Multi-Scale Attention Based Channel Estimation for RIS-Aided Massive MIMO Systems

Jian Xiao, Ji Wang, *Member, IEEE*, Zhaolin Wang, *Graduate Student Member, IEEE*, Wenwu Xie, and Yuanwei Liu, *Senior Member, IEEE*

***Abstract*—A multi-scale attention based channel estimation framework is proposed for reconfigurable intelligent surface (RIS) aided massive multiple-input multiple-output systems, in which hardware imperfections and time-varying characteristics of the cascaded channel are investigated. By exploiting the spatial correlations of different scales in the RIS reflection element domain, we construct a Laplacian pyramid attention network (LPAN) to realize the high-dimensional cascaded channel reconstruction with limited pilot overhead. In LPAN, we leverage the multi-scale supervision learning to progressively capture the spatial correlations of the cascaded channel, where the attention mechanism based dual-branch architecture is designed. To balance network performance and complexity of LPAN, we further propose a lightweight LPAN-L architecture. In LPAN-L, the partial standard convolutional layers are decomposed into the group convolution, dilated convolution and point-wise convolution, which forms a sparse convolutional filter set to extract the channel feature with less computation cost. Furthermore, we leverage parameter sharing and recursion strategy to reduce the space complexity. Moreover, a selective fine-tuning strategy is developed to realize the domain adaption. Simulation results show that the proposed LPAN can achieve higher estimation accuracy than the existing estimation schemes, while the LPAN-L architecture with a close performance to LPAN efficiently reduces the network complexity[1].**

***Index Terms*—Reconfigurable intelligent surface, channel estimation, multi-scale attention, hardware impairments.**

I. Introduction

CONSIDERING the enormous communication bandwidth available at the high frequency band, millimeter wave (mmWave) has been regarded as a promising communication frequency for the future wireless communication system. However, the significant path loss of high-frequency electromagnetic waves limits the coverage of mmWave communication [2]. The intuitive solutions are to deploy denser access points (APs) or to integrate more antennas into communications equipment, e.g., massive multiple-input multiple-output (MIMO) communication, which will result in expensive hardware cost and much energy consumption. Fortunately, the reconfigurable intelligent surface (RIS), comprised of densely packed sub-wavelength units, provides new possibility to enhance mmWave communication with low cost and energy [3]. The electromagnetic response of each RIS unit is tunable by adjusting the size or spatial arrangement. By utilizing the unique electromagnetic properties, RISs have been applied in various communication scenarios to improve the system performance, e.g., wireless power transfer, mobile edge computing, and multi-hop Terahertz communications [4]. The most promising applications of RIS depend on the accurate channel state information to design the passive beamforming of RIS. However, the channel estimation is the key challenge for RIS-aided massive MIMO communication system [5].

Since the passive RIS is not equipped with radio frequency (RF) chains, channel estimation can only be carried out at the base station (BS) or the user equipment (UE). It has been verified in [6] that the performance gain of the RIS is superior to the traditional relay technology only when there are a large number of reflection elements on RIS. The increasing number of reflection units will increase the dimension of BS-RIS-UE cascaded channel matrix correspondingly. However, high-dimensional channel estimation requires more pilot overhead, which will significantly reduce communication spectrum efficiency. Besides, the hardware imperfection of communication devices also retrograde the accuracy of channel estimation in practical communication systems, i.e., the hardware impairments (HWIs) at the RIS and terminals [7].

A. Prior Works

To reduce the pilot overhead of channel estimation for RIS-aided communication system, many works have provided various design ideas, e.g., the semi-passive channel estimation by equipping with dedicated sensing devices in RIS [8]–[10], the compressed sensing (CS)-based sparse channel estimation by exploiting the sparsity of RIS channel [11]–[13], and the deep learning (DL)-based intelligent channel estimation scheme [15], [18], [19].

1) Semi-passive channel estimation schemes: In the semi-passive channel estimation scheme, limited RF chains are mounted with the RIS to process the received signal, so the BS/UE-RIS channel can be separately estimated [5], which can effectively reduce the complexity of channel estimation.

This work was supported in part by the National Natural Science Foundation of China under Grant 62101205, in part by the Natural Science Foundation of Hubei Province under Grant 2021CFB248, and in part by the Key Research and Development Program of Hubei Province under Grant 2023BAB061. Part of this work has been presented at the 2023 IEEE Wireless Communications and Networking Conference (IEEE WCNC'23), Glasgow, Scotland, UK, Mar. 2023 [1]. *(Corresponding author: Ji Wang.)*

Jian Xiao and Ji Wang 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).

Zhaolin Wang and Yuanwei Liu are with the School of Electronic Engineering and Computer Science, Queen Mary University of London, E1 4NS London, U.K. (e-mail: zhaolin.wang@qmul.ac.uk; yuanwei.liu@qmul.ac.uk).

Wenwu Xie is with the School of Information Science and Engineering, Hunan Institute of Science and Technology, Yueyang 414006, China (e-mail: gavinxie@hnist.edu.cn).

[1]The code is available at https://github.com/Holographic-Lab/LPAN

In [8], the cascaded channel estimation was divided into direction-of-arrival (DOA) and path gain estimation, where the DOA estimation was implemented by using the RF chains. The work of [9] proposed the sparse Bayesian learning-based channel reconstruction method and design an efficient data transmission strategy. In [10], an algebraic algorithm was designed to recover the multipath parameters of the BS/UE-RIS channel. However, it is necessary to configure cables or power supplies for the semi passive channel estimation, which limits the diversity of RIS application scenarios and also increases the energy consumption.

2) Sparse channel estimation schemes: Since wireless channels are often sparse in a certain transform domain, e.g., angular domain for the high-frequency communication [11]–[13], the CS has been widely used in the RIS channel estimation. The authors in [11] used the properties of Katri-Rao and Kronecker product to derive a sparse representation of cascaded channels. In [12], the double-structured orthogonal matching pursuit (OMP) was proposed by utilizing the common angle domain sparsity of multi-user cascaded channel. In [13], a two-step channel estimation scheme was proposed, in which the mmWave channel sparsity and multi-user correlations are leveraged to reduce the required pilot overhead. The CS-based channel estimation need found the pure sparse representation to avoid the grid mismatch, and then uses iterative method to approximate the solution. However, the authors in [14] believes that there is no theory can accurately prove that the CS model can obtain the most sparse representation of the channel especially for dynamic sparsity channel in the complex communication scenarios.

3) Intelligent channel estimation schemes: By leveraging the non-linear mapping ability of neural network, the channel estimation model can be constructed to realize the mapping from pilot signal to channel matrix. The authors in [15] first estimates a initial channel matrix using (least square) LS algorithm, and then obtains the accurate channel matrix using convolutional neural network (CNN). However, in this LS pre-estimation based channel estimation schemes, the required minimum pilot overhead was not reduced, which was equal to the LS estimation. Single image super-resolution (SR) reconstruction technologies provided another feasible framework for wireless channel estimation, whose theoretical foundation is the natural correlations of channel matrix, e.g., the correlations of time-frequency and spatial domain. In [16], super-resolution CNN (SRCNN) was applied to recover the complete time-frequency channel from partial channel of pilot subcarriers. However, the reconstruction performance of SRCNN was limited due to the simple network architecture. In [17], enhanced SR network (EDSR) was used to further improve the channel estimation accuracy by introducing the residual learning. Since the metamaterial units of RIS are generally integrated closely, the channels at the neighboring units are highly correlated in spatial domain. Hence, the design ideas in [16], [17] have been extend to the RIS-aided communication system. In [18], the low-dimensional cascaded channel matrix was obtained by opening partial RIS elements firstly, and then SRCNN was applied to recover the high-dimensional cascaded channel from the low-dimensional cascaded channel matrix. The work of [19] considered the part of cascaded channel estimation based on EDSR, where some active elements were equipped with RIS to acquire the initial channel information.

B. Motivations and Contributions

Against the above background, there are two main challenges for the existing channel estimation schemes based on the SR network. Firstly, for the SR model proposed in [16]–[19], the channel extrapolation was realized in merely one upsampling step, e.g., the pre-upsampling in the input layer of SRCNN [16], [18] or the post-upsampling in the output layer of EDSR [17], [19], which restricts the reconstruction precision of high-dimensional channel estimation due to the larger upscaling factor. Secondly, in the two-stage SR estimation model, the coarse low-dimensional channel matrix obtained by limited pilot overhead is used as the input of the network, which makes the reconstruction performance of the complete channel matrix depend on the accuracy of initial channel estimation. In particular, the imperfect hardware at the RIS and terminals will significantly reduce the initial channel estimation performance of the SR network due to the huge noise imposed on the input data.

Motivated by the above challenges, we propose a multi-scale attention based cascaded channel estimation framework for the RIS-aided multi-user massive MIMO communication system with the practical HWIs. The main contributions can be summarized as follows.

- We propose a Laplacian pyramid attention network (LPAN) to progressively reconstruct the cascaded channel matrix in a coarse-to-fine fashion, which can better capture the spatial correlations in high-dimensional reflection element domain of RIS. With the increase of network layers, the representation of the neural network will contain more high-frequency information. Hence, we introduce residual learning to fuse the high-frequency and low-frequency features of cascaded channel by designing the dual-branch architectures, i.e., feature extraction branch (FEB) and channel reconstruction branch (CRB).
- We integrate the attention mechanism into FEB in the LPAN, which effectively improve the channel feature learning ability of each spatial scale. Compared with the existing work [20] that applied the attention mechanism for the massive MIMO channel estimation fully following the Squeeze-and-Excitation Network (SENet) in computer vision [21], we rethink the specific characteristic of wireless channels and further design the improved channel attention mechanism. Furthermore, we merge the attention map of different spatial-scale channel matrices in the pyramid network, which are more suitable for the "divide-and-conquer" policy in the large-scale array communication system.
- We construct the lightweight version of LPAN, which is termed as LPAN-L, to reduce the parameters and the computational complexity of the proposed LPAN by exploiting efficient convolution operations and network backbone. Specifically, we combine group convolution, dilated convolution and point-wise convolution layers

to replace the standard convolutional layer in the classic CNN architecture. Furthermore, the recursion design within each pyramid level and parameter sharing strategy across pyramid levels is leveraged to reduce the network parameters. Moreover, by leveraging the multi-scale pyramid architecture of LPAN-L, we develop a selective fine-tuning based transfer learning framework to realize the cross-domain adaption of the LPAN-L model.

- Our numerical results show that the channel estimation performance of the proposed LPAN is superior to the existing classic algorithm and other DL models. Compared with LPAN, the further optimized LPAN-L can reduce approximately half of the complexity, while achieving a close performance to LPAN. The generalization and robustness of the LPAN-L are verified under different system setups, i.e., different degrees of HWIs and user mobility. The proposed transfer learning framework can be applied the LPAN-L model into different communication scenarios with limited target domain samples and training cost.

Note that compared with the conference version [1], this work further increases the contributions in terms of the system modeling and the network architecture design. Firstly, in the system model and problem formulation, we consider more practical RIS assisted mmWave systems, in which both hardware imperfections of the RIS/terminals and time-varying channel characteristics are investigated. Secondly, we introduce the attention mechanism into the Laplacian pyramid network to enhance the network representation ability and further exploit a lightweight LPAN-L architecture. Thirdly, we develop a transfer learning framework to deal with the domain mismatch problem in the practical deployment of the proposed LPAN-L model.

C. Organizations and Notations

Organizations: The remainder of the paper is organized as follows. Section II introduces the system model of RIS-aided multi-user mmWave communication system with HWIs. In Section III, we propose the LPAN to realize the progressive reconstruction of cascaded channel. In order to reduce the network complexity in the progressive reconstruction framework, we further design the low-complexity LPAN-L model in Section IV. Section V and VI provide numerical results and conclusions, respectively.

Notations: Lower-case and upper-case boldface letters $\mathbf{a}$ and $\mathbf{A}$ denote a vector and a matrix, respectively; $\mathbf{A}^T$, $\mathbf{A}^H$ and $\mathbf{A}^\dagger$ denote the transpose, conjugate transpose, and pseudo inverse of matrix $\mathbf{A}$, respectively; a^* denotes the conjugate of complex number a; diag($\mathbf{a}$) denotes the diagonal matrix with the vector $\mathbf{a}$ on its diagonal; $||\,.\,||_F$ denote the Frobenius norm; $\lfloor x \rceil$ denotes the smallest integer that is greater than or equal to x. Moreover, $\odot$ and $\otimes$ denotes the Hadamard product and convolution, respectively. $\mathbf{I}_a$ is the $a \times a$ identity matrix.

II. System Model and Problem Formulation

In this section, we first model the clustered mmWave channel for RIS-aided multi-user massive MIMO communication system. Then, we formulate the uplink channel estimation problem and present the challenge for classic channel estimation scheme, where the hardware impairments at the RIS, transmitter, and receiver are considered.

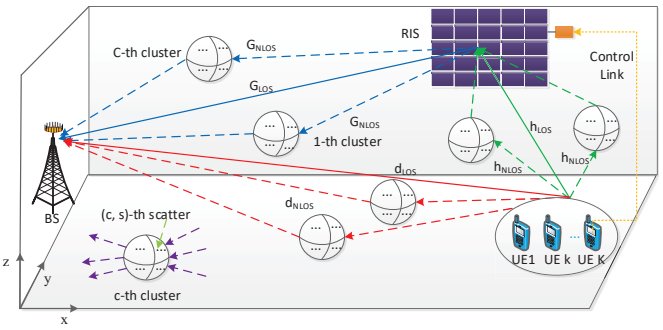


Fig. 1. The three-dimensional RIS-aided mmWave communication environment with random scattering elements.

A. Channel Model

As shown in Fig. 1, we consider the uplink of the RIS-aided multi-user mmWave communication system, where K single-antenna UEs at the x-y plane simultaneously communicate the BS with $M = M_1 \times M_2$ uniform planar array (UPA) antennas via the RIS with $N = N_1 \times N_2$ reflection elements at the x-z plane. Let $\mathbf{G}$ and $\mathbf{h}_k$ represent the RIS-BS channel and the UE$_k$-RIS channel, respectively. Following the 3rd generation partnership project (3GPP) standard [22] and the channel modeling in [23], the clustered statistical MIMO channel model is used to capture the dynamic variations of the environmental objects, e.g., a large number of randomly distributed scattering elements between the terminals and the RIS. The RIS-BS channel $\mathbf{G} = \mathbf{G}_{\text{NLOS}} + \mathbf{G}_{\text{LOS}} \in \mathbb{C}^{M \times N}$ can be represented as

$$\mathbf{G} = \underbrace{\sqrt{G_e\left(\varphi_{\text{LOS}}^{G_t}\right) L_{\text{LOS}}^{G_t}}\mathbf{b}\left(\phi_{\text{LOS}}^{G_r}, \varphi_{\text{LOS}}^{G_r}\right)\mathbf{a}^T\left(\phi_{\text{LOS}}^{G_t}, \varphi_{\text{LOS}}^{G_t}\right)}_{\mathbf{G}_{\text{LOS}}} + \underbrace{\bar{\gamma}\sum_{c=1}^{\bar{C}}\sum_{s=1}^{\bar{S}_c}\bar{\beta}_{c,s}\sqrt{G_e\left(\varphi_{c,s}^{G_t}\right) L_{c,s}^{G_t}}\mathbf{b}\left(\phi_{c,s}^{G_r}, \varphi_{c,s}^{G_r}\right)\mathbf{a}^T\left(\phi_{c,s}^{G_t}, \varphi_{c,s}^{G_t}\right)}_{\mathbf{G}_{\text{NLOS}}}, \tag{1}$$

where $\bar{C}$ and $\bar{S}_c$ denote the total number of clusters and scatters in the c-th cluster between BS and RIS for non-line of sight (NLOS) component, respectively. The parameter $\bar{\gamma} = \sqrt{\frac{1}{\sum_{c=1}^{\bar{C}} S_c}}$ is a normalization factor in the clustered channel model. The parameter $\bar{\beta}_{c,s} \sim \mathcal{CN}(0,1)$ is the propagation path gain of the scatter (c,s). The parameter $G_e\left(\varphi_{c,s}^{G_t}\right) = 2(2\xi+1)\cos^{2\xi}\left(\varphi_{c,s}^{G_t}\right)$ denotes the RIS elements pattern for the scatter (c,s), where ξ determines the gain of the element [24]. The path loss $L_{c,s}^{G_t}$

in the (c,s)-th scatter can be expressed as [25]

$$L_{c,s}^{G_t} = -20\log_{10}\left(\frac{4\pi}{\lambda}\right) - 10n\left(1+b_0\left(\frac{f_c-f_0}{f_0}\right)\right)\log_{10}\left(d_{c,s}\right) - X_{\sigma_x}, \quad (2)$$

where λ, n_0, b_0, f_c and f_0 stand for the carrier wavelength, path loss exponent, model parameter, carrier and reference frequency, respectively. $d_{c,s}$ represents the ray path length of the (c,s)-th scatter and $X_{\sigma_x} \sim \mathcal{CN}(0,{\sigma_x}^2)$ is a shadow factor.

$\phi_{c,s}^{G_t}$ ($\varphi_{c,s}^{G_t}$) and $\phi_{c,s}^{G_r}$ ($\varphi_{c,s}^{G_r}$) represent the azimuth (elevation) angle of departure at the RIS, and the azimuth (elevation) angle of arrival at the BS for the (c,s)-th path, respectively. The azimuth departure angles ($\phi_{c,s}^{G_t}, s=1,\ldots,\bar{S}_c$) follow the conditional Laplacian distribution $\phi_{c,s}^{G_t} \sim \mathcal{L}\left(\phi_c^{G_t},\sigma_\phi\right)$, where $\phi_c^{G_t}$ follows a uniform distribution $\phi_c^{G_t} \sim \mathcal{U}[-\pi/2,\pi/2]$ and σ_ϕ denotes a constant angular spread [26]. Similarly, the elevation departure angles are given by $\varphi_{c,s}^{G_t} \sim \mathcal{L}\left(\varphi_c^{G_t},\sigma_\varphi\right)$, where $\varphi_{c,s}^{G_t} \sim \mathcal{U}[-\pi/4,\pi/4]$ and σ_φ denotes angular spread. $\mathbf{a}(\phi,\varphi) \in \mathbb{C}^{N\times 1}$ and $\mathbf{b}(\phi,\varphi) \in \mathbb{C}^{M\times 1}$ denote the array response at the RIS and the BS, respectively. Specifically, the UPA array response $\mathbf{a}(\phi,\varphi)$ at the RIS can be represented as

$$\mathbf{a}(\phi,\varphi) = \Big[1,\cdots e^{j2\pi d(xsin\varphi+y\sin\phi\cos\varphi)/\lambda}, \cdots, e^{j2\pi d((N_1-1)sin\varphi+(N_2-1)\sin\phi\cos\varphi)/\lambda}\Big]^T, \quad (3)$$

where $0 \le x \le N_1-1$ and $0 \le y \le N_2-1$. The scalar d denotes the antenna spacing.

Similarly, the UE$_k$-RIS channel $\mathbf{h}_k = \mathbf{h}_{\mathrm{NLOS}}^k + \mathbf{h}_{\mathrm{LOS}}^k \in \mathbb{C}^{N\times 1}$ can be represented as

$$\mathbf{h}_k = \underbrace{\sqrt{G_e\left(\varphi_{\mathrm{LOS}}^{r,k}\right)L_{\mathrm{LOS}}^{r,k}}\mathbf{a}\left(\phi_{\mathrm{LOS}}^{r,k},\varphi_{\mathrm{LOS}}^{r,k}\right)}_{\mathbf{h}_{\mathrm{LOS}}^k} + \underbrace{\widehat{\gamma}\sum_{c=1}^{\widehat{C}}\sum_{s=1}^{\widehat{S}_c}\widehat{\beta}_{c,s}\sqrt{G_e\left(\varphi_{c,s}^{r,k}\right)L_{c,s}^{r,k}}\mathbf{a}\left(\phi_{c,s}^{r,k},\varphi_{c,s}^{r,k}\right)}_{\mathbf{h}_{\mathrm{NLOS}}^k}, \quad (4)$$

where $\widehat{C}$ and $\widehat{S}$ represent the total number of clusters and scatters in the c-th cluster between the RIS and the UE$_k$, respectively. The normalization factor $\widehat{\gamma}$ satisfies $\widehat{\gamma} = \sqrt{\frac{1}{\sum_{c=1}^{\widehat{C}}\widehat{S}_c}}$ and $\widehat{\beta}_{c,s} \sim \mathcal{CN}(0,1)$ denotes path gain. The parameters $G_e\left(\varphi_{c,s}^{r,k}\right)$ and $L_{c,s}^{r,k}$ represent the RIS element gain and path loss, respectively. $\phi_{c,s}^{r,k}$ ($\varphi_{c,s}^{r,k}$) denotes the azimuth (elevation) angle at the RIS.

Based on above statistical MIMO channel model, we further consider RIS assisted high-mobility communication scenarios, in which UE$_k$ is moving with speed v at the x-y plane. Due to the severe Doppler effect caused by the user mobility, the time-varying UE$_k$-RIS channel $\mathbf{h}_{k,n}$ at the n-th time block can be expressed as

$$\mathbf{h}_{k,n} = \underbrace{\zeta_{\mathrm{LOS}}^{r,k}\mathbf{a}\left(\phi_{\mathrm{LOS}}^{r,k},\varphi_{\mathrm{LOS}}^{r,k}\right)e^{j2\pi(nT_s f_{\mathrm{LOS}}^d - f_c\tau_{\mathrm{LOS}})}}_{\mathbf{h}_{\mathrm{LOS}}^k} + \underbrace{\widehat{\gamma}\sum_{c=1}^{\widehat{C}}\sum_{s=1}^{\widehat{S}_c}\widehat{\beta}_{c,s}\zeta_{(c,s)}^{r,k}\mathbf{a}\left(\phi_{c,s}^{r,k},\varphi_{c,s}^{r,k}\right)e^{j2\pi(nT_s f_{c,s}^d - f_c\tau_{c,s})}}_{\mathbf{h}_{\mathrm{NLOS}}^k}, \quad (5)$$

where T_s is sampling period, $\zeta_u^{r,k} = \sqrt{G_e\left(\varphi_u^{r,k}\right)L_u^{r,k}}$ with the indicator $u \in \{\mathrm{LOS},(c,s)\}$. Parameters τ_u and f_u^d denote the delay and Doppler frequency shift of the LOS path or scatter (c,s) path, respectively, in which f_u^d is given by

$$f_u^d = v\cos(\phi_n^k)\cos(\varphi_n^k)/\lambda, \quad (6)$$

where ϕ_n^k (φ_n^k) denote the azimuth (elevation) angle at UE$_k$ at the n-th time block, respectively. The maximum Doppler frequency is $f_{\max}^d = v f_c/c$, in which c is the speed of light.

B. Problem Formulation

By turning off the all reflection elements for RIS-aided communication system, the direct channel estimation from the UE to the BS is similar with conventional communication system. As such, we mainly focus on the high-dimensional cascaded channel estimation problem. Let $\boldsymbol{\theta} = [\beta_1 e^{j\theta_1},\beta_2 e^{j\theta_2},\cdots,\beta_N e^{j\theta_N}]^T \in \mathbb{C}^{N\times 1}$ denote the RIS reflecting coefficients, where $\theta_i (i=1,2,\cdots,N)$ and $\beta_i \in \{0,1\}$ denote the phase shift and the ON/OFF state of the i-th RIS element. We consider practical restrictions for ON/OFF reflection modes of the RIS [27], which can be expressed as

$$\beta_i = \begin{cases} 1-\epsilon_1 & \text{ON} \\ 0+\epsilon_0 & \text{OFF}, \end{cases} \quad (7)$$

where non-negative constants ϵ_1 and ϵ_0 model these realistic implementation errors in ON and OFF modes, respectively. Except the amplitude control error of reflection elements, there is the reflection phase error due to the intrinsic hardware imperfection of the passive reflectors, e.g., the reflection phase quantization noise $\bar{\theta}_i$ [28]. Specifically, the practical reflecting phase shift $\hat{\theta}_i$ satisfies $\hat{\theta}_i = \theta_i + \bar{\theta}_i$, in which $\bar{\theta}_i \sim \mathcal{U}[-2^{-b}\pi, 2^{-b}\pi]$ and b denotes the phase quantization bits. In the q-th ($q=1,2,\cdots,Q$) pilot slots, the received signal $\mathbf{y}_q \in \mathbb{C}^{M\times 1}$ at the BS is given by

$$\mathbf{y}_q = \sum_{k=1}^{K}\mathbf{G}\mathrm{diag}(\boldsymbol{\theta}_q)\mathbf{h}_k s_{q,k} + \mathbf{w}_q = \sum_{k=1}^{K}\mathbf{G}\mathrm{diag}(\mathbf{h}_k)\boldsymbol{\theta}_q s_{q,k} + \mathbf{w}_q, \quad (8)$$

where $\boldsymbol{\theta}_q = [\beta_{1,q}e^{j\hat{\theta}_{1,q}},\beta_{2,q}e^{j\hat{\theta}_{2,q}},\cdots,\beta_{N,q}e^{j\hat{\theta}_{N,q}}]^T \in \mathbb{C}^{N\times 1}$, $s_{q,k}$ denotes the pilot sent by the k-th UE with $\mathbb{E}[s_{q,k}s_{q,k}^*] = p_k$, and $\mathbf{w}_q \sim \mathcal{CN}(0,\sigma_n^2 I_M)$ stands for Gaussian noise. Let $\mathbf{H}_k = \mathbf{G}\mathrm{diag}(\mathbf{h}_k) \in \mathbb{C}^{\mathrm{M}\times\mathrm{N}}$ be denoted as the cascaded channel.

We consider the residual hardware impairments at the BS and the UE due to the non-ideality of the hardware in practical communication system, which can be modeled as the additive Gaussian distribution [7]. Moreover, the multiplicative phase drift $\varepsilon_q = e^{j\psi_q}$ caused by the local oscillator at the receiver is also considered, in which $\psi_q \sim \mathcal{N}(\psi_{q-1}, \delta_o)$ follows the Wiener process and δ_o denotes the oscillator quality. In this case, we rewrite (8) as

$$\widetilde{\mathbf{y}}_q = \varepsilon_q \sum_{k=1}^{K} \mathbf{H}_k \boldsymbol{\theta}_q (s_{q,k} + \eta_{q,k}^t) + \mathbf{w}_q + \eta_q^r, \tag{9}$$

where $\eta_{q,k}^t \sim \mathcal{CN}(0, \rho_{t,k}^2 p_k)$ denotes the distortion of transmitted signal caused by HWIs at UE$_k$. $\eta_q^r \sim \mathcal{CN}(0, \rho_r^2 \mathbf{p}_r)$ denotes the HWIs at the BS with $\mathbf{p}_r = \sum_{k=1}^{K} \left(p_k \mathbf{I}_M \odot (\mathbf{H}_k \boldsymbol{\theta}_q)(\mathbf{H}_k \boldsymbol{\theta}_q)^H\right)$. The $\rho_{t,k}$ and ρ_r stands for the error vector magnitudes (EVM) at UE$_k$ and BS [27], respectively.

Define $\widetilde{s}_{q,k} = s_{q,k} + \eta_{q,k}^t$ and $\widetilde{\mathbf{w}}_\mathbf{q} = \mathbf{w}_q + \eta_q^r$. After Q time slots of pilot transmission, we can collect the $M \times Q$ observation matrix $\mathbf{Y} = [\widetilde{\mathbf{y}}_1, \widetilde{\mathbf{y}}_2, \cdots, \widetilde{\mathbf{y}}_Q]$ at the BS, which is given by

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

where $\mathbf{s}_k = \left[\widetilde{s}_{k,1}, \widetilde{s}_{k,2}, \cdots, \widetilde{s}_{k,Q}\right] \in \mathbb{C}^{Q\times 1}$ and $\mathbf{s}_k^H \mathbf{s}_k = p_k Q$. The joint noise matrix $\mathbf{W} = [\widetilde{\mathbf{w}}_1, \widetilde{\mathbf{w}}_2, \cdots, \widetilde{\mathbf{w}}_Q] \in \mathbb{C}^{M\times Q}$.

The orthogonal pilot transmission strategy is widely used to realize the multi-user channel estimation based on DL framework [15], i.e., $\mathbf{s}_{k_1}^H \mathbf{s}_{k_2} = 0$ for $1 \le k_1, k_2 \le K$ and $k_1 \ne k_2$. Consequently, we can separate the received pilot signal of different users at the BS, which can be expressed as

$$\widetilde{\mathbf{Y}}_k = \frac{1}{p_k Q} \mathbf{Y} \mathbf{s}_k = \mathbf{H}_k \boldsymbol{\theta} + \widetilde{\mathbf{W}}_k, \tag{11}$$

where $\widetilde{\mathbf{W}}_k = \frac{1}{p_k Q} \mathbf{W} \mathbf{s}_k$.

In the classic LS estimator [27], the estimated cascaded channel can be expressed as

$$\widehat{\mathbf{H}}_k = \widetilde{\mathbf{Y}}_k \boldsymbol{\theta}^\dagger, \tag{12}$$

where $\boldsymbol{\theta}^\dagger = \boldsymbol{\theta}^H \left(\boldsymbol{\theta}\boldsymbol{\theta}^H\right)^{-1}$.

Remark 1: Due to the constraint of full-rank condition in (10), the required pilot overhead satisfy $Q \ge N$ for the conventional LS estimator, which causes intractable training overhead for the RIS with a large number of reflection elements. An alternative is to take advantage of the sparsity of $\mathbf{H}$ in a specific transform domain $\boldsymbol{\varphi}$. For example, in the angular domain, the channel can be represented as $\mathbf{H} = \boldsymbol{\varphi}\mathbf{H}^{\mathrm{a}}$, where $\mathbf{H}^{\mathrm{a}}$ is a sparsity matrix with $k \ll M \times N$ non-zero elements. However, the correlation of the wireless channel in practical communication scenarios is hardly to be confined to a single transform domain $\boldsymbol{\varphi}$ that fully represent the internal sparse structure of $\mathbf{H}$ [14]. In addition, the HWIs of RIS and communication devices will significantly affect the channel estimation performance for the mathematic model-based deterministic schemes.

III. Multi-Scale Attention-Aided Laplacian Pyramid Attention Network (LPAN)

In this section, we first design the dataset construction for the progressive channel estimation scheme. Then, we present the channel attention mechanism, Laplacian pyramid framework, and the detailed LPAN architecture with dual branch. Lastly, we design the multi-scale supervised training method to realize the cascaded channel reconstruction under different scales.

A. Dataset Construction

The basic idea of dataset construction follows the SR-based channel estimation scheme, which can be divided into two substage. In the first stage, we utilize the conventional channel estimator to obtain the partial channel matrix with limited pilot overhead. Then a SR network is designed to recover the complete channel matrix. In contrast to the existing SR-based channel estimation schemes, we proposed a progressive channel reconstruction scheme to better capture the spatial correlations in the cascaded channel, where the extrapolation of channel matrix is carried out under different scales.

In the channel pre-estimation stage, we adopt the LS pre-estimation presented in (12) to obtain the low-dimensional partial cascaded channel matrix $\widehat{\mathbf{H}}_k^{\mathcal{P}} \in \mathbb{R}^{M\times P}$. Specifically, we select $\mathcal{P} = \{1, \Gamma+1, \cdots, (P-1)\times\Gamma+1\}$ $\left(P = \left\lfloor \frac{N-1}{\Gamma} + 1 \right\rfloor\right)$ RIS elements with the interval $\Gamma = 2^S (0 \le S \le \log_2 N)$ as a subset of whole RIS elements, and then estimate the partial cascaded channel matrix by controlling the reflection vector of subset elements. We resort to the discrete Fourier transform (DFT) protocol in [30] to control the reflection vector of subset elements in the channel estimation stage, i.e., $\boldsymbol{\theta}_q = [1, \cdots, \theta_{i=q} = e^{-j2\pi(q-1)(i-1)/Q}, \cdots, \theta_{i=P} = e^{-j2\pi(q-1)(P-1)/Q}]^T$ at the q-th slot. Due to the phase quantization noise and hardware imperfection, the practical RIS reflection coefficients are given by $\boldsymbol{\theta}_q = [\beta_1, \cdots, \theta_{i=q} = \beta_i e^{-j(2\pi(q-1)(i-1)/Q + \bar{\theta}_i)}, \cdots, \theta_{i=P} = \beta_P e^{-j(2\pi(q-1)(P-1)/Q + \bar{\theta}_P)}]^T$ at the q-th slot for the dataset construction.

In the dataset construction, we consider two cases of quasi-static channel and time-varying channel estimation. For the quasi-static channel channel estimation, the flat-fading channel $\mathbf{H}_k$ remains approximately constant within each frame. Hence, the estimated channel at the pilot block can be used into the data transmission stage in the same frame. In this case, we define $\bar{\mathbf{H}}^{\mathcal{P}} \in \mathbb{R}^{M\times P\times 2}$ as the input data of channel extrapolation network, and $\bar{\mathbf{H}}_{m,p,1}^{\mathcal{P}} = \mathrm{Re}(\widehat{\mathbf{H}}^{\mathcal{P}})$ and $\bar{\mathbf{H}}_{m,p,2}^{\mathcal{P}} = \mathrm{Im}(\widehat{\mathbf{H}}^{\mathcal{P}})(1 \le m \le M)$, as the real and imaginary components, respectively. We design the label group $\bar{\mathbf{H}} = (\bar{\mathbf{H}}_1, \bar{\mathbf{H}}_2, \cdots, \bar{\mathbf{H}}_S)$ to achieve the progressive reconstruction of the cascaded channel, where $\bar{\mathbf{H}}_S$ represents the complete cascaded channel matrix and $\bar{\mathbf{H}}_s \in \mathbb{R}^{M\times 2^s P\times 2}(1 \le s \le S)$ is the spatial sampling with scaling factor s of the complete cascaded channel.

For the time-varying channel estimation, the channels of consecutive time blocks within a frame may vary due to the short channel coherence time T_c. Fig. 2 shows the specific frame structure with B time blocks for the time-varying channel estimation, which is divided into B^{p} pilot blocks and B^{d}

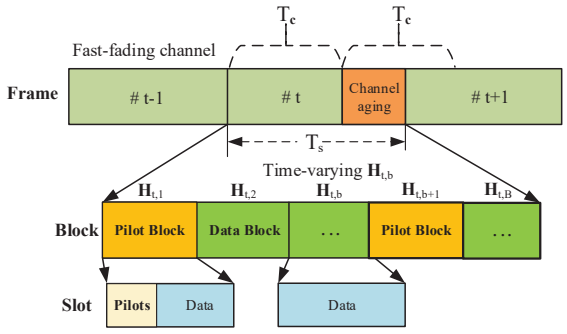


Fig. 2. The specific frame structure for time-varying channel estimation.

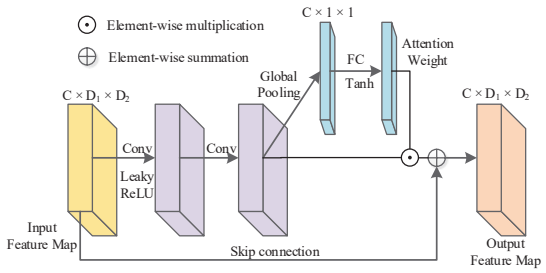


Fig. 3. The improved channel attention block.

data blocks, i.e., $B^{\mathrm{p}}+B^{\mathrm{d}} = B$. The channel at the $n(1 \le n \le B)$-th data block need to be predicted from the estimated channels at pilot blocks. In this case, the input tensor $\widetilde{\mathbf{H}}^{\mathcal{P}} \in \mathbb{R}^{M\times P\times 2B^{\mathrm{p}}}$ of the channel estimation network is the concatenation of pre-estimated channels at B^{p} pilot blocks. Accordingly, the output tensor $\widehat{\mathbf{H}} \in \mathbb{R}^{M\times N\times 2B}$ of the network denotes the concatenation of cascaded channels of all time blocks within a frame. By constructing similar label group with the quasi-static channel estimation, i.e., $\widetilde{\mathbf{H}} = (\widetilde{\mathbf{H}}_1, \widetilde{\mathbf{H}}_2, \cdots, \widetilde{\mathbf{H}}_S)$ with the spatial sampling channel $\widetilde{\mathbf{H}}_s \in \mathbb{R}^{M\times 2^s P\times 2B}$, the multi-scale supervision training framework can be developed for the cascaded channel estimation.

B. Channel Attention Mechanism

The attention mechanism has been widely applied in numerous DL tasks, which can enhance local useful features and suppress other useless information. Fig. 3 shows the designed channel attention block (AB) based on classic SENet architecture [21], which sets adaptive weights for different channels in the feature map. Note that some more advanced attention mechanism have been proposed in the DL field. Compared with other attention mechanisms, the architecture of AB is more simple and concise, which only introduces an extra branch to learn a set of attention weights compared with the classic residual block [32] Besides, the adaptive learning of AB is efficient for channel estimation, which conform to the "divide-and-conquer" policy in the traditional large-scale array communications [20]. As such, the improved AB is exploited to the RIS channel estimation in the following.

Let $\mathbf{X}_i \in \mathbb{R}^{C\times D_1\times D_2}$ denote the input feature map in the i-th AB, where C, D_1 and D_2 denote the channel, height and width of feature map $\mathbf{X}_i$, respectively. Firstly, we stack two convolution layer with C filters to obtain the feature $\mathbf{F}_i = [\mathbf{f}_1, \mathbf{f}_2, \cdots \mathbf{f}_C] \in \mathbb{R}^{C\times D_1\times D_2}$. Then, we adopt the global average pooling to shrink $\mathbf{F}_i$ through spatial dimensions $D_1 \times D_2$. Let $\mathbf{z}_i = [z_1, z_2, \cdots z_c, \cdots, z_C]^T \in \mathbb{R}^{C\times 1}$ denote the channel statistic of $\mathbf{F}_i$, where z_c is given by

$$z_c = \frac{1}{D_1 \times D_2} \sum_{d_1=1}^{D_1} \sum_{d_2=1}^{D_2} \mathbf{f}_c(d_1, d_2). \tag{13}$$

In the learning process of attention weight for the original SENet, two fully connected (FC) layers are designed to capture non-linear cross-channel interaction, which involves dimensionality reduction for controlling model complexity, and the attention architecture also be adopted in the existing works for channel estimation of massive MIMO systems [20]. However, the dimensionality reduction between two FC layers destroys the direct correspondence between channel and its weight [33]. Consequently, we adopt a FC layer with C neurons to realize the direct connection between channel and weights, which can capture the channel-wise dependencies. Moreover, the Sigmoid activation function is used to obtain the attention weight $\boldsymbol{\alpha} = \delta(\mathbf{W}_{\mathrm{FC}}\mathbf{z}_i) \in {}^{C\times 1}$, where $\mathbf{W}_{\mathrm{FC}} \in {}^{C\times C}$ denotes the weight of the FC layer and $0 \le \delta(x) = \frac{1}{1+e^{-x}} \le 1$. In general, the Sigmoid activation function confines the attention weight α to the range of $(0, 1)$, which is suitable for *positive real*-valued pixel in the computer vision. However, the communication data in the considered scenario is *complex*-valued, whose amplitude and phase information can not be well characterized by the Sigmoid function.

In the proposed AB, we use the hard Tanh activation function gating mechanism to generate the attention weight, i.e., $-1 \le \delta(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \le 1$, which can not only adjust the each channel intensity of the feature map, but can also control the direction of the output feature. Moreover, the Tanh function is centrally symmetric with a mean value of 0. Therefore, it can still map the Gaussian distribution $\mathcal{N}(0, 1)$ to a distribution that maintains the characteristic of zero mean value. We rescale the $\mathbf{F}_i$ with attention weight α to obtain the weighted feature map by adopting the channel-wise multiplication, and then skip connection is used to fuse the semantic information between the original feature and the weighted feature. Based on the above mechanism, the output $\mathbf{A}_i$ of AB can be expressed as

$$\mathbf{A}_i = \mathbf{X}_i + \mathbf{F}_i \odot \boldsymbol{\alpha}. \tag{14}$$

C. Laplacian Pyramid

The backbone and information flow of the proposed LPAN follow the Laplacian pyramid framework that is the improvement of Gauss pyramid by introducing the residual coefficients [34]. In the Gauss pyramid, the original resolution image at the bottom of pyramid is sequentially downsampled, which forms a set of images arranged from top to bottom in the shape of a pyramid according to the size of image resolution. However, this sampling operation will lose high-frequency information of images. Let $\mathcal{G}(\Xi) = [\boldsymbol{\Xi}_0, \boldsymbol{\Xi}_1, \cdots, \boldsymbol{\Xi}_S]$ denote a Gauss pyramid with S levels, where $\boldsymbol{\Xi}_s (0 < s < S)$ denotes the

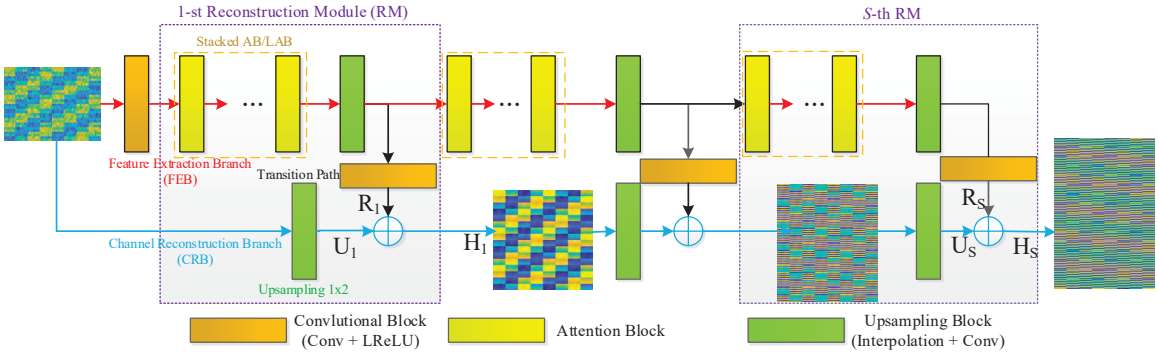


Fig. 4. The proposed Laplacian pyramid attention network (LPAN) architecture.

s-th level image of the pyramid. The s-th level of Laplacian pyramid can be represented as

$$\mathbf{R}_s = \mathcal{G}_s(\mathbf{\Xi}) - u\left(\mathcal{G}_{s+1}(\mathbf{\Xi})\right) = \mathbf{\Xi}_s - u\left(\mathbf{\Xi}_{s+1}\right), \tag{15}$$

where $u\left(\mathbf{\Xi}_{s+1}\right)$ denotes the upsampling image of $\mathbf{\Xi}_{s+1}$, and the residual coefficient $\mathbf{R}_s$ represents the high-frequency information of image.

In the DL-based channel estimation model, the feature map $\mathbf{F}$ compose of high-frequency feature $\mathbf{F}_H$ and low-frequency feature $\mathbf{F}_L$, that is $\mathbf{F} = \mathbf{F}_H + \mathbf{F}_L$. When we use neural network to extract the feature of data, the feature map will represent more high-frequency information for deeper network layer, while $\mathbf{F}_L$ is the important component for the reconstruction of $\mathbf{F}$. Consequently, we can design the progressive cascaded channel estimation model by imitating the Laplacian pyramid architecture, where the upsampling operator $u(.)$ and the Laplacian coefficients $\mathbf{R}_s$ is designed by neural network.

D. The Dual-Branch and Multi-Scale Architecture of LPAN

Fig. 4 shows the proposed Laplacian pyramid attention network (LPAN) architecture with S reconstruction modules (RMs), which progressively upscale the lower-dimensional channel matrix by a scale of 2 in the reflection element domain of RIS. The s-th RM can be divided into two branches, namely FEB and CRB, which learn the high-frequency information $\mathbf{R}_s$ and the low-frequency information $\mathbf{U}_s$ of cascaded channel matrix, respectively. Note that the image size is decreasing with the increase of the pyramid level s, while the dimension of the channel matrix is increasing with the increase of the number of RM s, i.e., $\mathbf{\Xi}_S = \tilde{\mathbf{H}}^{\mathcal{P}}$ and $\mathbf{\Xi}_0 = \tilde{\mathbf{H}}_S$ in the channel estimation.

In the FEB of the s-th RM, we first use a convolutional block (CB), which is composed of convolutional layer and Leaky Rectified Linear Unit (LeakyReLU) activation functional layer, to boost the number of channel of the input feature map $\tilde{\mathbf{H}}^{\mathcal{P}}$. Next, J ABs are stacked to extract the more representative features. Generally, normalization layers are used to stabilize the training process of deep neural network, e.g., batch normalization (BN) layer. In the SR-based channel estimation model, the input low-resolution channel matrix has a similar space distribution to the complete channel matrix, while BN will change the original data distribution. In the proposed LPAN, we adopt weight normalization (WN) to reparameterize the weight vector of the network instead of normalizing the mini-batch data of each layer in BN [35]. Specifically, WN decouples the original network weight $\mathbf{w}$ into a parameter vector $\mathbf{v}$ and a scalar parameter g as follows:

$$\mathbf{w} = \frac{g}{\|\mathbf{v}\|}\mathbf{v}, \tag{16}$$

where $\mathbf{v}/\|\mathbf{v}\|$ denotes the identity vector of $\mathbf{w}$. Let $\nabla_{\mathbf{w}}\mathcal{L}$, $\nabla_g\mathcal{L}$, and $\nabla_{\mathbf{v}}\mathcal{L}$ represent the gradients of loss function $\mathcal{L}$ with respect to $\mathbf{w}$, g, and $\mathbf{v}$, respectively. In the process of network training, the optimization of $\mathbf{w}$ is transformed into the optimization of g and $\mathbf{v}$, which are given by

$$\nabla_g\mathcal{L} = \nabla_g\mathbf{w}(\nabla_{\mathbf{w}}\mathcal{L})^T = \frac{\nabla_{\mathbf{w}}\mathcal{L}v^T}{\|\mathbf{v}\|}, \tag{17}$$

$$\nabla_{\mathbf{v}}\mathcal{L} = \frac{g}{\|\mathbf{v}\|}\nabla_{\mathbf{w}}\mathcal{L} - \frac{g\nabla_g\mathcal{L}}{\|\mathbf{v}\|^2}\mathbf{v} = \frac{g}{\|\mathbf{v}\|}M_{\mathbf{w}}\nabla_{\mathbf{w}}\mathcal{L}, \tag{18}$$

where $M_{\mathbf{w}} = \mathbf{I} - \mathbf{w}\mathbf{w}^T/\|\mathbf{w}\|^2$ is a projection matrix that projects onto the complement of the vector $\mathbf{w}$. Compared the with initial $\nabla_{\mathbf{w}}\mathcal{L}$, $\nabla_{\mathbf{v}}\mathcal{L}$ scales the weight gradient by $g/\|\mathbf{v}\|$, and it projects the gradient away from the current weight vector $\mathbf{w}$, which can stabilize the training of the network and accelerate the network convergence.

Compared with BN, the performance of WN is not related with the batch size and data, and the memory and computation overhead is lower. Moreover, the SR-based channel estimation is sensitive to the learning rate η with a small value, e.g., $\eta = 10^{-4}$ [19]. The training loss of the network without WN layer will explode for a larger η, while the small learning rate is easy to cause overfitting. The WN can provide a wider range of η in the training, which improve the estimation accuracy in the test phase.

After the feature extraction of J ABs in the s-th RM, we use an upsampling block (UB) to scale the feature map to a desired dimension of the channel matrix, e.g., $\tilde{\mathbf{H}}^{\mathcal{P}} \in \mathbb{R}^{M\times P\times 2} \rightarrow \tilde{\mathbf{H}}_1 \in \mathbb{R}^{M\times 2P\times 2}$ or $\widetilde{\mathbf{H}}^{\mathcal{P}} \in \mathbb{R}^{M\times P\times 2B^{\mathrm{P}}} \rightarrow \tilde{\mathbf{H}}_1 \in \mathbb{R}^{M\times 2P\times 2B}$ in the first RM. In the UB, we adopt the nearest interpolation and the convolutional layer to increase the size of the feature map, which can avoid check artifacts in the upsampling [36]. Since the cascaded channel is represented as the real-valued matrix with two channels, the output of FEB reduces the number of channels of the feature map to $\mathbf{R}_s$ through a CB with 2 filters,

which is termed as the transition path (TP).

In the second branch of the proposed LPAN, we design the CRB to characterize the information flow of low-frequency components in high-dimensional cascaded channel estimation, which is equal to the function $u(\cdot)$ in the Laplacian pyramid. In the CRB of the s-th RM, the lower-dimensional channel matrix is directly upscaled to $\mathbf{U}_s$ by the UB. Hence, the output of the s-th RM can be expressed as

$$\widehat{\mathbf{H}}_s = \mathbf{U}_s + \mathbf{R}_s. \tag{19}$$

E. Multi-Scale Supervision

To realize the progressive construction for high-dimension cascaded channel, we adopt multi-scale supervised learning to generate the cascaded channel matrix with different scales. The normalized mean squared error (NMSE) is widely used as the performance metric of channel estimation, which is defined as $\text{NMSE} = \mathbb{E}[||\widehat{\mathbf{H}} - \mathbf{H}||_F^2 / ||\mathbf{H}||_F^2]$. Intuitively, the L_2 loss with Euclidean distance can directly reflect the NMSE metric. In fact, for the SR-based channel estimation task, L_1 loss function with Manhattan distance can achieve better performance compared with L_2 loss [17]. However, the gradient of L_1 loss will jump at the extreme point, e.g., zero value, and a small difference will also lead to a large gradient. As a remedy, we adopt the Charbonnier loss function to optimize the whole network [34], which is a differentiable variant of L_1 loss. The multi-scale supervision based loss function is given by

$$\begin{aligned}\mathcal{L}(\bar{\mathbf{H}}, \widehat{\mathbf{H}}) &= \frac{1}{\mathcal{B}} \sum_{i=1}^{\mathcal{B}} \sum_{s=1}^{S} \rho\left(\bar{\mathbf{H}}_s^{(i)} - \widehat{\mathbf{H}}_s^{(i)}\right) \\ &= \frac{1}{\mathcal{B}} \sum_{i=1}^{\mathcal{B}} \sum_{s=1}^{S} \rho\left(\bar{\mathbf{H}}_s^{(i)} - \mathbf{U}_s^{(i)} - \mathbf{R}_s^{(i)}\right),\end{aligned} \tag{20}$$

where $\rho(\mathbf{X}) = \sqrt{\mathbf{X}^2 + \varepsilon^2}$ is the Charbonnier penalty function, ε is a regularization parameter, and $\mathcal{B}$ is the number of training sample in each batch. For the case of time-varying channel estimation, the label $\bar{\mathbf{H}}$ in (20) is replaced as $\widetilde{\mathbf{H}}$.

Remark 2: Compared with existing SR network-based channel estimation models, the architecture of the proposed LPAN has two unique characteristics: 1) progressive upsampling strategy along the depth direction; and 2) dual-branch pipeline along the width direction, which increase the network capability and realize the fine channel reconstruction. However, such a structure may cause a more complex network in terms of parameters and floating point of operations (FLOPs). Specifically, the computation of neural network is proportional to the dimension of computed feature map, while the progressive upsampling strategy in the LPAN will introduce more FLOPs compared with the post-upsampling SR model, e.g., EDSR. Besides, the dual-branch architecture of the LPAN introduces more parameters. To address the aforementioned challenges, we design a lightweight version of LPAN in the following, which can achieve good balance between performance and complexity.

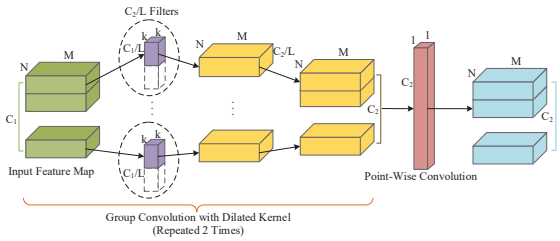


Fig. 5. The proposed lightweight convolution module in LPAN-L.

IV. LPAN-L: The Low-Complexity Architecture Designing of LPAN

In this section, we propose a computationally efficient LPAN-L model for the cascaded channel estimation in RIS systems, where the network complexity of LPAN is further optimized from the basic convolution operation and the whole network architecture. We also develop a transfer learning framework to efficiently deal with the domain mismatch problem in the practical deployment of the proposed LPAN-L model. Finally, we present the model parameters and computational complexity analysis for the proposed LPAN and LPAN-L model.

A. Lightweight Convolutional Module

In the original AB, two standard convolutional layers are used to obtain semantic features of the cascaded channel. To reduce the complexity of AB, we proposed a lightweight attention block (LAB) by redesigning the two convolutional layers in the original AB. Fig. 5 shows the designed convolutional module in the LAB, which consists of group convolution, dilation convolution, and point-wise convolution.

1) Group convolution: Let $\mathbf{X} \in \mathbb{R}^{C_1 \times D_1 \times D_2}$ and $\mathbf{O} \in \mathbb{R}^{C_2 \times D_1 \times D_2}$ denote the input feature map and output feature map of convolutional layer, respectively. The $\mathbf{\Omega} \in \mathbb{R}^{C_2 \times C_1 \times k_1 \times k_2}$ denotes the filter set of convolutional layer, where $k_1 \times k_2$ denotes the size of convolutional kernel. In the standard convolutional layer, each element of O_{c_2} is obtained by the convolutional operation between all elements of $\mathbf{X}$ and $\mathbf{\Omega}_{c_2}$ $(1 \le c_2 \le C_2)$, where the number of parameters and FLOPs are $\Upsilon_p = C_1 \times C_2 \times k_1 \times k_2$ and $\Upsilon_f = D_1 \times D_2 \times C_1 \times C_2 \times k_1 \times k_2$, respectively. Consequently, the training of large convolutional networks is difficult for memory limited hardware, e.g., massive mobile terminals. By designing the group strategy based on standard convolution layers, the group convolution is an efficient alternative of dense convolution operations for the CNN architecture. In the group convolution layer, the tensor $\mathbf{X}$ is divided into L grouped feature map $\mathbf{X}_l \in \mathbb{R}^{C_1/L \times D_1 \times D_2}$ along the dimension of channel. Similarly, we also divide the tensor $\mathbf{\Omega}$ into L group subfilters $\mathbf{\Omega}_l \in \mathbb{R}^{C_2/L \times C_1/L \times k_1 \times k_2}$. Then, the convolution operation is carried out between $\mathbf{X}_l$ and $\mathbf{\Omega}_l$. The convolution results is defined as $\mathbf{O}_l = \mathbf{X}_l \otimes \mathbf{\Omega}_l \in \mathbb{R}^{C_2/L \times D_1 \times D_2}$. Lastly, we concatenate all $\mathbf{O}_l$ along the dimension of the feature channel

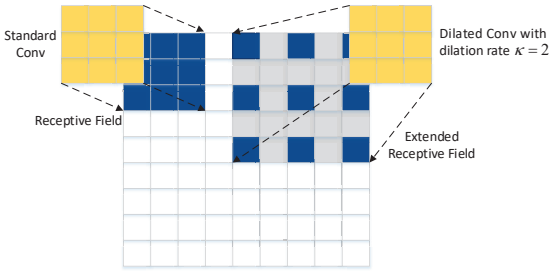


Fig. 6. The receptive field of dilated convolution in LPAN-L.

to obtain the output feature map $\mathbf{O} \in \mathbb{R}^{C_2 \times D_1 \times D_2}$, which is given by

$$\mathbf{O} = \begin{bmatrix} \mathbf{X}_1 \otimes \mathbf{\Omega}_1 \\ \mathbf{X}_2 \otimes \mathbf{\Omega}_2 \\ \vdots \\ \mathbf{X}_L \otimes \mathbf{\Omega}_L \end{bmatrix} = \begin{bmatrix} \mathbf{O}_1 \\ \mathbf{O}_2 \\ \vdots \\ \mathbf{O}_L \end{bmatrix}. \tag{21}$$

In the group convolution, the number of parameters and FLOPs will be reduced with the increase of group number L, which can be expressed as Υ_p/L and Υ_f/L compared with standard convolution, respectively.

2) Dilated convolution: Considering the acceptable complexity of neural network, the CNN typically use convolutional layers with small-size convolutional kernels for feature extraction. With the development global attention mechanism [31], the insufficient ability to extract global information of CNN has been amplified, while large convolution kernel is desired to obtain enough receptive field for CNN architecture. Since the large convolution kernel will introduce more parameters and thereby results in high computation complexity, we use the method of dilated convolution to expand the receptive field without increasing the additional computation complexity. Fig. 6 compares the receptive field of the dilated convolution with the standard convolution, where the solid and shadow areas represent the effective convolutional operations and the receptive field, respectively. Let an integer κ and $k_c = k_1 = k_2$ denote the dilation rate and the size of the original kernel, respectively. The size of dilated convolution kernel is $k_e = k_c + (k_c - 1)(\kappa - 1)$, where the constant zero is filled in the dilated location of the dilated convolution kernel. Further, the receptive field ς_i of the i-th dilated convolutional layer can be expanded as

$$\varsigma_i = \varsigma_{i-1} + (k_e - 1) \prod_{j=1}^{i-1} \text{Stride}_j, \tag{22}$$

where Stride_j denotes the strides of the j-th convolutional layer. Since the convolution kernel k_e is expanded by the zero padding, the dilated convolution can obtain the larger receptive field to capture the long-range dependency of the cascaded channel feature.

3) Point-wise convolution: In the LAB, we first use two group convolutions to replace two standard convolutional layers, which reduces the computations and parameters of the LPAN model. In the first group convolution, the dilated convolution kernel is used to expand the receptive field. Then, we use the point-wise convolution with multiple 1×1 convolution kernels to realize cross channel information interaction of the feature map obtained by group convolution. The parameters and FLOPs of Point-wise convolution is $C_1 \times C_2$ and $D_1 \times D_2 \times C_1 \times C_2$, respectively, which is much less than that in the general convolutional layer.

B. Deep Recursion and Parameter Sharing

Next, we will reduce the network parameters by decoupling the architecture characteristic of LPAN. In the each RM of FEB, we stack J ABs to learn the high-frequency component of the cascaded channel, where the number of network layers of each AB is the same. Consequently, we adopt the architecture of recursive layers to replace the original multi-layers network [37], where a AB carries out $J^{\text{re}}(0 \le J^{\text{re}} \le J)$ recursion operations to substitute for J^{re} ABs. The receptive field of the convolution layer increases once after each recursion, while the number of parameters of the network is fixed with the increase of network depth.

Similarly, the network structure is the same in each RM for the CRB, which realizes the upsampling mapping from the low-dimensional channel matrix to the high-dimensional channel matrix. Since all UBs learn the spatial correlation of the cascaded channel in CRB, the parameter values of the network layer are very close. Consequently, we share the network parameters of the CRB across different pyramid levels, i.e., $S^{\text{sh}}(0 \le S^{\text{sh}} \le S)$ RMs in the LPAN-L. Thus, a single parameter set can construct the multi-level CRB under different upsampling scales.

C. Transfer Learning Framework for Domain Adaption

As a data-driven channel estimation scheme, the performance of DL-based estimator depends on the matched sampling space between the training stage and the test stage. Specifically, the involved datasets in the DL estimator are divided into the source domain data in the training stage and target domain data in the test stage. In the idea case, the data distribution in source domain and target domain is similar. However, in the practical deployment of the DL estimator, the trained model may need to be applied into the new communication environments [38]. Moreover, for RIS systems, the cascaded channel modeling is related to dynamic parameters, e.g., the RIS location, scatterers distribution, and carrier frequency. Hence, we develop an efficient transfer learning framework to realize the cross-domain adaption based on the proposed LPAN-L architecture.

Firstly, the LPAN-L model is trained in the source domain $\mathcal{D}^{\text{s}} = \{\mathcal{F}^{\text{s}}, P(\bar{\mathbf{H}}^{\text{s},\mathcal{P}})\}$ with N^{s} paired samples, in which $\mathcal{F}^{\text{s}}$ denotes the the feature space of the source domain, and $P(\bar{\mathbf{H}}^{\text{s},\mathcal{P}})$ denotes the marginal probability distribution with $\bar{\mathbf{H}}^{\text{s},\mathcal{P}} \in \mathcal{F}^{\text{s}}$. The channel estimation task in $\mathcal{D}^{\text{s}}$ can be expressed as $\wp^{\text{s}} = \{\Omega^{\text{s}}, P(\bar{\mathbf{H}}^{\text{s}}|\bar{\mathbf{H}}^{\text{s},\mathcal{P}})\}$, where Ω^{s} represents the label space of $\mathcal{D}^{\text{s}}$ and $P(\bar{\mathbf{H}}^{\text{s}}|\bar{\mathbf{H}}^{\text{s},\mathcal{P}})$ denotes the posterior probability distribution with $\bar{\mathbf{H}}^{\text{s}} \in \Omega^{\text{s}}$. In fact, the trained model can be regarded to learn the distribution $P(\bar{\mathbf{H}}^{\text{s}}|\bar{\mathbf{H}}^{\text{s},\mathcal{P}})$ based on the source domain data $\mathcal{D}^{\text{s}}$. In the test stage of the LPAN-L model, we defined the target domain as $\mathcal{D}^{\text{t}} = \{\mathcal{F}^{\text{t}}, P(\bar{\mathbf{H}}^{\text{t},\mathcal{P}})\}$ composed

Algorithm 1 Transfer Learning: Selective Fine-tuning

Initialization:
  $i^{\mathrm{s}} = 0$, $i^{\mathrm{t}} = 0$,
  $f = \{f_1, \cdots, f_s, \cdots, f_S\}$ with random weights
Pre-training in Source Domain $\mathcal{D}^{\mathrm{s}}$:
  Construct source domain task $\wp^{\mathrm{s}}$ with N^{s} samples
  while $i^{\mathrm{s}} \leq E^{\mathrm{s}}$ **do**
    update f with the gradient of loss function $\mathcal{L}(\bar{\mathbf{H}}^{\mathrm{s}}, \widehat{\mathbf{H}}^{\mathrm{s}})$
    $i^{\mathrm{s}} = i^{\mathrm{s}} + 1$
  end while
Transfer Learning in Target Domain $\mathcal{D}^{\mathrm{t}}$:
  Construct target domain task $\wp^{\mathrm{t}}$ with N^{t} samples
  freeze S^{f} RMs parameters $f^{\mathrm{f}} = \{f_1, \cdots, f_{S^{\mathrm{f}}}\}$
  while $i^{\mathrm{t}} \leq E^{\mathrm{t}}$ **do**
    only update $f^{\mathrm{t}} = \{f_{S-S^{\mathrm{f}}}, \cdots, f_S\}$ with the gradient of loss function $\mathcal{L}(\bar{\mathbf{H}}^{\mathrm{t}}, \widehat{\mathbf{H}}^{\mathrm{t}})$
    $i^{\mathrm{t}} = i^{\mathrm{t}} + 1$
  end while
Online Estimation in Target Domain:
  $\widehat{\mathbf{H}} = (\widehat{\mathbf{H}}_1, \widehat{\mathbf{H}}_2, \cdots, \widehat{\mathbf{H}}_S) = f(\bar{\mathbf{H}}^{\mathcal{P}})$

of $N^{\mathrm{t}} \ll N^{\mathrm{s}}$ samples, in which $\mathcal{F}^{\mathrm{t}}$ denotes the the feature space of the source domain, and $P(\bar{\mathbf{H}}^{\mathrm{t},\mathcal{P}})$ denotes the marginal probability distribution. Accordingly, the channel estimation task in $\mathcal{D}^{\mathrm{t}}$ can be denoted as $\wp^{\mathrm{t}} = \{\Omega^{\mathrm{t}}, P(\bar{\mathbf{H}}^{\mathrm{t}}|\bar{\mathbf{H}}^{\mathrm{t},\mathcal{P}})\}$, where Ω^{t} represents the label space of $\mathcal{D}^{\mathrm{t}}$ and $P(\bar{\mathbf{H}}^{\mathrm{t}}|\bar{\mathbf{H}}^{\mathrm{t},\mathcal{P}})$ denotes the posterior probability distribution. According to the inductive and homogeneous transfer learning theory, the feature space in the source domain and target domain satisfies $\{F^{\mathrm{s}}, F^{\mathrm{t}}\} \in \mathcal{F}$, while the probability distributions present $P(\bar{\mathbf{H}}^{\mathrm{s},\mathcal{P}}) \neq P(\bar{\mathbf{H}}^{\mathrm{t},\mathcal{P}})$ and $P(\bar{\mathbf{H}}^{\mathrm{s}}|\bar{\mathbf{H}}^{\mathrm{s},\mathcal{P}}) \neq P(\bar{\mathbf{H}}^{\mathrm{t}}|\bar{\mathbf{H}}^{\mathrm{t},\mathcal{P}})$ due to different communication environments.

Although the difference of the source domain and target domain restricts the online deployment of the trained model, the learned knowledge in the source domain can be transfered into the target domain by utilizing the transfer learning framework. In this work, we exploit a selective fine-tuning strategy-based transfer learning model, which leverages the multi-scale hierarchical characteristics of the proposed LPAN-L architecture. Suppose the proposed LPAN-L model is defined as f, in which the s-th RM in LPAN-L is denoted as f_s, i.e., $f = \{f_1, \cdots, f_s, \cdots, f_S\}$. For the trained model in the source domain, we selectively freeze the network parameters of $S^{\mathrm{f}}(S^{\mathrm{f}} < S)$ RMs in the LPAN-L model, e.g., from the 1-st RM to the S^{f}-th RM. Then, we fine-tune the network parameters of $S - S^{\mathrm{f}}$ RMs in the LPAN-L model by utilizing the limited target domain samples. The specific fine-tuning procedures of the transfer learning are provided in **Algorithm 1**, where the pre-training epochs E^{s} are larger than the fine-tuning epochs E^{t}. Compared with the pre-training stage, the training cost in the fine-tuning stage can be reduced because only partial parameters of the LPAN-L model need to be updated. By utilizing the proposed transfer learning framework, we only use limited fine-tuning samples to realize the domain adaption for the proposed LPAN-L model, which avoids the re-training process with a large number of target domain samples.

D. Parameters and Computational Complexity Analysis

Suppose the number of the convolutional filters is w and the filter size is $k_c \times k_c$ for two convolutional layers in the AB, the parameters and FLOPs of a AB are $w^2(2k_c^2 + 1)$ and $2w^2(k_c^2 D_1 D_2 + 1)$, where D_1 and D_2 denote the height and width of feature map in the AB. Hence, the time complexity of FEB, CRB, and TP in the s-th RM is $O\left(2^s PMw^2 {k_c}^2 (J+1)\right)$, $O\left(2^{s+2} PMk_c{}^2\right)$ and $O\left(2^{s+1} PMw{k_c}^2\right)$. For the quasi-static channel estimation, the total time complexity of LPAN is $O\left(\sum_{s=1}^{S} 2^s PMk_c{}^2(w^2(J+1) + 2w + 4)\right)$, while the time complexity in the time-varying channel estimation is $O\left(\sum_{s=1}^{S} 2^s PMk_c{}^2(w^2(J+1) + Bw + B^2)\right)$. The space complexity of the quasi-static and time-varying channel estimation are $O\left(\sum_{s=1}^{S} k_c{}^2(w^2(2J+1) + 2w + 4)\right)$ and $O\left(\sum_{s=1}^{S} k_c{}^2(w^2(2J+1) + Bw + B^2)\right)$, respectively.

In the LPAN-L architecture, the parameters and FLOPs of a LAB are $w^2 k_c{}^2(1/g_1 + 1/g_2 + 1)$ and $w^2 k_c^2 D_1 D_2(1/g_1 + 1/g_2 + 2)$, where g_1 and g_2 denote the group number of the first and second group convolution layer in the LAB, respectively. Let J_s^{lw} and J_s^{re}, $0 \leq J_s^{\mathrm{lw}}, J_s^{\mathrm{re}} \leq J$, denote the number of LABs and recursion operations in the s-th RM, respectively. For the case of the quasi-static channel estimation, the time complexity of FEB in the s-th RM is reduced to $O\left(2^{s-1} PMw^2 k_c{}^2(2(J - J_s^{\mathrm{lw}}) + (1/g_1 + 1/g_2)J_s^{\mathrm{lw}} + 2)\right)$ in LPAN-L. For the time complexity of CRB and TP in the s-th RM, the LPAN-L model is the same with LPAN. The space complexities of FEB and CRB are reduced to $O\left(\sum_{s=1}^{S} w^2 k_c{}^2(2(J - J_s^{\mathrm{re}}) + (1/g_1 + 1/g_2)J_s^{\mathrm{re}})\right)$ and $O\left((S - S^{\mathrm{sh}})4k_c{}^2\right)$, respectively. The similar complexity reduction can be obtained for the time-varying channel estimation. Hence, both parameters and computation complexity of the proposed LPAN-L model are efficiently reduced compared to the LPAN model.

V. Numerical Results

In this section, we first present the simulation setting, including the system parameters of the RIS-aided mmWave communications and hyper-parameters adopted for the network training. Then, we provide the numerical results to verify the channel estimation performance of the proposed LPAN in terms of estimation accuracy, convergence speed, and robustness.

A. Simulation Setup

In the simulation, we set $M = 8 \times 8$, $N = 16 \times 16$ and $K = 6$ for the RIS-aided multi-user massive MIMO communication system. The mmWave communication frequency is set to $f_c =$ 28 GHz and the parameter λ_p is set to $\lambda_p = 1.8$ [39]. The angular spread is set to $\sigma = \sigma_\varphi = \sigma_\phi = 5°$. The Poisson distribution and uniform distribution are used to model the cluster $C \sim \max\{P(\lambda_p), 1\}$ and scatters of each cluster $S_c \sim \mathcal{U}[1, U]$, respectively. Unless otherwise specified, we set $U =$ 30, $n_0 = 3.19$, $b_0 = 0.06$, $\sigma_x = 8.29$ dB, and $f_0 = 24.2$ GHz for NLOS path in the path loss model, while $n_0 = 1.73$, $b_0 =$

TABLE I
THE HYPER-PARAMETERS FOR THE BASELINE LPAN

Hyper-Parameter	Value
The number of RM S	3
The number of AB J in each RM	4
The number of LAB $J_s^{\mathrm{lw}}, (1 \le s \le S-1)$	2
The number of recursion $J_s^{\mathrm{re}}, (1 \le s \le S-1)$	2
The number of LAB J_S^{lw} in the S-th RM	4
The number of recursion J_S^{re} in the S-th RM	4
The number of the first group g_1 in LAB	16
The number of the second group g_2 in LAB	4
The number of shared UB S^{sh} in CRB	2
The number of filters w in each LAB/AB	96
The size of standard kernel (k_c, k_c)	$(3,3)$
The dilation rate of kernel κ	2
The total epochs E^{s}	100
The initial learning rate η_0	1×10^{-3}
The end of learning rate η_1	5×10^{-6}
The regularization parameter ε	10^{-4}
The training batchsize $\mathcal{B}$	64

0.06, and $\sigma_x = 3.02$ dB for LOS path [23]. In the source domain scenario, the three-dimensional coordinates of BS and RIS are set to $(x^{\mathrm{BS}}, y^{\mathrm{BS}}, z^{\mathrm{BS}}) = (0, 25, 2)$, $(x^{\mathrm{RIS}}, y^{\mathrm{RIS}}, z^{\mathrm{RIS}}) = (40, 50, 2)$, respectively. To mitigate the serve multiplicative fading effect for cascaded reflection link in RIS systems, the coordinates of UE are randomly distributed in the 1 m height with a horizontal radius of 8 m centered on RIS. According to the 3GPP LTE-A standard [29], we set the EVM $\rho = \rho_t = \rho_r = 0.1$ and the error of ON/OFF mode $\epsilon = \epsilon_1 = \epsilon_0 = 0.01$ [27]. We define $r = \frac{P}{N} = \frac{1}{k}$ as the ratio of the number of the activated RIS elements to the total elements, where P pilots are used for the LS pre-estimation.

In the pre-training dataset construction, we generate $N_k = 5\times10^3$ paired samples for each user to construct the dataset, i.e., total samples of $N_{\mathrm{s}} = KN_k = 3\times10^4$. The range of training SNR is set to [0, 20] dB with the interval of 5 dB to generate the received pilot signal $\mathbf{Y}$. In the training process, we adopt the cosine learning rate decay schedule to avoid converge directly to a poor local minimum point, where the learning rate η_i at the i-th training epoch is given by

$$\eta_i = \eta_0 + \frac{1}{2}(\eta_1 - \eta_0)\left(1 + \cos\left(\frac{i}{E^{\mathrm{s}}}\pi\right)\right), 0 \le i \le E^{\mathrm{s}}, \quad (23)$$

where η_0, η_1 and E^{s} represent initial learning rate, final learning rate and the total number of epochs, respectively. Table I shows the detailed training hyper-parameters of the baseline LPAN.

B. Performance Comparison for Different Estimation Schemes

In Fig. 7, we compare the NMSE performance of the proposed LPAN with the traditional estimators, i.e., binary reflection protocol-based LS estimator [27], PARAllel FACtor decomposition-based alternating LS (ALS) estimator [40],

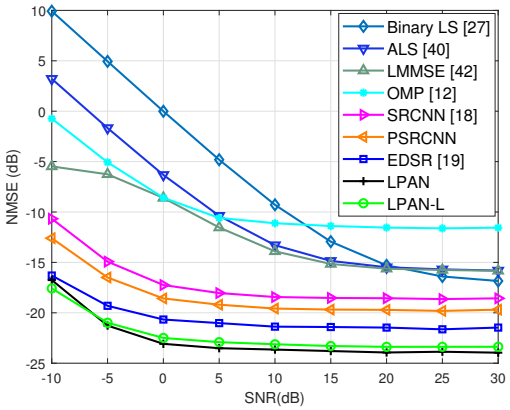


Fig. 7. NMSE performance for different channel estimation schemes.

[41], empirical linear minimum mean square error (LMMSE) estimator [15], [42], and OMP algorithms [12], as well as other SR networks, i.e., SRCNN [18] and EDSR [19]. In the traditional algorithms, the required pilot overhead is set to $Q_{\mathrm{LS}} = Q_{\mathrm{LMMSE}} = N$ and $Q_{\mathrm{OMP}} = N/2$, respectively, while the required pilot overhead is $Q_{\mathrm{DL}} = P = N/2^S = N/8$ for DL-based channel estimation networks. In the SRCNN and EDSR, the UB is designed at the input and output layers of the network, respectively. To demonstrate the impact of upsampling strategy on channel estimation, we modify the single-step up-sampling in the SRCNN to the asymptotic sampling with 2 times factor, which is termed as PSRCNN in Fig. 7. For the fair comparison of different networks, the number of filters and the depth of network layers are set to the same values for EDSR and LPAN.

As the classic linear estimator, the estimation accuracy of LS and LMMSE algorithm is non-ideal for unacceptable noise and HWIs. Note that the required second order statistics of the LMMSE estimator are replaced by the Monte Carlo-based empirical correlation matrix with training samples in Fig. 7. In the clustered statistical MIMO channel modeling of RIS systems, the sparsity of the cascaded channel is variable and relatively large due to the extensive scatters, which limit the estimation performance of OMP. The SR-based channel estimation is related to the method and location of the upsampling. In the SRCNN, the single-step upsampling, i.e., $\bar{\mathbf{H}}^{\mathcal{P}} \in \mathbb{R}^{M\times P\times 2} \rightarrow \bar{\mathbf{H}}_S \in \mathbb{R}^{M\times N\times 2}$, will introduce serve interpolation errors in the input layer, which results in limited recovery effect of the subsequent network, and the high-dimensional input also increases the computational complexity of the network. The PSRCNN is an improved model from the SRCNN, which progressively upscaling the low-dimensional input tensor to the complete dimension of $M\times N\times 2$ with 2 times upsampling factor, and thereby reduces the interpolation error of input data. In the EDSR, a large number of residual blocks are stacked before upsampling, and then the extracted efficient features are used to the final reconstruction. This post-upsampling architecture can reduce the computational complexity and improve the reconstruction performance. However, the post-upsampling layer is difficult

TABLE II
TRAINING OVERHEAD FOR DIFFERENT NETWORKS

	LPAN	LPAN-L	LPAN-L (small)	LPAN-L (medium)	LPAN-L (large)	EDSR
Model size (MB)	**18.1**	**9.33**	3.2	4.9	11.4	9.46
Parameters (M)	**2.37**	**1.09**	0.483	0.659	1.45	2.25
FLOPs (G)	**11.953**	**6.06**	3.72	4.46	8.53	6.49
Average NMSE (dB)	**-22.7**	**-22.3**	-20.1	-20.9	-22.4	-19.6

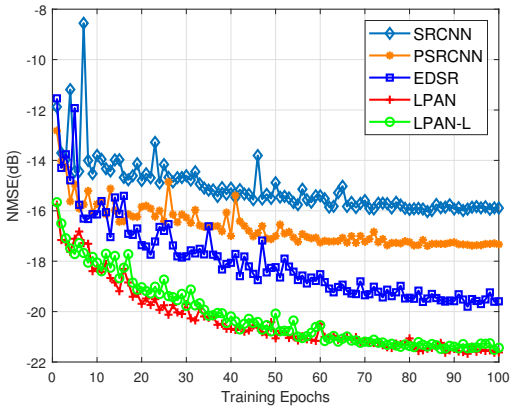


Fig. 8. Convergence performance with training epochs E of DL estimators.

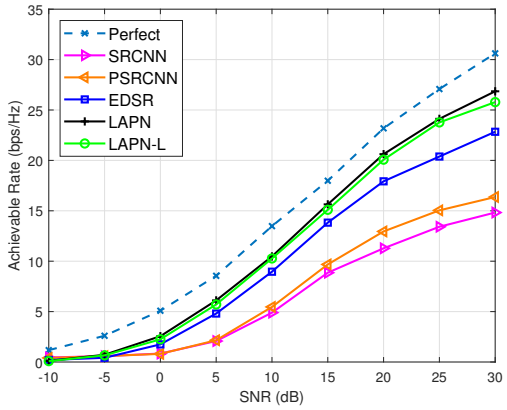


Fig. 9. Achievable sum-rate performance of DL estimators.

to recover the high-resolution cascaded channel matrix directly when the upsampling factor is large. In the proposed LPAN with superior channel estimation performance, we further optimize the location of up-sampling by introducing the multi-scale supervision, where the UB is embedded in the network from low dimension to high dimension to realize progressive reconstruction of the cascaded channel matrix. Furthermore, the lightweight LPAN-L architecture is developed to reduce the network complexity with sight performance loss.

Table II compares the required memory, parameters and FLOPs for different channel estimation schemes, in which we provide the NMSE performance of different LPAN-L variants with different scales of network parameter, i.e., the small-size, medium-size and large-size LPAN-L by controlling hyper-parameters g_1, g_2, J_s^{lw} and J_s^{re} in Table I. In the original LPAN architecture, we design the CRB to obtain the low-frequency component of cascaded channel, which introduces more network parameters compared with EDSR. In addition, the progressive upsampling operation in each RM increases the computational complexity because the size of feature map is enlarged. Compared with the proposed LPAN model, the LPAN-L adopts the group convolution operation and the parameter sharing strategy to reduce half of the parameters and the computation complexity, while providing a close performance to LPAN. We observe that the performance gap between the LPAN and the LPAN-L will be progressively reduced by increasing the network size of the LPAN-L. Moreover, the small-size LPAN-L model is still superior the EDSR model in terms of channel estimation accuracy and network complexity.

In Fig. 8, we show the convergence speed of different channel estimation models, where we use the average NMSE of validation set as the performance evaluation metric. Compared with the existing schemes, the convergence of the proposed schemes is more stable and fast with the increase of training epochs E. Based on the cascaded channel matrix estimated by different estimation schemes, we further compares the achievable sum-rate performance of different DL estimators in Fig. 9. Suppose $\mathbf{v}_k \in \mathbb{C}^{M\times 1}$ is the normalized precoding vector at the BS for k-th UE, the signal-to-interference-plus-distortion-noise-ratio for the UE$_k$ can be expressed as

$$\gamma_k = \frac{p_k\left|\mathbf{v}_k^T\mathbf{H}_k\boldsymbol{\theta}\right|^2}{\bar{\rho}p_k\left|\mathbf{v}_k^T\mathbf{H}_k\boldsymbol{\theta}\right|^2 + p_k\sum\limits_{i=1,i\neq k}^{K}(1+\bar{\rho})\left|\mathbf{v}_i^T\mathbf{H}_k\boldsymbol{\theta}\right|^2 + \delta_n^2}, \quad (24)$$

where $\bar{\rho} = \rho_t^2 + \rho_r^2$. Furthermore, the achievable sum-rate of RIS systems can be calculated by $R = \sum_{k=1}^{K}\log_2(1+\gamma_k)$.

Following the work of [43], we adopt the cross-entropy optimization method to determine the precoding matrix $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_K]$ at the BS and reflecting vector $\boldsymbol{\theta}$ at the RIS, where we set $b = 2$ bits discrete reflection phase shift considering the hardware constraint, i.e., $\theta_i \in \{+1, -1, +1j, -1j\}$. Fig. 9 shows the achievable sum-rate of RIS-aided communication system by utilizing the estimated cascaded channel of different models. The LPAN-based channel estimation scheme can achieve better achievable sum-rate compared with other estimation schemes, while the achievable sum-rate performance of both LPAN and LPAN-L is very close.

C. Robustness Analysis for the Proposed LPAN-L Model

In Fig. 10, we presents the NMSE performance of the

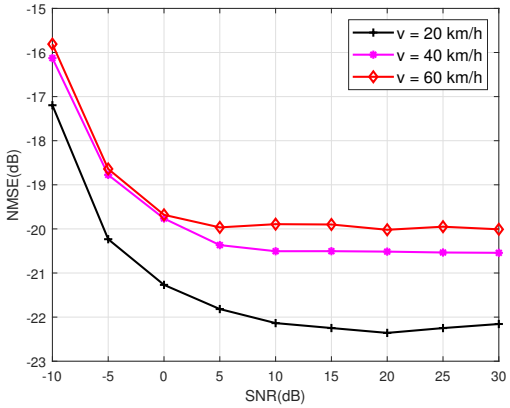


Fig. 10. NMSE performance of LPAN-L for different mobility speeds v.

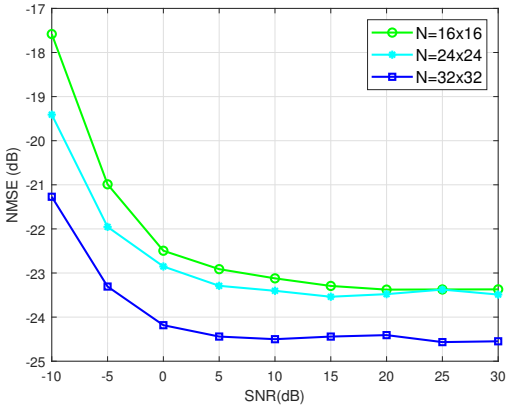


Fig. 12. NMSE performance of LPAN-L under different number of refection elements N.

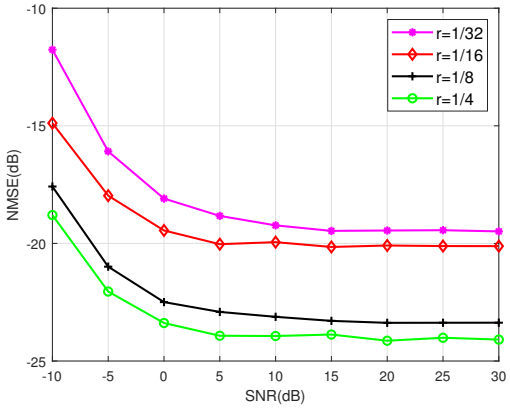


Fig. 11. NMSE performance of LPAN-L under different pilot overhead r.

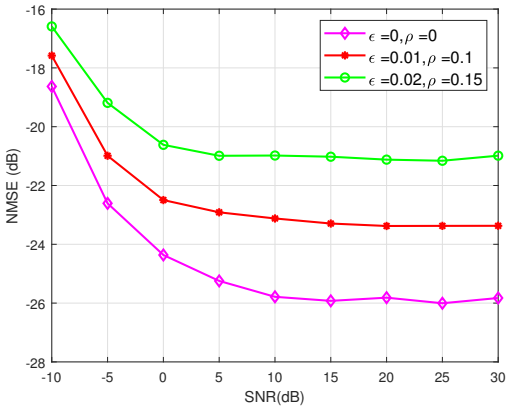


Fig. 13. NMSE performance of LPAN-L under different HWIs (ϵ, ρ).

proposed LPAN-L under different mobility speeds v, in which the number of pilot block is set to $B^{\mathrm{p}} = 2$ within a frame with $B = 6$ blocks, and the sampling period T_b of each time block is fixed as $T_b = 1/(4f^d_{\max}) \approx 0.24$ ms. With the increase of v, the coherence time T_c will be shorter and the channel variation of consecutive time blocks within a frame will be faster. Hence, the channel estimation accuracy of the proposed LPAN-L will be sightly decreased, while a stable NMSE performance can be obtained even for the high-speed scenario with $v = 60$ km/h. Note that we pretrain the LPAN-L with $N_T = 3 * 10^4$ samples for the case of $v = 60$ km/h in the network training, while the transfer learning is used to fine-tune the LPAN-L model with limited samples for the cases of $v = 20$ and $v = 40$ km/h.

Fig. 11 shows the NMSE performance of the proposed LPAN-L under different pilot overhead ratios r. The baseline LPAN-L model in simulation composes of 3 RMs, each of which realize 2 times upsampling based on the input tensor. The pilot overhead for baseline LPAN-L is $P = \frac{1}{2^3}N = \frac{1}{8}N$. To reduce the training overhead of LPAN-L under different pilot lengths, we use the pretrained model of the baseline LPAN-L to initialize the network weights. Specifically, if $r < \frac{1}{8}$, we only increase the UBs in the last RM based on baseline LPAN-L, e.g., adding 1 UB when $r = \frac{1}{16}$, while the network weights of $s(1 \le s \le S-1)$ th RM are initialized by baseline LPAN. Conversely, we delete partial RMs for larger r, e.g., deleting 1 UB when $r = \frac{1}{4}$. By leveraging the pretrained model, we only use half of the sample size and training epochs for other pilot length, i.e., $r = \frac{1}{16}$ or $\frac{1}{4}$. With the decrease of r, the required upsampling dimension will be larger, so the high-dimensional channel reconstruction becomes more challenging. Nevertheless, LPAN can achieve satisfactory channel estimation accuracy even with small pilot overhead, e.g., $P = Nr = 8$.

Fig. 12 presents the NMSE performance of the proposed LPAN-L under different number of RIS elements N. For the DFT protocol-based LS estimator, the NMSE performance of channel estimation can be improved with the increase of N [15]. In this case, the more accurate pre-estimated input tensor can be obtained for the DL model. Hence, the channel estimation accuracy of the LPAN-L model is also improved. Thanks to the multi-scale pyramid architecture of the LPAN-L model, the same LPAN-L architecture is compatible with the RIS with different sizes. Note that the network complexity will be increased for the larger-size RIS because the operating dimension of the feature map in LPAN-L is boosted.

In Fig. 13, we study the NMSE performance of the proposed LPAN-L for different degrees of HWIs, where the LPAN-L is trained under the fixed HWIs sets $(\epsilon, \rho) = (0.01, 0.1)$.

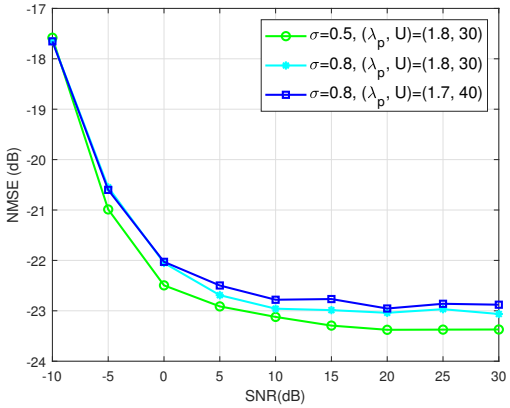


Fig. 14. NMSE performance of LPAN-L under different angular spreads σ and scattering distribution (λ_p, U).

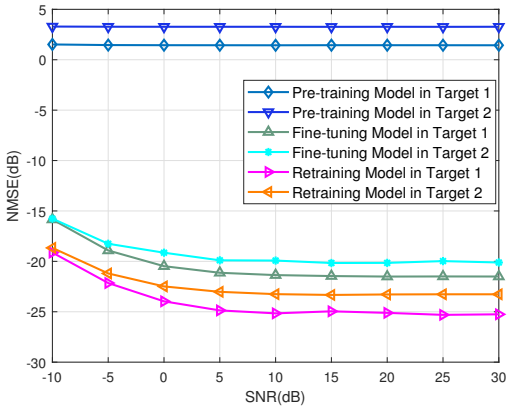


Fig. 15. NMSE performance of the transfer learning-based LPAN-L model.

With the increase of levels of HWIs (ϵ, ρ), the estimation accuracy of the LS pre-estimation will be decreased, which introduces more estimation error into the input tensor of the DL model and results in the performance degradation of the LPAN-L model. However, since the neural network is robust for a certain degree of disturbance of input data, even under severe HWIs, i.e., $(\epsilon, \rho) = (0.02, 0.15)$, the LPAN-L can still achieve satisfactory performance.

Fig. 14 shows the NMSE performance of the proposed LPAN-L under different scattering distributions, in which the LPAN-L model is trained with the given scattering distribution, i.e., angular spread $\sigma = 0.5$ and the scattering distribution $(\lambda_p, U) = (1.8, 30)$ in the training stage. We observe that the trained LPAN-L model can work well under different scattering distributions in the test stage, in which the LPAN-L model has better robustness for the scattering parameters (λ_p, U). This stable performance benefits from the various channel samples generation in the training dataset construction, in which we adopt the dynamic cascaded channel modeling with randomly distributed scatters.

D. Domain Adaption Performance for Transfer Learning

In Fig. 15, we provide the transfer learning performance of the LPAN model for two cases of the spatial-varying target domain. For the case of target domain 1, the RIS is placed at the x-z plane (the opposite-wall for users) instead of the y-z plane (the side-wall for users) in the source domain. On the basis of target domain 1, the scenario of target domain 2 further considers the case of different cells, in which the carrier frequencies of neighbor cells are different to avoid the inter-cell interference, and the path loss model in equation (2) has different system parameters. Due to the difference of data distribution, the pre-trained model in the source domain can not be directly applied to the channel estimation in the target domain. By utilizing the proposed selective fine-tuning strategy, the fine-tuning model can obtain stable NMSE performance in the source domain. We also provide a completely retrained LPAN model in the target domain as the performance lower bound. The required training samples and epochs are $N^{\mathrm{r}} = N_k \times K$ and $E^{\mathrm{r}} = E^{\mathrm{s}}$ for the retraining scheme. However, in Fig. 15, the proposed transfer learning framework only needs $N^{\mathrm{t}} = N^{\mathrm{r}}/10$ samples and $E^{\mathrm{t}} = E^{\mathrm{r}}/5$ epochs.

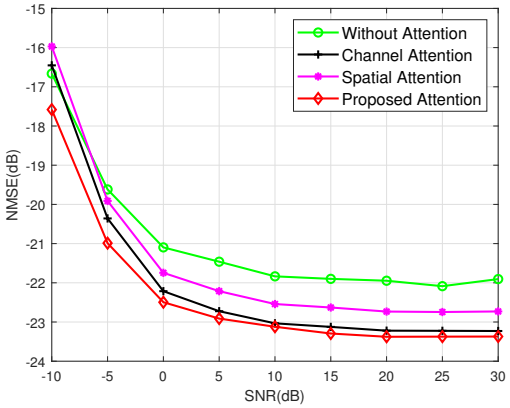


Fig. 16. NMSE performance for different attention modules in LPAN-L.

E. Ablation Experiment for the Proposed Attention Block

Fig. 16 shows the ablation experiment to verify the effectiveness of the proposed AB, in which we provide three benchmarks for the attention mechanism variants. Specifically, the convolution-based residual module without attention mechanism is used as the basic benchmark [17], [32]. We also compare the existing channel attention and spatial channel modules [21], [44], which generates the channel attention weight $z^{\mathrm{c}} \in \mathbb{R}^{C\times 1\times 1}$ from the feature channel dimension and the spatial attention weight $z^{\mathrm{s}} \in \mathbb{R}^{1\times D_1\times D_2}$ from the spatial dimension of the feature map, respectively. We observe that the channel estimation accuracy can be improved by introducing the attention mechanism into the FEB of the LPAN-L model. Compared with the excitation operation of the existing attention modules, the proposed AB can retain the direct correspondence between channel of feature map and attention weight by adding the single FC layer. Furthermore, the Tanh activation can restrict the attention weight to a more reasonable range. Consequently, the proposed AB can achieve better NMSE performance with more simple architecture.

VI. Conclusions

In this paper, we have proposed a progressive cascaded channel reconstruction strategy by utilizing the multi-scale supervised learning for RIS-aided multi-user mmWave communication systems. In contrast to the one-step reconstruction used in previous works, we have designed the pyramid network to implement channel extrapolation hierarchically. The proposed LPAN with dual branch architecture separately extract the high frequency and low frequency information of the cascaded channel matrix, and then the residual learning with attention mechanism is used to realize information fusion. Moreover, we have designed the efficient convolution operation and parameter sharing strategy to construct the lightweight LPAN-L model. Numerical results show that the proposed LPAN and LPAN-L with limited pilot overhead is superior to existing channel estimation schemes, and have good robustness for different system setups. The developed transfer learning framework provides a domain adaptive solution for the practical deployment of the proposed channel estimation model. In the future works, we will extend the multi-scale pyramid architecture to higher-dimensional channel estimation scenarios, e.g., cooperative communications of multi-hop RISs [4] and Holographic intelligent surfaces [45].

References

[1] J. Xiao, J. Wang, W. Xie, X. Wang, C. Wang and H. Xu, "Multi-scale supervised learning-based channel estimation for RIS-aided communication systems," in *Proc. IEEE Wireless Commun. Netw. Conf. (IEEE WCNC'23)*, 2023, pp. 1-6.

[2] S. A. Busari, K. M. S. Huq, S. Mumtaz, L. Dai, and J. Rodriguez, "Millimeter-wave massive MIMO Communication for future wireless systems: A survey," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 2, pp. 836-869, Secondquarter 2018.

[3] Y. Liu, X. Liu, X. Mu., T. Hou, J. Xu, M. Di Renzo, and N. Al-Dhahir, "Reconfigurable intelligent surfaces: Principles and opportunities," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 3, pp. 1546-1577, thirdquarter 2021.

[4] C. Huang, Z. Yang, G. C. Alexandropoulos, K. Xiong, L. Wei, C. Yuen, Z. Zhang and M. Debbah, "Multi-hop RIS-empowered terahertz communications: a DRL-based hybrid beamforming design",*IEEE J. Sel. Areas Commun.*, vol. 39, no. 6, pp. 1663-1677, Jun. 2021.

[5] 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, Secondquarter 2022.

[6] E. Björnson, Ö. Özdogan, and E. G. Larsson, "Intelligent reflecting surface versus decode-and-forward: How large surfaces are needed to beat relaying?," *IEEE Wireless Commun. Lett.*, vol. 9, no. 2, pp. 244-248, Feb. 2020.

[7] Z. Xing, R. Wang, J. Wu, and E. Liu, "Achievable rate analysis and phase shift optimization on intelligent reflecting surface with hardware impairments," *IEEE Trans. Wireless Commun.*, vol. 20, no. 9, pp. 5514-5530, Sept. 2021.

[8] X. Chen, J. Shi, Z. Yang, and L. Wu, "Low-complexity channel estimation for intelligent reflecting surface-enhanced massive MIMO," *IEEE Wireless Commun. Lett.*, vol. 10, no. 5, pp. 996-1000, May 2021.

[9] M. Jian and Y. Zhao, "A modified off-grid SBL channel estimation and transmission strategy for RIS-assisted wireless communication systems," in *Proc. Int. Wireless Commun. and Mobile Comput. (IWCMC)*, 2020, pp. 1848-1853.

[10] Y. Lin, S. Jin, M. Matthaiou, and X. You, "Tensor-based algebraic channel estimation for hybrid IRS-assisted MIMO-OFDM," *IEEE Trans. Wireless Commun.,* vol. 20, no. 6, pp. 3770-3784, June 2021.

[11] 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.

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

[13] G. Zhou, C. Pan, H. Ren, P. Popovski and A. L. Swindlehurst, "Channel estimation for RIS-aided multiuser millimeter-wave systems," *IEEE Trans. Signal Process.*, vol. 70, pp. 1478-1492, 2022.

[14] E. Balevi, A. Doshi, A. Jalal, A. Dimakis, and J. G. Andrews, "High dimensional channel estimation using deep generative networks," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 1, pp. 18–30, Jan. 2021.

[15] 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 Trans. Wireless Commun.,* vol. 21, no. 2, pp. 898-912, Feb. 2022.

[16] M. Soltani, V. Pourahmadi, A. Mirzaei, and H. Sheikhzadeh, "Deep learning-based channel estimation," *IEEE Commun. Lett.,* vol. 23, no. 4, pp. 652-655, April 2019.

[17] L. Li, H. Chen, H. -H. Chang, and L. Liu, "Deep residual learning meets OFDM channel estimation," *IEEE Wireless Commun. Lett.,* vol. 9, no. 5, pp. 615-618, May 2020.

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

[19] Y. Jin, J. Zhang, X. Zhang, H. Xiao, and B. Ai, D. W. K. Ng, "Channel estimation for semi-passive reconfigurable intelligent surfaces with enhanced deep residual networks," *IEEE Trans. Veh. Technol.,* vol. 69, no. 8, pp. 9223–9228, Jun. 2020.

[20] J. Gao, M. Hu, C. Zhong, G. Y. Li, and Z. Zhang, "An attention-aided deep learning framework for massive MIMO channel estimation," *IEEE Trans. Wireless Commun.*, vol. 21, no. 3, pp. 1823-1835, March 2022.

[21] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-andexcitation networks," Sep. 2017.

[22] "3GPP TR 38.901 V16.1.0 - Study on channel model for frequencies from 0.5 to 100 GHz," Dec. 2019.

[23] 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. Commun.,* vol. 69, no. 12, pp. 8600-8611, Dec. 2021.

[24] P. Nayeri, F. Yang, and A. Z. Elsherbeni, Reflectarray antennas: theory, designs, and applications. *USA: Wiley,* 2018.

[25] "5G channel model for bands up to 100 GHz," [Online]. Available: http://www.5gworkshops.com/5GCMSIG White.

[26] O. E. Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. W. Heath, "Spatially sparse precoding in millimeter wave MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 13, no. 3, pp. 1499–1513, Mar. 2014.

[27] D. Mishra and H. Johansson, "Channel estimation and low-complexity beamforming design for passive intelligent surface assisted MISO wireless energy transfer," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP).*, May 2019.

[28] M.-A. Badiu and J. P. Coon, "Communication through a large reflecting surface with phase errors," *IEEE Wireless Commun. Lett.*, vol. 9, no. 2, pp. 184-188, 2019.

[29] H. Holma and A. Toskala, "LTE for UMTS: Evolution to LTE-Advanced, 2nd ed," *IEEE Commun. Mag.*, Jan. 2011.

[30] B. Zheng and R. Zhang, "Intelligent reflecting surface-enhanced OFDM: Channel estimation and reflection optimization," IEEE Wireless Commun. Lett.,, vol. 9, no. 4, pp. 518-522, Apr. 2020.

[31] G. Brauwers and F. Frasincar, "A general survey on attention mechanisms in deep learning," *IEEE Trans. Knowl. Data Eng.*, 2021. doi: 10.1109/TKDE.2021.3126456.

[32] K He, X Zhang, S Ren, and J Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. on Comput. Vis. and Pattern Recognit (CVPR).*, 2016, pp. 770-778.

[33] Q. Wang, B. Wu, P Zhu, P, Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks", in *Proc. IEEE Conf. on Comput. Vis. and Pattern Recognit. (CVPR)*, 2020.

[34] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep laplacian pyramid networks for fast and accurate super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit(CVPR).,* 2017, pp. 624–632.

[35] T. Salimans and D. P. Kingma. "Weight normalization: A simple reparameterization to accelerate training of deep neural networks," in *Proc. Neural Informat. Process. Syst. (NIPS)*, 2016, pp. 901–909.

[36] A. Odena, V. Dumoulin, and C. Olah. "Deconvolution and checkerboard artifacts," *Distill*, 2016.

[37] J. Kim, J. K. Lee, and K. M. Lee, "Deeply-recursive convolutional network for image super-resolution," in *Proc. IEEE Conf. on Comput. Vis. and Pattern Recognit. (CVPR)*, 2016, pp. 1637-1645.

[38] W. Alves, I. Correa, N. González-Prelcic and A. Klautau, "Deep transfer learning for site-specific channel estimation in low-resolution mmWave MIMO," *IEEE Wireless Commun. Lett.*, vol. 10, no. 7, pp. 1424-1428, July 2021.

[39] I. A. Hemadeh, K. Satyanarayana, M. El-Hajjar, and L. Hanzo,"Millimeter-wave communications: Physical channel models, design considerations, antenna constructions, and link-budget," *IEEE Commun.Surveys Tuts.*, vol. 20, no. 2, pp. 870-913, 2nd Quart. 2018.

[40] L. Wei, C. Huang, G. C. Alexandropoulos, C. Yuen, Z. Zhang and M. Debbah "Channel estimation for RIS-empowered multi-user MISO wireless communications", *IEEE Trans. Commun.*, vol. 69, no. 6, pp. 4144-4157, June 2021.

[41] 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 J. Sel. Top. Signal Process.*, vol. 15, no. 3, pp. 789-802, Apr. 2021.

[42] 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, Mar. 2021.

[43] C. Hu, L. Dai, S. Han, and X. Wang, "Two-timescale channel estimation for reconfigurable intelligent surface aided wireless communications," *IEEE Trans. Commun.*, vol. 69, no. 11, pp. 7736-7747, Nov. 2021.

[44] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11211 LNCS, pp. 3–19, 2018.

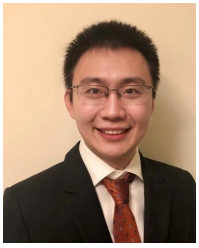
[45] C. Huang, S. Hu, G. C. Alexandropoulos, A. Zappone, C. Yuen, R. Zhang, M. D. Renzo, M. Debbah, "Holographic MIMO surfaces for 6G wireless networks: opportunities, challenges, and trends", *IEEE Wireless Commun.*, vol. 27, no. 5, pp. 118-125, Oct. 2020.

Jian Xiao received the B.Eng. degree and the M.Sc. degree from the Hunan Institute of Science and Technology, Yueyang, China, in 2019 and 2022, respectively. He is currently working toward the Ph.D. degree with Central China Normal University. His research interests include reconfigurable intelligent surface and machine learning.

Ji Wang received the B.S. degree from the School of Electronic Information and Communications, Huazhong University of Science and Technology, China, in 2008, and the Ph.D. degree from the School of Information and Communications Engineering, Beijing University of Posts and Telecommunications, China, in 2013. He is currently an Associate Professor with the Department of Electronics and Information Engineering, Central China Normal University, China. Prior to that, he held postdoctoral positions with the School of Electronic Information and Communications, Huazhong University of Science and Technology, and the Department of Electrical Engineering, Columbia University, USA. His research interests include 5G/6G networks and machine learning.

Zhaolin Wang received the first B.Eng. degree from the Beijing University of Posts and Telecommunications, China, the second B.Eng. degree (Hons.) from the Queen Mary University of London, U.K., in 2020, and the M.Sc. degree (Distinction) from Imperial College London, U.K., in 2021. He is currently pursuing the Ph.D. degree with the Queen Mary University of London. His research interests include near-field communications, integrated sensing and communications, reconfigurable intelligent surface, and optimization theory. He is the recipient of the Best Student Paper Award in IEEE VTC2022-Fall and the 2023 IEEE Daniel E. Noble Fellowship Award.

Wenwu Xie received the B.S., M.S., and Ph.D. degrees in communication engineering from Huazhong Normal University, Wuhan, China, in 2004, 2007, and 2017, respectively. He is currently an Associate Professor with the School of Information Science and Engineering, Hunan Institute of Science and Technology, Yueyang, Hunan, China. His research interests include communication algorithms and control algorithms.



Yuanwei Liu (S'13-M'16-SM'19) received the PhD degree in electrical engineering from the Queen Mary University of London, U.K., in 2016. He was with the Department of Informatics, King's College London, from 2016 to 2017, where he was a Post-Doctoral Research Fellow. He has been a Senior Lecturer (Associate Professor) with the School of Electronic Engineering and Computer Science, Queen Mary University of London, since Aug. 2021, where he was a Lecturer (Assistant Professor) from 2017 to 2021. His research interests include non-orthogonal multiple access, reconfigurable intelligent surface, integrated sensing and communications, and machine learning.

He is a Web of Science Highly Cited Researcher since 2021, an IEEE Communication Society Distinguished Lecturer, an IEEE Vehicular Technology Society Distinguished Lecturer, the academic Chair for the Next Generation Multiple Access Emerging Technology Initiative, the rapporteur of ETSI Industry Specification Group on Reconfigurable Intelligent Surfaces, and the UK representative for the URSI Commission C. He was listed as one of 35 Innovators Under 35 China in 2022 by MIT Technology Review. He received IEEE ComSoc Outstanding Young Researcher Award for EMEA in 2020. He received the 2020 IEEE Signal Processing and Computing for Communications (SPCC) Technical Committee Early Achievement Award, IEEE Communication Theory Technical Committee (CTTC) 2021 Early Achievement Award. He received IEEE ComSoc Outstanding Nominee for Best Young Professionals Award in 2021. He is the co-recipient of the Best Student Paper Award in IEEE VTC2022-Fall, the Best Paper Award in ISWCS 2022, and the 2022 IEEE SPCC-TC Best Paper Award. He serves as the Co-Editor-in-Chief of IEEE ComSoc TC Newsletter, an Area Editor of IEEE Communications Letters, an Editor of IEEE Communications Surveys & Tutorials, IEEE Transactions on Wireless Communications, IEEE Transactions on Vehicular Technology, and IEEE Transactions on Network Science and Engineering. He serves as the (leading) Guest Editor for Proceedings of the IEEE on Next Generation Multiple Access, IEEE JSAC on Next Generation Multiple Access, IEEE JSTSP on Intelligent Signal Processing and Learning for Next Generation Multiple Access, and IEEE Network on Next Generation Multiple Access for 6G. He serves as the Publicity Co-Chair for IEEE VTC 2019-Fall, Symposium Co-Chair for Cognitive Radio & AI-Enabled Networks for IEEE GLOBECOM 2022 and Communication Theory for IEEE GLOBECOM 2023. He serves as the chair of Special Interest Group (SIG) in SPCC Technical Committee on Signal Processing Techniques for Next Generation Multiple Access, the vice-chair of SIG in SPCC Technical Committee on Near Field Communications for Next Generation Mobile Networks, and the vice-chair of SIG WTC on Reconfigurable Intelligent Surfaces for Smart Radio Environments.