HCLN: A Hierarchical Collaborative Learning Network for UAV-STAR-RIS-Assisted Wireless Communications

Junjie Xu[1], Bin Li[2], Yinqiu Liu[3], Ruichen Zhang[3], Dongdong Yang[2], Dusit Niyato[3]

[1]Waterford Institute, Nanjing University of Information Science and Technology, Nanjing 210044, China
[2]School of Computer Science, Nanjing University of Information Science and Technology, Nanjing 210044, China
[3]The College of Computing and Data Science, Nanyang Technological University, Singapore 639798
202383890027@nuist.edu.cn, bin.li@nuist.edu.cn, yinqiu001@ntu.edu.sg, ruichen.zhang@ntu.edu.sg, 202312200024@nuist.edu.cn, dniyato@ntu.edu.sg.

Abstract

The integration of simultaneously transmitting and reflecting reconfigurable intelligent surfaces (STAR-RIS) with unmanned aerial vehicles (UAVs) exhibits great potential to extend coverage and enhancing connectivity in space–air–ground 6G networks. However, aerial STAR-RIS (ASTAR-RIS) performance is highly sensitive to UAV orientation fluctuations, which distort the effective channel. Moreover, the joint design of base station beamforming, ASTAR-RIS phase shifts, and amplitude coefficients results in a high-dimensional and nonconvex problem. Classical convex formulations exhibit high complexity and significant resource costs, while existing deep learning methods exhibit limited robustness in dynamic environments. To address these challenges, we propose a hierarchical collaborative learning network (HCLN) that combines deep reinforcement learning (DRL) for outer-layer orientation control with manifold-based meta-gradient learning (M2GL) for inner-layer beamforming, phase shifts, and amplitude optimization. Simulation results indicate that HCLN improves the weighted sum rate by up to 6.7% over non-learning optimization approaches and by approximately 10.1% over traditional learning-based methods. Furthermore, it delivers more stable and robust performance across diverse scenarios, confirming its effectiveness for dynamic ASTAR-RIS-assisted communications.

Introduction

The evolution of sixth-generation (6G) wireless networks calls for programmable environments that can flexibly reconfigure electromagnetic propagation. Reconfigurable intelligent surfaces (RIS) have emerged as a promising enabler, while conventional RIS are fundamentally limited to half-space operation (Liu et al. 2021). To overcome this restriction, simultaneously transmitting and reflecting RIS (STAR-RIS) enables full-space coverage. However, the potential of terrestrial STAR-RIS systems is still limited by their fixed placement and the unpredictability of ground propagation. Unmanned aerial vehicles (UAVs), on the other hand, possess intrinsic mobility and altitude agility, enabling the establishment of stable line-of-sight (LoS) links and real-time adaptation to network dynamics. Motivated by these capabilities, the integration of STAR-RIS with UAV platforms has led to the development of the aerial STAR-RIS (ASTAR-RIS) architecture. Through spatial repositioning, ASTAR-RIS adapts its operation to user distribution and channel variations, effectively improving coverage and link robustness while preserving efficient resource utilization within space-air-ground integrated networks (Wang et al. 2025).

Despite this potential, ASTAR-RIS deployments face two critical challenges. First, due to aerodynamic disturbances and control latency, ASTAR-RIS orientation exhibits small yet persistent Euler angle drift, commonly termed self rotation. These fluctuations perturb the angles of arrival and departure as well as phase coherence, thereby reducing communication quality (Zeng et al. 2021). Second, since end-to-end link gain is jointly determined and mutually constrained by the ASTAR-RIS orientation, the base station (BS) beamforming, and ASTAR-RIS coefficients, any perturbation to one variable reshapes the equivalent channel, which invalidates divide and conquer designs. Therefore, a unified joint optimization framework is essential. However, as the antenna array scales, the decision space grows combinatorially, combined with ASTAR-RIS constraints, the objective becomes nonconvex. As a result, conventional convex-optimization becomes high cost, while existing deep learning strategies tend to break down in dynamic settings due to limited robustness (Zhu et al. 2024).

To address these issues, we propose a hierarchical collaborative learning network (HCLN). Specifically, HCLN consists of an outer decision-making layer and an inner optimization layer. The former follows an agent–environment interaction loop, formulates continuous orientation control as a Markov decision process (MDP), and solves it via deep reinforcement learning (DRL), whereby the agent maps observed states to continuous attitude updates. The latter ingests the outer layer's controls and presents a novel manifold-based meta-gradient learning (M2GL) (Mai, Chen, and Du 2021). Specifically, it applies manifold learning for dimensionality reduction, recasts the resulting low-dimensional representation into gradient-form signals, and feeds them to a meta-learning routine to rapidly refine BS beamforming and STAR-RIS coefficients. The inner-layer performance gains are fed back to the outer layer to shape rewards, thereby enabling a closed-loop collaboration between different learning layers. Our main contributions are threefold:

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

- We model an ASTAR-RIS-assisted communication system in which UAV self-rotation inevitably causes orientation fluctuations. To fully capture these effects, we formulate a joint optimization problem over ASTAR-RIS orientation, BS beamforming, ASTAR-RIS phase shifts, and amplitude coefficients.
- To address this high-dimensional and non-convex problem, we introduce a novel HCLN. In this design, the outer layer transforms the continuous orientation optimization into an MDP, which is solved by DRL. The resulting orientation decisions, together with instantaneous channel states, are then passed to the inner layer, where M2GL refines the beamforming, phase shifts, and amplitude coefficients.
- Extensive simulations show that HCLN outperforms DDPG + M2GL and the greedy policy by approximately 10.1% and 6.7%, respectively. Furthermore, it achieves stable and robust performance across diverse scenarios, underscoring its superiority in efficiency reliability.

Related Work

UAV-Assisted RIS/STAR-RIS Communications

Integrating RIS/STAR-RIS into UAV communications has emerged as a promising approach to enhance coverage and connectivity. (Liu et al. 2022a) optimizes aerial RIS trajectory and power allocation to improve average downlink throughput, while (Liu et al. 2024) extends this by incorporating elevation angle effects into the joint design of resource allocation, RIS phase shifts, and UAV trajectory to enhance fairness. Moreover, (Zhao et al. 2025) introduces ASTAR-RIS and demonstrates its superiority over conventional RIS, and (Guo et al. 2024), building on this, examines mobile scenarios and shows that ASTAR-RIS improves secrecy energy efficiency against eavesdropping. In parallel, (Xiao et al. 2024) proposes a STAR-RIS-UAV scheme for mobile edge computing networks to enable simultaneous task offloading and communication, thereby improving overall system efficiency. Although ASTAR-RIS achieves full-space coverage, most works focus on trajectory and coefficient optimization while neglecting orientation variations that crucially impact communication performance.

Orientation Control for RIS/STAR-RIS

RIS orientation has been shown to critically influence system performance, since the effective reflecting area directly affects the signal-to-noise ratio (Cui et al. 2021). To this end, orientation optimization strategies have been proposed to maintain specular reflection and maximize received power (Cheng et al. 2022a). More recently, RIS orientation has also been studied in low earth orbit satellite communications, where joint optimization of passive beamforming and orientation significantly boosts downlink throughput (Zheng et al. 2022). Despite recent advances, the orientation control of ASTAR-RIS has received limited attention, even though UAV self-rotation inevitably induces significant channel variations and degrade system performance.

Optimization Theories for ASTAR-RIS

The joint optimization of BS beamforming and STAR-RIS coefficients is inherently a high-dimensional and non-convex problem, which has been shown to be NP-hard (Abeywickrama, Zhang, and Yuen 2020). Conventional approaches include fractional programming (Guo et al. 2020), alternating optimization (Wu and Zhang 2019), and semidefinite relaxation (Zhang et al. 2024). Although effective, these methods suffer from high complexity and significant computational overhead. To address these limitations, deep learning has been introduced for STAR-RIS optimization (Lin and Zhu 2020). Neural networks (NNs) can replace parts of the iterative process, significantly reducing computational overhead while maintaining satisfactory performance in high-dimensional optimization. Nevertheless, these methods are heavily dependent on training data and often lack robustness in heterogeneous scenarios (Liu et al. 2022b). To further improve adaptability, meta-learning has been proposed to enable rapid model updates and reduce retraining overhead in dynamic STAR-RIS-assisted networks (Su, Wang, and Zhang 2025). However, despite these advances, existing methods still struggle with the multiple coupled variables in high-dimensional spaces and lack robustness to ASTAR-RIS orientation variations.

System Model and Problem Formulation

We consider an ASTAR-RIS-assisted wireless communication system consisting of an M antenna BS, K single-antenna user devices (UDs), and a UAV-mounted STAR-RIS with N elements, as illustrated in Fig. 1. For tractable analysis, the UAV is assumed to fly along a fixed circular trajectory at a constant altitude H. The total operation time L is equally divided into T slots. During each slot, the orientation of the ASTAR-RIS and the positions of the UDs jointly determine whether a UD falls into the transmission or reflection region, thereby ensuring full coverage of all K UDs. Specifically, $\mathcal{K}_{\mathrm{t}}$ and $\mathcal{K}_{\mathrm{r}}$ are the sets of UDs in the transmission and reflection regions, respectively. They are disjoint and cover all UDs as follows:

$$\mathcal{K}_{\mathrm{t}} \cap \mathcal{K}_{\mathrm{r}} = \varnothing, \qquad \mathcal{K}_{\mathrm{t}} \cup \mathcal{K}_{\mathrm{r}} = \mathcal{K}. \tag{1}$$

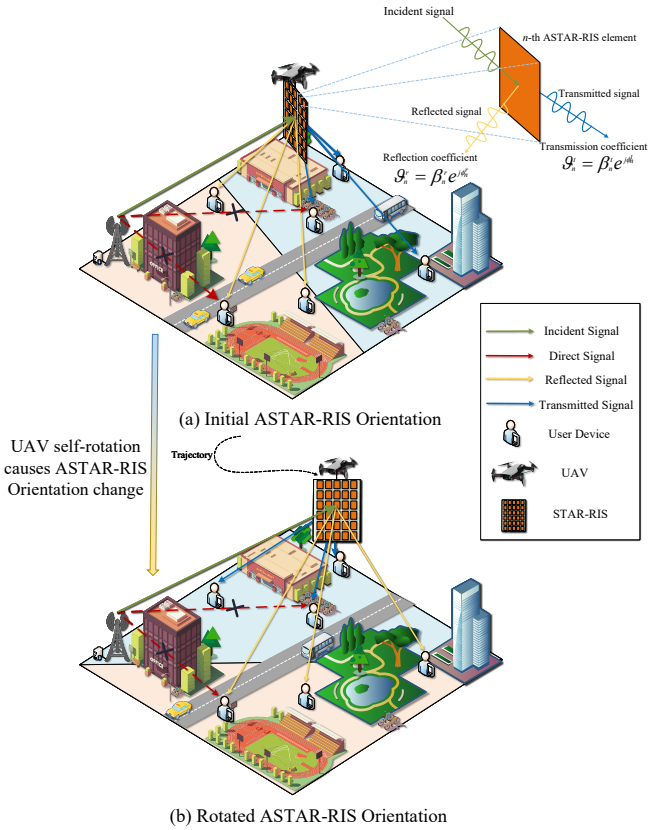
For each element of the ASTAR-RIS, we introduce the coefficient matrices to parameterize the surface as follows:

$$\mathbf{\Theta}^{\kappa} = \mathrm{diag}\left(\beta_1^{\kappa} e^{j\phi_1^{\kappa}}, \ldots, \beta_N^{\kappa} e^{j\phi_N^{\kappa}}\right), \kappa \in \{\mathrm{t/r}\}. \tag{2}$$

Considering that STAR-RIS introduces two coupled coefficients, namely the amplitudes and phase shifts, direct joint optimization is challenging. Therefore, we decompose its coefficient matrices to decouple these effects and enable tractable optimization, i.e.,

$$\mathbf{\Theta}^{\kappa} = \mathbf{B}^{\kappa}\mathbf{\Phi}^{\kappa}, \tag{3}$$

where $\mathbf{B}^{\kappa} = \mathrm{diag}\big(\beta_1^{\kappa}, \ldots, \beta_N^{\kappa}\big) \in \mathbb{R}^{N\times N}$ and $\mathbf{\Phi}^{\kappa} = \mathrm{diag}\big(e^{j\,\phi_1^{\kappa}}, \ldots, e^{j\,\phi_N^{\kappa}}\big) \in \mathbb{C}^{N\times N}$, the reflection/transmission coefficients of the n-th element is denoted as $\vartheta_n^{\kappa} = \beta_n^{\kappa} e^{j\,\phi_n^{\kappa}}$.



(a) Initial ASTAR-RIS Orientation

(b) Rotated ASTAR-RIS Orientation

Figure 1: Illustration of ASTAR-RIS assisted wireless communications under orientation variations.

Moreover, the UAV introduces a yaw angle ψ, which modifies the effective azimuth of the panel and, in turn, influences the cascaded channel gains through its effect on the angle of arrival (AOA) and angle of departure (AOD).

The BS-ASTAR-RIS and ASTAR-RIS-UD links follow Rician fading (Thangasamy et al. 2022). For the n-th element, we obtain

$$\mathbf{h}_n = \sqrt{\rho_{b,n}}\left(\sqrt{\tfrac{K_1}{1+K_1}}\bar{\mathbf{h}}_n + \sqrt{\tfrac{1}{1+K_1}}\tilde{\mathbf{h}}_n\right), \tag{4}$$

$$v_n = \sqrt{\rho_{n,k}}\left(\sqrt{\tfrac{K_2}{1+K_2}}\bar{v}_n + \sqrt{\tfrac{1}{1+K_2}}\tilde{v}_n\right), \tag{5}$$

where $\bar{\mathbf{h}}_n$ and $\bar{v}_n$ denote LoS components, while the non-line-of-sight (NLoS) components $\tilde{\mathbf{h}}_n$ and $\tilde{v}_n$ follow a circularly symmetric complex Gaussian distribution with zero mean and unit variance. Additionally, K_1 and K_2 are the Rician factors. To capture angular dependence, the path gain $\rho_{b,n}$ and $\rho_{n,k}$ are modeled as

$$\rho = \frac{\rho_0}{d^z} G_1\big(\theta^{\mathrm{AOD}}, \varphi^{\mathrm{AOD}}\big)\, G_2\big(\theta^{\mathrm{AOA}}, \varphi^{\mathrm{AOA}}\big). \tag{6}$$

Here d is the link distance, ρ_0 is the path-loss constant at the reference distance $d = 1$ m, $(\theta^{\mathrm{AOD}}, \varphi^{\mathrm{AOD}})$ and $(\theta^{\mathrm{AOA}}, \varphi^{\mathrm{AOA}})$ represent the azimuth and elevation angles of AODs and AOAs, respectively. $G_1(\cdot)$ and $G_2(\cdot)$ represent the directional gains at the transmitter and receiver (Cheng et al. 2022b).

As the direct BS-UD link is blocked, ASTAR-RIS is employed to assist all transmissions. The resulting effective cascaded channel for UD k, denoted by $\mathbf{g}_k$, accounts for both propagation hops and the STAR-RIS coefficients. The signal-to-interference-plus-noise ratio (SINR) of UD k is given by

$$\mathrm{SINR} = \frac{|\mathbf{g}_k \mathbf{w}_k|^2}{\sum_{i\neq k} |\mathbf{g}_k \mathbf{w}_i|^2 + \sigma^2}, \tag{7}$$

with $\mathbf{w}_k$ denoting k-th column of the BS beamforming matrix $\mathbf{W} \in \mathbb{C}^{M\times K}$. The weighted sum rate across T slots is

$$R(\mathbf{W}, \mathbf{B}^{\kappa}, \mathbf{\Phi}^{\kappa}, \mathbf{\Psi}) = \sum_{t=1}^{T}\sum_{k=1}^{K} \bar{\omega}_k \log_2\big(1 + \mathrm{SINR}\big). \tag{8}$$

We aim to maximize this rate by jointly optimizing BS beamforming $\mathbf{W}$, the amplitude $\mathbf{B}^{\kappa}$ and phase shifts $\mathbf{\Phi}^{\kappa}$ of the ASTAR-RIS, and its orientation $\mathbf{\Psi}$. The problem can be formulated as

$$\max_{\mathbf{W},\mathbf{B}^{\kappa},\mathbf{\Phi}^{\kappa},\mathbf{\Psi}} \quad R(\mathbf{W}, \mathbf{B}^{\kappa}, \mathbf{\Phi}^{\kappa}, \mathbf{\Psi}) \tag{9a}$$

$$\text{s.t.} \quad \mathrm{Tr}(\mathbf{W}^H\mathbf{W}) \le P_{\mathrm{BS}}^{\max}, \tag{9b}$$

$$(\beta_n^{\mathrm{r}})^2 + (\beta_n^{\mathrm{t}})^2 = 1,\ \forall n \in \mathcal{N}, \tag{9c}$$

$$\phi_n^{\mathrm{r}}, \phi_n^{\mathrm{t}} \in [0, 2\pi)\ \forall n \in \mathcal{N}, \tag{9d}$$

$$\psi \in [-\tfrac{\pi}{2}, \tfrac{\pi}{2}], \tag{9e}$$

where (9b) enforces the transmit power budget, with $P_{\mathrm{BS}}^{\max}$ denoting the maximum total power, (9c) bounds the transmission and reflection amplitude coefficients of each ASTAR-RIS element, (9d) specifies the admissible range of phase shifts, and (9e) defines the feasible set of ASTAR-RIS orientation.

Proposed HCLN Framework

Motivation and Hierarchical Idea

Solving the problem defined in (9) is highly challenging. This difficulty stems from the high dimensionality of $\mathbf{W}$, the coupled and nonconvex structure of $(\mathbf{B}^{\kappa}, \mathbf{\Phi}^{\kappa})$, and the dynamic nature of the $\mathbf{\Psi}$. Traditional convex formulations exhibit high complexity and significant resource costs, while existing deep learning methods exhibit limited robustness in dynamic environments. Consequently, there is a pressing need for a low-complexity, highly efficient, and robust solution framework.

We propose the HCLN, a two-layer network that jointly integrates multiple learning paradigms. The *outer layer* formulates the ASTAR-RIS orientation decision for $\mathbf{\Psi}$ as an MDP and solves it via DRL, capturing long-term effects on the cascaded channels. Conditioned on this decision, the *inner layer* employs the M2GL that updates $(\mathbf{W}, \mathbf{B}^{\kappa}, \mathbf{\Phi}^{\kappa})$.

Outer Layer: DRL

In the outer layer, we formulate the orientation control as an MDP $\langle \mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{P}\rangle$. The state $s \in \mathcal{S}$ is defined by the

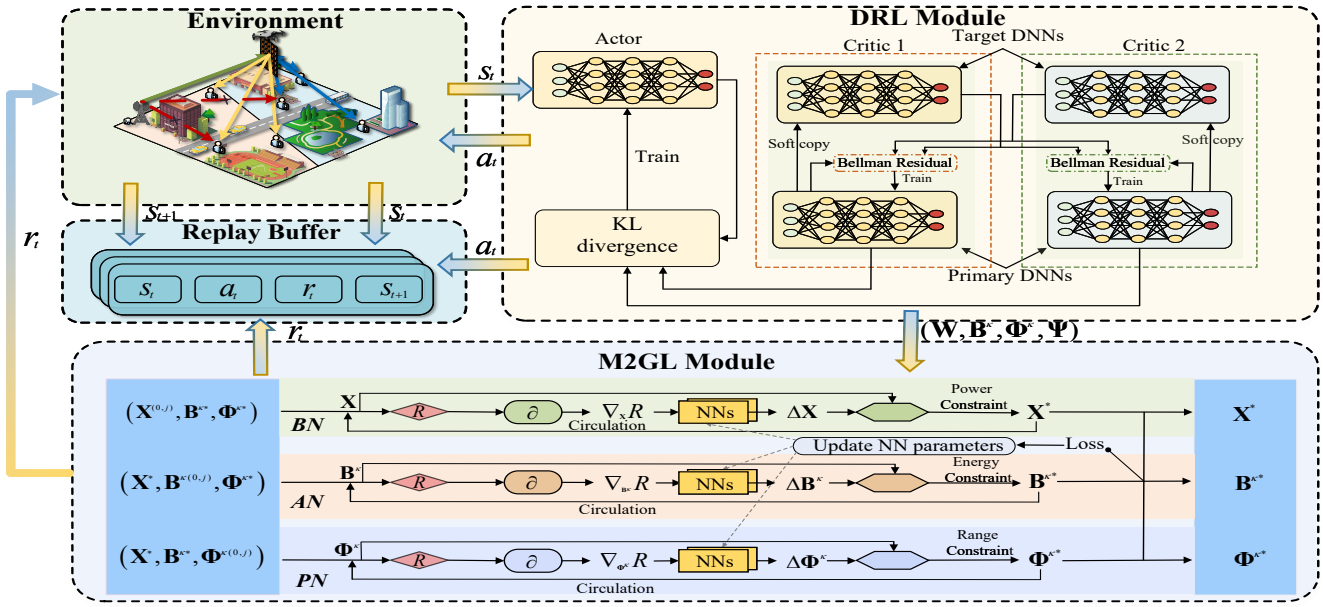


Figure 2: System Flow of the Proposed HCLN for Communication Optimization.

cascaded channels and the previous orientation, whereas the action $a \in \mathcal{A}$ corresponds to the control of the yaw angle ψ, and the reward $r \in \mathcal{R}$ is the weighted sum rate after inner optimization, i.e.,

$$r = R(\mathbf{W}, \mathbf{B}^{\kappa}, \mathbf{\Phi}^{\kappa}, \mathbf{\Psi}). \tag{10}$$

This formulation explicitly captures the sequential dependence of orientation decisions on channel evolution and throughput.

After formulating ASTAR-RIS orientation control as an MDP, we adopt soft actor-critic (SAC) as the outer decision making solver (Alajmi et al. 2023). The agent learns a stochastic policy $\pi(a|s)$ that maximizes the entropy-regularized return given as follows:

$$J_{\alpha}(\pi) = \max_{\pi} \ \mathbb{E}_{(s,a)\sim\pi}\left[\sum_{l=t}^{T} \gamma\Big(r(s,a) + \alpha\mathcal{H}(\pi(\cdot|s))\Big)\right], \tag{11}$$

where $\gamma \in (0,1)$ is the discount factor, $\alpha > 0$ is the temperature parameter, and $\mathcal{H}(\cdot)$ is the entropy of the policy.

SAC algorithm adheres to a policy iteration framework, which alternates between policy evaluation and policy improvement phases. During the policy evaluation phase, the soft Q-value is determined by employing the Bellman expectation backup, a key component in the estimation process

$$Q_{\pi}(s,a) = r(s,a) + \gamma\,\mathbb{E}_{s'\sim P}\big[V_{\pi}(s')\big], \tag{12}$$

where $V_{\pi}(s)$ is the soft state-value function. To enhance the efficiency of sampling, SAC algorithm utilizes a dual-network architecture comprising an actor network and a pair of critic networks. These critic networks are tasked with estimating the Q-values through the minimization of the temporal difference (TD) loss, which can be expressed as

$$\mathcal{L}_{Q}(\omega) = \mathbb{E}_{(s,a)\sim\mathcal{D}}\Big[\big(Q_{\omega}(s,a) - \hat{Q}_{\tilde{\omega}}(s,a)\big)^2\Big], \tag{13}$$

where $\mathcal{D}$ is the replay buffer, $\hat{Q}_{\tilde{\omega}}(s,a)$ denotes the soft target evaluated by the target critic $\tilde{\omega}$ using the entropy-augmented backup.

Given the parameter of the policy-network, the actor improves the policy by minimizing

$$\mathcal{L}_{\pi}(\varsigma) = \mathbb{E}_{s\sim\mathcal{D},\, a\sim\pi_{\varsigma}}\Big[\alpha \log \pi_{\varsigma}(a|s) - Q_{\omega}(s,a)\Big], \tag{14}$$

typically using the reparameterization trick to obtain low-variance gradients.

Through alternating critic and actor updates, SAC agent outputs the real time orientation ψ, which then conditions the *inner-layer* M2GL optimization.

Inner Layer: M2GL

The M2GL serves as a learnable inner solver that adapts the physical variables $(\mathbf{W}, \mathbf{B}^{\kappa}, \mathbf{\Phi}^{\kappa})$ to the orientation action ψ produced by the outer SAC agent. As shown in Fig. 2, distinct from manually designed alternating routines or black-box regressors, M2GL constructs three lightweight subnetworks: a *beamforming network (BN)*, an *amplitude network (AN)*, and a *phase network (PN)*. These subnetworks take the gradients of the weighted sum rate as inputs and produce increments that satisfy the imposed constraints. Specifically, it employs two modules. First, manifold learning performs dimensionality reduction on the matrices to obtain a compact representation. Second, after dimensionality reduction, the weighted sum rate gradients with respect to the decision variables are supplied to a meta learning optimizer to refine the updates.

Manifold Learning Module. Let $\mathbf{g}_k \in \mathbb{C}^{1\times M}$ denote the ASTAR-RIS-assisted effective channel for UD k, and stack them into the cascaded channel matrix

$$\mathbf{G}_c \triangleq [\mathbf{g}_1 \quad \mathbf{g}_2 \quad \cdots \quad \mathbf{g}_K]^T \in \mathbb{C}^{K\times M}. \tag{15}$$

Algorithm 1: Proposed HCLN

Initialize SAC actor ς, critics ω_1, ω_2, replay buffer $\mathcal{D}$
Initialize M2GL subnetworks: *BN*, *AN*, *PN*
for *each episode* **do**
  for *each time slot* **do**
    Execute action a to update the ASTAR-RIS orientation
    Build $\mathbf{G}_c$ from s
    for $l = 1$ **to** N_e **do**
      $\bar{\mathcal{L}} \leftarrow 0$
      for $j = 1$ **to** N_o **do**
        Initialize $\mathbf{X}, \mathbf{B}^{\kappa}, \mathbf{\Phi}^{\kappa}$
        for $i = 1$ **to** N_i **do**
          $\Delta\mathbf{X} \leftarrow BN_{\theta_{\mathbf{X}}}(\nabla_{\mathbf{X}} R)$
          $\mathbf{X}^{(i,j)} \leftarrow \mathbf{X}^{(i-1,j)} + \Delta\mathbf{X}$
        $\mathbf{W}^{(i,j)} \leftarrow \mathbf{G}_c^H \mathbf{X}^{(i,j)}$
        for $i = 1$ **to** N_i **do**
          $\Delta\mathbf{B}^{\kappa} \leftarrow AN_{\theta_{\mathbf{B}^{\kappa}}}(\nabla_{\mathbf{B}^{\kappa}} R)$
          $\mathbf{B}^{(i,j)} \leftarrow \mathbf{B}^{(i-1,j)} + \Delta\mathbf{B}^{\kappa}$
        for $i = 1$ **to** N_i **do**
          $\Delta\mathbf{\Phi}^{\kappa} \leftarrow PN_{\theta_{\mathbf{\Phi}^{\kappa}}}(\nabla_{\mathbf{\Phi}^{\kappa}} R)$
          $\mathbf{\Phi}^{(i,j)} \leftarrow \mathbf{\Phi}^{(i-1,j)} + \Delta\mathbf{\Phi}^{\kappa}$
        Update $\mathbf{X}^*, \mathbf{B}^*, \mathbf{\Phi}^*$ by (18), (19), (20)
        $\bar{\mathcal{L}} \leftarrow \mathcal{L}^j \leftarrow -R$
      $\bar{\mathcal{L}} \leftarrow \frac{1}{N_o}\bar{\mathcal{L}}$
      Update $(\theta_{\mathbf{X}}, \theta_{\mathbf{B}^{\kappa}}, \theta_{\mathbf{\Phi}^{\kappa}})$
    Transmit to next state, compute reward, and store transition tuple in $\mathcal{D}$
    for *each gradient step* **do**
      Sample a mini-batch from $\mathcal{D}$
      Compute target Q-value (12)
      Update critics/actor by (13), (14)
      Update target networks
return $\mathbf{W}, \mathbf{B}^{\kappa}, \mathbf{\Phi}^{\kappa}, \mathbf{\Psi}$

Directly optimizing $\mathbf{W} \in \mathbb{C}^{M\times K}$ is costly when M is large. We use the following stationary properties: *(i)* Any nontrivial stationary point saturates the BS power constraint; *(ii)* At any nontrivial stationary point the beamformer lies in the range space of $\mathbf{G}_c^H$ (Zhu et al. 2024), i.e.,

$$\mathbf{W} = \mathbf{G}_c^H \mathbf{X}, \quad \mathbf{X} \in \mathbb{C}^{K\times K}. \tag{16}$$

Substituting this form yields the reduced problem

$$\max_{\mathbf{X}} \; R\big(\mathbf{G}_c^H \mathbf{X},\ \mathbf{B}^{\kappa},\ \mathbf{\Phi}^{\kappa},\ \mathbf{\Psi}\big) \tag{17a}$$

$$\text{s.t.} \;\; \mathrm{Tr}\big(\mathbf{X}^H \mathbf{G}_c \mathbf{G}_c^H \mathbf{X}\big) \le P_{\mathrm{BS}}^{\max}, \tag{17b}$$

with the amplitude and phase-shift constraints inherited from the original formulation.

Gradient as Input. Given the reduced representation $\mathbf{W} = \mathbf{G}_c^H \mathbf{X}$, M2GL evaluates the weighted sum rate gradients with respect to the decision variables, i.e., $\nabla_{\mathbf{X}} R$, $\nabla_{\mathbf{B}^{\kappa}} R$, and $\nabla_{\mathbf{\Phi}^{\kappa}} R$. As shown in Fig. 2, these gradients are supplied to the light-weighted NNs, respectively. The NNs then output the increments $(\Delta\mathbf{X}, \Delta\mathbf{B}^{\kappa}, \Delta\mathbf{\Phi}^{\kappa})$, which refine the current variables. This gradient formulation explicitly exploits structural information while retaining neural adaptability, improving both interpretability and convergence efficiency.

Meta-Gradient Learning. Conventional data-driven meta-learning methods typically demand extensive offline pretraining followed by online adaptation. However, they are sensitive to channel-estimation errors and distributional shift and impose considerable computational overhead, limiting their viability in latency-critical, dynamic settings. To address this challenge, we propose a pretraining-free, model-driven meta-learning framework. The framework employs a three-level nested cyclic optimization comprising epoch, outer, and inner iterations.

Inner iteration sequentially updates each subnetwork's target variables, initializing one set and inheriting the other from its counterpart. This cyclic refinement enforces consistent interaction among subnetworks, and the overall update in the j-th outer iteration can be formulated as

$$\mathbf{X}^* = BN(\mathbf{X}^{(0,j)}, \mathbf{B}^{\kappa *}, \mathbf{\Phi}^{\kappa *}), \tag{18}$$

$$\mathbf{B}^{\kappa *} = AN(\mathbf{X}^*, \mathbf{B}^{\kappa(0,j)}, \mathbf{\Phi}^{\kappa *}), \tag{19}$$

$$\mathbf{\Phi}^{\kappa *} = PN(\mathbf{X}^*, \mathbf{B}^{\kappa *}, \mathbf{\Phi}^{\kappa(0,j)}), \tag{20}$$

where $(\cdot)^{(0,j)}$ denotes the initialization of the j-th outer iteration and $(\cdot)^*$ the updated variable after the cyclic refinement.

Outer iteration accumulates the loss. As shown in Fig. 2, the three subnetworks complete their respective N_i inner iterations to produce updates. The loss at the j-th outer iteration is defined as the negative weighted sum rate, thereby ensuring a self supervised objective, i.e.,

$$\mathcal{L}^j = -R(\mathbf{G}_c^H \mathbf{X}^*, \mathbf{B}^{\kappa *}, \mathbf{\Phi}^{\kappa *}, \mathbf{\Psi}). \tag{21}$$

Epoch iteration updates the subnetworks' parameters and contains N_o outer iterations per epoch. Following the N_o outer iterations, the losses are accumulated and averaged as

$$\bar{\mathcal{L}} = \frac{1}{N_o}\sum_{j=1}^{N_o} \mathcal{L}^j. \tag{22}$$

The parameters of *BN/AN/PN* are updated by backpropagation with the Adam optimizer using learning rates $\alpha_{\mathbf{X}}, \alpha_{\mathbf{B}^{\kappa}}, \alpha_{\mathbf{\Phi}^{\kappa}}$.

A key design is the inheritance mechanism. Taking equation (16) as an example, the task of the *BN* is to optimize the beamforming matrix, as illustrated in Fig. 2. At the beginning of each outer iteration, a randomly initialized beamforming matrix $\mathbf{X}^{(0,j)}$ is provided as the input to *BN*. In this setting, the beamforming variable does not inherit the result from the previous outer iteration, whereas the amplitude coefficients and the phase-shift matrix preserve the optimized outcomes obtained in earlier iterations. This design enables the *BN* to optimize the beamforming variable from scratch in each outer iteration, while still leveraging the continuity

of the other variables. By analogy, the *AN* and *PN* adopt the same inheritance mechanism, where one variable is re-optimized from scratch in its corresponding network, while the others retain the results from prior iterations. This alternating update strategy balances exploration and stability, with inner iterations acting as successive refinement steps within each outer iteration. As a result, the optimization trajectory is progressively improved, yielding a more effective control policy.

Simulation & Results

We evaluate the proposed HCLN in a continuous-control environment designed for ASTAR-RIS. The scenario is a BS equipped with M=8 antennas. A UAV flies at 50 m altitude along a circular trajectory of radius 80 m, carrying a STAR-RIS with $N = 25$ elements. The number of UDs is $K = 4$, uniformly and randomly distributed within a 200×200 m area. Channels follow Rician fading with factor 10, and large-scale fading is modeled as d^{-2} attenuation with directional gains. UAV trajectory contains $T = 15$ time slots. For HCLN, the outer decision-making layer adopts the SAC algorithm, where the policy and Q-networks are implemented as two-layer MLPs with hidden sizes $[256, 256]$. The learning rate is 5×10^{-4}, the discount factor γ is 0.99, and the soft update coefficient τ is 0.005. The replay buffer size is 5×10^{5} with a mini-batch size of 128, and the agent updates every 10 steps while target networks are updated every 320 steps. For M2GL, the learning rates are set to $\alpha_{\mathbf{X}} = 1 \times 10^{-3}$, $\alpha_{\mathbf{B}^{\kappa}} = 5 \times 10^{-3}$, and $\alpha_{\mathbf{\Phi}^{\kappa}} = 5 \times 10^{-3}$. The inner iterations are $N_i = 1$, the outer iterations are $N_0 = 5$, and the epoch are iterations $N_e = 400$. To evaluate the effectiveness of the proposed method, the following benchmark schemes are considered:

- **DDPG + M2GL.** This baseline replaces SAC with DDPG as the outer decision-making module while retaining M2GL for inner optimization, used to assess the benefit of entropy-regularized exploration.
- **Fixed orientation.** This scheme maintains a constant ASTAR-RIS orientation, with beamforming, amplitude, and phase shifts still optimized by M2GL, serving to isolate the gain from adaptive orientation control.
- **Random phase.** In this baseline, ASTAR-RIS phase shifts are randomly assigned, while orientation, beamforming, and amplitude coefficients are optimized as in HCLN, highlighting the contribution of phase-shift optimization.
- **Greedy policy.** At each slot, the orientation is chosen to maximize the instantaneous reward without considering long-term effects, used to demonstrate the advantage of sequential learning in HCLN over myopic strategies.

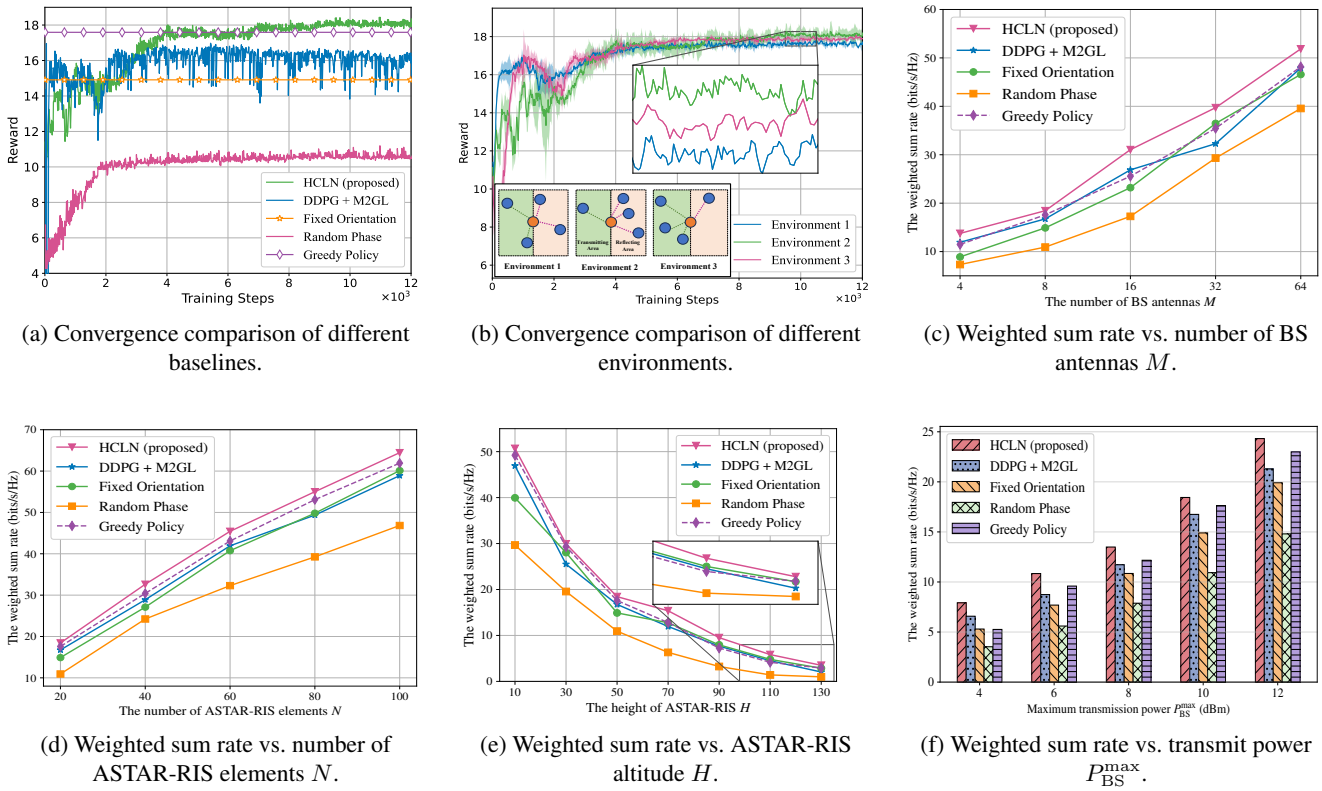
Fig. 3(a) compares the convergence behaviors of different baselines. Leveraging the entropy-regularized exploration of SAC and the efficient inner optimization of M2GL, HCLN converges rapidly and stably to the highest reward, highlighting the advantage of its hierarchical design. At convergence, HCLN outperforms DDPG + M2GL and greedy policy by approximately 10.1% and 6.7%, respectively. In contrast, DDPG + M2GL, lacking a closed-loop coupling between outer-layer orientation policy and inner-layer parameter optimization, suffers from limited exploration and converges to lower levels. Fixed orientation and random phase schemes, constrained by the absence of adaptive mechanisms, remain flat or stagnant, while the greedy policy exhibits fast initial improvement but quickly plateaus due to its myopic nature. These observations collectively underscore the importance of adaptivity and long-term planning in this setting.

Fig. 3(b) illustrates the training performance of HCLN under three environments with randomly distributed UDs. In Environment 1, the positions of the UDs were represented as (x, y) coordinates in meters: (16.0, 129.5), (82.6, 21.2), (136.3, 168.4), and (182.3, 90.6), with half in the transmission region and half in the reflection region. Environment 2 includes three UDs in the reflection region and one in the transmission region, while Environment 3 is configured oppositely. Across different training environments, HCLN consistently demonstrates superior performance with strong robustness and adaptability. This robustness stems from its meta-gradient learning, which focuses on the evolution of the optimization trajectory rather than individual variable updates. Combined with an alternating update strategy, this design ensures stable convergence and improves HCLN's generalization across diverse conditions.

Fig. 3(c) shows the weighted sum rate versus the number of BS antennas M. As M increases, all schemes exhibit performance improvement due to enhanced spatial degrees of freedom. HCLN consistently achieves the highest rates across all antenna settings, demonstrating the effectiveness of its hierarchical optimization. This advantage stems from the manifold learning module, which reduces the dimensionality of the beamforming matrix and enables the BS to more effectively exploit additional array elements. In particular, at $M = 64$, HCLN outperforms DDPG + M2GL and greedy policy by approximately 8.2% and 7.7%, respectively. Furthermore, in multi-user systems, a larger number of antennas allows more efficient beam allocation across individual UDs, thereby enhancing the system capacity.

Fig. 3(d) illustrates the impact of the number of ASTAR-RIS elements N on the weighted sum rate. As N increases, all schemes improve steadily because a larger ASTAR-RIS enhances reflection and transmission capabilities and provides greater spatial degrees of freedom for beamforming. HCLN consistently achieves the best performance, owing to meta-gradient learning that performs constraint-compliant joint optimization of ASTAR-RIS amplitudes and phase shifts, the system achieves phase alignment and enhanced end-to-end link gain. Notably, at $N = 100$, HCLN achieves a weighted sum rate that is approximately 9.5% and 5.9% greater than that of DDPG + M2GL and the greedy policy, highlighting the benefits of entropy-regularized exploration and trajectory-level optimization. In contrast, fixed orientation and random phase exhibit significantly inferior performance, underscoring the importance of adaptive orientation and phase-shift optimization in large scale ASTAR-RIS configurations.

Fig. 3(e) illustrates the impact of the ASTAR-RIS altitude



(a) Convergence comparison of different baselines.

(b) Convergence comparison of different environments.

(c) Weighted sum rate vs. number of BS antennas M.

(d) Weighted sum rate vs. number of ASTAR-RIS elements N.

(e) Weighted sum rate vs. ASTAR-RIS altitude H.

(f) Weighted sum rate vs. transmit power $P_{\mathrm{BS}}^{\max}$.

Figure 3: Results compare the proposed framework with benchmark models for different metrics.

on the weighted sum rate. Overall, the system performance degrades as the altitude increases, since larger distances attenuate both the BS-to-ASTAR-RIS and ASTAR-RIS-to-UD links, thereby reducing the overall channel gains. The proposed HCLN framework consistently achieves superior performance over all baselines. Although DDPG + M2GL and the greedy policy exhibit competitive results, they remain inferior to HCLN because its DRL-based outer layer provides entropy-regularized exploration and stable policy improvement, avoiding the limited exploration of DDPG and the myopic adaptation of the greedy baseline. Importantly, at higher altitudes, where path loss dominates, the performance gain from orientation control becomes marginal, and precise phase-shift optimization emerges as the key performance factor. For instance, at $H = 110$ m, the fixed orientation scheme outperforms DDPG + M2GL and the greedy policy by approximately 6.4% and 15.1%, respectively. These results indicate that under severely attenuated channels, system throughput is primarily impacted by accurate phase-shift design rather than orientation adaptation.

Fig. 3(f) demonstrates the weighted sum rate of different schemes under identical antenna numbers but varying maximum transmit powers. As the transmit power increases, the performance gain of the proposed HCLN scheme becomes progressively more pronounced compared to benchmark approaches. Specifically, at $P_{\mathrm{BS}}^{\max} = 12$ dBm, HCLN outperforms DDPG + M2GL and the greedy policy by approximately 14.2% and 5.8%, respectively. Furthermore, by jointly optimizing the ASTAR-RIS orientation, amplitude coefficients, and phase shifts, HCLN effectively translates the additional transmit power into substantial improvements in spectral efficiency, underscoring the value of hierarchical and integrated optimization in power-limited environments.

Conclusion

This paper proposes HCLN, a hierarchical collaborative learning network. The outer layer models ASTAR-RIS orientation as an MDP and applies DRL for sequential control, while the inner layer employs M2GL to jointly optimize beamforming, phase shifts, and amplitude coefficients. By tackling high dimensionality and nonconvexity, HCLN achieves stable and adaptable performance across diverse initializations and network settings. Extensive experiments validate the effectiveness of the proposed framework, demonstrating robust performance and significant performance gains over competitive baselines. The results show that the proposed scenario and HCLN address high dimensional control challenges and scale to deployment, strengthening system robustness and service accessibility. They also provide long-term strategic value for digitizing urban services, and for enabling key 6G applications.

References

Abeywickrama, S.; Zhang, R.; and Yuen, C. 2020. Intelligent Reflecting Surface: Practical Phase Shift Model and Beamforming Optimization. In *Proceedings of the IEEE*

International Conference on Communications (ICC), 1–6. Dublin, Ireland: IEEE.

Alajmi, A.; Ahsan, W.; Fayaz, M.; and Nallanathan, A. 2023. Intelligent Resource Allocation in Backscatter-NOMA Networks: A Soft Actor-Critic Framework. *IEEE Transactions on Vehicular Technology*, 72(8): 10119–10132.

Cheng, X.; Lin, Y.; Shi, W.; Li, J.; Pan, C.; Shu, F.; Wu, Y.; and Wang, J. 2022a. Joint Optimization for RIS-Assisted Wireless Communications: From Physical and Electromagnetic Perspectives. *IEEE Transactions on Communications*, 70(1): 606–620.

Cheng, Y.; Peng, W.; Huang, C.; Alexandropoulos, G. C.; Yuen, C.; and Debbah, M. 2022b. RIS-Aided Wireless Communications: Extra Degrees of Freedom via Rotation and Location Optimization. *IEEE Transactions on Wireless Communications*, 21(8): 6656–6671.

Cui, Z.; Guan, K.; Zhang, J.; and Zhong, Z. 2021. SNR Coverage Probability Analysis of RIS-Aided Communication Systems. *IEEE Transactions on Vehicular Technology*, 70(4): 3914–3919.

Guo, H.; Liang, Y.-C.; Chen, J.; and Larsson, E. G. 2020. Weighted Sum-Rate Maximization for Reconfigurable Intelligent Surface Aided Wireless Networks. *IEEE Transactions on Wireless Communications*, 19(5): 3064–3076.

Guo, L.; Jia, J.; Chen, J.; and Wang, X. 2024. Secure Communication Optimization in NOMA Systems With UAV-Mounted STAR-RIS. *IEEE Transactions on Information Forensics and Security*, 19: 2300–2314.

Lin, T.; and Zhu, Y. 2020. Beamforming Design for Large-Scale Antenna Arrays Using Deep Learning. *IEEE Wireless Communications Letters*, 9(1): 103–107.

Liu, X.; Yu, Y.; Li, F.; and Durrani, T. S. 2022a. Throughput Maximization for RIS-UAV Relaying Communications. *IEEE Transactions on Intelligent Transportation Systems*, 23(10): 19569–19574.

Liu, Y.; Duo, B.; Wu, Q.; Yuan, X.; Li, J.; and Li, Y. 2024. Elevation Angle-Dependent 3D Trajectory Design for Aerial RIS-Aided Communication. *IEEE Transactions on Intelligent Transportation Systems*, 25(3): 2696–2702.

Liu, Y.; Hu, Q.; Cai, Y.; Yu, G.; and Li, G. Y. 2022b. Deep-Unfolding Beamforming for Intelligent Reflecting Surface Assisted Full-Duplex Systems. *IEEE Transactions on Wireless Communications*, 21(7): 4784–4800.

Liu, Y.; Zhang, X.; Mu, X.; Hou, T.; Xu, J.; Di Renzo, M.; Schober, R.; Al-Dhahir, N.; Poor, H. V.; and Letaief, K. B. 2021. Reconfigurable intelligent surfaces: Principles and opportunities. *IEEE Communications Surveys & Tutorials*, 23(3): 1546–1577.

Mai, Z.; Chen, Y.; and Du, L. 2021. A Novel Blind mmWave Channel Estimation Algorithm Based on ML-ELM. *IEEE Communications Letters*, 25(5): 1549–1553.

Su, Y.; Wang, F.; and Zhang, X. 2025. Meta-Learning-Based STAR-RIS for Dynamic Multi-Mobile-User Downlink Communications Over 6G Mobile Wireless Networks. In *Proceedings of the IEEE International Conference on Communications (ICC)*, 4565–4570. IEEE.

Thangasamy, V.; Singh, I.; Karishma; Yaswanth; and Sai, J. 2022. Analysis of Capacity and Outage Probability for NOMA-Based Cellular Communication over Rician Fading Channel. In *Proceedings of the 2022 First International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT)*, 1–5.

Wang, Q.; Guo, S.; Wu, C.; Xing, C.; Zhao, N.; Niyato, D.; and Karagiannidis, G. K. 2025. STAR-RIS aided covert communication in UAV air–ground networks. *IEEE Journal on Selected Areas in Communications*, 43(1): 245–259.

Wu, Q.; and Zhang, R. 2019. Intelligent Reflecting Surface Enhanced Wireless Network via Joint Active and Passive Beamforming. *IEEE Transactions on Wireless Communications*, 18(11): 5394–5409.

Xiao, H.; Hu, X.; Wang, W.; Su, Z.; Wong, K.-K.; and Yang, K. 2024. STAR-RIS and UAV Combination for Simultaneous Task Offloading and Communications in MEC Networks. In *Proceedings of the 16th International Conference on Wireless Communications and Signal Processing (WCSP)*, 857–862.

Zeng, S.; Zhang, H.; Di, B.; Han, Z.; and Song, L. 2021. Reconfigurable Intelligent Surface (RIS) Assisted Wireless Coverage Extension: RIS Orientation and Location Optimization. *IEEE Communications Letters*, 25(1): 269–273.

Zhang, Y.; Lin, J.; Zhao, Y.; Huang, L.; Gao, Z.; and Xu, G. 2024. Active RIS-Assisted Integrated Sensing and Communication Systems: Joint Receive-Transmit Beamforming and Reflection Design. In *Proceedings of the 2024 IEEE 100th Vehicular Technology Conference (VTC2024-Fall)*, 1–5. IEEE.

Zhao, J.; Xu, Q.; Mu, X.; Liu, Y.; and Zhu, Y. 2025. Aerial Active STAR-RIS-Aided IoT NOMA Networks. *IEEE Internet of Things Journal*, 12(8): 9525–9538.

Zheng, Z.; Jing, W.; Lu, Z.; and Wen, X. 2022. RIS-Enhanced LEO Satellite Communication: Joint Passive Beamforming and Orientation Optimization. In *Proceedings of the 2022 IEEE Globecom Workshops (GC Wkshps)*, 874–879.

Zhu, F.; Wang, X.; Huang, C.; Yang, Z.; Chen, X.; Al Hammadi, A.; Zhang, Z.; Yuen, C.; and Debbah, M. 2024. Robust Beamforming for RIS-Aided Communications: Gradient-Based Manifold Meta Learning. *IEEE Transactions on Wireless Communications*, 23(11): 15945–15956.