U-MLP-Based Hybrid-Field Channel Estimation for XL-RIS Assisted Millimeter-Wave MIMO Systems

Jian Xiao, Ji Wang, Zhao Chen, *Member, IEEE*, and Guangming Huang

***Abstract*—In this letter, we study the hybrid-field cascaded channel estimation in the extremely large-scale RIS (XL-RIS) assisted multi-user millimeter wave systems, where the cascaded channel is composed of far-field and near-field radiation components, and has spatial non-stationarity caused by visibility regions. We propose a U-shaped network based on the dedicated multilayer perceptron (MLP) architecture, termed as U-MLP, to realize the high-dimensional channel reconstruction with limited pilot overhead. In U-MLP, a basic feature extraction module-*Permutator* is designed to capture the long-range dependency of non-stationary channel, while the U-shaped backbone is constructed to exploit effective latent representation of the high-dimensional cascaded channel. Numerical results show that the proposed U-MLP outperforms existing channel estimation benchmarks with less pilot overhead.**

***Index Terms*—Reconfigurable intelligent surface, channel estimation, multilayer perceptron, attention mechanism.**

I. Introduction

IN PASSIVE reconfigurable intelligent surface (RIS) enabled wireless systems, the high-dimensional channel estimation is an intrinsic challenge [1], where the dimension of cascaded channel (transmitter-RIS-receiver channel) increases with the number of reflection elements. To reduce the required pilot overhead, many works have provided various channel estimation schemes, e.g., the compressed sensing (CS)-based and deep learning (DL)-based channel estimation [2], [3]. Note that most of the channel estimation schemes consider the conventional far-field communication system, where the uniform plane wave is commonly used to model the wireless channel. As the reflection elements of RIS further increase, the conventional RIS evolves to extremely large-scale RIS (XL-RIS) [4], and existing far-field assumptions are no longer valid. In this case, the near-field propagation likely happens in XL-RIS systems, which is decided by the Rayleigh distance [5].

In near-field communication, several new channel characteristics need to be considered, e.g., the spherical wavefront radiation and spatial channel non-stationarity caused by visibility region (VR) [6]. In the near-field channel estimation for the extremely large-scale multiple input multiple output (XL-MIMO) systems, the sparsity of near-field channel has been exploited by utilizing various mathematical models [7]. For the XL-RIS assisted near-field communication system, the work of [6] designed a compressed sensing (CS) algorithm to reconstruct the channel multipath parameters. CS-based channel estimation schemes need to exhibit the sparse representation of wireless channel in a particular transform domain, e.g., angular domain for the far-field channel in [2] and polar-domain for the near-field channel in [7]. Nevertheless, certain sparse transform bases are hard to fully represent the sparse structure of dynamic wireless channel.

It has been shown in [5] that both far-field and near-field signal components may coexist in practical communication systems, which constitutes a hybrid-field communication scenario. In [8], two different transform matrices were used to serially estimate the far-field and near-field components for the XL-MIMO systems, respectively. In this estimation scheme, the accuracy of near-field path estimation heavily relies on the far-field path estimation, which causes inevitable error propagation. In [9], a fixed point network-based channel estimator was proposed for hybrid-field Terahertz XL-MIMO systems, which has linear convergence guarantee and adaptive computational complexity. To the best of our knowledge, there are few works that study the hybrid-field channel estimation in the XL-RIS system. In fact, the hybrid-field channel distribution in XL-RIS communication is more complex than conventional XL-MIMO systems, where the far-field and near-field components are mixed by the product form instead of elements-wise summation.

In this letter, we study the cascaded channel estimation for XL-RIS assisted millimeter-wave (mmWave) MIMO communications, where the hybrid-field electromagnetic radiation and the different VR types are fully considered. To realize the high-dimensional and non-stationary channel reconstruction with limited pilot overhead, we proposed a U-shaped multilayer perceptron (U-MLP) network. Specifically, we design a *Permutator* module based on the dedicated MLP architecture and split attention mechanism to capture the long-range dependency of channel features [10], [11], which realizes the non-local feature extraction of spatial non-stationary channel. Furthermore, we fuse the *Permutator* module into a U-shaped hierarchical network backbone [12], which can learn the effective latent representation of the hybrid-field cascaded channel.

II. System Model and Problem Formulation

In Fig. 1, we consider a three-dimensional indoor mmWave communication environment with randomly clustered scatters,

Manuscript received 21 February 2023; revised 16 March 2023; accepted 16 March 2023. Date of publication 21 March 2023; date of current version 9 June 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62101205; in part by the Key Research and Development Program of Hubei Province under Grant 2021BAA170; and in part by the Natural Science Foundation of Hubei Province under Grant 2021CFB248. The associate editor coordinating the review of this article and approving it for publication was Y. C. Wu. *(Corresponding author: Ji Wang.)*

Jian Xiao, Ji Wang, and Guangming Huang are with the Department of Electronics and Information Engineering, College of Physical Science and Technology, Central China Normal University, Wuhan 430079, China (e-mail: jianx@mails.ccnu.edu.cn; jiwang@ccnu.edu.cn; gmhuang@ccnu.edu.cn).

Zhao Chen is with the Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100084, China (e-mail: zhao_chen@tsinghua.edu.cn).

Digital Object Identifier 10.1109/LWC.2023.3259465

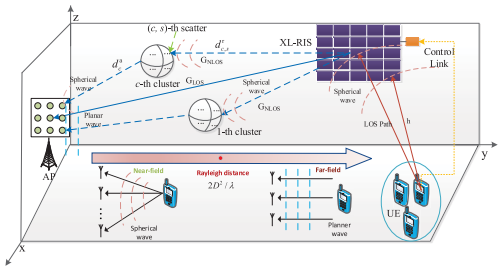
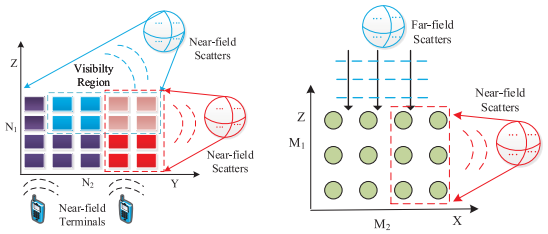


Fig. 1. XL-RIS assisted hybrid-field mmWave communications.



(a) Near-field radiation for RIS (b) Hybrid-field radiation for AP

Fig. 2. Radiation field and VR distribution for XL-RIS and AP.

where K single-antenna user equipments (UEs) communicate with a wireless access point (AP) with $M = M_1 \times M_2$ uniform planar array (UPA) antennas, aided by the XL-RIS with $N = N_1 \times N_2$ UPA elements. Following the mmWave physical channel modeling [13], we assume all scatters in RIS-AP link are grouped into C_s clusters, each having $S_c (c = 1, 2, \ldots, C_\mathrm{s})$ scatters. Considering the existing multiplicative fading effect in RIS reflection link, the XL-RIS is located at the sidewall closed to the UEs for a clear line-of-sight (LOS) path [13], while the VRs are distinct between the XL-RIS and different UEs [6]. The cascaded channel of the k-th UE is composed of UE$_k$-RIS and RIS-AP links, where the near-field region is determined by the harmonic mean of the RIS-scatter (c, s) distance $d_{c,s}^\mathrm{R}$ and UE$_k$-RIS distance d_k^UR, satisfying [5]

$$\frac{d_{c,s}^\mathrm{R} d_k^\mathrm{UR}}{d_{c,s}^\mathrm{R} + d_k^\mathrm{UR}} < Z = \frac{2D^2}{\lambda}, \tag{1}$$

where λ and D denote the carrier wavelength and equivalent array aperture of XL-RIS systems, respectively. According to (1), XL-RIS systems will operate in the near-field area when any of $d_{c,s}^\mathrm{R}$ and d_k^UR is shorter than the Rayleigh distance Z. In this letter, we consider a practical communication environment between AP and scatters, where far-field and near-field signal components coexist, constituting a hybrid-field communication scenario [5]. Fig. 2 shows the different radiation fields and VRs, in which hybrid-field scatters are distributed around the AP. According to the clustered statistical MIMO channel model [13], the RIS-AP channel $\mathbf{G} \in \mathbb{C}^{M \times N}$ can be decomposed into the LOS component $\mathbf{G}_\mathrm{LOS}$ and non-LOS (NLOS) component $\mathbf{G}_\mathrm{NLOS}$, i.e., $\mathbf{G} = I(d^\mathrm{RA})\mathbf{G}_\mathrm{LOS} + \mathbf{G}_\mathrm{NLOS}$, in which $I(d^\mathrm{RA})$ denotes the LOS possibility related to the RIS-AP distance d^RA, and $\mathbf{G}_\mathrm{NLOS}$ is given by

$$\mathbf{G}_\mathrm{NLOS} = \gamma \sum_{c=1}^{C_\mathrm{s}} \sum_{s=1}^{S_c} \beta_{c,s} \sqrt{R_{c,s}^{G_\mathrm{r}} L_{c,s}^{G_\mathrm{r}}} \mathbf{a}_{c,s} \mathbf{b}_{c,s}^T, \tag{2}$$

where $\gamma = \sqrt{\frac{1}{\sum_{c=1}^{C_\mathrm{s}} S_c}}$ is a normalization factor. Parameters $\beta_{c,s}$, $R_{c,s}$ and $L_{c,s}^{G_\mathrm{r}}$ denote the propagation complex gain, RIS elements pattern and path fading model for scatter (c, s), respectively. In conventional far-field radiation, the array response vector $\mathbf{a} \in \mathbb{C}^{M \times 1}$ at the AP and $\mathbf{b} \in \mathbb{C}^{N \times 1}$ at the RIS only depend on the identical angle of departure/arrival of scatters. The far-field array response can be expressed as

$$\mathbf{a}^\mathrm{f}\left(\phi_{c,s}^\mathrm{A}, \varphi_{c,s}^\mathrm{A}\right) = \Big[1 \cdots e^{j2\pi d_\mathrm{p}(x \sin\varphi_{c,s}^\mathrm{A} + y \sin\phi_{c,s}^\mathrm{A} \cos\varphi_{c,s}^\mathrm{A})/\lambda} \cdots e^{j2\pi d_\mathrm{p}((M_1-1)\sin\varphi_{c,s}^\mathrm{A} + (M_2-1)\sin\phi_{c,s}^\mathrm{A}\cos\varphi_{c,s}^\mathrm{A})/\lambda}\Big], \tag{3}$$

where $0 \le x \le M_1 - 1$, $0 \le y \le M_2 - 1$, and d_p is the antenna spacing. Parameters $\phi_{c,s}^\mathrm{A}$ and $\varphi_{c,s}^\mathrm{A}$ represent the azimuth and elevation angle of arrival for the (c, s)-th scatter path, respectively.

For near-field communication with a spherical wavefront, the array response is related to not only the incident angle but also the distance $d_{c,s}^\mathrm{A}$ between the AP and scatter (c, s) across different array antennas [7]. To determine the location of scatters for convenience, we assume S_c scatters in a given cluster c are distributed at the same distance d_c^A with the center of AP [13], and then the distance $d_{c,s}^\mathrm{R}$ between XL-RIS center and scatter (c, s) can be obtained. The near-field array response at the AP is given by [4]

$$\mathbf{a}^\mathrm{n}\left(d_{c,s}^\mathrm{A}\right) = \Big[e^{j2\pi d_{c,s}^\mathrm{A}(0,0)/\lambda}, \ldots, e^{j2\pi d_{c,s}^\mathrm{A}(0,M_2-1)/\lambda}, \ldots, e^{j2\pi d_{c,s}^\mathrm{A}(M_1-1,0)/\lambda}, \ldots, e^{j2\pi d_{c,s}^\mathrm{A}(M_1-1,M_2-1)/\lambda}\Big], \tag{4}$$

where $d_{c,s}^\mathrm{A}(m_1, m_2)$ represents the distance from scatter (c, s) to the (m_1, m_2)-th AP antenna, which depends on parameters $\phi_{c,s}^\mathrm{A}$, $\varphi_{c,s}^\mathrm{A}$, and d_c^A with $d_c^\mathrm{A} = d_{c,s}^\mathrm{A}(0, 0)$. Similarly, the near-field transmitting array response at the XL-RIS is given by

$$\mathbf{b}\left(d_{c,s}^\mathrm{R}\right) = \Big[e^{j2\pi d_{c,s}^\mathrm{R}(0,0)/\lambda}, \ldots, e^{j2\pi d_{c,s}^\mathrm{R}(0,N_2-1)/\lambda}, \ldots, e^{j2\pi d_{c,s}^\mathrm{R}(N_1-1,0)/\lambda}, \ldots, e^{j2\pi d_{c,s}^\mathrm{R}(N_1-1,N_2-1)/\lambda}\Big], \tag{5}$$

where $d_{c,s}^\mathrm{R}(n_1, n_2)$ represents the distance from the (n_1, n_2)-th RIS element to the (c, s)-th scatter, and $d_{c,s}^\mathrm{R} = d_{c,s}^\mathrm{R}(0, 0)$.

Since the energy distribution across array elements is not constant in near-field propagation, we consider two types of VRs in Fig. 2: cluster VR $\Omega_c (c = 1, 2, \ldots, C_\mathrm{s})$ [14], and user VR $\Psi_k (k = 1, 2, \ldots, K)$ [6]. The cluster VR Ω_c is defined as elements (antennas) region on the XL-RIS (AP) array that is visible to the given cluster c, and is identified by Ω_c's center $(V_c^\mathrm{x}, V_c^\mathrm{y})$ and length $(V_l^\mathrm{x}, V_l^\mathrm{y})$, i.e., $\Omega_c = \{[V_c^\mathrm{x} - V_l^\mathrm{x}, V_c^\mathrm{x} + V_l^\mathrm{x}], [V_c^\mathrm{y} - V_l^\mathrm{y}, V_c^\mathrm{y} + V_l^\mathrm{y}]\}$. The VR length V_l follows the Lognormal distribution $V_l \sim \mathcal{LN}(\mu_l, \sigma_l)$. The VR cover vector $p(\Omega_c) \in \mathbb{C}^{N \times 1}$ for cluster c is given by

$$[p(\Omega_c)]_n = \begin{cases} 1, & \text{if } n \in \Omega_c, \\ 0, & \text{else.} \end{cases} \tag{6}$$

Consequently, the hybrid-field NLOS channel $\mathbf{G}_\mathrm{NLOS}$ can be rewritten as

$$\mathbf{G}_\mathrm{NLOS} = \gamma \sum_{c=1}^{C_\mathrm{s}} \sum_{s=1}^{S_c} \upsilon_{c,s} \mathbf{a}_{c,s} \mathbf{b}^T\left(d_{c,s}^\mathrm{R}\right) \odot p(\Omega_c^\mathrm{R}), \tag{7}$$

where the sign $\odot$ denotes the Hadamard product, $\upsilon_{c,s} = \beta_{c,s}\sqrt{R_{c,s}^{G_r} L_{c,s}^{G_r}}$, and $p(\Omega_c^{\mathrm{R}})$ is the VR cover vector at the XL-RIS. The value of $\mathbf{a}_{c,s}$ depends on the comparison between d_c^{A} and Rayleigh distance Z, which is given by

$$\mathbf{a}_{c,s} = \begin{cases} \mathbf{a}^{\mathrm{f}}\left(\phi_{c,s}^{\mathrm{A}}, \varphi_{c,s}^{\mathrm{A}}\right), & \text{if } d_c^{\mathrm{A}} > Z, \\ \mathbf{a}^{\mathrm{n}}\left(d_{c,s}^{\mathrm{A}}\right) \odot p\left(\Omega_c^{\mathrm{A}}\right), & \text{otherwise.} \end{cases} \tag{8}$$

where $p(\Omega_c^{\mathrm{A}})$ denotes the VR cover vector at the AP.

Since UEs communicate with the XL-RIS in the near-field region, we adopt the spherical wavefront to model the receiving array response at the XL-RIS for the UE_k-RIS link. For the definition of the k-th user's VR Ψ_k, we refer to the modeling framework in [6] to directly determine the effective VR of UE_k for the LOS channel. The UE_k-RIS channel $\mathbf{h}_k \in \mathbb{C}^{N\times 1}$ can be expressed as

$$\mathbf{h}_k = \sqrt{R_k^{\mathrm{r}} L_k^{\mathrm{r}}} e^{j\eta_k} \mathbf{u}_k \odot p(\Psi_k), \tag{9}$$

where R_k^{r} denotes the radiation of RIS elements, L_k^{r} is the path fading, $\mathbf{u}_k \in \mathbb{C}^{N\times 1}$ represents the near-field receiving array response at the XL-RIS and $\eta_k \sim \boldsymbol{\mathcal{U}}[0, 2\pi]$. Vector $p(\Psi_k) \in \mathbb{C}^{N\times 1}$ denotes the n-th element of UE_k's VR cover vector.

Let $\boldsymbol{\theta} = [\beta_1 e^{j\theta_1}, \beta_2 e^{j\theta_2}, \ldots, \beta_N e^{j\theta_N}]^T \in \mathbb{C}^{N\times 1}$ denote the RIS reflecting vector, where $\theta_i (i = 1, 2, \ldots, N)$ and $\beta_i \in \{0, 1\}$ denote the phase shift and the ON/OFF state at i-th RIS element, respectively. In the q-th $(q = 1, 2, \ldots, Q)$ transmission slot, the received pilot signal $\mathbf{y}_q \in \mathbb{C}^{M\times 1}$ at the AP can be expressed as

$$\mathbf{y}_q = \sum_{k=1}^{K} \mathbf{G}\,\mathrm{diag}(\boldsymbol{\theta}_q)\mathbf{h}_k s_{q,k} + \mathbf{w}_{q,k},$$

where $s_{q,k}$ is the pilot sent by UE_k, and $\mathbf{w}_q \sim \mathcal{CN}(0, {\sigma_n}^2 \mathbf{I}_M)$ stands for complex Gaussian noise. Let $\mathbf{H}_k = \mathbf{G}\,\mathrm{diag}(\mathbf{h}_k) \in \mathbb{C}^{M\times N}$ denote the cascaded channel. After Q time slots of pilot transmission, we can obtain the $M \times Q$ observation matrix $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_Q]$ at the AP, which can be represented as

$$\mathbf{Y} = \sum_{k=1}^{K} \mathbf{H}_k \boldsymbol{\Theta} \mathbf{s}_k + \mathbf{W}_k, \tag{10}$$

where $\mathbf{s}_k = [s_{1,k}, s_{2,k}, \ldots, s_{Q,k}]^T \in \mathbb{C}^{Q\times 1}$. $\boldsymbol{\Theta} = [\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_Q] \in \mathbb{C}^{N\times Q}$, and $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_Q] \in \mathbb{C}^{M\times Q}$. In this letter, we adopt the widely used orthogonal pilot transmission strategy to realize the multi-user channel estimation [2], [6], i.e., $\mathbf{s}_{k_1}^H \mathbf{s}_{k_2} = 0$ for $1 \le k_1, k_2 \le K$ and $k_1 \ne k_2$. Hence, we can obtain the received pilot signal $\mathbf{Y}_k$ for the k-th UE at the AP.

III. Proposed Method

In existing DL-enabled channel estimation works, the convolutional neural network (CNN) with spatial modeling ability is widely used as the network backbone. However, the local convolution operations of CNN limit the effective feature extraction ability for the non-stationary high-dimensional channel. In this section, we will improve channel estimation model by exploiting the basic feature extraction module and network backbone, respectively.

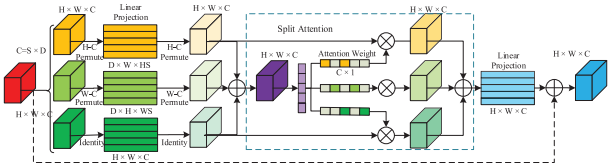


Fig. 3. The multi-branch MLP architecture-based *Permutator* module.

A. MLP-Based Feature Extraction Module

To efficiently capture the spatial non-stationary feature of the hybrid-field cascaded channel, we resort to classic MLP architecture with global receptive field to realize the long-range dependency modeling of cascaded channel. However, one-dimension MLP architecture can hardly model the spatial correlations of high-dimensional cascaded channel estimation with acceptable computational complexity. Fig. 3 shows the designed spatial domain modeling module based on MLP architecture, termed as *Permutator* [10], where different dimensions of cascaded channel feature are separately encoded by designing the multi-branch architecture.

Suppose $\mathbf{F} \in \mathbb{R}^{H\times W\times C}$ is the input tensor of the *Permutator* module, where H, W and C represent the dimensions of height, width and channel of $\mathbf{F}$, respectively. We first split $\mathbf{F}$ into $S = C/D$ segments along the channel dimensions, yielding $[\mathbf{F}_1, \mathbf{F}_2, \ldots, \mathbf{F}_S]$, satisfying $\mathbf{F}_i \in \mathbb{R}^{H\times W\times D} (1 \le i \le S)$ and $\mathbf{F} \in \mathbb{R}^{H\times W\times DS}$. In the first branch of *Permutator*, we carry out a height-channel (H-C) permutation operation with respect to each segment $\mathbf{F}_i$, and then feature maps are concatenated along the channel dimensions to obtain the feature representation $\mathbf{F}_{\mathcal{H}}^1$, i.e., $\mathbf{F} \in \mathbb{R}^{H\times W\times DS}$ is conversed to $\mathbf{F}_{\mathcal{H}}^1 \in \mathbb{R}^{D\times W\times HS}$. Further, a fully connected (FC) layer with weight $\mathbf{U}_{\mathcal{H}} \in \mathbb{R}^{HS\times HS}$ is used to interact information in the height dimensions. Lastly, we carry out an H-C permutation operation once again to recover the original dimensions $\mathbf{F}_{\mathcal{H}} \in \mathbb{R}^{H\times W\times C}$. In the second branch, the basic operations are similar to the first branch, but we carry out a width-channel (W-C) permutation operation for $\mathbf{F}$, i.e., $\mathbf{F}_{\mathcal{W}}^1 \in \mathbb{R}^{D\times H\times WS}$, and then obtain the feature $\mathbf{F}_{\mathcal{W}} \in \mathbb{R}^{H\times W\times C}$ by utilizing similar operations. In the third identity branch, we directly project $\mathbf{F}$ into the feature $\mathbf{F}_{\mathcal{H}} \in \mathbb{R}^{H\times W\times C}$ along the channel dimensions by connecting with an FC layer.

B. Split Attention for Feature Fusion

Considering different semantic information among feature branches in the *Permutator* module, we leverage the split attention mechanism to realize adaptive weighted aggregation of different feature branches [11]. For the obtained feature map $\mathbf{F}_f, \forall f \in \{\mathcal{H}, \mathcal{W}, \mathcal{C}\}$, we first compute the initial fused feature $\overline{\mathbf{F}} = \mathbf{F}_{\mathcal{H}} + \mathbf{F}_{\mathcal{W}} + \mathbf{F}_{\mathcal{C}}$ by carrying out elements-wise additions. Then, we utilize the global average pooling operation to shrink $\overline{\mathbf{F}}$ through spatial dimensions $H \times W$, and obtain the feature vector $\mathbf{z} = [z_1, \ldots, z_c, \ldots, z_C]^T \in \mathbb{R}^{C\times 1} (1 \le c \le C)$, in which z_c is given by

$$z_c = \frac{1}{H \times W} \sum_{h=1}^{H} \sum_{w=1}^{W} \overline{\mathbf{F}}_c(h, w), \tag{11}$$

where $\overline{\mathbf{F}}_c \in \mathbb{R}^{H\times W}$ denotes the c-th channel of feature map $\overline{\mathbf{F}}$.

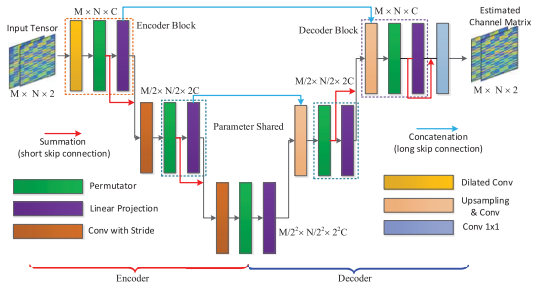


Fig. 4. The proposed U-MLP network backbone.

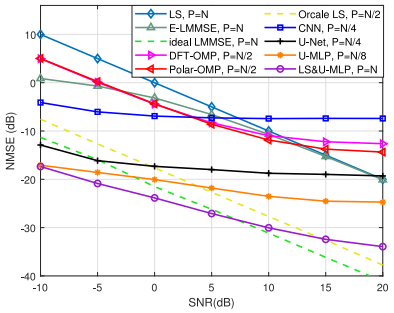


Fig. 5. NMSE v.s. SNR for different algorithms.

We first leverage an FC layer with weight $\mathbf{U}_\alpha \in \mathbb{R}^{C\times 3C}$ to obtain the feature vector $\mathbf{v} = \mathbf{z}^T\mathbf{U}_\alpha = [\mathbf{v}_\mathcal{H}, \mathbf{v}_\mathcal{W}, \mathbf{v}_C] \in \mathbb{R}^{3C}$, yielding $\mathbf{v}_f \in \mathbb{R}^C$. Then, the Softmax function is used to activate the specific attention weight $\boldsymbol{\alpha} = [\boldsymbol{\alpha}_\mathcal{H}, \boldsymbol{\alpha}_\mathcal{W}, \boldsymbol{\alpha}_C] \in \mathbb{R}^{3C}$ with respect to $\mathbf{F}_f$, i.e., $\boldsymbol{\alpha}_f^c = \frac{e^{\mathbf{v}_f^c}}{\sum_{c=1}^{C} e^{\mathbf{v}_f^c}}$ and $\mathbf{v}_f^c$ denotes the c-th elements of $\mathbf{v}_f$. We rescale $\overline{\mathbf{F}}_f$ with attention weight α_f by channel-wise multiplications and different feature branches are fused by elements-wise additions, i.e., $\overline{\mathbf{F}}_\alpha = \sum_{f=\mathcal{H}}^{C} \mathbf{F}_f \odot \boldsymbol{\alpha}_f$. Lastly, we connect $\overline{\mathbf{F}}_\alpha$ with an FC layer with weight $\mathbf{U}_l \in \mathbb{R}^{C\times C}$ and then the skip connections are designed to fuse the semantic information between original features and weighted features, i.e., $\widehat{\mathbf{F}} = \overline{\mathbf{F}}_\alpha\mathbf{U}_l + \mathbf{F}$.

C. U-Shaped Backbone Architecture for Channel Estimation

To design intelligent channel estimation models compatible with different pilot lengths, we refer to the deep CS framework to construct the dataset [3]. The input tensor of the network is designed as $\overline{\mathbf{Y}}_k = \{\mathrm{Re}(\kappa\mathbf{Y}_k\boldsymbol{\Theta}^T), \mathrm{Im}(\kappa\mathbf{Y}_k\boldsymbol{\Theta}^T)\} \in \mathbb{R}^{M\times N\times 2}$, where κ is a scaling constant to control the value range of sample data. Fig. 4 shows the proposed U-shaped network backbone, which can be regarded as the encoder-decoder architecture with skip connections [12]. In the encoding stage, we design B encoder blocks to compress the input signal $\overline{\mathbf{Y}}_k$ into the low-rank tensor $\boldsymbol{\chi} \in \mathbb{R}^{M/2^B\times N/2^B\times 2^B C}$, where C denotes the number of filters in the first convolutional layer. Specifically, the encoder block is composed of a convolutional layer, a *Permutator*, and a linear projection layer. In the first encoder block, we use the dilated convolution to expand the receptive field without increasing extra computation. For the subsequent encoder blocks, we adopt convolutional layers with stride 2 to downsample the spatial size and increase the channel number of feature map, i.e., the feature map $\mathbf{F}_b^{\mathrm{e}} \in \mathbb{R}^{M/2^{b-1}\times N/2^{b-1}\times 2^{b-1}C}$ is conversed into $\mathbf{F}_{b+1} \in \mathbb{R}^{M/2^b\times N/2^b\times 2^b C} (1 \le b \le B)$.

In the decoding stage, we adopt $B-1$ decoder blocks to realize the upsampling of the compressed tensor $\boldsymbol{\chi}$, i.e., $\boldsymbol{\chi} \in \mathbb{R}^{M/2^B\times N/2^B\times 2^B C}$ is conversed into the estimated channel matrix $\widehat{\mathbf{H}}_k$ with dimensions of $M \times N \times 2$. In the decoder block, we adopt the nearest interpolation and convolutional layer in series to increase the spatial size of feature map $\mathbf{F}_b^{\mathrm{u}}$. In the U-shaped architecture, we design the long skip connections between encoder blocks and decoder blocks, which can enhance the desired information flow in the process of channel reconstruction. Moreover, we find the architecture similarity of encoder and decoder blocks in the U-MLP, and the major drawback of MLP architecture is parameter redundancy caused by dense connection. Therefore, we share network parameters of the *Permutator* and linear projection layer between the b-th decoder block and the $(B-b)$-th encoder block to reduce the required memory of U-MLP.

IV. Numerical Results

In our simulation, we set $M = 4 \times 8$, $N = 4 \times 128$, $K = 4$, $B = 3$, $C = 48$, and carrier frequency $f_c = 73$ GHz. The detailed environment scatters distribution and path loss parameters follow the setting of [13]. In the dataset construction, we collect $S = 5000$ paired samples for each user, i.e., the total samples are $T = KS = 2 \times 10^4$. The normalized mean squared error (NMSE) is used as the performance evaluation metric $\mathrm{NMSE} = \mathbb{E}\{||\widehat{\mathbf{H}}_k - \mathbf{H}_k||_F^2/||\mathbf{H}_k||_F^2\}$, based on which we adopt the differentiable variant of L_1 loss function to optimize the proposed U-MLP network

$$L(\widehat{\mathbf{H}}_k, \mathbf{H}_k) = \frac{1}{\mathcal{B}}\sum_{i=1}^{\mathcal{B}} \sqrt{\left(\widehat{\mathbf{H}}_k^i - \mathbf{H}_k^i\right)^2 + \boldsymbol{\varepsilon}^2}, \qquad (12)$$

where $\varepsilon = 1\times 10^{-3}$ is a regularization parameter, and $\mathcal{B} = 16$ is the number of training samples in each batch.

In Fig. 5, we compare the estimation performance of the proposed U-MLP with state-of-the-art benchmarks, where the SNR is defined as $\Gamma_k = \mathbb{E}\{||\mathbf{H}_k\Theta||_F^2/||\mathbf{W}_k||_F^2\}$ for UE_k-RIS-AP link. Specifically, we provide the practical and ideal linear estimators based on a deterministic statistic model, i.e., least square (LS) and linear minimum mean square error (LMMSE) estimator [15], in which the required pilot overhead is set to $P_{\mathrm{LS}} = P_{\mathrm{LMMSE}} = N$. In empirical LMMSE (E-LMMSE) estimator, the required cascaded channel correlation matrix is a statistical correlation matrix based on Monte Carlo method, while the correlation matrix in ideal LMMSE estimation is assumed to be perfectly known. Furthermore, the estimation performance of dominated far-field and near-field channel estimation schemes are provided, i.e., DFT-OMP [2] and Polar-OMP algorithms [7] with $P_{\mathrm{OMP}} = N/2$. Note that ideal LMMSE and Oracle LS estimator provide the performance bounds for linear estimators and CS methods, respectively. We also present the performance of fully convolution-based networks, i.e., CNN and U-Net with $P_{\mathrm{CNN}} = P_{\mathrm{U\text{-}Net}} = N/4$ [3], while the pilot overhead is set to $P_{\mathrm{U\text{-}MLP}} = N/8$ for the proposed U-MLP model. To fairly compare with the DL models, the depth and width of both networks are set to be close, e.g., the number of network layers and neural nodes, while the *Permutator* modules in U-MLP are replaced by the residual convolutional blocks in CNN and U-Net. Moreover, we provide an LS&U-MLP model with better estimation accuracy, where the LS pre-estimation is utilized and hence the

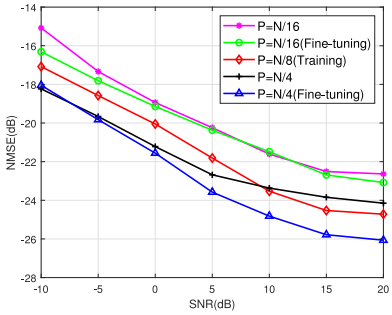


Fig. 6. NMSE v.s. SNR for different P.

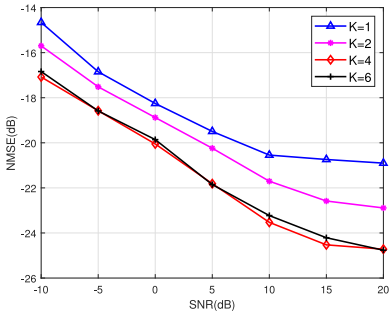


Fig. 7. NMSE v.s. SNR for different K.

TABLE I
THE TRAINING OVERHEAD FOR DIFFERENT NETWORKS

	CNN	U-Net	U-MLP
FLOPS (G)	42.25	18.79	**11.18**
Parameters (M)	**2.578**	6.347	12.79
Inference times (ms)	6.938	5.371	**4.843**

pilot overhead is equal to P_{LS}. As shown in Fig. 5, the proposed U-MLP with less pilot overhead can obtain better channel estimation accuracy than practical channel estimation schemes. Compared with ideal estimators under the high SNR, the performance gap of U-MLP will be obvious due to the inherent shortcoming of the universal approximator.

In Fig. 6 and Fig. 7, the generalization performance of U-MLP is presented under different pilot overhead P and the number of users K. In U-MLP, the trained model under fixed pilot length $P = N/8$ can realize the satisfactory estimation of $\mathbf{H}_k$ with different P in the test stage. Furthermore, we offer more accurate estimation results by utilizing the fine-tuning strategy, where pilots with different lengths are added to the training set and are trained by few epochs, i.e., $E_{\mathrm{fine}} = E/10$. Since the AP can collect more training samples with the increase of K, the estimation accuracy of the U-MLP model can be improved due to the data augmentation.

Table I compares the floating point of operations (FLOPs), parameters and inference times for different channel estimation networks, where the NVIDIA GeForce RTX 3090 is used as the training and inference platform. Although the U-MLP architecture has more redundant parameters than the CNN, the MLP only relies on basic matrix multiplication routines and realizes the faster inference speed. According to the complexity analysis in [15], the complexity of LS and the conventional CNN/MLP is quadratic in M and N, while LMMSE is cubic in M and N. The proposed *Permutator* is cubic in M and is linear in NS, ($S \leq N$). Compared with CS methods, the time-consuming with iterative operations in sparsity reconstruction can be avoided for the inference of DL models, and the high-performance GPU provides the significant acceleration.

V. CONCLUSION

In this letter, we have proposed an effective cascaded channel estimation scheme with limited pilot overhead for the XL-RIS assisted mmWave MIMO systems, where the U-MLP architecture is designed to realize the high-dimensional hybrid-field channel reconstruction. We have exploited the dedicated feature extraction module *Permutator* to capture the long-range dependency feature of spatial non-stationary cascaded channel. Furthermore, we have constructed the U-shaped network backbone to learn low-dimensional representations caused by the rank deficiency of cascaded channel. In the future works, we will consider possible cooperative XL-RIS communication scenarios, where the dimension of cascaded channel will be further increased.

REFERENCES

[1] B. Zheng, C. You, W. Mei, and R. Zhang, "A survey on channel estimation and practical passive beamforming design for intelligent reflecting surface aided wireless communications," *IEEE Commun. Surveys Tuts.*, vol. 24, no. 2, pp. 1035–1071, 2nd Quart., 2022.

[2] 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 Commun. Lett.*, vol. 25, no. 5, pp. 1403–1407, May 2021.

[3] W. Xie, J. Xiao, C. Yu, and L. Yang, "Deep compressed sensing-based cascaded channel estimation for RIS-aided communication systems," *IEEE Wireless Commun. Lett.*, vol. 11, no. 4, pp. 846–850, Apr. 2022.

[4] X. Wei, L. Dai, Y. Zhao, G. Yu, and X. Duan, "Codebook design and beam training for extremely large-scale RIS: Far-field or near-field?" *China Commun.*, vol. 19, no. 6, pp. 193–204, Jun. 2022.

[5] M. Cui, Z. Wu, Y. Lu, X. Wei, and L. Dai, "Near-field MIMO communications for 6G: Fundamentals, challenges, potentials, and future directions," *IEEE Commun. Mag.*, vol. 61, no. 1, pp. 40–46, Jan. 2023.

[6] Y. Han, S. Jin, C.-K. Wen, and T. Q. S. Quek, "Localization and channel reconstruction for extra large RIS-assisted massive MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 5, pp. 1011–1025, Aug. 2022.

[7] M. Cui and L. Dai, "Channel estimation for extremely large-scale MIMO: Far-field or near-field?" *IEEE Trans. Commun.*, vol. 70, no. 4, pp. 2663–2677, Apr. 2022.

[8] X. Wei and L. Dai, "Channel estimation for extremely large-scale massive MIMO: Far-field, near-field, or hybrid-field?" *IEEE Commun. Lett.*, vol. 26, no. 1, pp. 177–181, Jan. 2022.

[9] W. Yu, Y. Shen, H. He, X. Yu, J. Zhang, and K. B. Letaief, "Hybrid far- and near-field channel estimation for THz ultra-massive MIMO via fixed point networks," in *Proc. IEEE GLOBECOM*, 2022, pp. 5384–5389.

[10] Q. Hou, Z. Jiang, L. Yuan, M.-M. Cheng, S. Yan, and J. Feng, "Vision permutator: A permutable MLP-like architecture for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 1328–1334, Jan. 2023.

[11] H. Zhang et al., "ResNeSt: Split-attention networks," in *Proc. IEEE/CVF CVPR*, 2022, pp. 2736–2746.

[12] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Med. Image Comput. Comput. Assist. Interv.*, 2015, pp. 234–241.

[13] E. Basar, I. Yildirim, and F. Kilinc, "Indoor and outdoor physical channel modeling and efficient positioning for reconfigurable intelligent surfaces in mmWave bands," *IEEE Trans. Wireless Commun.*, vol. 69, no. 12, pp. 8600–8611, Dec. 2021.

[14] A. Amiri, S. Rezaie, C. N. Manchon, and E. De Carvalho, "Distributed receiver processing for extra-large MIMO arrays: A message passing approach," *IEEE Trans. Wireless Commun.*, vol. 21, no. 4, pp. 2654–2667, Apr. 2022.

[15] N. K. Kundu and M. R. McKay, "Channel estimation for reconfigurable intelligent surface aided MISO communications: From LMMSE to deep learning solutions," *IEEE Open J. Commun. Soc.*, vol. 2, pp. 471–487, 2021.