatenation layer that combines the position of IRS elements groups with the graph embedding. An attention layer is added after the concatenation layer that applies the attention

mechanism to the embedding of known IRS channel groups to generate the key matrix $\mathbf{K}_{(e)}$, query matrix $\mathbf{Q}_{(e)}$, and value matrix $\mathbf{V}_{(e)}$ using three different linear transformations. Within the attention layer, the attention matrix $\mathbf{A}_{(e)}$ is then calculated by applying the softmax function to the dot product of the key matrix transposed $\mathbf{K}_{(e)}^T$ and the query matrix $\mathbf{Q}_{(e)}$, scaled by the square root of the dimension d.

Finally, the output of the attention layer is obtained by taking a weighted sum of the value matrix $\mathbf{V}_{(e)}$ using the attention matrix $\mathbf{A}_{(e)}$ as weights. The weighted sum is computed as a matrix multiplication between the attention matrix and the value matrix.

To prevent the problem of vanishing gradients, residual connections, and layer normalization are added after the attention layer through the add and normalize layer of the encoder. The residual connection allows the gradient to flow through the attention layer more easily, while layer normalization helps to stabilize the activations of the layer. The output of the add and normalize layer is given as:

$$\mathbf{Z}_{(e)}^1 = \mathrm{NZ}\left(\mathbf{E}_{(e)} + \mathbf{V}_{(e)}\mathrm{Softmax}\left(\frac{\mathbf{K}_{(e)}^T\mathbf{Q}_{(e)}}{\sqrt{d}}\right)\right), \tag{16}$$

The output $\mathbf{Z}_{(e)}^1$ is passed to the FC layer. In the FC layer, we have the ReLU activation function. The FC layer is followed by another add and normalize layer to obtain the final output $\mathbf{Z}_{(e)}^2$.

In the transformer decoder, the embeddings of known groups represented as $\mathbf{e}_2, \mathbf{e}_3, ..., \mathbf{e}_{P-P'}$, obtained from the encoder's GNN, are fed to the input layer of the decoder. Additionally, zero padding groups are included in the input layer of the decoder to represent the channels of unknown groups. The mask attention mechanism is applied to the decoder's embeddings, generating the key $\mathbf{K}(d)$, query $\mathbf{Q}(d)$, and value $\mathbf{V}(d)$. This mask attention ensures that the decoder does not attend to the zero padding groups, focusing only on the known groups. To mask the last k groups from the input layer, we can create a mask matrix $\mathbf{R}$ of shape (n_1, n_1) where n_1 is the number of the input groups, and set the last k rows of the mask matrix to $-\infty$. The mask matrix $\mathbf{R}$ can be expressed as follows

$$\mathbf{R}_{i,j} = \begin{cases} -\infty & \text{if } i \geq n_1 - k + 1, \\ 1 & \text{otherwise.} \end{cases} \tag{17}$$

The output of the mask attention layer is given as defined:

$$\mathbf{C}_{mask} = \mathbf{V}_{(d)}\mathrm{Softmax}\left(\left(\frac{\mathbf{K}_{(d)}^T\mathbf{Q}_{(d)}}{\sqrt{d}}\right)\mathbf{R}\right). \tag{18}$$

The output from the attention layer, denoted as $\mathbf{C}_{mask}$, is subsequently propagated through the add and normalize layer. The output of this layer is given as:

$$\mathbf{Z}_{(d)} = \mathrm{NZ}(\mathbf{E}_{(d)} + \mathbf{C}_{mask}). \tag{19}$$

The output $\mathbf{Z}_{(d)}$ and the output of the encoder $\mathbf{Z}^2_{(e)}$ is provided to the full attention layer of the decoder. Different from the mask attention layer operations, in this layer, $\mathbf{V}$ and $\mathbf{K}$ matrix are calculated based on encoder output $\mathbf{Z}_{(e)}$ and $\mathbf{Q}$ is calculated by $\mathbf{Z}_{(d)}$. Here, the full attention mechanism helps to establish the relationship between $\mathbf{Z}^2_{(e)}$ and $\mathbf{Z}_{(d)}$ by multiplying the attention matrix $\mathbf{A}$ and value matrix $\mathbf{V}$. The next layer adds $\mathbf{Z}_{(d)}$ to the output of the full attention layer, then normalize it. The next layer is the FC layer, followed by the add and normalize layer to generate the output as:

$$= \mathrm{NZ}(\mathbf{Z} + \mathrm{FC}(\mathbf{Z})), \tag{20}$$

where $\mathbf{Z} = \mathrm{NZ}(\mathbf{Z}_{(d)} + \mathbf{C})$. The final output of the decoder is obtained as $\widehat{\mathbf{G}} = [\widehat{\mathbf{G}}_2, \widehat{\mathbf{G}}_3, ..., \widehat{\mathbf{G}}_{P-3}, \widehat{\mathbf{G}}_{P-2}, \widehat{\mathbf{G}}_{P-1}, \widehat{\mathbf{G}}_P]$ by applying the FC layer at the end.

To train the transformer, the following loss function is minimized

$$\frac{\sum_{p=2}^{P}(\mathbf{G}_p - \widehat{\mathbf{G}}_p)}{\sum_{p=2}^{P}\mathbf{G}_p}, \tag{21}$$

where $\mathbf{G}_p$ is the ground truth of the cascaded channel and $\widehat{\mathbf{G}}_p$ is the cascaded channel of the p-th group estimated by the graph transformer. Further, the overall estimate of the cascaded channel is given as: $\widehat{\mathbf{G}} = [\tilde{\mathbf{G}}_1, \widehat{\mathbf{G}}_2, \widehat{\mathbf{G}}_3, ..., \widehat{\mathbf{G}}_P]$.

The proposed framework consists of two phases: offline training and estimation phase. During the offline training, the G-TIRC model is trained according to (21). To facilitate the training process, a training dataset is prepared, wherein each data point takes the following form: $[\tilde{\mathbf{G}}_1, \ldots, \tilde{\mathbf{G}}_{P-P'}, \mathbf{G}_2, \ldots, \mathbf{G}_P]$. In this dataset, each data point consists of the set of LS channel estimates of the $P - P'$ groups, along with the corresponding ground truth values for these groups. Additionally, the data point includes the ground truth of the remaining unknown groups. To train the G-TIRC model, we leverage the true knowledge of the channel conditions. This knowledge helps in mitigating the errors that may have occurred during the LS estimation for some groups. Only the estimation error related to the first group propagates through the training process since we do not employ the true values for this initial group during training. Once the transformer is successfully trained, it refines the LS channel estimates for all the input groups.

This refinement is achieved via learned knowledge to enhance the accuracy of the channel estimates. During the estimation phase, the trained G-TIRC model is deployed to predict the cascaded channel based on the LS estimate of some groups.

IV. Experimental settings and results

We consider a communication system operating at a carrier frequency of 28 GHz, where B = 8, I = 32, L = 4, and P = 8. We choose n = 5, n_1 = 7, and bandwidth (BW) = 50 MHz. The location of the BS in the cartesian coordinate system is (100, 100, 0). The first group of IRS elements is at location (0, 0, 0). The separation between each element is $\frac{\Delta}{2}$, where Δ is the carrier wavelength. The user is at the location (5, 35, -20). We assume channels from the user to the IRS and from the IRS to the BS consist of LoS and NLoS components. So, we model the channels $\mathbf{h}_p$ and $\mathbf{F}_p$ as Rician fading channel:

$$\mathbf{h}_p = \alpha_p \left(\sqrt{\frac{\zeta_{uI_p}}{\zeta_{uI_p}+1}} \bar{\mathbf{h}}_p^{L} + \sqrt{\frac{1}{\zeta_{uI_p}+1}} \breve{\mathbf{h}}_p^{NL} \right), \tag{22}$$

$$\mathbf{F}_p = \beta_p \left(\sqrt{\frac{\zeta_{I_pB}}{\zeta_{I_pB}+1}} \bar{\mathbf{F}}_p^{\,L} + \sqrt{\frac{1}{\zeta_{I_pB}+1}} \breve{\mathbf{F}}_p^{\,NL} \right), \tag{23}$$

where ζ_{uI_p} and ζ_{I_pB} represent the Rician factors of user-p-th group of IRS and p-th group of IRS-BS channel, respectively. Further, superscript L represents the LoS and NL represents the NLoS part of the channels. The path loss from the user to the p-th group of IRS elements and the path loss from p-th group of IRS elements to the BS is denoted as α_p = 30 + 22 $\log(\rho_{uI_p})$ and β_p = 30 + 22 $\log(\rho_{I_pB})$, respectively [35]. The distances of the user-p-th group and p-th group-BS are denoted as ρ_{uI_p} and ρ_{IpB}, respectively.

The LoS part of the channel $\mathbf{h}_p$ is also a function of an angle of arrival (AoA) at the IRS. Thus, it is given as:

$$\bar{\mathbf{h}}_p^{\,L} = \mathbf{a}_{\mathrm{IRS_p}}(\theta_p^{\mathrm{azi}}, \theta_p^{\mathrm{ele}}), \tag{24}$$

where θ_p^{azi} and θ_p^{ele} denote azimuth AoA and elevation AoA at the p-th group of IRS, respectively. The steering vector $\mathbf{a}_{\mathrm{IRS}_p}(\theta_p^{\mathrm{azi}}, \theta_p^{\mathrm{ele}})$ of the l-th element is given as:

$$[\mathbf{a}_{\mathrm{IRS}_p}(\theta_p^{\mathrm{azi}}, \theta_p^{\mathrm{ele}})]_i = e^{j\frac{\pi}{2}\{i_1(i) sin(\theta_p^{\mathrm{azi}}) cos(\theta_p^{\mathrm{azi}}) + i_2(l) sin(\theta_p^{\mathrm{ele}})\}}, \tag{25}$$

where $i_1(l)$ = mod$(l-1, 10)$ and $i_2(l) = \lfloor (l-1)/10 \rfloor$. The location of the p-th group of IRS is denoted as $(\rho_x^{IS_p}, \rho_y^{ISp}, \rho_z^{ISp})$ and the location of user is given as $(\rho_x^u, \rho_y^u, \rho_z^u)$, we have:

$$sin(\theta_p^{azi}) cos(\theta_p^{ele}) = \frac{\rho_y^u - \rho_y^{IS_p}}{\rho_{uI_p}}, \tag{26}$$

TABLE I

DETAILS FOR DATA GENERATION

Symbol	Parameters	Values
ζ_{IB}	Rician factor of IRS-BS	10
ζ_{uI}	Rician factor of user-IRS	0
$(\rho_x^{BS}, \rho_y^{BS}, \rho_z^{BS})$	location of BS	(100, 100, 0)
$(\rho_x^{I_pS}, \rho_y^{I_pS}, \rho_z^{I_pS})$	location of p-th group of IRS	(0, 0, 0)
$(\rho_x^u, \rho_y^u, \rho_z^u)$	location of user	(5, 35, -20)

$$sin(\theta_p^{ele}) = \frac{\rho_z^u - \rho_z^{IS_p}}{\rho_{uI_p}}, \tag{27}$$

Similarly, the LoS part of the channel $\mathbf{F}_p$ is given as:

$$\bar{\mathbf{F}_p}^L = \mathbf{a}_{\mathrm{BS}}(\phi_p^{azi}, \phi_p^{ele})\mathbf{a}_{\mathrm{IRS_p}}^H(\psi_p^{\mathrm{azi}}, \psi_p^{\mathrm{ele}}). \tag{28}$$

The steering vector $\mathbf{a}_{\mathrm{BS}}(\phi_p^{azi}, \phi_p^{ele})$ is given as:

$$\mathbf{a}_{\mathrm{BS}}(\phi_p^{azi}, \phi_p^{ele}) = [1, ..., e^{j\pi(B-1)\lambda cos(\phi_p^{ele})cos(\phi_p^{azi})}], \tag{29}$$

where ϕ_p^{azi} and ϕ_p^{ele} are the azimuth and elevation AoA, respectively at the BS. The azimuth AoD and elevation AoD from the p-th group of IRS are given as ψ_p^{azi} and ψ_p^{ele}, respectively. We have:

$$cos(\psi_p^{azi})cos(\psi_p^{ele}) = \frac{\rho_x^{IS_p} - \rho_x^{BS}}{\rho_{I_pB}}. \tag{30}$$

The location of BS is denoted as $(\rho_x^{BS}, \rho_y^{BS}, \rho_z^{BS})$. The NLoS components, $\breve{\mathbf{h}}_p^{NL}$ and $\breve{\mathbf{F}}_p^{NL}$ are modeled as i.i.d. Gaussian distribution. Table I, illustrates the details of the various parameters for data generation. A dataset consisting of 20,000 samples is created by performing experiments as per the channel and system settings mentioned above. From the 20,000 samples, a validation set and a testing set are created, each containing 4000 samples. To train the graph transformer, the loss function described in (21) is minimized using the Adam optimizer with a learning rate of 0.002.

We assume that the IRS's elements are very close to each other, and the phase shift of each element is correlated [36]. To generate a correlated phase shift matrix for a 4x8 IRS rectangular grid, we followed these steps:

- Define the desired correlation matrix $\boldsymbol{\Upsilon}(i,j)$ with size 32 $\times$ 32. This matrix represents the correlation among the phase shift values of each pair of IRS elements and can be computed using a desired correlation function. One common choice is the exponential correlation function [37], which is given as $\boldsymbol{\Upsilon}(i,j) = \exp(-\frac{\delta(i,j)}{\lambda})$ where $\delta(i,j)$ is the Euclidean distance between elements i and j, and λ is a parameter that controls the correlation. A larger value of λ corresponds to a smoother correlation function, while a smaller value corresponds to a more rapidly decaying function.
- Compute the Cholesky decomposition of the correlation matrix $\boldsymbol{\Upsilon}$.
- Compute the correlated phase shift values by multiplying the Cholesky factor $\boldsymbol{\Psi}$ with the uncorrelated phase shift vector $\boldsymbol{\varphi}$. That is, $\mathbf{s} = \boldsymbol{\Psi}\boldsymbol{\varphi}$.

A. Performance comparison

The normalized mean squared error (NMSE) is a performance metric used to evaluate the performance of the G-TIRC method. It is defined as follows:

$$NMSE = \frac{E\|\widehat{\mathbf{G}} - \mathbf{G}\|_F^2}{E\|\mathbf{G}\|_F^2}, \tag{31}$$

where $\widehat{\mathbf{G}}$ is the estimated cascaded channel matrix and $\mathbf{G}$ is the true cascaded channel matrix. To analyze the impact of λ on the NMSE of channel estimation, we generated a plot depicting signal-to-noise ratio (SNR) values against different values of λ, as shown in Fig. 3. This plot provides a visual representation of how the correlation among the phase shift values of IRS elements influences the performance of channel estimation. The plot reveals that as λ increases (indicating a smoother correlation function), the NMSE tends to decrease as well. The reason behind this observation is that a smoother correlation function fosters stronger coherence among IRS elements, resulting in a more accurate estimation of the channel. From Fig. 3, it is evident that a λ = 0.8 yields minimized NMSE, indicating that it is well-suited for achieving optimal channel estimation performance.

To showcase the efficacy of the G-TIRC method, a comparative analysis is conducted against the benchmark schemes: Oracle-OMP and Oracle-LS [23]. Additionally, we compare the G-TIRC method with the following schemes:

1) OMP [23]: By treating the channel as a sparse signal, we utilize the OMP algorithm to efficiently identify and recover the significant components of the channel.

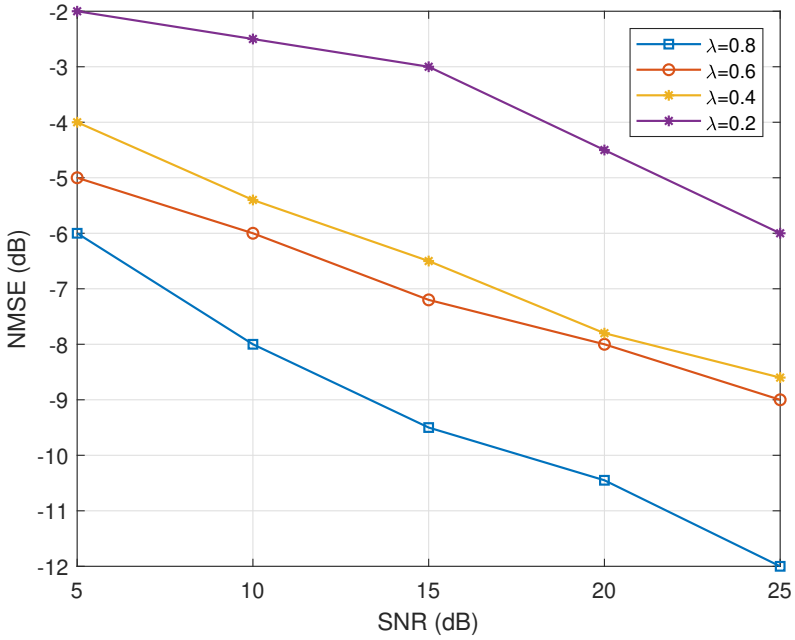

Fig. 3. NMSE vs SNR for B = 8, I = 32, and n = 5 for different values of λ.

2) SOMP [24]: By applying the structured OMP algorithm, we aim to reconstruct the sparse channel representation from limited measurements.
3) SOMP+CNN [25]: Sparse recovery of the cascaded channel is done by the SOMP algorithm. The estimate obtained from SOMP is further given to CNN to improve estimation accuracy.
4) LS and LMMSE [16]: Full rank cascaded channel estimates are obtained, sparsity is not assumed.
5) CDRN [18]: LS estimate of the full rank-cascaded channels is an input to deep residual network to predict the more accurate estimate of the cascaded channels.

We evaluate the performance of the G-TIRC method by analyzing the variation of the NMSE with respect to the SNR. The obtained results are presented in Fig. 4, where we can observe that the NMSE decreases with an increase in the SNR. The decrease in NMSE with increasing SNR can be attributed to the fact that a higher SNR results in a higher quality of the received signal. The reduced presence of noise in the received signal leads to more accurate channel estimation. The proposed G-TIRC method outperforms the other existing methods by achieving a significantly lower NMSE. Moreover, it performs closely to the oracle-OMP scheme with a gap of 1.5 dB at

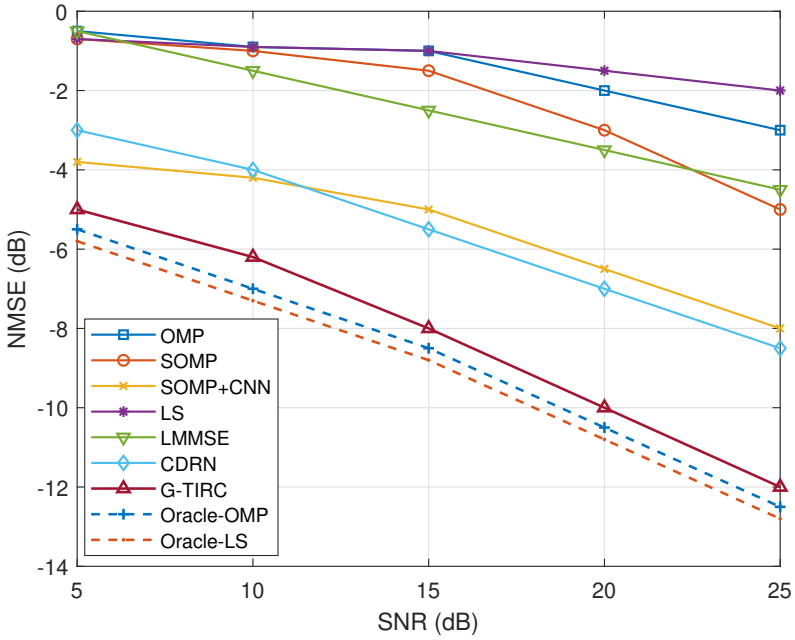

Fig. 4. NMSE vs SNR for B = 8, I = 32, and $\lambda = 0.8$.

SNR = 25 dB. Fig. 5 indicates the scalability of the proposed approach, indicating the G-TIRC

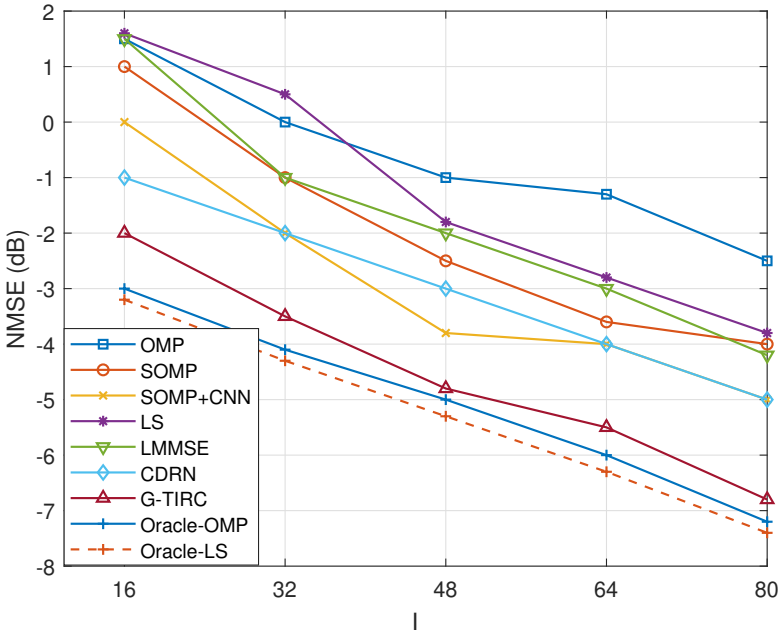

Fig. 5. NMSE vs number of IRS elements for B = 8, SNR = 5 dB, and $\lambda = 0.8$.

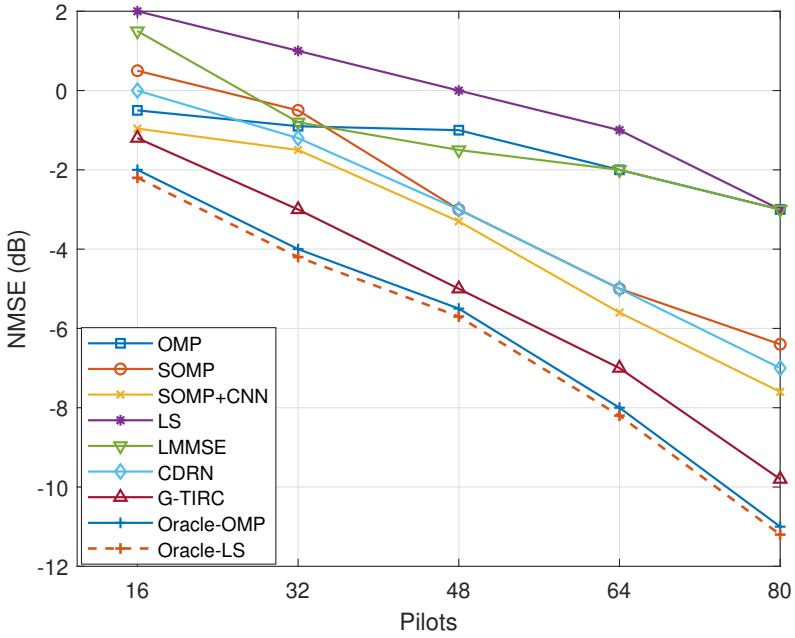

Fig. 6. NMSE vs pilots for B = 8, I = 32, and $\lambda = 0.8$ at SNR = 5 dB.

model maintains consistent performance as the system complexity increases (number of IRS elements). Furthermore, we investigate the impact of the number of pilots used in the channel estimation. As depicted in Fig. 6, the G-TIRC model requires fewer pilots than other existing methods. The G-TIRC method gives -1 dB NMSE for pilot count = 16 and the CDRN method gives the same NMSE for a pilot count = 32. Moreover, to achieve -3 dB NMSE, LMMSE requires 80 pilots but the G-TIRC achieves this NMSE with just 32 pilots. The pilot overhead reduction is 60% in comparison to LMMSE and 50% in comparison to the CDRN. This gain can be attributed to the utilization of pilots for estimating the channel in some groups, while for the remaining groups, we exploit the correlation information obtained from the known groups. By leveraging the correlation among different groups, the G-TIRC model achieves accurate channel estimation with reduced pilot overhead. The ability to estimate the channel for multiple groups based on the correlation information from a subset of known groups allows us to effectively utilize the available resources and minimize the number of required pilots.

To evaluate the robustness of our method, we perform experiments by varying the SNR values in the test dataset while keeping the SNR value in the training dataset fixed. Fig. 7 clearly demonstrates that when there is a substantial disparity between the training and testing SNR, our proposed method exhibits higher NMSE compared to scenarios where the training and testing

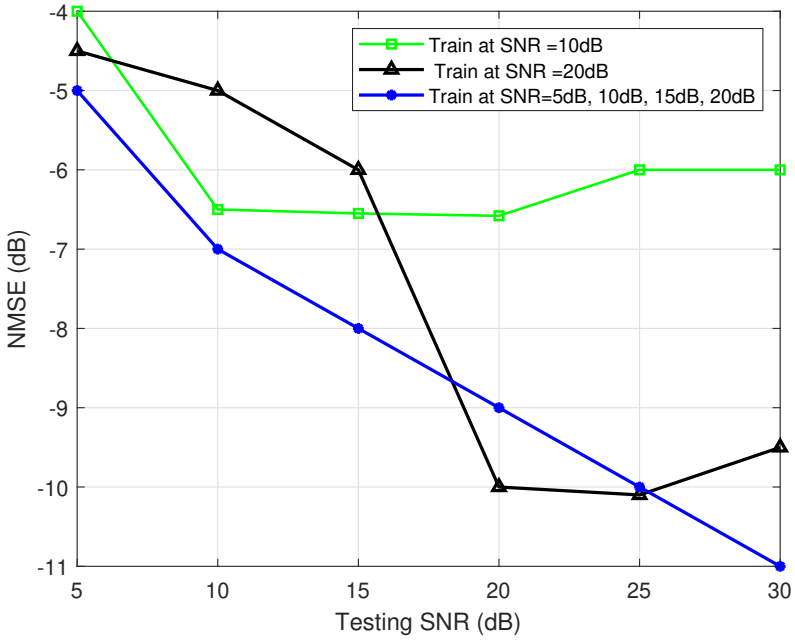

Fig. 7. NMSE vs SNR for B = 8, I = 32, and λ = 0.8.

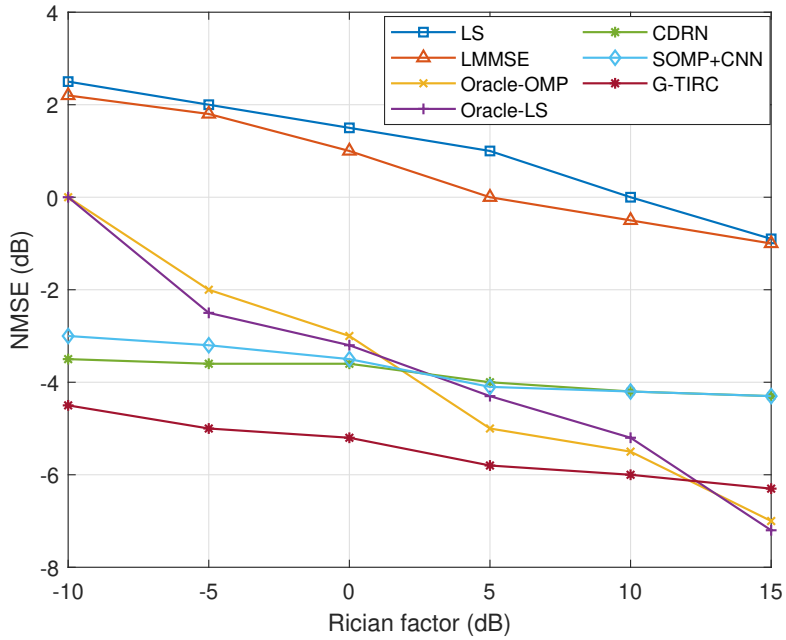

Fig. 8. NMSE vs Rician factor for B = 8, I = 32, and λ = 0.8.

SNR values are similar. To enhance the robustness of our method, we introduce mixed SNR values in the training dataset. As depicted in the figure, when the training dataset contains a mixture of SNRs, our proposed method demonstrates consistent and resilient performance across different SNR values. This indicates that by training on a diverse range of system settings, the G-TIRC model becomes more adaptable and capable of handling variations in SNR during testing, thus improving its overall robustness.

To depict the generalizability of the proposed model in different channel conditions, we calculate NMSE for different Rician factors ($\zeta = \zeta_{uI} = \zeta_{IB}$). A higher Rician factor signifies a stronger LoS component relative to NLoS components. In Fig. 8, we observe that in NLoS scenarios, traditional LS and LMMSE estimators tend to perform poorly. The complexity of NLoS channels poses challenges for these conventional methods. In contrast, deep learning-based methods, CDRN, SOMP+CNN, and G-TIRC methods demonstrate fair performance in NLoS scenarios. This is due to their skill to learn the complicated structure of the channel. Furthermore, the G-TIRC method consistently gives the least NMSE even in challenging NLoS conditions.

B. System-Level performance

During the data transmission phase, the received signal at the BS is given as:

$$y = \sqrt{\eta}\mathbf{G}\mathbf{s}x_d + n_d, \tag{32}$$

where x_d is the transmitted signal from the user and $n_d \sim CN(0, \sigma^2)$. Consider a time duration T that includes channel estimation and data transmission, SE can be expressed as:

$$SE = \left(1 - \frac{T_p}{T}\right) log_2 \left(1 + \frac{\eta}{\sigma^2} \left|\mathbf{w}^T \mathbf{G}\mathbf{s}\right|^2\right), \tag{33}$$

where $\mathbf{w} \in \mathbb{C}^{B\times 1}$ is beamforming vector at the BS. T_p is the time allocated for pilot transmission, given as $T_p = \tau_t T_0$. τ_t and T_o denote the total pilot count and T_0 is the duration of each pilot, respectively. The EE is given as:

$$EE = BW \frac{SE}{\eta_t}, \tag{34}$$

where η_t is the total power consumption in the duration T, which is expressed as:

$$\eta_t = \eta_E + \frac{T - T_p}{T}\eta + \eta_c, \tag{35}$$

since η is the power consumed for duration $T - T_p$. $\eta_E = \frac{T_p}{T}\eta_0$ is the power required during the estimation phase with η_0 is the power of each pilot, and η_c is the static power consumed by

the hardware. We conducted a study to assess the influence of our proposed channel estimation scheme, which incorporates reduced pilot overhead, on both the uplink SE and EE of the system. Initially, the channel estimates derived from our novel scheme are employed to optimize the beamforming of the IRS, as outlined in [13]. Subsequently, data transmission takes place. It is important to note that the overhead associated with channel estimation does have an impact on SE and EE, as detailed in (33) and (34).

In Fig. 9, we examine the SE while varying the number of IRS elements, across different channel estimation schemes for fixed transmitting SNR = 20 dB. The SE initially increases with a growing number of IRS elements for all schemes. However, there exists a point at which the SE reaches its maximum value. Beyond this point, further increments in I lead to a decline in SE for LS,

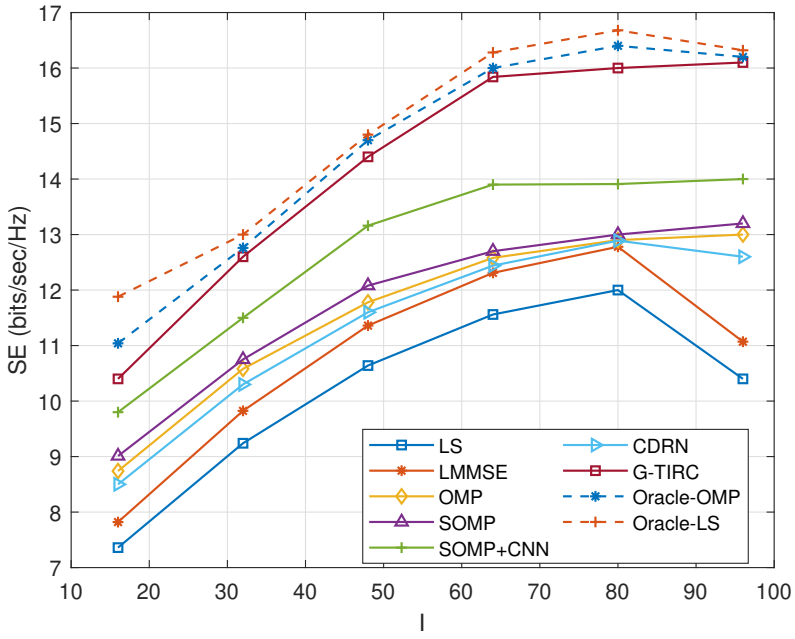

Fig. 9. SE for $T_0 = 1\mu s, T = 200\mu s$, B = 8, and I = 32.

LMMSE, and CDRN methods. This decline in SE is primarily due to the increasing channel estimation overhead as I rises, negatively impacting the system's overall performance. Notably, the G-TIRC method, designed to work with reduced pilot overhead, maintains its efficiency even with an increased number of IRS elements.

In Fig. 10, we compare the EE of the system, with the reduced pilot overhead required by the G-TIRC method, with other methods. The power required during the estimation phase is

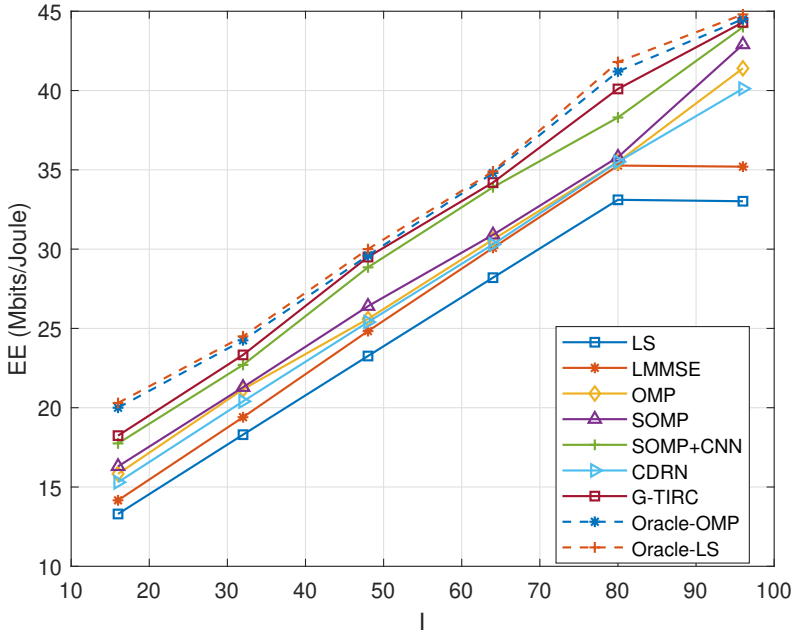

Fig. 10. EE for $T_0 = 1\mu s, \eta_0 = 0.3W, \eta = 30$W, B = 8, and I = 32.

given as $\eta_E = T_p\eta_0$, where $\eta_0 = 0.3$ W, and $\eta, \eta_c = 30$ W. Given the constraint of limited power at the user end, it becomes important to strike a delicate balance between the allocation of power for pilot signals and data transmission. When a higher number of pilots are engaged for channel estimation, a significant portion of the available power is channeled into the pilot transmission. Consequently, the power allocated for data transmission becomes restricted. This limitation directly impacts SE and ultimately influences the overall EE of the system. In Fig. 11, we depict the bit error rate (BER) attained by the IRS-assisted system employing binary phase-shift keying (BPSK) signaling, plotted against the SNR. In the context of channel estimation errors, the G-TIRC method outperforms the CDRN method, demanding 5 dB less SNR to achieve an approximate BER of 10^{-4}. This observation aligns with our earlier findings from the NMSE analysis, where we demonstrated that the G-TIRC estimates consistently yield a lower NMSE.

C. Complexity Analysis

We compute the complexity of the graph transformer in the following steps: The first two layers of the encoder are GNN layers. For the GNN layers complexity is given as $O(nd_el)$, where n is the number of groups of IRS channels and d_e is the dimension of embedding and l

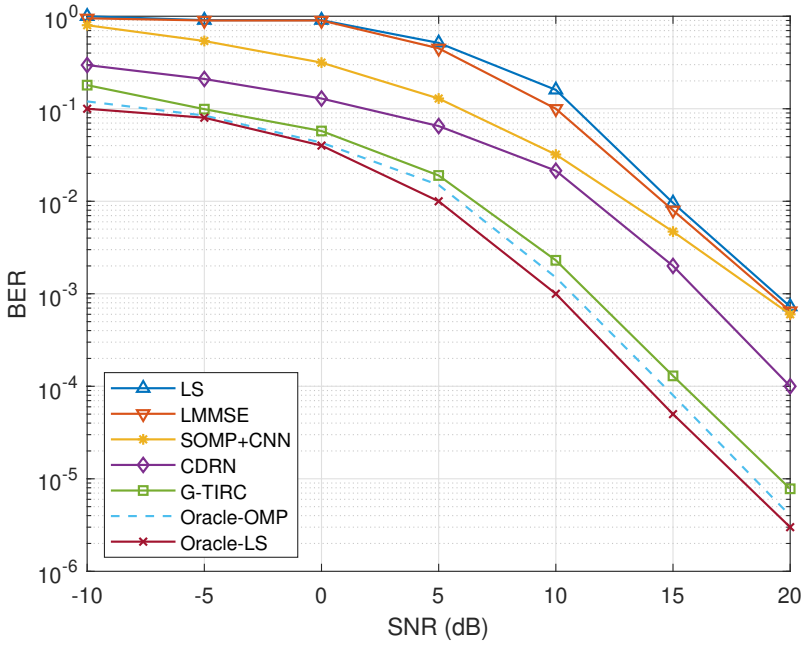

Fig. 11. BER vs SNR for B = 8 and I = 32.

is the number of GNN layers.

The complexity in multiplication of the embedding matrix with the $\mathbf{\Theta}^k$, $\mathbf{\Theta}^k$, and $\mathbf{\Theta}^k$ is $O(dun), O(dun)$, and $O(u^2n)$, respectively. The complexity in obtaining $\mathbf{A}$ is $O(udn)$ and then to obtain $\mathbf{C}$ is $O(un^2)$. It is important to highlight that the dimension d is carefully configured to ensure consistency with the dimension n. Thus, overall complexity from the attention layer is reduced to $O(u^2n)$ + $O(un^2)$. The operation in the FC layer adds complexity in order of $O(u^2n)$. Thus, the overall complexity of the encoder can be approximated as $O(u^2n) + O(un^2) + O(nd_el)$.

The mask attention layer utilized in the decoder exhibits a similar level of computational intensity as the full attention layer employed in the encoder. It is denoted as $O(u^2n_1)+O(un_1{}^2)$, where n_1 is the number of groups that are given as input to the decoder. The complexity of the full attention layer of the decoder is $O(u^2n_1) + O(u^2n) + O(un_1n)$. The complexity of the decoder can be refined as $O(u^2n_1)+O(un_1{}^2)+O(un_1n)+O(u^2n)$. The overall complexity of the proposed model can be rewritten as $O(u^2n)+O(un^2)+O(u^2n_1)+O(un_1n)+O(ndl)$. Since the input to the G-TIRC is $\tilde{\mathbf{G}}_p \in \mathbb{R}^{2BL\times 1}$, the complexity can be rewritten as:$O(B^2L^2n)+O(BLn^2)+O(B^2L^2n_1)+O(BLnn_1)+O(BLnl)$, which can be further summarized to $O(B^2L^2n_1)+O(BLn^2)$ as $n_1 > n$.

The time complexity of the proposed method is further compared in Table II with other methods.

TABLE II

COMPUTATIONAL COMPLEXITY OF DIFFERENT METHODS

Model	Computational Complexity	Estimation time (milliseconds)
LS	$O(BIlog_2(BI))$	42
MMSE	$O(B^3(I+1)^3)$	97
SOMP and OMP	$O(BI^2I_t)^\star$	71
SOMP+CNN	$O(BI^2I_t) + O\left(BI\sum_{c=1}^{N_c} n_{c-1}\cdot f_l^2\cdot n_c\right)^{\star\star}$	151
CDRN	$O\left(BI\sum_{c=1}^{N_c} n_{c-1}\cdot f_l^2\cdot n_c\right)$	90
G-TIRC	$O(B^2L^2n_1) + O(BLn^2)$	82

* I_t is the number of iterations.

** Here, f_l is the spatial size of the filter of the c-th Conv layer, $c \in \{1, 2, .., N_c\}$ and n_c represents the channels of neural network.

To simplify the comparison we compare the estimation time of each method. All the experiments are performed under the same computing power (i7-7700 at 3.6 GHz). It is noted, that the computational complexity of the G-TIRC is lower than the MMSE, CDRN, and SOMP+CNN approaches. The computational complexity of the proposed method is lower than LMMSE, CDRN, and SOMP+CNN. The complexity of CDRN and SOMP+CNN increases due to multiple convolutional layers, each involving numerous additions and multiplications in their operations. The estimation time of the proposed channel estimator is a little higher than the OMP and LS estimators. However, the estimation time can be greatly reduced by deploying a high-computing GPU at the BS.

V. CONCLUSION

In this paper, a novel technique for channel estimation in IRS-governed communication systems is proposed. We have successfully addressed the challenge of achieving accurate channel estimation while reducing the training overhead associated with IRS elements by partitioning the IRS elements into groups and leveraging the spatial correlations among them. Through the use of group-based channel estimation, we are able to estimate the channel response of the entire IRS using a reduced number of pilot symbols. The proposed technique incorporates a graph transformer, in which GNN layers extract valuable correlation information among groups. Further, the attention layer of the graph transformer effectively extracts the important correlations among different groups, to predict the channel of unknown groups, enabling more accurate and

reliable channel prediction. The experimental findings not only validate the effectiveness of the proposed technique but also its superiority in terms of channel estimation performance, showcasing reductions in training overhead when compared to other methods. Additionally, we illustrate the impact of the channel estimates obtained via the G-TIRC model on the uplink SE, EE, and BER of the system.

REFERENCES

[1] S. Gong, X. Lu, D. T. Hoang, D. Niyato, L. Shu, D. I. Kim, and Y.-C. Liang, "Toward smart wireless communications via intelligent reflecting surfaces: A contemporary survey," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 4, pp. 2283–2314, 2020.

[2] L. Dong and H.-M. Wang, "Enhancing secure MIMO transmission via intelligent reflecting surface," *IEEE Transactions on Wireless Communications*, vol. 19, no. 11, pp. 7543–7556, 2020.

[3] Y. Han, W. Tang, S. Jin, C.-K. Wen, and X. Ma, "Large intelligent surface-assisted wireless communication exploiting statistical CSI," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 8, pp. 8238–8242, 2019.

[4] M. Cheng, J.-B. Wang, H. Zhang, J.-Y. Wang, M. Lin, and J. Cheng, "Impact of finite-resolution precoding and limited feedback on rates of IRS based mmWave networks," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 5, pp. 5172–5186, 2022.

[5] Z. Sun and Y. Jing, "On the performance of multi-antenna irs-assisted noma networks with continuous and discrete irs phase shifting," *IEEE Transactions on Wireless Communications*, vol. 21, no. 5, pp. 3012–3023, 2022.

[6] S. Singh, A. Trivedi, and D. Saxena, "Unsupervised LoS/NLoS identification in mmwave communication using two-stage machine learning framework," *Physical Communication*, p. 102118, 2023.

[7] G. Yu, X. Chen, C. Zhong, D. W. K. Ng, and Z. Zhang, "Design, analysis, and optimization of a large intelligent reflecting surface-aided B5G cellular internet of things," *IEEE Internet of Things Journal*, vol. 7, no. 9, pp. 8902–8916, 2020.

[8] Z. Kang, C. You, and R. Zhang, "IRS-aided wireless relaying: Deployment strategy and capacity scaling," *IEEE Wireless Communications Letters*, vol. 11, no. 2, pp. 215–219, 2022.

[9] G.-H. Li, D.-W. Yue, and F. Qi, "Joint beamforming and power allocation for intelligent reflecting surface-aided millimeter wave MIMO systems," *Wireless Networks*, vol. 28, no. 5, pp. 1935–1947, 2022.

[10] C. Huang, A. Zappone, M. Debbah, and C. Yuen, "Achievable rate maximization by passive intelligent mirrors," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 3714–3718.

[11] H. Liu, X. Yuan, and Y.-J. A. Zhang, "Matrix-calibration-based cascaded channel estimation for reconfigurable intelligent surface assisted multiuser MIMO," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 11, pp. 2621–2636, 2020.

[12] Z.-Q. He and X. Yuan, "Cascaded channel estimation for large intelligent metasurface assisted massive MIMO," *IEEE Wireless Communications Letters*, vol. 9, no. 2, pp. 210–214, 2019.

[13] D. Mishra and H. Johansson, "Channel estimation and low-complexity beamforming design for passive intelligent surface assisted MISO wireless energy transfer," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 4659–4663.

[14] T. L. Jensen and E. De Carvalho, "An optimal channel estimation scheme for intelligent reflecting surfaces based on a minimum variance unbiased estimator," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 5000–5004.

[15] Z. Wang, L. Liu, and S. Cui, "Channel estimation for intelligent reflecting surface assisted multiuser communications: Framework, algorithms, and analysis," *IEEE Transactions on Wireless Communications*, vol. 19, no. 10, pp. 6607–6620, 2020.

[16] N. K. Kundu and M. R. McKay, "Channel estimation for reconfigurable intelligent surface aided MISO communications: From LMMSE to deep learning solutions," *IEEE Open Journal of the Communications Society*, vol. 2, pp. 471–487, 2021.

[17] G. T. de Araújo, A. L. F. de Almeida, and R. Boyer, "Channel estimation for intelligent reflecting surface assisted MIMO systems: A tensor modeling approach," *IEEE Journal of Selected Topics in Signal Processing*, vol. 15, no. 3, pp. 789–802, 2021.

[18] C. Liu, X. Liu, D. W. K. Ng, and J. Yuan, "Deep residual learning for channel estimation in intelligent reflecting surface-assisted multi-user communications," *IEEE Transactions on Wireless Communications*, vol. 21, no. 2, pp. 898–912, 2021.

[19] A. Taha, M. Alrabeiah, and A. Alkhateeb, "Enabling large intelligent surfaces with compressive sensing and deep learning," *IEEE access*, vol. 9, pp. 44 304–44 321, 2021.

[20] Y. Wang, H. Lu, and H. Sun, "Channel estimation in IRS-enhanced mmwave system with super-resolution network," *IEEE Communications Letters*, vol. 25, no. 8, pp. 2599–2603, 2021.

[21] P. Wang, J. Fang, H. Duan, and H. Li, "Compressed channel estimation for intelligent reflecting surface-assisted millimeter wave systems," *IEEE Signal Processing Letters*, vol. 27, pp. 905–909, 2020.

[22] X. Wei, D. Shen, and L. Dai, "Channel estimation for RIS assisted wireless communications—part ii: An improved solution based on double-structured sparsity," *IEEE Communications Letters*, vol. 25, no. 5, pp. 1403–1407, 2021.

[23] P. Wang, J. Fang, H. Duan, and H. Li, "Compressed channel estimation for intelligent reflecting surface-assisted millimeter wave systems," *IEEE signal processing letters*, vol. 27, pp. 905–909, 2020.

[24] Y. You, L. Zhang, M. Yang, Y. Huang, X. You, and C. Zhang, "Structured OMP for IRS-assisted Mmwave channel estimation by exploiting angular spread," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 4, pp. 4444–4448, 2022.

[25] S. Liu, Z. Gao, J. Zhang, M. D. Renzo, and M.-S. Alouini, "Deep denoising neural network assisted compressive channel estimation for mmWave intelligent reflecting surfaces," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 8, pp. 9223–9228, 2020.

[26] Z.-Q. He and X. Yuan, "Cascaded channel estimation for large intelligent metasurface assisted massive MIMO," *IEEE Wireless Communications Letters*, vol. 9, no. 2, pp. 210–214, 2020.

[27] L. Wei, C. Huang, Q. Guo, Z. Yang, Z. Zhang, G. C. Alexandropoulos, M. Debbah, and C. Yuen, "Joint channel estimation and signal recovery for RIS-empowered multiuser communications," *IEEE Transactions on Communications*, vol. 70, no. 7, pp. 4640–4655, 2022.

[28] C. Ruan, Z. Zhang, H. Jiang, J. Dang, L. Wu, and H. Zhang, "Approximate message passing for channel estimation in reconfigurable intelligent surface aided MIMO multiuser systems," *IEEE Transactions on Communications*, vol. 70, no. 8, pp. 5469–5481, 2022.

[29] Y. Yang, B. Zheng, S. Zhang, and R. Zhang, "Intelligent reflecting surface meets OFDM: Protocol design and rate maximization," *IEEE Transactions on Communications*, vol. 68, no. 7, pp. 4522–4535, 2020.

[30] B. Zheng and R. Zhang, "Intelligent reflecting surface-enhanced OFDM: Channel estimation and reflection optimization," *IEEE Wireless Communications Letters*, vol. 9, no. 4, pp. 518–522, 2019.

[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[32] A. Gillioz, J. Casas, E. Mugellini, and O. Abou Khaled, "Overview of the transformer-based models for NLP tasks," in *2020 15th Conference on Computer Science and Information Systems (FedCSIS)*. IEEE, 2020, pp. 179–183.

[33] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu *et al.*, "A survey on vision transformer," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 1, pp. 87–110, 2022.

[34] H. Jiang, M. Cui, D. W. K. Ng, and L. Dai, "Accurate channel prediction based on transformer: Making mobility negligible," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 9, pp. 2717–2732, 2022.

[35] T. Jiang, H. V. Cheng, and W. Yu, "Learning to reflect and to beamform for intelligent reflecting surface with implicit channel estimation," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 7, pp. 1931–1945, 2021.

[36] E. Björnson and L. Sanguinetti, "Rayleigh fading modeling and channel hardening for reconfigurable intelligent surfaces," *IEEE Wireless Communications Letters*, vol. 10, no. 4, pp. 830–834, 2021.

[37] T. Van Chien, A. K. Papazafeiropoulos, L. T. Tu, R. Chopra, S. Chatzinotas, and B. Ottersten, "Outage probability analysis of IRS-assisted systems under spatially correlated channels," *IEEE Wireless Communications Letters*, vol. 10, no. 8, pp. 1815–1819, 2021.