Multi-Task Learning for Near/Far Field Channel Estimation in STAR-RIS Networks

Jian Xiao, Ji Wang, Zhaolin Wang, Jun Wang, Wenwu Xie, and Yuanwei Liu, *Fellow, IEEE*

***Abstract*—A joint cascaded channel estimation scheme is proposed for simultaneously transmitting and reflecting reconfigurable intelligent surface (STAR-RIS) systems with hardware imperfections. In particular, the practical hybrid near- and far-field electromagnetic radiation with spatial non-stationarity is investigated. By exploiting the cascaded channel correlations between different users and between different STAR-RIS elements, a multi-task learning (MTL)-based channel estimation framework is proposed. This framework is capable of estimating the cascaded channels for transmission and reflection simultaneously based on noisy observations of the mixture channel. Following the design guideline of the proposed MTL framework, an efficient multi-task network (MTN) is developed to reconstruct the high-dimensional channels with limited pilot overhead. In the proposed MTN architecture, a mixed convolution and multilayer perception module is exploited to capture the effective hybrid-field channel features. This module integrates the locality bias modeling of the channel-wise convolution and the long-range dependency modeling of multilayer perception, which finely learns both local spatial correlations and specific spatial non-stationarity of the hybrid-field cascaded channels. Numerical results show that the proposed MTN achieves superior channel estimation accuracy with less training overhead compared with the existing state-of-the-art benchmarks, in terms of required pilots, computations, and network parameters[1].**

***Index Terms*—Channel estimation, multi-task learning, reconfigurable intelligent surfaces, simultaneous transmission and reflection.**

I. Introduction

METAMATERIA-based reconfigurable intelligent surface (RIS) has been regarded as a promising multiple-input multiple-output candidate to construct *smart radio environments* (SREs) with low cost and energy consumption [2]. The typical reflection-only RISs only reflect the incident signal to desired user equipments at the same side (referred to as UE^r), which only enables a *half-space* SRE. To break the limitation of reflection-only RISs and achieve the *full-space* SREs, the novel concept of *simultaneously transmitting and reflecting* RISs (STAR-RISs) has been proposed to facilitate the full-space SRE [3]–[5]. More particularly, the signal imping on the STAR-RIS is divided into two parts with the law of energy conservation. One part of the signal is reflected to UE^r at the same side as the incident wave, while the other part is transmitted to users at the opposite side (referred to as UE^t).

To satisfy the demand of innovative applications supported by the sixth generation communications, e.g., virtual reality, holographic projections, etc, the antenna array will be further scaled up to empower the extremely large-scale antenna array (ELAA) communications [6]. However, as the number of antennas or RIS elements grows large and the communication frequency becomes high, the widely used far-field radiation assumptions are no longer valid. Instead, near-field propagation is more likely to occur, due to the expansion of the array aperture and the increase of frequencies [7]. Note that the boundary of near-field region in RIS systems is more strict compared with conventional ELAA systems, which is determined by the harmonic mean of the transmitter-RIS distance and the RIS-receiver distance [6]. The general transmission schemes toward future near-field communications are becoming the new research branch. In particular, the accurate channel estimation is one of the most fundamental research problems [8].

In passive RIS systems, the estimation of high-dimensional cascaded channels is an inherent barrier due to the unacceptable pilot overhead. Compared with the channel estimation in reflection-only RIS systems, the channel estimation in STAR-RIS systems necessitates the consideration of both transmission and reflection channels, along with the practical operating protocols [9]. In near-field communications, specific channel characteristics, e.g., the spherical wavefront, variations angle of arrival/departure (AoA/AoD) across array elements, and spatial channel non-stationarity [10], should be taken into account. Moreover, a practical case of radiation field, i.e., hybrid far- and near-field, is highly likely to happen in practical ELAA systems [6], [11]. Specifically, the authors in [6] presented two typical hybrid-field communication scenarios. Firstly, in the communication environment with dynamic scatterers, some scatterers are far away from ELAA equipments, while others may exist in the near-field region. Secondly, in ultra-wideband systems, the signal at low frequencies is propagating in the far-field region, while others at high frequencies may operate in the near-field region. Consequently, the hybrid-field communications is practical and crucial, prompting a compelling need for a comprehensive investigation into cascaded channel estimation in hybrid-field

This work was supported in part by the National Natural Science Foundation of China under Grant 62101205, and in part by the Key Research and Development Program of Hubei Province under Grant 2023BAB061. Part of this work has been presented by the 2023 IEEE Global Communications Conference (IEEE Globecom'23), Kuala Lumpur, Malaysia, Dec. 2023 [1]. [DOI: 10.1109/GLOBECOM54140.2023.10437505]. *(Corresponding authors: Ji Wang, Yuanwei Liu.)*

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).

Jun Wang is with the CICT Mobile Communication Technology Co., Ltd., Wuhan 430010, China. (e-mail: jwang@cictmobile.com).

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/WiCi-Lab/MTN.

STAR-RIS systems.

A. Related Works

1) Far-field channel estimation in RIS systems: Previous studies on channel estimation in RIS systems primarily focus on the far-field electromagnetic wave radiation. These studies have yielded diverse design concepts aimed at mitigating the pilot overhead requirements for high-dimensional cascaded channel estimation [8]. For instance, a semi-passive channel estimation framework was conceived in [12] and [13], where a limited number of RF chains were equipped at the RIS to carry out some specific channel estimation tasks, such as AoA acquisition. To reduce the pilot overhead and enhance the channel estimation accuracy for pure-passive RISs, the authors of [14] and [15] proposed leveraging the nonlinear mapping capabilities of deep learning (DL) models to establish a data-driven channel estimation framework. Furthermore, the authors of [16] and [17] developed compressed sensing (CS) algorithms that capitalize on the sparsity properties exhibited by cascaded channels within specific transform domains, such as the angular domain, thus reducing the pilot overhead.

2) Near-field channel estimation in RIS systems: When comes to near-field communications, the practical spherical wavefront radiation restricts the effectiveness of the existing far-field channel estimation schemes. For instance, the widely used CS algorithms based on the far-field channel sparsity in the angular domain are not applicable to the near-field channel estimation, because the sparsity of near-field channels in the angular domain no longer holds due to the severe energy spreading effect. As a remedy, the authors of [18] and [19] redesigned the CS algorithms for near-field RIS systems, where the sparsity of near-field channels in the polar domain was leveraged to recover the channels. The above channel estimation schemes all adopted the CS-based sparsity channel estimation framework, which heavily depends on the pure sparsity of the wireless channel in a specific transform domain, e.g., the polar-domain presentation. However, the channel sparsity may exist between different domains in practice, which makes these given signal-independent sparse basis difficult to adequately capture the complex sparse structure within the channel. These correlation and sparse structures become even more complex when the practical hybrid-field channels is considered for RIS systems, which make it more challenging to determine the appropriate basis that guarantees an acceptable channel reconstruction accuracy.

3) Hybrid-field channel estimation in ELAA systems: Recently, some researches have studied the hybrid-field channel estimation in ELAA systems. In [20], two different channel transform matrices, i.e., the angular-domain and polar-domain transform matrix, were designed to individually estimate the near- and far-field path components successively for ELAA systems. However, in this framework, the near-field path estimation relies on the prior far-field path estimation, thus resulting in the inevitable error propagation between the near- and far-field channel estimation. As an innovative contribution, a model-driven fixed point network was proposed to estimate the hybrid-field Terahertz channel in [21], which avoids the successive path estimation between near- and far-field path components. Considering the hybrid-field cascaded channel estimation for RIS systems in [22], a U-shaped multilayer perceptron (MLP) network is proposed to improve the high-dimensional channel estimation performance.

B. Motivations and Contributions

In contrast to the channel estimation in RIS systems, the channel estimation design in STAR-RIS systems is related to the dedicated operating protocol of STAR-RIS. Specifically, time switching (TS) and energy splitting (ES) are dominated operating protocols for the STAR-RIS [9]. In the TS protocol, the STAR-RIS periodically switches all elements between the transmitting mode and the reflecting mode in different orthogonal time slots. Hence the channel estimation for the TS protocol is similar to that in reflecting-only RIS systems. In the ES protocol, the incident signal on each element of the STAR-RIS can be reflected and transmitted with an ES ratio at the same time slots, which can provide higher communication degree of freedom. Since the ES strategy reduces the received signal strength at $\mathrm{UE}^f (\forall f \in \{t, r\})$ and the practical phase shift model may be coupled, the channel estimation accuracy can be relatively lower than the TS protocol [23]. Nevertheless, the simultaneously transmitting and reflecting signal transmission in the ES protocol provides the potentiality for reducing pilot overhead of the multi-user channel estimation, which has not been exploited well in STAR-RIS systems. Compared to TS protocol-based STAR RIS systems, the received pilot signal in ES protocol-based STAR-RIS systems consists of the transmitting and reflecting signal from UE_k^t and UE_k^r at the same transmission slot, which can support the realization of the joint transmitting and reflecting channel estimation by constructing an end-to-end deep learning model.

Although the channel estimation has been widely investigated for reflection-only RIS systems, the design of channel estimation schemes in STAR-RIS systems is still at a preliminary stage due to the aforementioned unique challenges, especially for the hybrid-field communications. In [23], a least square (LS)-based channel estimation scheme was derived for STAR-RIS systems, which was applied to both the TS and ES protocols. However, as a classic linear estimator, the performance of the LS estimation is limited, especially for the severe communication noise and the non-linear hardware imperfections in practical communication systems. And, even more crucially, the required pilot overhead of the LS estimator is expensive for the extremely large-scale STAR-RIS. Specifically, the minimum pilot overhead is KN for in [23], where N and K denote the number of STAR-RIS elements and UEs in a paired user group (UG), respectively. Moreover, in hybrid-field STAR-RIS systems, the specific sparse representation of the cascaded channel is hard to obtain due to the hybrid-field radiation and spatial non-stationarity. Hence, the channel estimation performance of the existing CS algorithms will be degraded [18], [19]. In our previous work [22], we proposed a U-shaped MLP architecture to capture the channel spatial non-stationarity in reflection-only RIS systems, which has better channel reconstruction performance than conventional channel estimation networks, such as convolutional network architectures in [15]. However, the required pilot overhead in this framework is still proportional to K for multi-user RIS systems. In addition, the network parameters of the U-shaped MLP may be redundant due to the dense connections of neurons in the MLP architecture.

Against the above background, we investigate the hybrid-field channel estimation for STAR-RIS systems. In particular, to address the limitation of the existing schemes, we propose a multi-task learning (MTL)-based joint channel estimation framework. Our contributions are summarized as follows.

- We study the hybrid-field cascaded channel estimation for STAR-RIS aided multi-user millimeter-wave (mmWave) systems with hardware imperfections. We characterize the unique cascaded hybrid-field radiation and present the spatial non-stationarity caused by the different types of visibility regions (VRs), i.e., clustered VRs and user VRs. In contrast to the hybrid-field channel modeling in conventional ELAA systems, the specific cascaded channel characteristics in STAR-RIS systems are revealed.

- We propose an MTL-based joint cascade channel estimation framework, which leverages the hybrid-field cascaded channel correlations between different users and between different STAR-RIS elements. To realize the loss balancing of multi-task optimization in the MTL framework, we design an adaptive joint loss function to alleviate the multi-task competition between different subtasks, which not only utilizes the ES prior information but also introduces an learnable scalar to allocate the adaptive weight for different subtasks.

- Based on the design guideline of the proposed MTL framework, we exploit an efficient multi-task network (MTN) to precisely reconstruct hybrid-field cascaded channels. In the proposed MTN architecture, a mixed convolution and multilayer perception (ConvMLP) module is exploited to capture the effective features of the hybrid-field channels. Specifically, we first design the channel-wise convolution module to model the locality spatial correlations of the STAR-RIS channel. Then, the axial MLP architecture is constructed to carry out the global spatial modeling of the non-stationary cascaded channels. Furthermore, we design the hierarchical network backbone to learn the implicit sparsity of the mmWave channels.

- Compared with the existing traditional estimator [23], the required pilot overhead of the proposed MTN architecture is reduced to N/Γ, in which $\Gamma \geq 1$ is a sampling interval in the STAR-RIS element domain. For the widely used single-task learning (STL) models in RIS channel estimation, K independent STL networks are required to estimate K cascaded channels. However, the proposed MTN architecture can jointly estimate the transmitting and reflecting channels, which significantly reduces the training overhead of the multi-user cascaded channel estimation and improves the channel estimation accuracy.

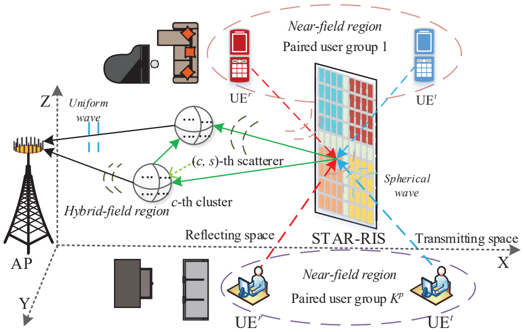


Fig. 1. STAR-RIS assisted hybrid-field multi-user communications.

C. Organizations and Notations

Organizations: The remainder of this paper is organized as follows. Section II introduces the hybrid-field channel modeling and system model of STAR-RIS systems. In Section III, we propose the MTL framework to realize the joint cascaded channel estimation. Based on the proposed MTL framework, we further design the efficient MTN architecture in Section IV. Section V provides numerical results of the proposed channel estimation scheme. Lastly, Section VI summarizes this work and looks forward the future research direction.

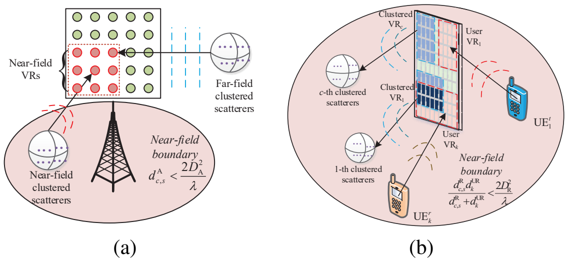
Notations: $\mathbf{A}^T$ and $\mathbf{A}^H$ denote the transpose and conjugate transpose of matrix $\mathbf{A}$, respectively; $\lfloor x \rfloor$ denotes the smallest integer that is greater than or equal to x; a^* denotes the conjugate of complex number a; diag($\mathbf{a}$) denotes the diagonal matrix with vector $\mathbf{a}$; $\mathbf{I}_a$ is the $a \times a$ identity matrix; $|\cdot|$, $\|\cdot\|$, and $\|\cdot\|_F$ denote the ℓ_1, ℓ_2, and Frobenius norm, respectively; $\propto$ denotes the proportionality relation. $\odot$ and $\otimes$ denote the Hadamard product and convolution, respectively. $\Re(\mathbf{A})$ and $\Im(\mathbf{A})$ denote the real and imaginary components of the complex-value matrix $\mathbf{A}$.

II. System Model and Problem Formulation

In this section, we first provide the field boundary in STAR-RIS systems. Then, the hybrid-field cascaded channel modeling is presented. Lastly, we formulate the cascaded channel estimation problem in the ES protocol, in which the practical signal model with hardware imperfection is investigated.

A. Field Boundary in STAR-RIS systems

As shown in Fig. 1, an N^{s}-element ($N^{\mathrm{s}} = N_1^{\mathrm{s}} \times N_2^{\mathrm{s}}$) STAR-RIS operating in ES mode and equipped with a uniform planar array (UPA) is deployed to enhance an indoor communication system, wherein there exists an M-element ($M = M_1 \times M_2$) wireless access point (AP) equipped with a UPA and K^{u} single-antenna UEs. Considering the high channel correlations between the sub-wavelength metamaterial elements, the typical element grouping strategy is commonly adopted to reduce the required control and training overhead in metasurface-based communication systems [8], [23], in which the adjacent metamaterial elements are grouped into a sub-surface to share a common phase-shift. In this work, the STAR-RIS elements are divided into $N = N_1 \times N_2$ sub-surfaces, each of which consists of $\nu = (N_1^{\mathrm{s}}/N_1) \times (N_2^{\mathrm{s}}/N_2)$ adjacent elements. Furthermore,


Near-field VRs
Far-field clustered scatterers
Near-field clustered scatterers
Near-field boundary
$d^{\mathrm{A}}_{c,s} < \frac{2D_{\mathrm{A}}^2}{\lambda}$
Clustered VR_c
User VR_1
c-th clustered scatterers
Clustered VR_1
UE^t_1
1-th clustered scatterers
User VR_k
Near-field boundary
$\frac{d^{\mathrm{R}}_{c,s}d^{\mathrm{UR}}_k}{d^{\mathrm{R}}_{c,s}+d^{\mathrm{UR}}_k} < \frac{2D_{\mathrm{R}}^2}{\lambda}$
UE^r_k

(a) (b)

Fig. 2. Electromagnetic radiation fields and VRs distribution. (a) Hybrid-field radiation at the AP, in which the marked regions with light red color denote the clustered VRs in the scatterer (c, s)→AP link; (b) Near-field radiation at the STAR-RIS, in which the marked regions with light red and green color denote the user VRs in UE_k→STAR-RIS and clustered VRs in STAR-RIS→scatterer (c, s) links, respectively, while the marked regions with dark green color denote the overlapping regions of multiple VRs at the STAR-RIS.

K^{u} UEs are equally divided into K^{p} paired user groups (UGs), composed of $K = K^{\mathrm{u}}/K^{\mathrm{p}}$ UEs. In the k^{p}-th $(1 \leq k^{\mathrm{p}} \leq K^{\mathrm{p}})$ UG, the number of UE^ts in the transmitting space of the STAR-RIS is K^t, while $K^r = K - K^t$ UE^rs are located on the reflecting space. To mitigate the inter-group interference in STAR-RIS systems, the nearest UE^f $(\forall f \in \{t, r\})$ are paired into the same UG according to geometric locations.

Both AP and STAR-RIS lie on the plane perpendicular to the xy-plane, whose array center coordinate are set to $\mathbf{c}^{\mathrm{A}} = \left(x^{\mathrm{A}}, y^{\mathrm{A}}, z^{\mathrm{A}}\right)$ and $\mathbf{c}^{\mathrm{R}} = \left(x^{\mathrm{R}}, y^{\mathrm{R}}, z^{\mathrm{R}}\right)$, respectively. Let $\Delta m = \Delta m_1 = \Delta m_2$ and $\Delta n = \Delta n_1 = \Delta n_2$ denote the distance between two adjacent antennas (elements) at the AP and the STAR-RIS, respectively. Hence, the coordinate of the AP antenna (m_1, m_2) is $\mathbf{c}^{\mathrm{A}}_{m_1,m_2} = \left(x^{\mathrm{A}}, y^{\mathrm{A}} + (m_1 - \frac{M_1+1}{2})\Delta m, z^{\mathrm{A}} + (m_2 - \frac{M_2+1}{2})\Delta m\right)$. Accordingly, the coordinate of the STAR-RIS element (n_1, n_2) is $\mathbf{c}^{\mathrm{R}}_{n_1,n_2} = \left(x^{\mathrm{R}}, y^{\mathrm{R}} + (n_1 - \frac{N_1+1}{2})\Delta n, z^{\mathrm{R}} + (n_2 - \frac{N_2+1}{2})\Delta n\right)$. The coordinate of UE_k $\mathbf{c}^{\mathrm{U}}_k = \left(x^{\mathrm{U}}_k, y^{\mathrm{U}}_k, z^{\mathrm{U}}_k\right)$ are randomly distributed around the STAR-RIS. The coordinate of scatterer s $(1 \leq s \leq S_c)$ in cluster c $(1 \leq c \leq C_{\mathrm{s}})$ is denoted as $\mathbf{c}^{\mathrm{S}}_{c,s} = (x^{\mathrm{S}}_{c,s}, y^{\mathrm{S}}_{c,s}, z^{\mathrm{S}}_{c,s})$ between the STAR-RIS and the AP. To alleviate the severe multiplicative fading effect of the cascaded link, the STAR-RIS is deployed on the wall near the UEs forming the line-of-sight (LOS) channel [24]. Hence, UEs are likely communicating in the near-field region of the STAR-RIS, which is determined by the Rayleigh distance Z. According to the near-field criterion in [6], the near-field region for RIS-aided systems is given by

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

where $d^{\mathrm{R}}_{c,s}$ and d^{UR}_k denote the distance from the STAR-RIS to scatterer (c, s) and the distance from the UE^f_k to the STAR-RIS, respectively. Parameter λ is the carrier wavelength and D is the equivalent array aperture of STAR-RIS systems.

According to (1), it can be further implied that as long as any of $d^{\mathrm{R}}_{c,s}$ and d^{UR}_k is shorter than the Rayleigh distance Z, the communication link is operating in the near-field region. On the other hand, As illustrated in Fig. 2(a), the environmental scatterers may be distributed in the near- or far-filed region of the AP [20]–[22], respectively. Hence, the near- and far-field signal components will coexist in the practical hybrid-field STAR-RIS systems.

B. Hybrid-Field Channel model

Following the 3GPP standard in the indoor mmWave communications [25], we adopt the general clustered statistical multiple-input multiple-output (MIMO) channel modeling framework, in which the STAR-RIS→AP channel $\mathbf{G} \in \mathbb{C}^{M \times N}$ from the STAR-RIS to the AP is given by

$$\mathbf{G} = \gamma \sum_{c=1}^{C_{\mathrm{s}}} \sum_{s=1}^{S_c} \varsigma_{c,s} \sqrt{R^{G_{\mathrm{r}}}_{c,s} L^{G_{\mathrm{r}}}_{c,s}} \mathbf{a}_{c,s} \mathbf{b}^T_{c,s}, \tag{2}$$

where the number of clusters C_{s} and scatterers S_c in the cluster c are characterized by Poisson distribution $C_{\mathrm{s}} \sim \max\{P(\lambda_p), 1\}$ and the uniform distribution $S_c \sim \mathcal{U}[1, \bar{s}_c]$, respectively. Parameter λ_p is related to the communication frequency f_c. $\gamma = \sqrt{\frac{1}{\sum_{c=1}^{C_{\mathrm{s}}} S_c}}$ is a normalization factor. The complex gain $\varsigma_{c,s}$ follows $\varsigma_{c,s} \sim \mathcal{CN}(0, 1)$. Parameters $L^{G_{\mathrm{r}}}_{c,s}$ and $R_{c,s}$ denote the path loss and STAR-RIS element gain for scatterer path (c, s), which follow the 5G mmWave path loss model in Indoor Hotspot environment and the reflectarray radiation pattern [24], respectively. The array response $\mathbf{b}_{c,s} \in \mathbb{C}^{N \times 1}$ and $\mathbf{a}_{c,s} \in \mathbb{C}^{M \times 1}$ denote the near-field transmitting response at the STAR-RIS and the hybrid-field receiving response at the AP, respectively. Specifically, the generic near-field array response $\mathbf{b}^n_{c,s}$ without the spatial non-stationarity at the STAR-RIS can be expressed as [22], [26]

$$\begin{aligned}\mathbf{b}^{\mathrm{n}}_{c,s}\left(d^{\mathrm{R}}_{c,s}\right) = \Big[&e^{-j2\pi d^{\mathrm{R}}_{c,s}(1,1)/\lambda}, \cdots, e^{-j2\pi d^{\mathrm{R}}_{c,s}(1,N_2)/\lambda}, \\ &\cdots, e^{-j2\pi d^{\mathrm{R}}_{c,s}(N_1,1)/\lambda}, \cdots, e^{-j2\pi d^{\mathrm{R}}_{c,s}(N_1,N_2)/\lambda}\Big],\end{aligned} \tag{3}$$

where $d^{\mathrm{R}}_{c,s}(n_1, n_2) = \left\|\mathbf{c}^{\mathrm{R}}_{n_1,n_2} - \mathbf{c}^{\mathrm{S}}_{c,s}\right\|$ denotes the distance from scatterer (c, s) to the (n_1, n_2)-th STAR-RIS element.

As shown in Fig. 2(b), different parts of the STAR-RIS elements may view different scatterers (terminals) due to the limitation of VRs, which results in the spatial non-stationarity of the hybrid-field STAR-RIS channel. The general VR definition at the UPA can be expressed as $\Omega = \left\{[\vec{c}^1 - \vec{l}^1/2, \vec{c}^1 + \vec{l}^1/2], [\vec{c}^2 - \vec{l}^2/2, \vec{c}^2 + \vec{l}^2/2]\right\}$, in which $(\vec{c}^1, \vec{c}^2)$ and $(\vec{l}^1, \vec{l}^2)$ denote the VR center and length at different directions, respectively. The VR length $(\vec{l}^1, \vec{l}^2)$ can be characterized by the Lognormal distribution [27], [28], i.e., $\vec{l}^1 \sim \mathcal{LN}(\bar{\mu}_1, \bar{\sigma}_1)$ and $\vec{l}^2 \sim \mathcal{LN}(\bar{\mu}_2, \bar{\sigma}_2)$, where parameters $\bar{\mu}$ and $\bar{\sigma}$ denote the mean and standard deviation of logarithmic values, respectively. For the VRs at the STAR-RIS, two different types of VRs, i.e., cluster VRs Ω^{R}_c caused by the near-field scatterers [22] and user VRs Ω_k for different UE_k [10], are comprehensively studied. Specifically, the VR Ω^{R}_c of cluster c in STAR-RIS→scatterer (c, s) link is given by $\Omega^{\mathrm{R}}_c = \left\{[\vec{c}^{\mathrm{R},1}_c - \vec{l}^{\mathrm{R},1}_c/2, \vec{c}^{\mathrm{R},1}_c + \vec{l}^{\mathrm{R},1}_c/2], [\vec{c}^{\mathrm{R},2}_c - \vec{l}^{\mathrm{R},2}_c/2, \vec{c}^{\mathrm{R},2}_c + \vec{l}^{\mathrm{R},2}_c/2]\right\}$, in which the VR center $(\vec{c}^{\mathrm{R},1}_c, \vec{c}^{\mathrm{R},2}_c)$ follows the independent uniform distribution, i.e., $\vec{c}^{\mathrm{R},1}_c \sim \mathcal{U}[\vec{l}^{\mathrm{R},1}/2, N_1 - \vec{l}^{\mathrm{R},1}/2]$ and $\vec{c}^{\mathrm{R},2}_c \sim \mathcal{U}[\vec{l}^{\mathrm{R},2}/2, N_2 - \vec{l}^{\mathrm{R},2}/2]$. Furthermore, the VR cover vector $v(\Omega_c^{\mathrm{R}}) \in \mathbb{C}^{N\times 1}$ at the STAR-RIS for cluster c can be expressed as

$$\left[v(\Omega_c^{\mathrm{R}})\right]_{(n_1,n_2)} = \begin{cases} 1, \text{ if } (n_1,n_2) \in \Omega_c^{\mathrm{R}}, \\ 0, \text{ else}. \end{cases} \tag{4}$$

Considering the spatial non-stationarity caused by the VR cover vector, the equivalent array response at the STAR-RIS is given by $\mathbf{b}_{c,s} = \mathbf{b}_{c,s}^{\mathrm{n}} \odot v(\Omega_c^{\mathrm{R}})$.

Since the far- and near-field scatterers coexist in the around of the AP, the array response $\mathbf{a}_{c,s}$ depends on the distance $d_{c,s}^{\mathrm{A}}$ from scatterer (c,s) to the AP [22], which is given by

$$\mathbf{a}_{c,s} = \begin{cases} \mathbf{a}_{c,s}^{\mathrm{f}}\left(\phi_{c,s}^{\mathrm{A}}, \varphi_{c,s}^{\mathrm{A}}\right), & \text{if } d_{c,s}^{\mathrm{A}} > Z, \\ \mathbf{a}_{c,s}^{\mathrm{n}}\left(d_{c,s}^{\mathrm{A}}\right) \odot v\left(\Omega_c^{\mathrm{A}}\right), & \text{otherwise}. \end{cases} \tag{5}$$

where the definition of the near-field array response $\mathbf{a}_{c,s}^{n}$ is similar to $\mathbf{b}_{c,s}^{\mathrm{n}}$ in (3), and $v(\Omega_c^{\mathrm{A}})$ denotes the VR cover vector at the AP for cluster c. The definition of VR Ω_c^{A} of cluster c in scatterer (c,s)→AP link is given by $\Omega_c^{\mathrm{A}} = \left\{[\vec{c}_c^{\,\mathrm{A},1} - \vec{l}_c^{\,\mathrm{A},1}/2, \vec{c}_c^{\,\mathrm{A},1} + \vec{l}_c^{\,\mathrm{A},1}/2], [\vec{c}_c^{\,\mathrm{A},2} - \vec{l}_c^{\,\mathrm{A},2}/2, \vec{c}_c^{\,\mathrm{A},2} + \vec{l}_c^{\,\mathrm{A},2}/2]]\right\}$, in which the VR center follows ($\vec{c}_c^{\,\mathrm{A},1} \sim \mathcal{U}[\vec{l}^{\,\mathrm{A},1}/2, M_1 - \vec{l}^{\,\mathrm{A},1}/2], \vec{c}_c^{\,\mathrm{A},2} \sim \mathcal{U}[\vec{l}^{\,\mathrm{A},2}/2, M_2 - \vec{l}^{\,\mathrm{A},2}/2]$). The far-field response $\mathbf{a}^{\mathrm{f}}$ at the AP is given by [24]

$$\mathbf{a}_{c,s}^{\mathrm{f}}\left(\phi_{c,s}^{\mathrm{A}}, \varphi_{c,s}^{\mathrm{A}}\right) = \Big[1, \cdots, e^{j2\pi\Delta m(x sin\varphi_{c,s}^{\mathrm{A}} + y \sin\phi_{c,s}^{\mathrm{A}} \cos\varphi_{c,s}^{\mathrm{A}})/\lambda}, \\ \cdots, e^{j2\pi\Delta m((M_1-1) sin\varphi_{c,s}^{\mathrm{A}} + (M_2-1)\sin\phi_{c,s}^{\mathrm{A}} \cos\varphi_{c,s}^{\mathrm{A}})/\lambda}\Big], \tag{6}$$

where $0 \le x \le M_1 - 1$, $0 \le y \le M_2 - 1$, $\phi_{c,s}^{\mathrm{A}}$ and $\varphi_{c,s}^{\mathrm{A}}$ denotes the azimuth and elevation of AoA for scatterer (c,s) at the AP, respectively. In the conventional far-field radiation assumption, the imping signals can be approximated as the uniform plane wave, in which $\mathbf{a}_{c,s}^{\mathrm{f}}$ only depends on the identical AoA/AoD.

For the LOS dominated UE$_k$→STAR-RIS link, the receiving array response $\mathbf{u}_k$ at the STAR-RIS is expressed as

$$\mathbf{u}_k\left(d_k^{\mathrm{UR}}\right) = \Big[e^{-j2\pi d_k^{\mathrm{UR}}(1,1)/\lambda}, \cdots, e^{-j2\pi d_k^{\mathrm{UR}}(1,N_2)/\lambda}, \\ \cdots, e^{-j2\pi d_k^{\mathrm{UR}}(N_1,1)/\lambda}, \cdots, e^{-j2\pi d_k^{\mathrm{UR}}(N_1,N_2)/\lambda}\Big], \tag{7}$$

where $d_k^{\mathrm{UR}}(n_1,n_2) = \left\|\mathbf{c}_{n_1,n_2}^{\mathrm{R}} - \mathbf{c}_k^{\mathrm{U}}\right\|$ denotes the distance from the UE$_k$ to the (n_1,n_2)-th STAR-RIS element.

The definition of user VR in UE$_{\mathrm{k}}$→STAR-RIS link is given by $\Omega_k = \left\{[\vec{c}_k^{\,1} - \vec{l}_k^{\,1}/2, \vec{c}_k^{\,1} + \vec{l}_k^{\,1}/2], [\vec{c}_k^{\,2} - \vec{l}_k^{\,2}/2, \vec{c}_k^{\,2} + \vec{l}_k^{\,2}/2]\right\}$, in which the VR center follows ($\vec{c}_k^{\,1} \sim \mathcal{U}[\vec{l}_k^{\,1}/2, N_1 - \vec{l}_k^{\,1}/2], \vec{c}_k^{\,2} \sim \mathcal{U}[\vec{l}_k^{\,2}/2, N_2 - \vec{l}_k^{\,2}/2]$), and the VR length follows ($\vec{l}_k^{\,1} \sim \mathcal{LN}(\bar{\mu}_{k,1}, \bar{\sigma}_{k,1})$, $\vec{l}_k^{\,2} \sim \mathcal{LN}(\bar{\mu}_{k,2}, \bar{\sigma}_{k,2})$). Accordingly, the VR cover vector $v(\Omega_k)$ for UE$_k$→STAR-RIS channel is given by

$$[v(\Omega_k)]_{(n_1,n_2)} = \begin{cases} 1, \text{ if } (n_1,n_2) \in \Omega_k, \\ 0, \text{ else}. \end{cases} \tag{8}$$

Hence, the UE$_k$→STAR-RIS channel $\mathbf{h}_k$ is given by

$$\mathbf{h}_k = \sqrt{R_k^{\mathrm{h}} L_k^{\mathrm{h}}}\,\mathbf{u}_k \odot v(\Omega_k), \tag{9}$$

where R_k^{h} and L_k^{h} and denote the radiation gain of STAR-RIS, the path loss, respectively.

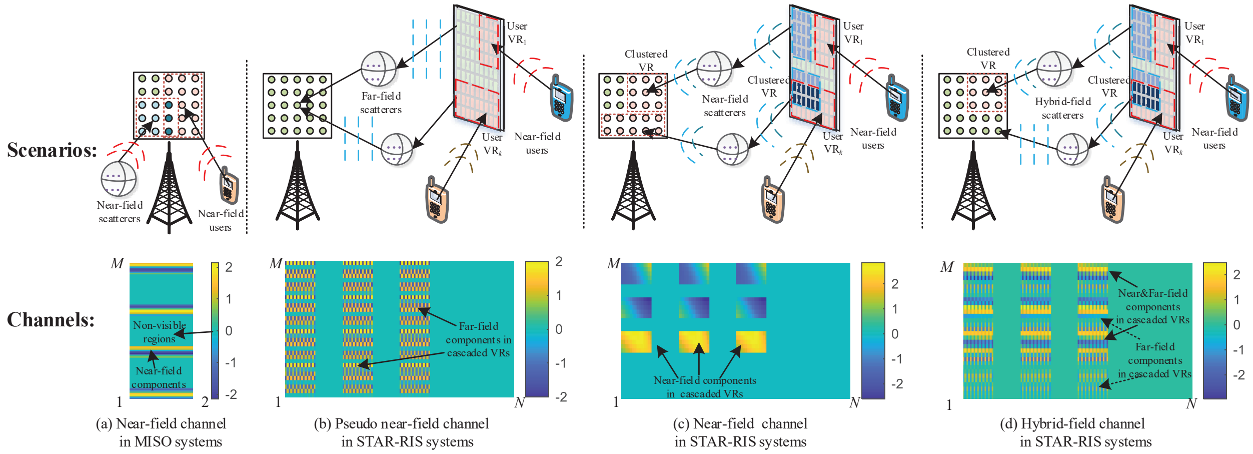
Remark 1: Compared with the hybrid-field channel modeling in conventional ELAA systems [20], [21], the hybrid-field cascaded channel for STAR-RIS systems has different characteristics. Firstly, the channel dimension is significantly increased due to numerous passive STAR-RIS elements. Secondly, the hybrid-field radiation in STAR-RIS communications is more complicated than ELAA systems, in which the near- and far-field path components are aggregated in a cascaded form instead of the addition form [22]. Finally, the spatial non-stationarity of the channel is further aggravated in STAR-RIS systems, where the VRs of UE$_k$→STAR-RIS link and STAR-RIS→scatterers links need to be specially considered except the VRs of scatterers→AP links. In Fig. 3, we present the channel visualization with the spatial non-stationarity for different radiation fields. Specifically, in the near-field MISO channel of Fig. 3(a), the value of partial channel elements is zero, which represents the non-visible region at the AP for the given scatterers/users. In the pseudo near-field cascaded channel of Fig. 3(b), the zero-value blocks of the cascaded channel matrix are randomly distributed along the N-dimension of STAR-RIS elements. This is because the near-field radiation with spatial non-stationarity is only adopted for UE$_k$→STAR-RIS link, while the STAR-RIS→AP link utilizes the far-field assumptions [10]. In the near-field cascaded channel of Fig. 3(c), since the near-field radiation with spatial non-stationarity is considered for both UE$_k$→STAR-RIS and STAR-RIS→AP links, the zero-value blocks of the cascaded channel matrix are randomly distributed along the N-dimension of STAR-RIS elements and M-dimension AP antennas. In the hybrid-field cascaded channel in Fig. 3(d), the far-field channel components in scatterers→AP paths introduce the non-zero channel elements in the whole M-dimension of AP antennas, while the near-field channel components in scatterers→AP paths are only distributed in the partial M-dimension of AP antennas due to the presence of VRs. Hence, we observe that the energy distribution of non-zero blocks in the hybrid-field cascaded channel presents significant difference along the M-dimension of AP antennas.

C. Problem Formulation

Let $\boldsymbol{\theta}^t = [\beta_1^t e^{j\theta_1^t}, \beta_2^t e^{j\theta_2^t}, \cdots, \beta_N^t e^{j\theta_N^t}]^T \in \mathbb{C}^{N\times 1}$ and $\boldsymbol{\theta}^r = [\beta_1^r e^{j\theta_1^r}, \beta_2^r e^{j\theta_2^r}, \cdots, \beta_N^r e^{j\theta_N^r}]^T \in \mathbb{C}^{N\times 1}$ denote the transmitting and reflecting vectors, respectively, in which the ES ratio β_n satisfies $(\beta_n^t)^2 + (\beta_n^r)^2 = 1 (n = 1, 2, \cdots, N)$ for the lossless STAR-RIS. We focus on the UE$_k^f$ →STAR-RIS→AP($\forall f \in \{t, r\}$) cascaded channel estimation, and the orthogonal pilot transmission strategy is adopted for different UGs. The received pilot signal $\mathbf{y}_q \in \mathbb{C}^{M\times 1}$ in the q-th time slot at the AP for the UG$_{k^{\mathrm{p}}}$ is given by

$$\mathbf{y}_q = \sum_{k=1}^{K} \mathbf{G}\mathrm{diag}(\boldsymbol{\theta}_q^k)\mathbf{h}_k s_{k,q} + \mathbf{w}_{k,q} \\ = \sum_{k=1}^{K} \mathbf{G}\mathrm{diag}(\mathbf{h}_k)\boldsymbol{\theta}_q^k s_{k,q} + \mathbf{w}_{k,q}, \tag{10}$$

where $\boldsymbol{\theta}^k = \boldsymbol{\theta}^t$ for $k \in \{1, \cdots, K^t\}$, while $\boldsymbol{\theta}^k = \boldsymbol{\theta}^r$ for $k \in \{K^t + 1, \cdots, K\}$. $s_{k,q}$ denotes the transmitted pilot signal and $\mathbb{E}[s_{k,q} s_{k,q}^*] = 1$. $\mathbf{w}_{k,q} \sim \mathcal{CN}(0, \sigma^2\mathbf{I}_M)$ represents the Gaussian noise. We define $\mathbf{H}_k = \mathbf{G}\mathrm{diag}(\mathbf{h}_k) \in \mathbb{C}^{M\times N}$ as the cascaded UE$_k$ → STAR-RIS → AP channel.



(a) Near-field channel in MISO systems

(b) Pseudo near-field channel in STAR-RIS systems

(c) Near-field channel in STAR-RIS systems

(d) Hybrid-field channel in STAR-RIS systems

Fig. 3. Visualization of spatial non-stationary channels for different radiation fields in specific communication scenarios. (a) Near-field MISO channel, (b) Pseudo near-field cascaded channel, (c) Near-field cascaded channel, and (d) Hybrid-field cascaded channel in STAR-RIS systems. In the channel visualization, the real part of the normalized channel vector/matrix is extracted. The element value of each channel entry is characterized based on RGB color space, in which the channel elements that are not within the VR have zero value. The VR lengths at the STAR-RIS and the AP are set to $\vec{l}_k^1 = \vec{l}_c^{\mathrm{R},1} = N_1/2$, $\vec{l}_k^2 = \vec{l}_c^{\mathrm{R},2} = N_2/2$, $\vec{l}_c^{\mathrm{A},1} = M_1/2$, and $\vec{l}_c^{\mathrm{A},2} = M_2/2$, respectively. Accordingly, the VR centers are set to $(\vec{c}_k^1, \vec{c}_k^2) = (\vec{c}_c^{\mathrm{R},1}, \vec{c}_c^{\mathrm{R},2}) = (\vec{l}_c^{\mathrm{R},1}/2+1, \vec{l}_c^{\mathrm{R},2}/2+1)$, and $(\vec{c}_c^{\mathrm{A},1}, \vec{c}_c^{\mathrm{A},2}) = (\vec{l}_c^{\mathrm{A},1}/2+1, \vec{l}_c^{\mathrm{A},2}/2+1)$, respectively.

In practical communication system, the hardware imperfections are non-negligible components due to the non-ideality of the hardware. For the considered STAR-RIS assisted mmWave systems, we investigate the coupled phase-shift model for purely passive STAR-RIS hardware at first [29]. Then, the hardware impairments (HWIs) imposed on both the transmitting signal at the UE$_k$ and the received signal at the AP are considered [30]. Specifically, the coupled STAR-RIS phase-shift model is given by [29]

$$\cos(\theta_n^t - \theta_n^r) = 0, n = 1, 2, \cdots, N. \tag{11}$$

Then, the residual HWIs at the UE and the AP are integrated to (10), which is given by

$$\tilde{\mathbf{y}}_q = \sum_{k=1}^{K} \mathbf{H}_k \boldsymbol{\theta}_q^k (s_{k,q} + \eta_{k,q}) + \mathbf{w}_q + \boldsymbol{\mu}_q, \tag{12}$$

where $\eta_{k,q} \sim \mathcal{CN}(0, (\rho_k^{\mathrm{u}})^2)$ denotes the transmitted distortion at the UE$_k$, $\boldsymbol{\mu}_q \sim \mathcal{CN}(0, (\rho^{\mathrm{a}})^2 \mathbf{p}_r)$ denotes the HWIs at the AP with $\mathbf{p}_r = \sum_{k=1}^{K} (\mathbf{H}_k \boldsymbol{\theta}_q^k)(\mathbf{H}_k \boldsymbol{\theta}_q^k)^H$, and parameters ρ_k^{u} and ρ^{a} denote the error vector magnitude at UE$_k$ and AP, respectively.

Let Q denote the pilot transmission slots, thus the overall received pilot signals at the AP is given by $\mathbf{Y} = [\tilde{\mathbf{y}}_1, \tilde{\mathbf{y}}_2, \cdots, \tilde{\mathbf{y}}_Q]^T \in \mathbb{C}^{Q \times M}$. Considering a two-user STAR-RIS systems, i.e., $K = 2$, a classic LS estimation of the cascaded channel was proposed in [23], which is given by

$$\hat{\mathbf{H}} = [\hat{\mathbf{H}}^t, \hat{\mathbf{H}}^r]^T = \mathbf{V}^H \left(\mathbf{V}\mathbf{V}^H\right)^{-1} \mathbf{Y}. \tag{13}$$

Here, $\hat{\mathbf{H}}^f \in \mathbb{C}^{M \times N} (\forall f \in \{t, r\})$ denotes the cascaded channel for the UE^f, and $\mathbf{V} \in \mathbb{C}^{Q \times KN}$ denotes the transmission pattern matrix. For the ideal case without hardware imperfections, the optimal $\mathbf{V}$ is the first KN columns of the $Q \times Q$ discrete Fourier transform (DFT) matrix in the minimum variance unbiased estimator.

Remark 2: In the existing LS estimator for STAR-RIS systems [23], each UG only supports a group paired UE^t and UE^r. In this case, the pilot overhead Q in the LS estimation is required to satisfy $Q = 2N$ due to the full-rank condition. For the general multi-user systems in this work, the total pilot overhead Q can be summarized as $Q \geq KN$ in the LS estimation, which is huge for STAR-RIS enabled ELAA systems. In contrast to the mathematical model-based estimator, a more intuitive data-driven solution can be provided for the STAR-RIS channel estimation in the ES protocol. According to (12), the collected pilot signals at the AP involve both transmitting and reflecting channels of all users. Hence, we can directly construct the mapping between the received mixture pilot signals and multi-user channels by exploiting a data-driven MTL framework, in which the required pilot overhead of the MTL-based channel estimation scheme is independent to K.

III. Multi-task Learning-based Joint Channel Estimation Framework

In this section, we first present the hybrid-field cascaded channel correlations between different users and between different STAR-RIS elements, which provides the theoretical foundation to realize the joint transmitting and reflecting channel estimation. Then, the MTL-based channel estimation framework is proposed by developing the joint adaptive optimization strategy, which simultaneously estimates the multi-user cascaded channels with limited pilot overhead.

A. Channel Correlations in STAR-RIS Systems

In STAR-RIS systems, the transmitting user UE$_k^t$ and reflecting user UE$_k^r$ communicate with the AP via the same STAR-RIS. Hence, the cascaded channels $\mathbf{H}_k^f (\forall f \in \{t, r\})$ associated with UE$_k^t$ and UE$_k^r$ share the same STAR-RIS→AP channel $\mathbf{G}$. In [31], the multi-user channel correlations is explicitly characterized as a scalar vector $\boldsymbol{\ell}_k =$

$[\boldsymbol{\ell}_{k,1}, \cdots, \boldsymbol{\ell}_{k,N}]^T \in \mathbb{C}^{N\times 1}$. Specifically, suppose $\mathbf{H}_{k,n} \in \mathbb{C}^{M\times 1}$ denote the cascaded channel from UE$_k$ to the AP via the n-th RIS element. Firstly, N pilots are used to estimate the cascade channel $\hat{\mathbf{H}}_1 \in \mathbb{C}^{M\times N}$ of the first user. Then, the cascaded channel $\hat{\mathbf{H}}_k (2 \le k \le K)$ of the other users is converted to estimate $\boldsymbol{\ell}_k$, satisfying $\hat{\mathbf{H}}_{k,n} = \boldsymbol{\ell}_{k,n}\hat{\mathbf{H}}_{1,n}$. In this work, we propose an MTL framework to implicitly exploit the multi-user correlations and directly realize the joint cascaded channel estimation by utilizing the common pilots, which can avoid the estimation overhead of the scalar vector $\boldsymbol{\ell}_k$. Compared with the STL-based channel estimation framework [13]–[15], the MTL-based channel estimation model also avoid the additional training overhead than the STL framework. Since the STL only support the one-to-one mapping, the transmitting and reflecting channels need to be independently estimated by different STL models. Moreover, the multi-task supervision effectively increases the sample space in the network training, which can attain the implicit data augmentation [32].

Furthermore, since the sub-wavelength units of the metasurface are arranged closely in hardware implementation, the channels at the neighboring elements of the STAR-RIS are highly correlated. However, the spatial non-stationarity caused by VRs will disrupt the conventional spatial correlations of the cascaded channel matrix in hybrid-field STAR-RIS systems, which motivates us to design a more efficient channel extrapolation model to reduce the pilot overhead. In the dataset construction of the proposed channel extrapolation model, we first select P STAR-RIS elements as a subset $\mathcal{P}$ of the whole STAR-RIS elements, satisfying $\mathcal{P} = \{1, \Gamma+1, \cdots, (P-1)\times\Gamma+1\}$ with the sampling interval $\Gamma = 2^U (0 \le U \le \log_2 N)$. Let $\mathbf{H}_k^P \in \mathbb{C}^{M\times P}$ denotes the cascaded channel matrix of the subset $\mathcal{P}$ for the UE$_k$. Then, we serially turn on each element in the subset $\mathcal{P}$, i.e., $\boldsymbol{\theta}_p^k = [0, \cdots, \theta_{n=p}^k = \beta_n^k, \cdots, 0]^T \in \mathbb{C}^{P\times 1}$ in the p-th pilot slot $(1 \le p \le P)$, to obtain the received signal $\mathbf{Y}^P \in \mathbb{C}^{M\times P}$ at the AP. Lastly, a channel extrapolation network is constructed to realize the mapping from $\mathbf{Y}^P$ to the complete channel matrix $\mathbf{H}_k \in \mathbb{C}^{M\times N}$. By utilizing the cascaded channel correlations in both user and spatial domain, an MTL-based joint channel estimation framework can be developed, which only requires N/Γ pilots to realize the precise cascaded channel reconstruction of K users in hybrid-field STAR-RIS systems.

Remark 3: In conventional reflection-only RIS systems, the LS estimator with P pilots can directly obtain the partial channel matrix $\hat{\mathbf{H}}_k^P \in \mathbb{C}^{M\times P}$ of the cascaded channel $\mathbf{H}_k$, which is used as the input tensor the channel extrapolation network. However, in multi-user STAR-RIS systems, KP pilots are required to obtain $\hat{\mathbf{H}}_k^P$ according to (13). To reduce the required pilot overhead in the channel pre-estimation, we resort to the typical ON/OFF protocol by serially operating the single STAR-RIS element [8], which can obtain the partial mixture sampling $\mathbf{Y}^P \in \mathbb{C}^{M\times P}$ of K cascaded channels with P pilots. Without loss of generality, the uplink pilot signal $s_{k,q}$ is assumed to be $s_{k,q} = 1$ for $q = 1, 2, \dots, Q$ within the same UG, while the mutual orthogonal time resources are assigned to transmit pilot sequences among different UGs. Hence, the p-th column of $\mathbf{Y}^P$ can be expressed as $\mathbf{Y}_p^P = \sum_{k=1}^K \beta^k \mathbf{H}_{k,p}^P (1+\eta_{k,p}) + \mathbf{N}_p^e$, in which $\mathbf{N}_p^e \in \mathbb{C}^{M\times 1}$ denotes the measurement error caused by the noise and hardware imperfections. Hence, the channel extrapolation difficulty in STAR-RIS systems is larger than the reflection-only RIS systems. Note that the drawbacks of the ON/OFF protocol have been discussed in the conventional reflection-only RIS channel estimation. Since only an RIS element is activated at the single pilot slot, the signal strength of received pilots will be reduced. Nevertheless, for the MTL-based STAR-RIS channel estimation framework, the advantage of this operating protocol is that it can provide the efficient training samples with fewer pilots, i.e., $Q = P$ in the proposed scheme instead of $Q = KP$.

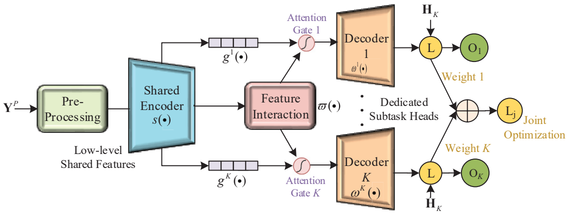


Fig. 4. Multi-task learning framework for joint multi-user channel estimation.

B. Multi-task Learning-based Joint Channel Estimation

As illustrated in Fig. 4, we present the proposed MTL framework for joint cascaded channel estimation, in which the cascaded estimation estimation of each UE is regarded as a subtask, i.e., K subtask heads are constructed in the MTL. The proposed MTL architecture is a low-level shared MTL framework, which can be divided into three parts, i.e., shared features extraction in the bottom of network, features interaction in the intermediate layers, and multi-task heads in the network output layers. In the classic shared-bottom MTL model with K subtasks [32], the individual real-value output $\mathbf{O}_k \in \mathbb{R}^{M\times N\times 2}$ for the k-th subtask $(1 \le k \le K)$ can be represented as

$$\mathbf{O}_k = \omega^k s(\mathbf{Y}^P), \tag{14}$$

where functions $s(\cdot)$ and $\omega^k(\cdot)$ denote the shared-bottom module and the k-th task-specific head, respectively.

For the classic MTL model in (14), each subtask head $\omega^k(\cdot)$ affects other subtasks by only updating common weight parameters in the shared layers $s(\cdot)$, which overlooks the subtask relationships and task-specific functionalities built upon shared representations. To model the task relationships and learn task-specific functionalities built upon shared representations, we introduce the attention gating $g^k(\cdot)$ into the proposed MTL framework, which is given by

$$\mathbf{O}_k = \omega^k \left(g^k \left(s(\mathbf{Y}^P) \right) \odot \varpi \left(s(\mathbf{Y}^P) \right) \right), \tag{15}$$

where function $\varpi(\cdot)$ denotes the feature interaction module. Both $\varpi(\cdot)$ and $g^k(\cdot)$ are designed to capture the specific shared task information for different perspectives, in which $g^k(\cdot)$ of different subtask k is generated by utilizing the attention mechanism. In the section IV, we will elaborate the detailed architecture of the attention mechanism.

C. Task Uncertainty-Based Joint Adaptive Optimization

In the multi-task optimization process, the loss balancing strategy of different subtasks need to be carefully designed to alleviate subtask competition and guarantee the stable convergence of each subtask. For the channel estimation in STAR-RIS system, the ES ratio $\beta_n^f\,(\forall f\in\{t,r\})$ will affect the estimation performance of the transmitting and reflecting cascaded channel. The corresponding channel estimation accuracy can be improved with larger β_n^f and vice versa. In the proposed joint optimization strategy, we utilize the prior knowledge of ES ratio β_n^f to balance the network training, in which the same ES ratio $\beta_n^f = \beta^f (1\le n\le N)$ is set for all STAR-RIS elements in the channel estimation stage. Furthermore, we leverage the task uncertainty to obtain a learnable scalar σ_k for subtask k [33]. Let $f(\cdot)$ denote the proposed MTL model. The DL-based channel estimation can be regarded as the regression task, in which the Gaussian likelihood of the MTL can be given by

$$p\left(\mathcal{H}\mid f(\mathbf{Y}^P)\right) = \mathcal{N}\left(f(\mathbf{Y}^P),\sigma^2\right), \tag{16}$$

where $\mathcal{H} = \{\mathbf{H}_1,\cdots,\mathbf{H}_K\}$, $p\left(\mathcal{H}\mid f(\mathbf{Y}^P)\right)$ denotes the marginal probability, $\sigma = \{\sigma_1,\cdots,\sigma_K\}$ is the set of observation noise scalar for each subtask, and $\mathcal{N}(\cdot)$ represents the Gaussian distribution. Since the output of each subtask is independent and identically distributed (i.i.d.) for the given sufficient statistics, the multi-task likelihood in (16) satisfies

$$p\left(\mathcal{H},\mid f(\mathbf{Y}^P)\right) = p\left(\mathbf{H}_1,\mid f^1(\mathbf{Y}^P)\right)\cdots p\left(\mathbf{H}_K,\mid f^K(\mathbf{Y}^P)\right). \tag{17}$$

The log likelihood of $p\left(\mathbf{H}_k,\mid f^k(\mathbf{Y}^P)\right)$ satisfies [34]

$$\log p\left(\mathbf{H}_k\mid f^k(\mathbf{Y}^P)\right) \propto -\frac{1}{2\sigma_k^2}\left\|\mathbf{H}_k - f^k(\mathbf{Y}^P)\right\|^2 - \log\sigma_k, \tag{18}$$

where $f^k(\cdot)$ denotes the output of subtask k in the proposed MTL model, i.e., $f^k(\mathbf{Y}^P) = \mathbf{O}_k$.

To maximize the Gaussian likelihood $p\left(\mathcal{H},\mid f(\mathbf{Y}^P)\right)$, we minimize the opposite objective function of $\log p\left(\mathcal{H},\mid f(\mathbf{Y}^P)\right)$, which is given by

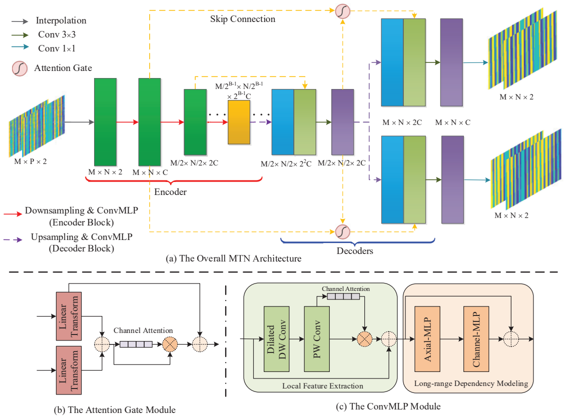
$$\begin{aligned}
&-\log p\left(\mathcal{H},\mid f(\mathbf{Y}^P)\right)\\
&= -\log\left(p\left(\mathbf{H}_1\mid f^1(\mathbf{Y}^P)\right)\cdots p\left(\mathbf{H}_K\mid f^K(\mathbf{Y}^P)\right)\right)\\
&\propto \sum_{i=k}^{K}\frac{1}{2\sigma_k^2}\left\|\mathbf{H}_k - f^k(\mathbf{Y}^P)\right\|^2 + \log\sigma_k,
\end{aligned}$$

where the loss function of each subtask can be set to ℓ_2-norm loss $\mathcal{L}_k = \left\|\mathbf{H}_k - f^k(\mathbf{Y}^P)\right\|^2$ [33].

Considering the unique characteristics of hybrid-field cascaded channel estimation in STAR-RIS systems, we replace ℓ_2-norm loss with ℓ_1-norm loss to achieve the better convergence performance, i.e., $\mathcal{L}_k = |\mathbf{H}_k - f^k(\mathbf{Y}^P)|$, wherein three reasons can be summarized for selecting ℓ_1 norm as the subtask loss function. Specifically, (1) ℓ_2 loss is sensitive to outliers in the spatial non-stationary channel matrix $\mathbf{H}^k$, e.g., the zero-value blocks and hybrid-field components in Fig. 3(d). (2) ℓ_2 loss is unfriendly to the small error between $\mathbf{H}_k$ and $f^k(\mathbf{Y}^P)$ as the network training progresses, resulting in a smaller training step and decreasing the convergence speed for the channel estimation task with the requirement of high precision. (3) The utilization of ℓ_2 loss in regression tasks will cause the losing of the details of target samples. The minimization of ℓ_2 loss leads to the suppression of high frequency details in $\mathbf{H}^k$, e.g., the overlapping intersection of VRs, entailing blurred and over-smoothed channel matrix. In fact, this limitations of ℓ_2 loss have been widely discussed in computer vision filed [35], as well as deep learning-based channel estimations models for various communication scenarios [36], [37], while ℓ_1 loss can provide preferable convergence performance according to the empirical observation. Hence, by aggregating the ES ratio β^f, the joint loss function of the proposed MTL framework can be expressed as

$$\mathcal{L}_{\mathrm{j}}\left(\sigma_k\right) \approx \sum_{i=k}^{K}\frac{1}{2\sigma_k^2}(2-\beta_k)\mathcal{L}_k + \log\sigma_k, \tag{19}$$

where β_k is set to $\beta_k = \beta^t$ if the subtask k is the transmitting channel estimation; otherwise β_k is set to $\beta_k = \beta^r$. Learnable parameter σ_k represents the homoscedastic task uncertainty of each subtask, in which the weight of $\mathcal{L}_k$ decreases as σ_k increases. The last term in (16), $\log\sigma_k$, is a regularization term to penalize the too large σ_k.



(a) The Overall MTN Architecture

(b) The Attention Gate Module

(c) The ConvMLP Module

Fig. 5. Network backbone of the proposed MTN architecture.

IV. Mixed Convolution and MLP-based Multi-task Network Architecture

The proposed MTL framework in Section III provides the basic design guideline for the joint STAR-RIS channel estimation, e.g., the input-to-output mapping relation and the construction of the joint loss function. In this section, we will elaborate the detailed network architecture, i.e., the MTN backbone and basic network components, for the high-dimensional cascaded channel reconstruction in hybrid-field STAR-RIS systems.

A. Overall Network Backbone of the Proposed MTN

Fig. 5 shows the overall network backbone of the proposed MTN architecture, whose design guidelines comply with the proposed MTL framework. In the pre-processing stage of the MTN, we first use the complex-to-real operation to obtain the real-value input tensor $\bar{\mathbf{Y}}^P = \{\Re(\mathbf{Y}^P), \Im(\mathbf{Y}^P)\} \in \mathbb{R}^{M\times P\times 2}$ and the output tensor $\bar{\mathbf{H}}_k = \{\Re(\mathbf{H}_k), \Im(\mathbf{H}_k)\} \in \mathbb{R}^{M\times N\times 2}$. Then, the bi-directional cubic interpolation method is used to carry out the pre-upsampling operation, in which $\bar{\mathbf{Y}}^P$ is upscaled to the feature map $\mathbf{F}^{\mathrm{u}} \in \mathbb{R}^{M\times N\times 2}$. In the bi-directional cubic interpolation operation, we first exhibit the coordination projection between the (X, Y)-th entry of the upsampling channel feature $\mathbf{F}^{\mathrm{u}}$ and the (x, y)-th entry of the low-dimensional $\bar{\mathbf{Y}}^P$. Suppose the upscaling factor is (u_1, u_2) in the horizontal and vertical direction of the feature map, i.e., $(u_1, u_2) = (M/M, N/P) = (1, \Gamma)$ in the pre-upsampling, and the coordination projection of the entry $\mathbf{F}^{\mathrm{u}}(X, Y)$ on $\bar{\mathbf{Y}}^P$ is $(x, y) = (X/u_1, Y/u_2)$. Then, the value of the entry $\mathbf{F}^{\mathrm{u}}(X, Y)$ is determined by the $t\times t$ nearest entries of $\bar{\mathbf{Y}}^P$, which can be expressed as

$$\mathbf{F}^{\mathrm{u}}(X, Y) = \sum_{i=-\frac{t}{2}}^{\frac{t}{2}-1} \sum_{j=-\frac{t}{2}}^{\frac{t}{2}-1} \bar{\mathbf{Y}}^P_{x+i,y+j} \cdot \varepsilon\left(i + \frac{t}{2}\right) * \varepsilon\left(j + \frac{t}{2}\right) \tag{20}$$

where the nearest range is set to $t \times t = 4 \times 4$ in the proposed MTN model. For the edge entries of $\bar{\mathbf{Y}}^P$, i.e., $x < \frac{t}{2}$ or $y < \frac{t}{2}$, we adopt the zero padding operation to compute the corresponding entries in $\mathbf{F}^{\mathrm{u}}$. The basis function $\varepsilon(\cdot)$ denotes the contributing weight of the entry $\bar{\mathbf{H}}^P(x, y)$ for the entry $\mathbf{F}^{\mathrm{u}}(X, Y)$, which is given by

$$\varepsilon(x) = \begin{cases} (a+2)|x|^3 - (a+3)|x|^2 + 1, & \text{for } |x| \le 1, \\ a|x|^3 - 5a|x|^2 + 8a|x| - 4a, & \text{for } 1 < |x| < 2, \\ 0, & \text{otherwise,} \end{cases} \tag{21}$$

where the solvable parameter a is set to $a = 0.5$ [38].

Although the cubic interpolation in the pre-upsampling operation can smoothly fit a given data point without losing the details of the feature maps, the interpolation error will be inevitably introduced for upsampling features. Hence, the encoder-decoder architecture is constructed to learn the low-rank latent representation of the upsampling features. This architecture implicitly characterize the sparsity of the hybrid-field cascaded channel and reduce the required network complexity compared with the flatten network architecture [15]. As shown in Fig. 5(a), the share encoder with B encoder blocks is designed to progressively compress $\mathbf{F}^{\mathrm{u}} \in \mathbb{R}^{M\times N\times 2}$ into the latent representation $\mathbf{F}^{\mathrm{e}} \in \mathbb{R}^{M/2^{B-1}\times N/2^{B-1}\times 2^{B-1}C} (B \le \lfloor log_2 M \rfloor)$, in which C represents the number of feature channels. In the encoder block, the convolutional operations with stride $(\iota_x^b, \iota_y^b) = (2, 2)$ are used to reduce the size of the feature map, i.e., the feature $\mathbf{F}^{\mathrm{e}}_b \in \mathbb{R}^{M/2^{b-1}\times N/2^{b-1}\times 2^{b-1}C}$ is converted into $\mathbf{F}^{\mathrm{e}}_{b+1} \in \mathbb{R}^{M/2^{b}\times N/2^{b}\times 2^{b}C} (1 \le b \le B-1)$. Note that the convolutional stride in the first encoder block is set to $(\iota_x^1, \iota_y^1) = (1, 1)$ so as to enrich the low-level features in the MTN architecture.

Accordingly, K subtask decoders with $B - 1$ decoder blocks are designed to recover the desired cascaded channels $\bar{\mathbf{H}}_k \in \mathbb{R}^{M\times N\times 2}$ from the feature representation set $\mathcal{F}^{\mathrm{e}} = \{\mathbf{F}^{\mathrm{e}}_1, \cdots, \mathbf{F}^{\mathrm{e}}_B\}$ obtained by the encoder. In the b-th decoder block, the cubic interpolation module with the upscaling factor $(u_1^b, u_2^b) = (2, 2)$ is used to carry out the upsampling operations of the feature map $\mathbf{F}^{\mathrm{d}}_b$. Then, a convolutional layer is used to reduce the number of feature channels of $\mathbf{F}^{\mathrm{d}}_b$, i.e., $\mathbf{F}^{\mathrm{d}}_b \in \mathbb{R}^{M/2^{B-b}\times N/2^{B-b}\times 2^{B-b}C}$ is converted into $\mathbf{F}^{\mathrm{d}}_{b+1} \in \mathbb{R}^{M/2^{B-b-1}\times N/2^{B-b-1}\times 2^{B-b-1}C}$. In the decoding stage, the parameters of the first decoder block are shared among S subtask decoders. This operation reduces the parameters and computations of the MTN model, and imposes different decoders to learn the common low-level representation.

Considering the effective information loss during the feature compression in the encoder, we introduce the feature skip connections between the shared encoder and multi-task decoders to concatenate feature maps through the channel dimension. Note that compared to the typical tensor summation operation-based feature skip connections [13], [15], the designed feature concatenation method in this work can preserve the original data distributions of all feature maps and increase the network capacity. Furthermore, we introduce the attention gating in the skip connections to enhance the effective representation obtained by encoder block, which is shown in Fig. 5(b). For the feature map $\mathbf{F}^{\mathrm{e}}_b$ obtained by the b-th encoder block $(2 \le b \le B)$, we use the element-wise addition to compute the fused feature $\bar{\mathbf{F}}_b = \mathbf{F}^{\mathrm{e}}_b + \mathbf{F}^{\mathrm{d}}_{b-1}$ between the encoder block b and the decoder block $b - 1$. Then, we design the improved channel attention mechanism to learn an adaptive weight α. Specifically, we use the global average pooling module to unbend $\bar{\mathbf{F}}_b$ through the spatial dimension so as to determine the feature vector $\mathbf{z} = [z_1, \cdots, z_c, \cdots, z_C] \in \mathbb{R}^{C\times 1}$. Here, the feature z_c can be expressed as

$$z_c = \frac{1}{M\times N} \sum_{m=1}^{M} \sum_{n=1}^{N} \bar{\mathbf{F}}_{b,c}(m, n), \tag{22}$$

where $\bar{\mathbf{F}}_{b,c} \in \mathbb{R}^{M\times N}$ denotes the feature matrix for feature channel c of $\bar{\mathbf{F}}_b$. Next, a linear layer with weight $\mathbf{W}_z \in \mathbb{R}^{C\times C}$ is used to obtain the feature vector $\bar{\mathbf{z}} \in \mathbb{R}^{C\times 1}$. We adopt Tanh activation to constrain the range of specific attention weight $\boldsymbol{\alpha} \in \mathbb{R}^{C\times 1}$, i.e., $\boldsymbol{\alpha}_c = \frac{e^{\bar{\mathbf{z}}_c} - e^{-\bar{\mathbf{z}}_c}}{e^{\bar{\mathbf{z}}_c} + e^{\bar{\mathbf{z}}_c}}$, where $\boldsymbol{\alpha}_c$ and $\bar{\mathbf{z}}_c$ denote the c-th element of $\boldsymbol{\alpha}$ and $\bar{\mathbf{z}}$, respectively. Furthermore, the feature $\bar{\mathbf{F}}_b$ is rescaled with $\boldsymbol{\alpha}$ by the channel-wise multiplication. Then, the residual connection is used to obtain the feature $\hat{\mathbf{F}}_b$ by fusing the low-level feature and the weighted feature, i.e., $\hat{\mathbf{F}}_b = \bar{\mathbf{F}}_b \odot \boldsymbol{\alpha} + \mathbf{F}^{\mathrm{e}}_b$. In this case, the concatenated feature in the b-th skip connection can be expressed as $\mathbf{F}^{\mathrm{c}}_b = \mathrm{Concat}\{\hat{\mathbf{F}}_b, \mathbf{F}^{\mathrm{u}}_b\} \in \mathbb{R}^{M/2^b\times N/2^b\times (2^{b+1}C)}$. Lastly, the convolutional layer with $2^b C$ filters is used to obtain the fused feature map $\mathbf{F}^{\mathrm{f}}_b \in \mathbb{R}^{M/2^b\times N/2^b\times 2^b C}$. For the last layer of the decoder, we use the convolutional layer with two filters to be in accord with the representation of channel matrix with two channel, where the filter size (S_1, S_2) is set to $(S_1, S_2) = (1, 1)$.

In Fig. 5(c), we present the basic feature extraction module in the encoder and decoder blocks, termed as ConvMLP, which is motivated by the network architectures with global feature modeling ability, e.g., Transformer [39] and sparse MLP [40]. The ConvMLP module is divided into the convolutional operations-based local feature extraction module and the MLP-

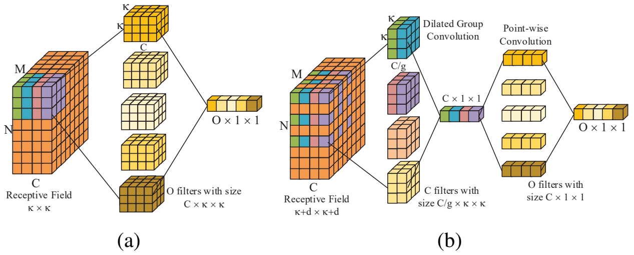


Fig. 6. Convolution operations for channel feature extraction. (a) The standard 2D convolutional layer with O filters; (b) The proposed dilated channel-wise convolution module.

based long-range dependency feature modeling module, which addresses to the local spatial channel correlation for ELAA communications, and the specific spatial non-stationarity for the hybrid-field STAR-RIS systems, respectively.

B. Channel-Wise Convolution-Based Local Feature Extraction

Since the wireless channels between densely packed STAR-RIS elements have obvious local spatial correlations, the two-dimensional convolutional neural network (CNN) is usually utilized as the basic architecture to capture the spatial correlations by utilizing local convolution operations. The standard convolutional (SD Conv) operation with C filters is defined as

$$(\mathbf{W}^{\mathrm{c}} \otimes \mathbf{F})_{m,n} = \sum_{s_1=1}^{S_1} \sum_{s_2=1}^{S_2} \sum_{c=1}^{C} \mathbf{W}^{\mathrm{c}}_{s_1,s_2,c} \cdot \mathbf{F}_{m+s_1,n+s_2,c}, \tag{23}$$

where $\mathbf{F} \in \mathbb{R}^{M\times N\times C}$ and $\mathbf{W}^{\mathrm{c}} \in \mathbb{R}^{S_1\times S_2\times C}$ denote the input feature map and the convolutional filter, respectively.

Fig. 6a visualizes the computing process of the SD Conv layer in the CNN, in which the all channels of the feature map need to participate the convolution operations, thus resulting in the high computing requirements for real-time applications. Moreover, the widely used small-size convolution kernel in the classic CNN architecture, e.g., 3×3 kernel, has a limited receptive field in the feature extraction. To improve the computing efficiency and increase the effective receptive field, we design a dilated channel-wise convolution (DC Conv) module to replace the SD Conv. As illustrated in Fig. 6b, the DC Conv module is divided into the dilated group convolution (DR Conv) and point-wise convolution (PW Conv). In the DR Conv, we design the channel group strategy and the dilated convolutional kernel to extract the preliminary features. The DR Conv operator $\otimes_{\mathrm{dr}}$ can be expressed as

$$(\mathbf{W}^{\mathrm{r}} \otimes_{\mathrm{dr}} \mathbf{F})_{m,n} = \sum_{s_1=1}^{S_1} \sum_{s_2=1}^{S_2} \sum_{c=1}^{C/g} \mathbf{W}^{\mathrm{r}}_{s_1,s_2,c} \cdot \mathbf{F}_{m+dl_1,n+dl_2,c}, \tag{24}$$

where g and d denote the number of convolutional groups and the dilated rate of convolutional kernel, respectively. Compared with the SD Conv, the DR Conv can reduce the computational cost by a factor C/g and the effective receptive field is expanded to $(S_1 + d)(S_2 + d)$.

Due to the non-overlapped feature channel group strategy in the DR Conv operation, the information interaction between channels of the feature map cannot be realized. Hence, we design the PW Conv operator $\otimes_{\mathrm{pw}}$ to aggregate the separable feature channels obtained by the DR Conv, which is given by

$$\left(\mathbf{W}^{\mathrm{p}} \otimes_{\mathrm{pw}} \mathbf{F}\right)_{m,n} = \sum_{c=1}^{C} \mathbf{W}^{\mathrm{p}}_{c} \odot \mathbf{F}_{m,n,c}, \tag{25}$$

where the spatial size of the filter $\mathbf{W}^{\mathrm{p}}$ is set to $(S_1, S_2) = (1, 1)$.

In the DC Conv module, we further introduce the channel attention mechanism to enhance local effective information and suppress other useless components of the feature map, whose network architecture is similar to the attention gating in Fig. 6c. Consequently, the output of the DC Conv module can be expressed as

$$\mathrm{DC}\left(\mathbf{W}^{\mathrm{p}}, \mathbf{W}^{\mathrm{r}}, \boldsymbol{\alpha}^{\mathrm{r}}, \mathbf{F}\right) = \mathbf{W}^{\mathrm{p}} \otimes_{\mathrm{pw}} \left(\mathbf{W}^{\mathrm{r}} \otimes_{\mathrm{dr}} \mathbf{F}\right) + \mathbf{F} \odot \boldsymbol{\alpha}^{\mathrm{r}}, \tag{26}$$

where $\boldsymbol{\alpha}^{\mathrm{r}} \in \mathbb{R}^{C\times 1}$ is an attention weight vector.

Complexity Comparison between SD Conv and DC Conv: In the SD Conv, the space and computational complexity are $O(C^2 S_1 S_2)$ and $O(MNC^2 S_1 S_2)$, respectively. The complexity of the DC Conv is composed of the PW Conv and the DR Conv, whose space and computational complexity can be summarized as $O(C^2(S_1S_2/g+1)$ and $O(MNC^2(S_1S_2/g+1))$, respectively. Since parameter g is set to $g \gg 1$, i.e., $g = C/8$ in the proposed MTN, the DR Conv can realize more efficient and lightweight network architecture than the SD Conv.

C. Axial MLP-Based Global Feature Modeling

In the hybrid-field STAR-RIS systems, the dynamic VRs results in the spatial channel non-stationarity. Hence, the local spatial correlation of the cascaded channel will be partly lost and has uneven distribution, which restricts the feature extraction ability for the local convolutional operations. Although we introduce the dilated convolution operation to increase the receptive field of convolutional layers, the long-range and non-local feature modeling ability is still limited for the DC Conv module. To extract the non-local features of the hybrid-field STAR-RIS channel, we first review a typical global modeling network, i.e., self-attention mechanism-based Transformer model [39]. In the Transformer model, the feature map $\mathbf{F}$ is flatten along the spatial dimension, termed as tokenization, i.e., $\mathbf{F} \in \mathbb{R}^{M\times N\times C} \rightarrow \mathbf{F}^{\mathrm{t}} \in \mathbb{R}^{L\times C}, (L = M \times N)$. Then, different linear transformations are applied to obtain the *Key* matrix $\mathbf{K} \in \mathbb{R}^{L\times D}$, *Query* matrix $\mathbf{Q} \in \mathbb{R}^{L\times D}$ and *Value* matrix $\mathbf{V} \in \mathbb{R}^{L\times D}$, which is given by

$$\mathbf{K} = \mathbf{F}^{\mathrm{t}}\mathbf{W}^{\mathrm{k}}, \tag{27a}$$

$$\mathbf{Q} = \mathbf{F}^{\mathrm{t}}\mathbf{W}^{\mathrm{q}}, \tag{27b}$$

$$\mathbf{V} = \mathbf{F}^{\mathrm{t}}\mathbf{W}^{\mathrm{v}}, \tag{27c}$$

where $\mathbf{W}^{d} \in \mathbb{R}^{C\times D} (\forall d \in \{\mathrm{k}, \mathrm{q}, \mathrm{v}\})$ are the weights of the linear layers. According to the scaled dot-product attention, the output $\mathbf{A} \in \mathbb{R}^{L\times D}$ of the self-attention module is given by

$$\mathbf{A} = \mathrm{Atten}(\mathbf{K}, \mathbf{Q}, \mathbf{V}) = \mathrm{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{T}}{\sqrt{D}}\right) \cdot \mathbf{V} = \mathbf{E} \cdot \mathbf{V}, \tag{28}$$

where $\mathbf{E}$ is termed as the attention matrix, the hyper-parameter D is generally set to $D = C$, and $\mathrm{Softmax}(\cdot)$ denotes the Softmax activation function, i.e., $\mathrm{Softmax}(\mathbf{v}) = \frac{e^{v_i}}{\sum_{i=1}^{L} e^{v_i}}$ for the

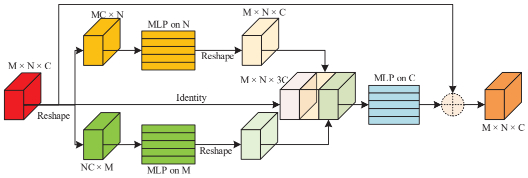


Fig. 7. Information flow in the axial MLP.

vector $\mathbf{v} \in \mathbb{R}^{L\times 1}$. According to (28), an intrinsic bottleneck to the self-attention mechanism is the quadratic complexity $O\left(L^2\right)$ with the token sequence length L in both computation and memory. This problem is particularly obvious for the Transformer empowered ELAA systems, where the channel with the dimension of $M \times N$ after tokenization will produce a very large L.

To realize the global modeling of the non-stationary channel with efficient space and computation complexity, we resort to an axial global modeling framework [40], [41], and design a axial MLP architecture with weaker inductive bias. As illustrated in Fig. 7, the axial MLP is divided into three branches to extract the feature extraction along different spatial directions, i.e., horizontal mixing path, vertical mixing path, and identity mapping path. In the horizontal mixing path, the feature tensor $\mathbf{F} \in \mathbb{R}^{M\times N\times C}$ is reshaped into the feature $\mathbf{F}^{\mathrm{h}} \in \mathbb{R}^{MC\times N}$, and a linear layer with weights $\mathbf{W}^{\mathrm{h}} \in \mathbb{R}^{MC\times N}$ is applied to each of the MC rows of $\mathbf{F}^{\mathrm{h}}$. The similar operation is applied to obtain the feature $\mathbf{F}^{\mathrm{v}} \in \mathbb{R}^{NC\times M}$ in the vertical mixing path, and the linear layer is characterized by weights $\mathbf{W}^{\mathrm{v}} \in \mathbb{R}^{NC\times M}$. Finally, the outputs of the three paths are fused together to produce the output tensor $\mathbf{A}^{\mathrm{a}} \in \mathbb{R}^{M\times N\times C}$. Specifically, in the axial MLP, we reconstruct the linear transformation module in (27) of the Transformer, which is given by

$$\mathbf{K}^{\mathrm{a}}{=}\mathbf{F}^{\mathrm{h}}\mathbf{W}^{\mathrm{h}}, \tag{29a}$$

$$\mathbf{Q}^{\mathrm{a}}{=}\mathbf{F}^{\mathrm{v}}\mathbf{W}^{\mathrm{v}}, \tag{29b}$$

$$\mathbf{V}^{\mathrm{a}}{=}\mathbf{F}^{\mathrm{h}}. \tag{29c}$$

Next, we drop the scaled dot-product attention operation in (28), while adopt the tensor concatenation operation to realize the multi-branch feature fusion as follows.

$$\mathbf{A}^{\mathrm{a}} = \mathrm{Concat}(\mathbf{K}, \mathbf{Q}, \mathbf{V}) \cdot \mathbf{W}^{\mathrm{a}}, \tag{30}$$

where $\mathbf{W}^{\mathrm{a}} \in \mathbb{R}^{3C\times C}$ is a linear layer. In this axial global modeling strategy, we carry out the MLP operations along the STAR-RIS elements domain (horizontal path) and AP antennas domain (vertical path) instead of the entire spatial cascaded channel matrix. The global spatial dependency of the feature map can be obtained by interacting twice with horizontal and vertical tokens [40], [41]. After the spatial modeling by constructing the axial MLP, we add the channel-mixing module to realize the information interaction between channel dimensions of the feature map, which is implemented by two linear layers with non-linear activation function. The feature tensor $\mathbf{A}^{\mathrm{a}}$ is reshaped into the feature $\mathbf{A}^{\mathrm{c}} \in \mathbb{R}^{MN\times C}$ at first. Then, two linear layers are used to obtain the feature $\mathbf{A}^{\mathrm{f}}$, which can be expressed as

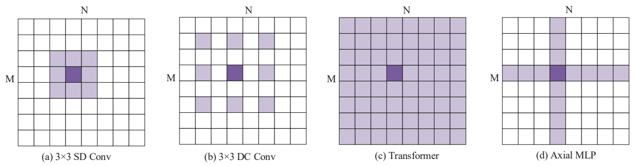


Fig. 8. Receptive field (connectivity pattern) of different network modules. The dark-purple grid denotes the operating target token, while the light-purple grids denote the connecting tokens in the whole feature map.

$$\mathbf{A}^{\mathrm{f}} = \mathrm{LR}(\mathbf{A}^{\mathrm{c}}\mathbf{W}_1) \cdot \mathbf{W}_2 + \mathbf{A}^{\mathrm{c}}, \tag{31}$$

where the first linear layer $\mathbf{W}_1 \in \mathbb{R}^{C\times \upsilon C}$ project the feature $\mathbf{A}^{\mathrm{c}}$ into the high-dimension representation space, and $\upsilon \geq 1$ is a hyper-parameter for ascending dimension of $\mathbf{A}^{\mathrm{c}}$. The second linear layer $\mathbf{W}_2 \in \mathbb{R}^{\upsilon C\times C}$ is used to recover the original channel dimension again. Between two linear layers, we use the LeakeyReLU activation function to provide the non-linearity of feature transformation, which is given by

$$\mathrm{LR}(x) = \begin{cases} x, & x \geq 0, \\ \frac{x}{a}, & x < 0, \end{cases} \tag{32}$$

where $a = (1, +\infty)$ is known as the Leakage value. Compared with the widely used ReLU activation in other channel estimation works [15], the LeakeyReLU can activate the negative components of the channel matrix, which enhances the representation and generalization abilities of ConvMLP.

Complexity Comparison between Transformer and Axial MLP: In the global spatial modeling of the feature map, the computation complexity of Transformer is $O((MN)^2C)$ [39]. The axial MLP adopts the multi-branch architecture to reduce the required computations for the entire $M \times N$ spatial modeling, whose complexity is summarized as $O(MNC(M+N+3C))$. While the computational complexity of the Transformer grows with $(MN)^2$, the computational complexity of the axial MLP grows with $MN\sqrt{MN}$, which is more computational friendly for ELAA communications with larger M and N.

Remark 4: In Fig. 8, we show the spatial receptive field of the cascaded channel feature for different network modules. As illustrated in Fig. 8(a) and Fig. 8(b), the convolutional modules focus on the locality bias of the feature map. By utilizing the dilated convolutional kernel, the proposed DC Conv can obtain larger receptive field without extra computations compared with the SD Conv. Fig. 8(c) shows that the Transformer architecture adopt the directly global spatial modeling strategy for the input feature map. For the axial MLP in Fig. 8(d), the operating target token only interacts with the light-purple tokens along the AP antennas and STAR-RIS elements direction, which significantly reduces the network complexity than the Transformer. In particular, the global receptive field can be obtained by executing twice axial MLP operations.

V. Numerical Results

In this section, we first introduce the simulation setups for the STAR-RIS communication scenarios, hyper-parameters of DL model, and the existing benchmarks. Then, the channel

TABLE I
HYPER-PARAMETERS FOR THE PROPOSED MTN

Hyper-Parameters	Value
Number of encoder block B	3
Number of filters C of the first Conv layer	48
Kernel size of the SD Conv layer $S_1 \times S_2$	3×3
Dilated rate of kernel size d	2
Ascending dimension factor υ	2
Leakey value a	5
Number of training epochs	50
Number of training batchsize	16
Initial learning rate	1×10^{-1}
Final learning rate	1×10^{-5}
Weight decay	5×10^{-4}

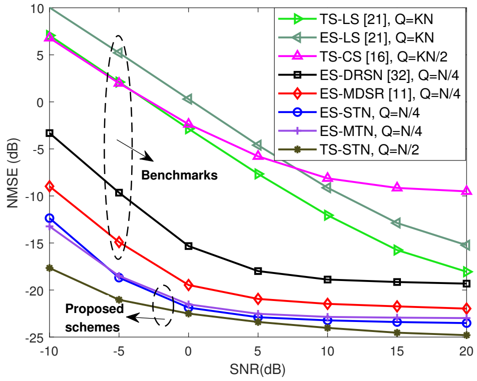


Fig. 9. NMSE performance for different channel estimation schemes.

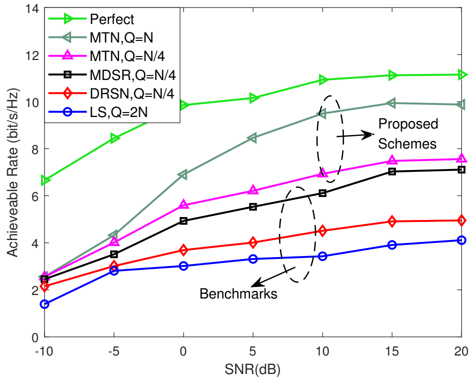


Fig. 10. Achievable sum-rate under the ES protocol for different channel estimation schemes.

estimation performance of the proposed MTN is evaluated under various parameter setups. We design the ablation studies to present the effects of each module in the proposed MTN.

A. Simulation Setups

In the simulation, we set $M = 4 \times 8$, $N^s = 8 \times 64$, $\nu = 2 \times 2$, $N = 4 \times 32$, and $f_c = 28$ GHz. The antenna spacing Δm at the AP and Δn at the STAR-RIS are set to $\Delta m = 2\lambda$ and $\Delta n = \lambda/2$ [24], [27], respectively. In the channel realization, the number of total clusters and scatterers in cluster c follow $C_{\mathrm{s}} \sim \max\{P(1.8), 1\}$ and $S_c \sim \mathcal{U}[1, 30]$, respectively. The VR lengths at the STAR-RIS follow $(\vec{l}_k^1, \vec{l}_k^2) = (\vec{l}_c^{\mathrm{R},1} \sim \mathcal{LN}(0.8, 0.2), \vec{l}_c^{\mathrm{R},2} \sim \mathcal{LN}(3, 0.2))$, while the VR lengths at the AP follow $(\vec{l}_c^{\mathrm{A},1} \sim \mathcal{LN}(0.8, 0.2), \vec{l}_c^{\mathrm{A},2} \sim \mathcal{LN}(1.5, 0.2))$ for cluster c. In the MTN-based joint channel estimation scheme, the required overhead is $Q = P = N/\Gamma$. Unless otherwise specified, we set $\Gamma = 4$, $\beta^t = \beta^r = 0.5$, $K = 2$ and $\rho_k^{\mathrm{u}} = \rho^{\mathrm{a}} = \rho = 0.1$. In the network training, the MTN is optimized with the stochastic gradient descent (SGD) method, in which the cosine learning rate decay schedule is arranged. Table I summarizes the main hyper-parameters for the proposed MTN model. We adopt normalized mean squared error (NMSE) as the performance evaluation metric of the channel estimation, i.e., $\mathrm{NMSE}_{\mathbf{H}_k} = \mathbb{E}\left\{||\hat{\mathbf{H}}_k - \mathbf{H}_k||_F^2 / ||\mathbf{H}_k||_F^2\right\}$. In this work, we compare the proposed MTN model with the following channel estimation benchmarks.

- **LS-based TS/ES protocol [23]:** A typical LS estimator is compared, in which the phase shift of the STAR-RIS are set to the DFT matrix in the pilot transmission stage. Due to the full-rank condition in matrix inversion and the orthogonal pilot transmission strategy, the required minimum pilot overhead is $Q = KN$ for K users.
- **CS-based TS protocol [18]:** A polar domain sparsity-based channel estimation scheme is extended to the hybrid-field channel estimation, in which the orthogonal matching pursuit algorithm is utilized to recover $\mathbf{H}_k$ from the observed signal. The required minimum pilot overhead is $Q = KT$, in which the observed pilot length T is set to $T = N/2$.
- **MDSR-based ES protocol [13]:** A multi-scale deep network (MDSR)-based channel extrapolation framework in RIS systems is extended to the STAR-RIS channel estimation, which adopts the convolution-based residual network architecture. For the DL estimators in the ES protocol, multi-user cascaded channels estimation in the same UG can be realized by utilizing $Q = P$ pilot transmission slots.
- **DRSN-based ES protocol [37]:** A deep residual shrinkage network (DRSN)-based channel estimation model is extended to STAR-RIS systems, which uses the residual shrinkage block as the basic feature extraction module. The required pilot overhead of the DRSN is $Q = P$.

Moreover, we also construct a single-task network (STN) model based on the MTN backbone to show the generalization of the proposed network architecture for the STL framework, where the multiple subtask heads in MTN is modified to a single-task head. In this case, K STN models are required for K cascaded channel estimation tasks. The required pilot overhead of STN is $Q = 2P$ for the TS protocol, while the pilot overhead is $Q = P$ for the ES protocol.

B. Performance Comparison for Different Estimation Schemes

In Fig. 9, we provide the NMSE performance for different channel estimation schemes. As a classic linear estimator, the required pilot overhead of LS estimator need to satisfy $Q \geq KN$, and the channel estimation performance is worse under lower signal-to-noise-ratio (SNR) condition. Since the

TABLE II
TRAINING OVERHEAD FOR DIFFERENT DL MODELS

	FLOPs (G)	**Parameters (M)**	**Pilots**
TS-STN	$1.270 \times K$	$1.180 \times K$	$P \times 2$
ES-STN	$1.270 \times K$	$1.180 \times K$	P
ES-DRSN	2.756	0.821	P
ES-MDSR	3.729	2.666	P
ES-MTN	1.808	1.357	P

effective sparse representation of the non-stationary hybrid-field cascaded channel is hard to characterize, the estimation accuracy of the CS estimator is non-ideal, in which a appropriate pilot overhead $Q = KN/2$ is adopted to provide a acceptable channel estimation accuracy. Thanks to the mighty non-linear mapping ability of the DL model, the DL estimators can obtain better channel estimation accuracy than traditional estimators. For deep learning estimators, i.e., DRSN, MDSR and the proposed MTN model, the pilot overhead is consistently set to $Q = P = KN/4$ to provide a fair comparison among deep learning estimators. Compared with the other DL estimators, the proposed MTN can achieve superior NMSE performance with less training overheads. In the same ES protocol and the network backbone, the MTN architecture can reach similar channel estimation accuracy with the STN model, which shows the good balancing in the multi-task optimization. Since the TS protocol avoids the power leakage effect in the ES protocol, the STN in the TS protocol outperforms other estimators. However, in the TS protocol, since the transmitting and reflecting signal transmission of STAR-RIS are realized at different slots, the required pilot overhead of the STN in the TS protocol is twice of that in the ES protocol.

In Fig. 10, we compare the achievable sum-rate of different channel estimation schemes under the ES protocol. Let $\mathbf{v}_k \in \mathbb{C}^{M\times 1}$ be the normalized precoding vector at the AP for the UE$_k$. The signal-to-interference-plus-distortion-noise-ratio (SIDNR) for the UE$_k$ can be expressed as

$$\gamma_k = \frac{\left|\mathbf{v}_k^T \mathbf{H}_k \boldsymbol{\theta}^k\right|^2}{\mathbf{U}\left|\mathbf{v}_k^T \mathbf{H}_k \boldsymbol{\theta}^k\right|^2 + \sum\limits_{i=1, i\neq k}^{K} (1+\mathbf{U})\left|\mathbf{v}_i^T \mathbf{H}_k \boldsymbol{\theta}^k\right|^2 + \sigma^2}. \quad (33)$$

where $\mathbf{U} = (\rho^{\mathrm{a}})^2 + (\rho_k^{\mathrm{u}})^2$. Accordingly, the achievable sum-rate is given by $\mathcal{R} = \sum_{k=1}^{K} \log_2(1+\gamma_k)$. In this work, the penalty-based alternative optimization framework in [29] is used to jointly optimize $\mathbf{v}_k$ and $\boldsymbol{\theta}_k$ according to the estimated cascaded channel $\hat{\mathbf{H}}_k$. Compared with traditional LS estimator and the other DL estimators, the proposed MTN with less pilot overhead can realize higher achievable sum-rate due to more accurate cascaded channel estimation. With the increase of available pilot overhead Q, the achievable sum-rate of the MTN progressively approaches the ideal sum-rate which is optimized by utilizing the perfect channel $\mathbf{H}_k$.

Table II summarizes the required training overhead for DL-based channel estimation schemes. In the STN estimator, the cascaded channel of each user is independently estimated. Hence, K independent networks need to be trained and saved, which results in more floating point of operations (FLOPs) and parameters than the proposed MTN. Compared with the fully convolution architecture-based DRSN and MDSR model, the proposed MTN introduces the axial MLP architecture to capture the non-local features of the non-stationary cascaded channel, which increases network parameters while improving channel estimation accuracy. However, the advantage of the MLP architecture is only composed of direct matrix multiplication routines. Consequently, the FLOPs of the proposed MTN is less than the DRSN and the MDSR.

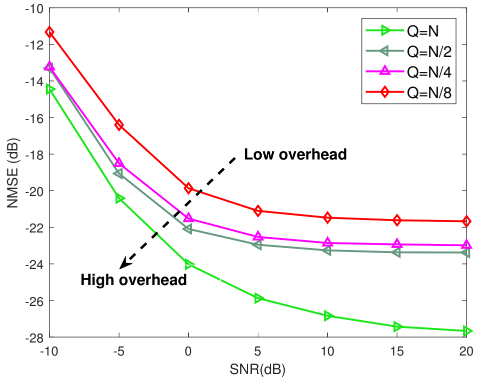


Fig. 11. NMSE performance of MTN for different pilot overhead Q.

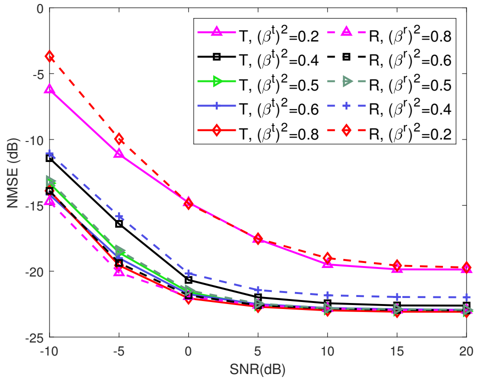


Fig. 12. NMSE performance of MTN under different ES ratios.

C. Generalization Performance for the Proposed MTN

In Fig. 11, the NMSE performance under different pilot overhead Q is shown, where we adjust the number of UBs U in the proposed MTN to match different Q. With the increase of Q, the input tensor $\mathbf{Y}^p$ contains more unknown entries of the cascaded channel matrix, which reduces the required upscaling factor Γ for the MTN model. Hence, the channel estimation accuracy of the proposed MTN is further improved. Note that when Γ is set to $\Gamma = 1$, i.e., $Q = N$, the multi-task channel extrapolation becomes a cascaded channel separation task from the mixture transmitting and reflecting channels with noise.

In Fig. 12, we provide the transmitting (T) and reflecting (R) channel estimation performance of the proposed MTN under different ES ratios $\beta^f\, (\forall f \in \{t, r\})$. When larger β^f is allocated to the transmitting or reflecting modes, the received

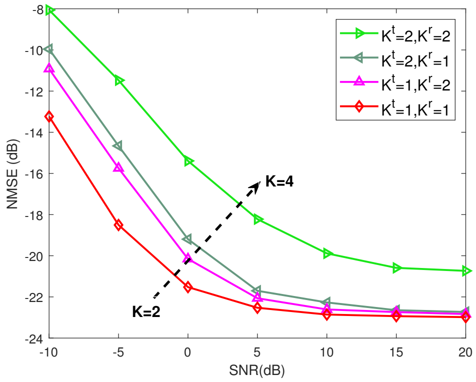


Fig. 13. NMSE performance of MTN for different number of users K.

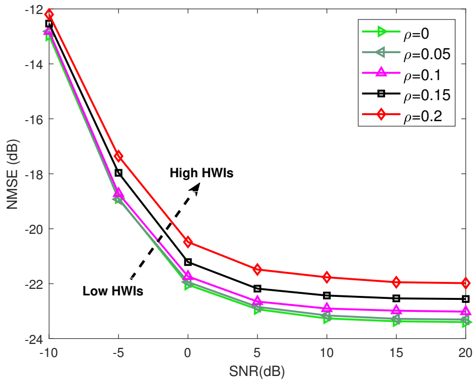


Fig. 14. NMSE performance of MTN under different levels of HWIs ρ.

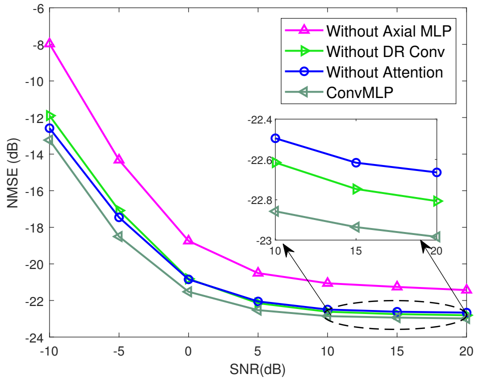


Fig. 15. NMSE performance for different variants of MTN.

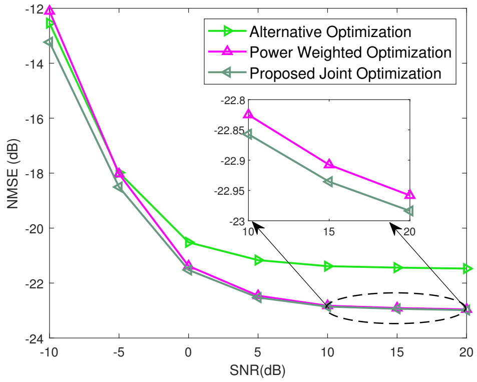


Fig. 16. NMSE performance for different optimization strategies.

pilot signal will involve more transmitting or reflecting signal components. Accordingly, the cascaded channel estimation performance of UE^f is also improved. However, it is worth noting that the equal ES ratio, i.e. $(\beta^r)^2 = (\beta^t)^2 = 0.5$, can ensure the high-quality channel estimation performance for both transmission and reflection. Note that in the beamforming stage of the STAR-RIS, the ES ratios β^f will be optimized.

In Fig. 13, we present the NMSE performance of the proposed MTN under different number of UE^f K in the same UG. As the increase of K, more cascaded channels need to be estimated simultaneously. Since the learning burden of the MTN is further increased, the overall channel estimation accuracy will be degraded. Due to the definitive capability of deep neural networks, the certain network capacity of the MTN is hard to perfectly match the increasing channel reconstruction tasks. Note that the reduced pilot overhead of the MTN is more pronounced than the STN estimator as the increase of K. When the total number of UEs $K = K^t + K^r$ is the same, the NMSE of channel estimation has slight differences for different K^t and K^r in a UG, such as the case of $(K^t = 1, K^r = 2)$ and $(K^t = 2, K^r = 1)$ in Fig. 13. The reason for this phenomenon is that the channel differences among the UEs at different geospatial locations change the correlations of different subtasks. However, the subtask correlations in the MTN will affect the overall channel estimation performance.

In Fig. 14, we provide the NMSE performance of the proposed MTN under different levels of HWIs ρ. To present the network generalization, the MTN model is trained with the fixed HWIs $\rho = 0.1$, while the trained MTN model is verified under different levels of HWIs in the test stage. By learning the latent representation of massive communication data, a robust nonlinear mapping between received pilots and desired cascaded channels is constructed for deep learning estimators. Consequently, the proposed MTN model has the robustness for the disturbances imposed on the input data, e.g., for the case of the varying HWIs, and hence shows satisfactory channel estimation performance even for the large HWIs $\rho = 0.2$.

D. Ablation Studies for the Proposed MTN

In Fig. 15, we present the NMSE performance of three variants of the proposed MTN by designing the ablation studies, which verifies the positive effects of each module for hybrid-field cascaded channel estimation in STAR-RIS systems. Compared with the DR Conv and the attention module in the MTN, the introduction of the axial MLP module can obtain more obvious performance improvements for channel estimation. Note that the similar insight has been presented in [22] for the MLP module. However, in the proposed MTN, we exploit the axial MLP architecture to significantly reduce the redundant parameters than the MLP in [22].

Fig. 16 compares the channel estimation performance of the proposed MTN under different optimization strategies, i.e., the alternative optimization, the power weighted optimization and the proposed joint optimization. In the alternative optimization, K subtask loss functions are independently optimized, while the loss function is aggregated with the fixed ES ratio β^f in the power weighted optimization. The proposed joint optimization strategy not only considers the ES prior information but also introduces a adaptive scalar δ_s for subtask s, which can balance the multi-task training process of the MTN. Consequently, the proposed joint optimization strategy outperforms other multi-task optimization strategies.

VI. Conclusions

By exploiting the ability to simultaneously tune transmission and reflection coefficients of the metasurface, the STAR-RISs provide a promising paradigm to realize the *full-space* SREs. In this work, we proposed an MTL-based joint cascaded channel estimation framework by utilizing the channel correlations between different users and between different STAR-RIS elements. Furthermore, based on the proposed MTL framework, an efficient MTN architecture was developed to realize the precise high-dimensional cascaded channel reconstruction. In the proposed MTN architecture, we exploited the ConvMLP module to capture the local spatial features and the long-range dependency of the hybrid-field cascaded channel. Compared to existing state-of-the-art benchmarks, the proposed MTN can realize superior channel estimation accuracy with less training overhead, whose pilot overhead is independent to the number of users within a UG. In future works, we will explore the enhanced MTN models with better scalability to deal with massive user scenarios in STAR-RIS systems.

References

[1] J. Xiao, J. Wang, Y. Liu, W. Xie, J. Wang, and S. Liu, “Multi-task learning based channel estimation for hybrid-field STAR-RIS systems,” in *Proc. IEEE GLOBECOM*, Dec. 2023.

[2] M. Di Renzo, A. Zappone, M. Debbah, M.-S. Alouini, C. Yuen, J. de Rosny, and S. Tretyakov, “Smart radio environments empowered by reconfigurable intelligent surfaces: How it works, state of research, and the road ahead,” *IEEE J. Sel. Areas Commun.*, vol. 38, no. 11, pp. 2450–2525, Nov. 2020.

[3] Y. Liu, X. Mu, J. Xu, R. Schober, Y. Hao, H. V. Poor, and L. Hanzo, “STAR: Simultaneous transmission and reflection for 360° coverage by intelligent surfaces,” *IEEE Wireless Commun.*, vol. 28, no. 6, pp. 102–109, May. 2021.

[4] X. Li, Y. Zheng, M. Zeng, Y. Liu, and O. A. Dobre, “Enhancing secrecy performance for STAR-RIS NOMA networks,” *IEEE Trans. Veh. Technol.*, vol. 72, no. 2, pp. 2684–2688, Feb. 2023.

[5] X. Li, Z. Tian, W. He, G. Chen, M. C. Gursoy, S. Mumtaz, and A. Nallanathan, “Covert communication of STAR-RIS aided NOMA networks,” *IEEE Trans. Veh. Technol.*, pp. 1–6, 2024.

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

[7] X. Mu, J. Xu, Y. Liu, and L. Hanzo, “Reconfigurable intelligent surface-aided near-field communications for 6G: Opportunities and challenges,” *IEEE Veh. Technol. Mag.*, vol. 19, no. 1, pp. 65–74, Mar. 2024.

[8] 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.

[9] X. Mu, Y. Liu, L. Guo, J. Lin, and R. Schober, “Simultaneously transmitting and reflecting (STAR) RIS aided wireless communications,” *IEEE Trans. Wireless Commun.*, vol. 21, no. 5, pp. 3083–3098, May 2022.

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

[11] J. Xu, X. Mu, and Y. Liu, “Exploiting STAR-RISs in near-field communications,” *IEEE Trans. Wireless Commun.*, vol. 23, no. 3, Mar. 2024.

[12] 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.

[13] Y. Jin, J. Zhang, X. Zhang, H. Xiao, B. Ai, and D. W. K. Ng, “Channel estimation for semi-passive reconfigurable intelligent surfaces with enhanced deep residual networks,” *IEEE Trans. Veh. Technol.*, vol. 70, no. 10, pp. 11 083–11 088, Oct. 2021.

[14] J. Xiao, J. Wang, Z. Wang, W. Xie, and Y. Liu, “Multi-scale attention based channel estimation for RIS-aided massive MIMO systems,” *IEEE Trans. Wireless Commun.*, Nov. 2023.

[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. 2021.

[16] J. Chen, Y.-C. Liang, H. V. Cheng, and W. Yu, “Channel estimation for reconfigurable intelligent surface aided multi-user mmWave MIMO systems,” *IEEE Trans. Wireless Commun.*, 2023.

[17] 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.

[18] J. Wu, S. Kim, and B. Shim, “Near-field channel estimation for RIS-assisted wideband Terahertz systems,” in *Proc. IEEE GLOBECOM*. IEEE, Jan. 2023, pp. 3893–3898.

[19] S. Yang, W. Lyu, Z. Hu, Z. Zhang, and C. Yuen, “Channel estimation for near-field XL-RIS-aided mmWave hybrid beamforming architectures,” *IEEE Trans. Veh. Technol.*, 2023.

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

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

[22] J. Xiao, J. Wang, Z. Chen, and G. Huang, “U-MLP based hybrid-field channel estimation for XL-RIS assisted millimeter-wave MIMO systems,” *IEEE Wireless Commun. Lett.*, vol. 12, no. 6, pp. 1042–1046, Jun. 2023.

[23] C. Wu, C. You, Y. Liu, X. Gu, and Y. Cai, “Channel estimation for STAR-RIS-aided wireless communication,” *IEEE Commun.Lett.*, vol. 26, no. 3, pp. 652–656, Mar. 2022.

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

[25] “3GPP TR 38.901 V16.1.0 - Study on channel model for frequencies from 0.5 to 100 GHz,” Dec. 2019.

[26] Y. Liu, Z. Wang, J. Xu, C. Ouyang, X. Mu, and R. Schober, “Near-field communications: A tutorial review,” *IEEE Open J. Commun. Soc.*, vol. 4, no. 1999-2049, Aug. 2023.

[27] V. C. Rodrigues, A. Amiri, T. Abrão, E. de Carvalho, and P. Popovski, “Low-complexity distributed XL-MIMO for multiuser detection,” in *Proc.IEEE ICC Workshops*, Jun. 2020, pp. 1–6.

[28] A. Amiri, S. Rezaie, C. N. Manchon, and E. De Carvalho, “Distributed receiver processing for extra-large MIMO arrays: A message passing approach,” *IEEE Trans. Wireless Commun.*, vol. 21, no. 4, pp. 2654–2667, Apr. 2021.

[29] Z. Wang, X. Mu, Y. Liu, and R. Schober, “Coupled phase-shift STAR-RISs: A general optimization framework,” *IEEE Wireless Commun. Lett.*, Feb. 2023.

[30] 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, Sep. 2021.

[31] Z. Wang, L. Liu, and S. Cui, “Channel estimation for intelligent reflecting surface assisted multiuser communications: Framework, algorithms, and analysis,” *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6607–6620, Oct. 2020.

[32] S. Ruder, “An overview of multi-task learning in deep neural networks,” *arXiv preprint arXiv:1706.05098*, 2017.

[33] R. Cipolla, Y. Gal, and A. Kendall, “Multi-task learning using uncertainty to weigh losses for scene geometry and semantics,” in *Proc. IEEE/CVF CVPR*, Dec. 2018, pp. 7482–7491.

[34] A. Albert and J. A. Anderson, “On the existence of maximum likelihood estimates in logistic regression models,” *Biometrika*, vol. 71, no. 1, pp. 1–10, Apr. 1984.

[35] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, “Loss functions for image restoration with neural networks,” *IEEE Trans. Comput. Imaging*, vol. 3, no. 1, pp. 47–57, Mar. 2017.

[36] 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.

[37] W. Xie, J. Xiao, P. Zhu, C. Yu, and L. Yang, “Multi-task learning-based channel estimation for RIS assisted multi-user communication systems,” *IEEE Commun.Lett.*, vol. 26, no. 3, pp. 577–581, Mar. 2022.

[38] N. A. Dodgson, “Quadratic interpolation for image resampling,” *IEEE Trans. Image Process.*, vol. 6, no. 9, pp. 1322–1326, Sep. 1997.

[39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Proc. NeurIPS*, vol. 30, Nov. 2017.

[40] C. Tang, Y. Zhao, G. Wang, C. Luo, W. Xie, and W. Zeng, “Sparse MLP for image recognition: Is self-attention really necessary?” in *Proc. AAAI*, vol. 36, no. 2, Jun. 2022, pp. 2344–2351.

[41] J. Ho, N. Kalchbrenner, D. Weissenborn, and T. Salimans, “Axial attention in multidimensional transformers,” *arXiv preprint arXiv:1912.12180*, 2019.

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 pursuing 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.

Jun Wang received the M.S., and Ph.D. degrees in communication engineering from Huazhong Normal University, Wuhan, China, in 2007, and 2016, respectively. He is currently a vice chief engineer of mobile communication business department, CICT Mobile Communication Technology Co., Ltd, Wuhan, China. His research interests include wireless communication algorithm, such as channel estimation, equalizer, encoding/decoding, etc.

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, China. His research interests include communication and control algorithms.

Yuanwei Liu (S’13-M’16-SM’19-F’24, http://www.eecs.qmul.ac.uk/~yuanwei) 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, near field communications, integrated sensing and communications, and machine learning.

Yuanwei Liu is a Fellow of the IEEE, a Fellow of AAIA, a Web of Science Highly Cited Researcher, an IEEE Communication Society Distinguished Lecturer, an IEEE Vehicular Technology Society Distinguished Lecturer, the rapporteur of ETSI Industry Specification Group on Reconfigurable Intelligent Surfaces on work item of “Multi-functional Reconfigurable Intelligent Surfaces (RIS): Modelling, Optimisation, and Operation”, and the UK representative for the URSI Commission C on “Radio communication Systems and Signal Processing”. 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, the 2022 IEEE SPCC-TC Best Paper Award, the 2023 IEEE ICCT Best Paper Award, and the 2023 IEEE ISAP Best Emerging Technologies 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, IEEE Transactions on Network Science and Engineering, and IEEE Transactions on Communications (2018-2023). 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, the Panel Co-Chair for IEEE WCNC 2024, Symposium Co-Chair for several flagship conferences such as IEEE GLOBECOM, ICC and VTC. He serves the academic Chair for the Next Generation Multiple Access Emerging Technology Initiative, vice chair of SPCC and Technical Committee on Cognitive Networks (TCCN).