Exploiting NOMA Transmissions in Multi-UAV-assisted Wireless Networks: From Aerial-RIS to Mode-switching UAVs

Exploiting NOMA Transmissions in Multi-UAV-assisted Wireless Networks: From Aerial-RIS to Mode-switching UAVs

Songhan Zhao, Shimin Gong, Bo Gu, Lanhua Li, Bin Lyu, Dinh Thai Hoang, and Changyan Yi

Abstract—In this paper, we consider an aerial reconfigurable intelligent surface (ARIS)-assisted wireless network, where multiple unmanned aerial vehicles (UAVs) collect data from ground users (GUs) by using the non-orthogonal multiple access (NOMA) method. The ARIS provides enhanced channel controllability to improve the NOMA transmissions and reduce the co-channel interference among UAVs. We also propose a novel dual-mode switching scheme, where each UAV equipped with both an ARIS and a radio frequency (RF) transceiver can adaptively perform passive reflection or active transmission. We aim to maximize the overall network throughput by jointly optimizing the UAVs' trajectory planning and operating modes, the ARIS's passive beamforming, and the GUs' transmission control strategies. We propose an optimization-driven hierarchical deep reinforcement learning (O-HDRL) method to decompose it into a series of sub-problems. Specifically, the multi-agent deep deterministic policy gradient (MADDPG) adjusts the UAVs' trajectory planning and mode switching strategies, while the passive beamforming and transmission control strategies are tackled by the optimization methods. Numerical results reveal that the O-HDRL efficiently improves the learning stability and reward performance compared to the benchmark methods. Meanwhile, the dual-mode switching scheme is verified to achieve a higher throughput performance compared to the fixed ARIS scheme.

Index Terms—Aerial reconfigurable intelligent surface, UAV-assisted wireless networks, NOMA, deep reinforcement learning.

I. INTRODUCTION

The reconfigurable intelligent surface (RIS) has been proposed to improve the transmission performance of future sixth-generation (6G) wireless networks [1]. The RIS is equipped with massive reflecting elements, which can reconfigure the propagation environment by inducing phase shifts on the incident signals. The RIS is attracting considerable attention from both academia and industry as a promising solution to improve the network performance in existing wireless networks, such as the RIS-assisted non-orthogonal multiple access (NOMA) networks [2]. The NOMA technology enables multiple users to share the same spectrum by allocating them into different power domains. Integrating the RIS into the NOMA networks can enhance the channel conditions for more flexible and controllable power adjustments [3]. This allows more NOMA users to access and thus achieves a higher transmission performance.

The unmanned aerial vehicle (UAV) as another promising technique is introduced into the RIS-assisted wireless networks to improve the network flexibility and capacity [4]. The UAV is able to provide the line-of-sight (LoS) connections from air to ground users (GUs), and thus improve the GUs' channel conditions [5]. The joint application of the UAV and the RIS has been shown to offer significant advantages to wireless networks [6]. Typically, the RIS can be deployed on the ground to facilitate the air-to-ground channel quality. In particular, deploying the RISs near GUs can help establish virtual LoS links between UAVs and GUs, which allows UAVs to avoid the long-distance flights to bypass obstacles. Besides the terrestrial RIS, the RIS can be also mounted on the UAV becoming the aerial-RIS (ARIS) [7]. The ARIS can collaborate with the UAVs to jointly design the trajectories to meet the GUs' dynamic traffic demands and network topology. As such, the ARIS shows superior adaptability and enhanced reliability in dynamic and complex network environments.

A. Motivations and Challenges

In the existing UAV- and ARIS-assisted networks, the UAVs and the ARISs generally keep the fixed roles (e.g., either passive reflection or active radio frequency (RF) communications) during their operations. However, their advantages will become limited when the wireless channels are severely attenuated by the long transmission distance. For example, when the ARIS is located far away from the UAVs involving in data transmission, the ARIS needs to approach to the UAVs for channel enhancement via passive reflection. The ARIS's mobility may increase the transmission delay and the UAVs' energy consumption, especially in the large-scale wireless networks. Conversely, when UAVs are in close proximity, they need to tolerate strong interference from each other. Although the NOMA method can be employed to cancel partial GUs' interference by leveraging the channel orthogonality, the performance will be limited when channel differences among GUs become very small. One typical solution to these limitations is to deploy multiple UAVs and ARISs across the wireless network to achieve the widespread coverage. However, as the number of UAVs and ARISs increases, it is challenging to coordinate them, especially in a decentralized wireless network. The fixed-mode UAVs and ARISs may not work efficiently when facing a dynamic and heterogeneous traffic demand over the service area. In particular, when the UAVs' traffic loads become unbalanced, the ARISs have to traverse among different UAVs to achieve the load balance by reshap-

Songhan Zhao, Shimin Gong, Bo Gu, and Lanhua Li are with the School of Intelligent Systems Engineering, Shenzhen Campus of Sun Yat-sen University, China (e-mail: zhaosh55@mail2.sysu.edu.cn, {gongshm5, gubo, lilh65}@mail.sysu.edu.cn). Bin Lyu is with the School of Communications and Information Engineering, Nanjing University of Posts and Telecommunications, China (e-mail: blyu@njupt.edu.cn). Dinh Thai Hoang is with the School of Electrical and Data Engineering, University of Technology Sydney, Australia (e-mail: hoang.dinh@uts.edu.au). Changyan Yi is with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, China (e-mail: changyan.yi@nuaa.edu.cn).

ing the UAVs' channel conditions and capacities. This will result in additional time delay and energy consumption, and thus reduce data transmission efficiency. If each UAV carries the RF transceiver and the passive ARIS simultaneously, it can dynamically adjust the operating mode for either data collection or signal reflection based on the current traffic demands and channel conditions. As such, we expect that the dual-mode UAVs' mode switching scheme can offer a more flexible solution to the challenges mentioned above.

Another challenge lies in the complexity of the optimization problem. Generally, the performance maximization of the ARIS-assisted wireless network can be formulated as a joint optimization of the passive beamforming and trajectory planning strategies [8]. These variables are coupled not only with each other but also across different time slots during the entire trajectory. One common approach is to decompose the original problem into multiple subproblems and solve them in an alternative manner [9]. The ARIS's passive beamforming and trajectory planning subproblems are often tackled by using the semidefinite relaxation (SDR) and successive convex approximation (SCA) methods, respectively. The approximation and alternating characteristics can obtain a feasible solution but may be far from the optimum of the original optimization problem. Besides the optimization methods, the deep reinforcement learning (DRL) provides another feasible approach. The DRL agent continuously adjusts the actions of all UAVs and GUs by interacting with the environment to improve the overall network performance. However, if the agent learns all control variables directly, the high-dimensional action space also leads to an unstable learning performance, making it difficult to learn the meaningful actions.

B. Solutions and Contributions

In this paper, we first explore an ARIS-assisted multi-UAV wireless network, where multiple UAVs serve as aerial access points (APs) to collect the data from the GUs and an ARIS moves around for channel enhancement. To improve the spectral efficiency and support the massive connectivity, we employ NOMA for the GUs' uplink transmissions. Multiple GUs connecting to the same UAV create an NOMA group. The successive interference cancellation (SIC) technique is employed to decode the GUs' signals within each NOMA group. However, the NOMA capacity will become limited if there is limited channel orthogonality in one NOMA group [10]. To tackle this difficulty, the ARIS is used to reshape the channel conditions via passive beamforming and thus induces a desirable channel orthogonality for the GUs' NOMA transmissions. Besides the AIRS's passive beamforming, we also exploit the trajectory planning of both UAVs and ARIS to create the spatial-temporal channel orthogonality for the GUs.

As the ARIS-assisted scheme may become inflexible for a large-scale wireless network, we further develop a novel dual-mode switching scheme aimed at enhancing both channel conditions and data transmission efficiency. Specifically, the dual-mode switching scheme allows each UAV to carry a passive ARIS and an active antenna system. Moreover, it enables the dual-mode UAVs to switch between the passive and the active modes to perform either the signal reflection or

the data transmission on demand. The passive UAVs reshape the air-to-ground channel conditions while the active UAVs serve as aerial APs for data transmissions. By switching the operating mode on demand, the UAVs can significantly improve the agility to meet the GUs' dynamic traffic demands. For example, when multiple UAVs are densely deployed, a portion of them can switch to the passive mode to reduce channel competition and signal interference. As the UAVs fly far apart, the interference among them gradually diminishes. This allows the UAVs to switch back to the active mode and thus serve the GUs more efficiently.

We aim to maximize the overall throughput by jointly optimizing the ARIS's passive beamforming, the UAVs' trajectories planning and operating modes, as well as the GUs' transmission control strategies. We design an optimization-driven hierarchical deep reinforcement learning (O-HDRL) framework, leveraging the benefits of both the optimization and DRL methods. Specifically, the original problem is firstly decomposed into two parts based on the characteristics of the control variables. The multi-agent deep deterministic policy gradient (MADDPG) algorithm is employed to adjust the trajectories of the UAVs and ARIS. The UAVs' mode switching optimization is a combinatorial problem, which will introduce the significant computational complexity as the number of the UAVs increases. Thus, the UAVs' mode switching strategy is also determined by the MADDPG method. After obtaining the UAVs' trajectories and the mode switching strategies, the optimization methods focus on solving the ARIS's passive beamforming and GUs' transmission strategies. This helps decouple different variables, and thus effectively reducing the complexity. The optimization method as the model-based approach can leverage partial model information from the wireless networks. As such, it can provide supervision to the model-free MADDPG, which is envisioned to further improve the MADDPG's learning performance [11]. The main contributions of this paper are outlined as follows:

- ARIS-assisted multi-UAV NOMA transmission: We develop a UAVs' collaborative scheme where multiple UAVs collect data from the GUs using the NOMA method. An ARIS is employed to improve the channel conditions for the GUs' NOMA transmissions by dynamically adjusting its trajectory and passive beamforming strategies. The overall network throughput is maximized by the joint optimization of the ARIS's passive beamforming, the UAVs' trajectories planning, as well as the GUs' transmission strategies.
- UAVs' mode switching for capacity enhancement: We develop a dual-mode switching scheme that allows each UAV to dynamically switch between the ARIS mode facilitating other UAVs' transmissions and the aerial AP mode for data transmission from GUs. By providing the UAVs more flexibility, the dual-mode switching scheme can improve the network throughput by 17.51% compared to the fixed-ARIS scheme.
- Optimization-driven hierarchical DRL algorithm: We propose a learning algorithm to solve the complex optimization problem efficiently by decoupling the control

variables into two parts. The MADDPG method can dynamically adjust the UAVs' trajectory planning and mode switching strategies. The ARIS's passive beamforming and GUs' transmission strategies are optimized by the SDR and SCA methods in each time slot. The proposed algorithm has been shown to reduce computational complexity and improve learning efficiency compared to the conventional learning methods.

The rest of this paper is outlined as follows. The related works are discussed in Section II. In Section III, we introduce the fixed-mode ARIS scheme and explain the design ideas behind the O-HDRL framework in Section IV. To further enhance the system's flexibility and transmission efficiency, we introduce a dual-mode switching scheme in Section V. Finally, we evaluate the O-HDRL framework and the dual-mode switching scheme in Section VI and present the conclusions in Section VII.

II. RELATED WORK

A. RIS-assisted NOMA Transmissions

Due to the passive characteristics, the RIS is easily employed in the wireless NOMA networks to augment the network capacity. The authors in [12] investigated the RIS in both NOMA and orthogonal multiple access (OMA) wireless networks. It is preferable to pair users with asymmetric rates and/or asymmetric deployment to improve the NOMA performance compared to the OMA. The integration of the RIS and NOMA has been shown to reduce energy consumption [13], improve transmission capacity [14], and enhance flexibility [15]. The RIS can also provide an additional channel gain for UAV-assisted wireless networks. The authors in [16] studied the RIS-assisted NOMA transmissions in UAV-assisted wireless networks by jointly optimizing the UAVs' trajectories and the RIS's passive beamforming strategies. The RIS and the UAV can jointly enhance the channel conditions to improve the GUs' NOMA transmission performance. The RIS can also be mounted on the UAVs to provide an extended service coverage for NOMA transmissions. However, deploying an RIS on a UAV can increase its energy consumption, making it challenging to work efficiently in energy-constrained scenarios. To address this, the authors in [17] explored the potential of using the solar energy to extend the ARIS's operational lifespan. The authors in [18] explored an ARIS-assisted wireless NOMA network, where the ARIS can dynamically adjust its trajectory and passive beamforming to meet the GUs' time-varying SIC demands. To further extend the service range, the authors in [19] investigated a multi-ARIS-assisted wireless network, where multiple ARISs can collaboratively generate passive beamforming to enhance the GUs' channel conditions. The above works demonstrated that the ARIS is capable of offering a more flexible 3D passive beamforming and providing the NOMA GUs with more access opportunities. However, the fixed-mode ARIS scheme may not perform efficiently in large-scale wireless networks. Since the ARIS's capability of channel enhancement is severely limited by the double fading effect, the ARIS is only effective in a limited coverage range when it is close to either the UAVs or GUs [20].

B. Multi-UAV Collaborative Data Transmission

By exploiting UAVs' mobilities, the UAV-assisted wireless networks can effectively overcome many limitations in terrestrial networks. The authors in [21] investigated the UAV-assisted backscatter communication networks, where the UAV operating as the relay assists backscatter users in forwarding signals to the base station. In addition, the UAV can serve as the aerial base station to receive and process computational tasks from the GUs in the mobile edge computing (MEC) system. The authors in [22] employed multiple UAVs in the MEC system, where each UAV either processes the offloaded tasks from the GUs or forwards them to the other UAVs for collaborative computation. The authors in [23] optimized the overall throughput in a RIS-assisted UAV network, where the UAV collects data from the GUs with the assistance of the RIS. Besides the data transmission, the authors in [24] investigated the multi-UAV-assisted wireless powered communication (W-PC) systems. The multiple UAVs first transfer RF energy to the GUs in the downlink, and then the GUs transmit uplink data to the associated UAVs by using the harvested energy. The multiple UAVs can also play different roles in wireless networks. The authors in [25] studied a dual-UAV cooperative wireless network, where an active UAV equipped with the transceiver transmits control signals to the GUs, while a passive UAV equipped with an ARIS enhances the air-to-ground channel conditions for the active UAV. However, the UAV and the ARIS keep the fixed operating modes during the trajectories, which makes it challenging to tackle the time-varying channel conditions and GUs' traffic demands. Differently, we focus on the collaborative transmissions and network performance optimization of multiple dual-mode UAVs, which dynamically switch the UAVs' working modes depending on the time-varying network environment.

C. Performance Maximization in RIS-assisted UAV Networks

The performance maximization for UAV- and RIS-assisted wireless networks is challenging due to the high dimensionality and deep coupling of different control variables. One typical approach is to simplify the original optimization problem into several subproblems and subsequently solve them individually by the optimization methods. The authors in [26] developed a joint optimization framework for the downlink multi-antenna RIS-assisted system. To minimize the transmit power of the base station, the active and passive beamforming are jointly optimized by using the SDR and alternating optimization (AO). To enhance the interpretability of the solution, the authors in [27] derived a closed-form solution to the sum-rate maximization problem based on the composite channel model in the RIS-assisted UAV systems. In the RIS-assisted WPC systems, the authors in [28] designed different RIS's passive beamforming strategies for the downlink energy transfer and uplink information transmissions, respectively. It was found that setting the same passive beamforming strategy for both energy and information transmission achieves considerable throughput performance while significantly reducing the computational complexity. Compared to the conventional optimization methods, the DRL methods are able to address more complex optimization problems by training the agent via
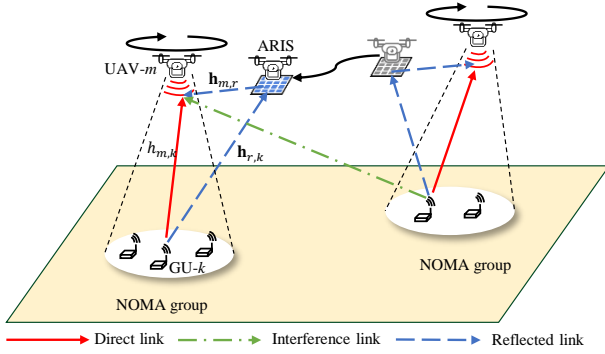
Fig. 1: ARIS-assisted multi-UAV wireless networks.

the interaction with the environment [29]. The authors in [30] designed both deep Q-network (DQN) and deep deterministic policy gradient (DDPG) algorithms for the UAV's trajectory planning. The results show that the DQN requires a shorter training time, while the DDPG has the superior learning performance in terms of reward. Besides the single-agent methods, the multi-agent DRL methods enable the centralized training of multiple agents, which are then distributed to wireless devices for distributed execution. This allows each device to make independent decisions according to its observation when it is deployed in the large-scale wireless networks. To maximize the overall throughput and energy efficiency, the authors in [31] designed a multi-agent learning approach that combines game theory with proximal policy optimization (PPO). Each agent acts as a game player that independently selects the transmission action based on its local observation. In this paper, we propose an O-HDRL framework combining the advantages of both the optimization and DRL methods. The optimization methods can improve the DRL's learning efficiency and reward performance by offering approximate solutions with partial system information.

III. SYSTEM MODEL

As illustrated in Fig. 1, we consider an ARIS-assisted multi-UAV wireless network consisting of M UAVs and an ARIS with a set of scattering elements $\mathcal{L} = \{1, \ldots, L\}$. The UAVs, represented by the set $\mathcal{M} = \{1, \ldots, M\}$, operate as mobile APs to collect and process data from the GUs, while the ARIS moves around to enhance channel conditions for the UAVs' data transmissions. Let $\mathcal{K} = \{1, \ldots, K\}$ represent the set of the GUs. We denote the m-th UAV and k-th GU as the UAV-m and GU-k, respectively. Each GU gathers sensing data from its surroundings and then transmits it to the UAVs. We consider that all UAVs and GUs are equipped with a single antenna and share the same frequency band. To improve the spectral efficiency, the GUs employ NOMA to transmit data to UAVs. The GUs connected to the same UAV form an NOMA group. By employing the NOMA, all UAVs can efficiently multiplex the wireless resources while tolerating a certain level of interference. However, simply increasing the number of UAVs may not improve the data collection performance, as it can increase the co-channel interference among the UAVs and introduce excessive deployment cost. Therefore, to improve the system's transmission efficiency, we employ a low-cost

ARIS to offer more flexible channel enhancements that adapts to the dynamic network conditions.

A. ARIS-enhanced Channel Model

The location of the GU-k is represented by the two-dimensional coordinate $\mathbf{z}_k$. All the UAVs and the ARIS work at the fixed altitude H. We consider a time-slotted frame structure for the UAVs' operations and let $\mathcal{N} = \{1, \ldots, N\}$ denote the set of all time slots with a unit duration τ. In the n-th time slot, the location of the UAV-m and the ARIS are denoted by $\mathbf{q}_m[n]$ and $\mathbf{q}_r[n]$, respectively. During the trajectory, all the UAVs and the ARIS need to ensure the safety distance between each other and satisfy the speed limit, which leads to the following mobility constraints:

$$d_i[n] \triangleq \|\mathbf{q}_i[n] - \mathbf{q}_i[n-1]\| \leq \tau V_{\max}, \tag{1a}$$

$$d_{i,i'}[n] \triangleq \|\mathbf{q}_i[n] - \mathbf{q}_{i'}[n]\| \geq D_{\min}, \tag{1b}$$

where $i \neq i'$ and $i, i' \in \mathcal{I} = \{1, \ldots, M+1\}$ represent the indexes of the UAVs and the ARIS. The distance of each UAV (or ARIS) traveled in adjacent time slots is represented by $d_i[n]$ and the distances among all the UAVs and the ARIS are represented by $d_{i,i'}[n]$. The constants $V_{\max}$ and $D_{\min}$ denote the maximum speed and safety distance, respectively.

The channels from the UAV-m to the ARIS, the ARIS to the GU-k, and the UAV-m to the GU-k are denoted by $\mathbf{h}_{m,r}[n] = \beta^{1/2} d_{m,r}^{-\alpha_0/2}[n] \mathbf{a}_{m,r}$, $\mathbf{h}_{r,k}[n] = \beta^{1/2} d_{r,k}^{-\alpha_0/2}[n] \mathbf{a}_{r,k}$ and $h_{m,k}[n] = \beta^{1/2} d_{m,k}^{-\alpha_0/2}[n] a_{m,k}$, respectively, where $d_{i,k}$ denotes the distance between the UAVs (or ARIS) and the GUs. The constant α_0 denotes the path loss exponent and β is the channel power gain at a reference distance of one meter. The antenna array response coefficients are denoted as $\mathbf{a}_{m,r}$, $\mathbf{a}_{r,k}$, and $a_{m,k}$, respectively [32]. The ARIS is capable of adjusting the phase-shift of air-to-ground links, enabling channel enhancement and interference mitigation. The phase-shift vector of the ARIS is denoted as $\boldsymbol{\theta}[n] = [e^{j\phi_l}[n], \ldots, e^{j\phi_L}[n]]^T$, where $e^{j\phi_l}[n]$ is the phase-shift induced by the ARIS's l-th scattering element. In the following, we omit the time index of each variable for brevity. The equivalent channel from the GU-k to the UAV-m is represented as:

$$\widehat{\mathbf{h}}_{m,k} = \mathbf{h}_{m,r} \text{diag}(\mathbf{h}_{r,k}) \boldsymbol{\theta} + h_{m,k} \triangleq \mathbf{H}_{m,k} \boldsymbol{\theta} + h_{m,k}, \tag{2}$$

where $\mathbf{H}_{m,k} = \mathbf{h}_{m,r} \text{diag}(\mathbf{h}_{m,k})$ and $\text{diag}(\mathbf{h}_{m,k})$ denotes the diagonal matrix with the diagonal vector $\mathbf{h}_{m,k}$. The equivalent channel in (2) can be dynamically controlled by the ARIS's passive beamforming and trajectory planning strategies.

B. UAV-assisted NOMA Transmissions

We consider that all GUs employ NOMA technique to transmit data to the UAVs, which is expected to achieve a higher transmission capacity than that of the conventional OMA schemes [33]. Note that the UAVs are not required to simultaneously connect all GUs, in each time slot, each UAV adaptively selects a subset of GUs for the NOMA transmissions depending on the GUs' channel conditions and traffic demands. Let $\rho_{m,k}$ denote the binary variable and $\rho_{m,k} = 1$ represents the association between the GU-k and

UAV-m. As each GU can transmit data to at most one UAV in each time slot, we have the following association constraint:

$$\sum_{m\in\mathcal{M}} \rho_{m,k} \leq 1 \text{ and } \rho_{m,k} \in \{0,1\}, \forall m \in \mathcal{M}, \forall k \in \mathcal{K}. \quad (3)$$

The interference to the GU-k contains the intra- and inter-UAV interference, respectively. The intra-UAV interference $I_{m,k}$ denotes the interference to GU-k caused by the other GUs associated with the same UAV-m. The inter-UAV interference I'_m represents the interference introduced by the GUs associated with the other UAVs. The SIC technique allows the UAVs to mitigate the intra-UAV interference by decoding and canceling the signals one by one.

The decoding order of the GUs can be determined by the GUs' channel conditions to the UAVs. However, the channel conditions are not only related to the distance between UAVs and the GUs but also correlated with the ARIS's passive beamforming strategy $\boldsymbol{\theta}$. Intuitively, we can formulate a joint optimization problem for the ARIS's passive beamforming and the NOMA decoding order. However, this can introduce additional computational overhead, which potentially hinders the real-time control of the UAVs and the ARIS. Instead, we propose a heuristic method to order the GUs' information decoding. We first consider an optimistic case in which the ARIS is capable of aligning all channels between GUs and the associated UAVs. As such, the channels between the GUs and the UAVs (denoted by $\widehat{\mathbf{h}}'_{m,k}$) are calculated as follows:

$$\|\widehat{\mathbf{h}}'_{m,k}\| = \sum_{l\in\mathcal{L}} |h_{m,r,l}h_{r,k,l}| + |h_{m,k}|, \quad (4)$$

where $h_{m,r,l}$ and $h_{r,k,l}$ are the l-th elements of $\mathbf{h}_{m,r}$ and $\mathbf{h}_{r,k}$, respectively. Let $\mathcal{K}^m = \{1,\ldots,D_m\}$ denote the set of the GUs associated with the UAV-m, where D_m is the number of associated GUs. Note that this method effectively mitigates the impact of the ARIS's passive beamforming on channel gains by assuming the upper bound of ARIS's reflection capability.

Then, the GUs' decoding order is determined by the sorting of their channel strengths as follows:

$$\|\widehat{\mathbf{h}}'_{m,1}\|^2 \geq \ldots \geq \|\widehat{\mathbf{h}}'_{m,d}\|^2 \geq \ldots \geq \|\widehat{\mathbf{h}}'_{m,D_m}\|^2, \forall d \in \mathcal{K}^m. \quad (5)$$

Following the ordering rule in (5), we denote $s_m(k) \in \mathcal{K}^m$ as the decoding order of the GU-k when it is associated with the UAV-m. According to the SIC principle, the intra-UAV interference to the GU-k is represented as follows:

$$I_{m,k} = \sum_{d>s_m(k),d\in\mathcal{K}^m} \rho_{m,d}p_G\|\widehat{\mathbf{h}}_{m,d}\|^2, \quad (6)$$

where p_G is the constant GUs' transmit power. The inter-UAV interference to the UAV-m is calculated as:

$$I'_m = \sum_{m'\neq m,m'\in\mathcal{M}} \sum_{k\in\mathcal{K}} \rho_{m',k}p_G\|\widehat{\mathbf{h}}_{m,k}\|^2. \quad (7)$$

Building on (6) and (7), the signal-to-interference-plus-noise ratio (SINR) for the GU-k is expressed as:

$$\text{SINR}_{m,k} = \frac{\rho_{m,k}p_G\|\widehat{\mathbf{h}}_{m,k}\|^2}{I_{m,k} + I'_m + \sigma^2}, \quad (8)$$

where σ^2 denotes the background noise power. For the last decoded GU, its SINR calculation removes all intra-UAV interference and only suffers from the inter-UAV interference. To ensure successful decoding, we have the following minimum SINR requirement constraint in each time slot:

$$\text{SINR}_{m,k} \geq \gamma, \forall m \in \mathcal{M}, \forall k \in \mathcal{K}, \quad (9)$$

where γ denotes the decoding threshold of the SIC receivers. To this end, the data rate from the GU-k to the UAV-m is represented by $r_{m,k} = \log(1 + \text{SINR}_{m,k})$.

In each time slot, all GUs consistently perform sensing the surrounding information. Each GU will offload the sensing data to one UAV when they establish an association. Let $D_k[n]$ denote the GU-k's data buffer size in the n-th time slot and $D_{\max}$ represent the maximum capacity of GUs' data queue. Thus, we can represent the GU-k's buffer dynamics as follows:

$$D_k[n] = \left[\left[D_k[n-1] + D_{s,k}[n]\right]^{D_{\max}} - \sum_{m\in\mathcal{M}} \tau r_{m,k}[n]\right]_0, \quad (10)$$

where $[\cdot]^{D_{\max}} \triangleq \min\{\cdot, D_{\max}\}$ and $[\cdot]_0 \triangleq \max\{\cdot, 0\}$. The GUs' sensing data at the beginning of each time slot has a random size, which is denoted as $D_{s,k}[n]$. The collected data by the UAV-m in each time slot is represented as follows:

$$R_m = \sum_{k\in\mathcal{K}} \min\{\tau r_{m,k}, D_k\}. \quad (11)$$

Note that the UAVs' trajectories and the ARIS's passive beamforming dynamically control the channel conditions for data transmission. Thus, we can jointly optimize them to improve the GUs' transmission performance.

IV. OPTIMIZATION-DRIVEN LEARNING SOLUTION TO THROUGHPUT MAXIMIZATION PROBLEM

We aim to maximize the overall network throughput by jointly optimizing all UAVs' trajectories $\{\mathbf{q}_i\}_{i\in\mathcal{I}}$, GUs' association $\boldsymbol{\rho} = \{\rho_{m,k}\}_{m\in\mathcal{M},k\in\mathcal{K}}$, and the ARIS's passive beamforming $\boldsymbol{\theta}$ strategies. The throughput maximization problem is formulated as follows:

$$\max_{\mathbf{q}_i,\boldsymbol{\rho},\boldsymbol{\theta}} \sum_{n\in\mathcal{N}} \sum_{m\in\mathcal{M}} R_m \text{ s.t. } (1)-(11). \quad (12)$$

The ARIS's passive beamforming dynamically reconfigures the channel conditions and thus influences the UAVs' decisions on their next positions. Meanwhile, the UAVs' trajectories determine the GUs' data transmissions. Typically, it is possible to use the classic AO method to tackle problem (12). However, it generally needs a central server to collect the global information to perform the optimizations, and then distribute the solutions to each wireless device, which will cause the huge signaling and computational overhead. Moreover, executing the optimization ahead of the UAVs' flight process relies on the future channel state information (CSI), which is difficult to obtain in dynamic network environments. Another feasible approach is to use the DRL method that is more flexible to handle the high-dimensional and complex problems. A straightforward application of the DRL method to solve problem (12) is to establish multiple agents to jointly
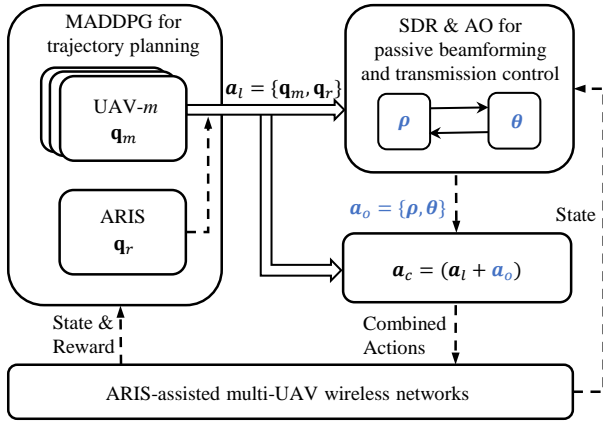
Fig. 2: The proposed O-HDRL framework.

adapt the UAVs' trajectories, the GUs' association, and the ARIS's passive beamforming strategies. However, the high-dimensional action space requires a large number of training samples to attain sufficient information for mature actions.

The intuition of the proposed O-HDRL framework is to decompose problem (12) into two parts, as shown in Fig. 2. The MADDPG agents first adjust all UAVs' next positions (denoted as $\mathbf{q_m}$ and $\mathbf{q_r}$ for the UAV-m and the ARIS, respectively) by fixing the ARIS's passive beamforming θ and GUs' association ρ. After obtaining the UAVs' trajectories, the ARIS's passive beamforming and GUs' association strategies are adapted by the optimization methods. The multiple agents execute the combined action from both DRL and optimization in the wireless environment. The state is updated and the feedback reward guides the agents' next round of learning. The optimization part can be also regarded as supervision to drive the DRL's learning process [34]. The above concept is expected to efficiently improve the DRL training rate and learning performance. Additionally, the UAVs can make online decisions based on the real-time CSI in each time slot without the need for the future CSI, which makes it more suitable for practical implementation. The MADDPG method can be replaced with other MADRL algorithms as needed. We focus on integrating the model-based optimization module into the model-free MADRL method to improve their learning efficiency and performance. In the sequel, we detail the O-HDRL framework for optimization and learning parts, respectively.

A. MADDPG for Trajectory Planning

We first reformulate the UAVs' trajectory planning optimization into a Markov decision process (MDP), which is characterized as a tuple $(\mathcal{S}, (\mathcal{A}_l, \mathcal{A}_o), \mathcal{R})$. The state space is denoted as $\mathcal{S}$, where $\mathbf{s} = \{\mathbf{s}_i\}_{i \in \mathcal{I}} \in \mathcal{S}$ denote the UAVs' states including the channel information and the UAVs' locations. The action space is divided into two subsets, i.e., the learning action space $\mathcal{A}_l$ and the optimization action space $\mathcal{A}_o$. The learning action $\mathbf{a}_l = \{\mathbf{a}_{l,i}\}_{i \in \mathcal{I}} \in \mathcal{A}_l$ is adapted by the MAD-DPG, which includes the UAVs' flying direction and speed in each time slot. The optimization action $\mathbf{a}_o \in \mathcal{A}_o$ consists of the ARIS's passive beamforming and GUs' association strategies obtained by the optimization method. We define the UAVs' combined actions from the MADDPG and the optimization

as $\mathbf{a}_c = \{\mathbf{a}_{c,i}\}_{i \in \mathcal{I}}$ with $\mathbf{a}_{c,i} = \{\mathbf{a}_{l,i}, \mathbf{a}_o\}, i \in \mathcal{I}$. The reward $R(\mathbf{s}_i, \mathbf{a}_{c,i}) \in \mathcal{R}$ evaluates the goodness of the state-action pair $(\mathbf{s}_i, \mathbf{a}_{c,i})$, and then drives each agent to make better decisions.

We design distinct reward functions for the agents of the UAVs and the ARIS, as they have the different objectives during their operations. In particular, the UAVs focus on maximizing data transmission from the GUs, while the ARIS aims to improve the channel conditions for the UAVs. Thus, we define the reward functions $R_{m,u}$ and R_r for the UAV-m and the ARIS, respectively, as follows:

$$R_{m,u} = \ell_u \sum_{k \in \mathcal{K}^m} \min\{\tau \log(1 + \text{SINR}_{m,k}), D_k\} - \jmath_u, \quad (13a)$$

$$R_r = \ell_r \sum_{m \in \mathcal{M}} \sum_{k \in \mathcal{K}} \rho_{m,k} \mathbf{h}_{m,r} \text{diag}(\mathbf{h}_{r,k}) \theta - \jmath_r, \quad (13b)$$

where ℓ_u and ℓ_r are the weighting coefficients of the UAV-m and the ARIS, respectively. The first part of $R_{m,u}$ denotes the UAV-m's collected data from the GUs. The first part of R_r represents the ARIS's channel improvement to all UAVs. The factors $\jmath_u$ and $\jmath_r$ are the penalty terms assigned to the UAVs and the ARIS, respectively, if the constraints in problem (12) do not hold. The MADDPG learning process continues until all agents's policies yield the stable rewards.

During the learning, each agent possesses a pair of actor and critic networks. These networks serve to approximate the policy and value function enabling the agents to adapt and evaluate actions within each training step [35]. The weighting parameters of the i-th agent's actor and critic networks are denoted as ϕ_i and ω_i, respectively. Given the current state $\mathbf{s}_i$, the actor network $\boldsymbol{\pi}(\mathbf{s}_i|\phi_i)$ generates a deterministic policy $\mathbf{a}_{l,i} = \boldsymbol{\pi}(\mathbf{s}_i|\phi_i)$ to maximize the value function $\mathcal{J}_i(\phi_i)$, which is evaluated by the i-th critic networks defined as follows:

$$\mathcal{J}_i(\phi_i) \approx \mathbb{E}_{\mathcal{D}}\left[Q_i(\mathbf{s}, \mathbf{a}_{c,i}|\omega_i)\right]. \quad (14)$$

The Q-value $Q_i(\mathbf{s}, \mathbf{a}_{c,i}|\omega_i)$ with weighting parameter ω_i is generated by the i-th critic network to evaluate the performance of state-action pair $(\mathbf{s}, \mathbf{a}_{c,i})$. Given the samples from the experience replay buffer $\mathcal{D}$, the expectation of the Q-value approximates $\mathcal{J}_i(\phi_i)$, which depends on both ϕ_i and ω_i from the i-th actor and critic networks, respectively. We update the weighting parameter ϕ_i of each actor network to enable the agents to choose more preferable actions. The actor network aims to maximize the value function $\mathcal{J}_i(\phi_i)$ evaluated from the critic network. Hence, we update the actor network's parameter ϕ_i by using the deterministic policy gradient method [36], and the derivative of $\mathcal{J}_i(\phi_i)$ is represented as follows:

$$\nabla \mathcal{J}_i(\phi_i) = \mathbb{E}_{\mathcal{D}}\left[\nabla_{\mathbf{a}_{l,i}}[Q_i(\mathbf{s}, \mathbf{a}_{c,i}|\omega_i)\nabla_{\phi_i}\boldsymbol{\pi}(\mathbf{s}_i)\right]. \quad (15)$$

The i-th critic network is updated by the temporal difference (TD) error. The TD error is computed as the difference between the real-time Q-value $Q_i(\mathbf{s}, \mathbf{a}_{c,i}|\omega_i)$ and its target value $y_t = R(\mathbf{s}_i, \mathbf{a}_{c,i}) + \gamma_d Q_i'$, where Q_i' denotes the Q-value estimated from the target critic network. The discount factor γ_d influences the agents' preference between the short-term and long-term rewards. To minimize the TD error, the loss function $L(\omega_i)$ of the i-th critic network is defined as follows:

$$L(\omega_i) = \mathbb{E}_{\mathcal{D}}[(y_t - Q_i(\mathbf{s}, \mathbf{a}_{c,i}|\omega_i))^2]. \quad (16)$$

By adjusting the weighting parameter ω_i for each critic network, it provides an increasingly accurate estimation of the Q-value to evaluate the agents' decision making. The iterative training of the actor and critic networks drives agents to take action with better reward in dynamic wireless environment.

B. Passive Beamforming and Transmission Control

Given the UAVs' trajectories, we focus on the ARIS's passive beamforming and GUs' association strategies, which are optimized by solving the following subproblem:

$$\max_{\boldsymbol{\rho},\boldsymbol{\theta}} \sum_{n\in\mathcal{N}} \sum_{m\in\mathcal{M}} R_m \quad \text{s.t.} \quad (2)-(11). \tag{17}$$

Problem (17) is challenging to solve due to the coupling feature between $\boldsymbol{\rho}$ and $\boldsymbol{\theta}$. Hence, we propose an AO method to decompose problem (17) into two subproblems (i.e., ARIS's passive beamforming and GUs' transmission control subproblems) and solve them alternately.

We first solve the ARIS's passive beamforming subproblem and reformulate the UAV-m's rate R_m^l as follows:

$$R_m^l = \log\Big(\sum_{d\in\mathcal{K}^m} \rho_{m,d} p_G \text{Tr}(\mathbf{Q}_{m,d}\boldsymbol{\Phi})$$

$$+ \sum_{m'\neq m,m'\in\mathcal{M}} \sum_{d\in\mathcal{K}^{m'}} \rho_{m',d} p_G \text{Tr}(\mathbf{Q}_{m,d}\boldsymbol{\Phi}) + \sigma^2 \Big)$$

$$- \log\Big(\sum_{m'\neq m,m'\in\mathcal{M}} \sum_{d\in\mathcal{K}^{m'}} \rho_{m',d} p_G \text{Tr}(\mathbf{Q}_{m,d}\boldsymbol{\Phi}) + \sigma^2 \Big). \tag{18}$$

The auxiliary variables $\{\mathbf{Q}_{m,d}\}_{m\in\mathcal{M},d\in\mathcal{K}}$ and $\boldsymbol{\Phi}$ introduced into (18) are represented as follows:

$$\mathbf{Q}_{m,d} = \begin{bmatrix} \mathbf{H}_{m,d}^H \mathbf{H}_{m,d} & \mathbf{H}_{m,d}^H h_{m,d} \\ h_{m,d}^H \mathbf{H}_{m,d} & h_{m,d}^H h_{m,d} \end{bmatrix}, \tag{19a}$$

$$\boldsymbol{\Phi} = [\boldsymbol{\theta};1][\boldsymbol{\theta};1]^H \succ 0 \text{ and } \Phi_{l,l}=1, \ l=1,\dots,L+1, \tag{19b}$$

where $\Phi_{l,l}$ denotes the diagonal element of $\boldsymbol{\Phi}$. The second term in (18) can be linearly approximated by introducing auxiliary variables $\{\ell_m\}_{m\in\mathcal{M}}$ given point $\ell_{m,0}$, which is reformulated as follows:

$$\log(\ell_m + \sigma^2) \geq \log(\ell_{m,0} + \sigma^2) + \frac{\ell_m - \ell_{m,0}}{\ell_{m,0} + \sigma^2}, \tag{20a}$$

$$\ell_m \geq \sum_{m'\neq m,m'\in\mathcal{M}} \sum_{d\in\mathcal{K}^{m'}} \rho_{m',d} p_G \text{Tr}(\mathbf{Q}_{m,d}\boldsymbol{\Phi}). \tag{20b}$$

By substituting (19) into the constraint (9), it is transformed into a convex form in terms of $\boldsymbol{\Phi}$ as follows:

$$\rho_{m,k} p_G \text{Tr}(\mathbf{Q}_{m,k}\boldsymbol{\Phi}) \geq \gamma\Big(\sum_{d>s_m(k),d\in\mathcal{K}^m} \rho_{m,d} p_G \text{Tr}(\mathbf{Q}_{m,d}\boldsymbol{\Phi})$$

$$+ \sum_{m'\neq m,m'\in\mathcal{M}} \sum_{d\in\mathcal{K}^{m'}} \rho_{m',d} p_G \text{Tr}(\mathbf{Q}_{m,d}\boldsymbol{\Phi}) + \sigma^2 \Big). \tag{21}$$

To this end, we drop the rank-one constraint and adapt passive beamforming matrix $\boldsymbol{\Phi}$ by solving the following subproblem:

$$\max_{\boldsymbol{\Phi}} \sum_{m\in\mathcal{M}} R_m^l \quad \text{s.t.} \quad (2),(4)-(5), \text{ and } (18)-(21). \tag{22}$$

Problem (22) is a standard SDR problem and thus is easily tackled by the standard SDR solver. The rank-one solution can be extracted by using the Gaussian randomization method [34].

Given the ARIS's passive beamforming $\boldsymbol{\theta}$, problem (17) becomes an integer programming problem with respect to the GUs' association strategies $\boldsymbol{\rho}$. We first relax $\boldsymbol{\rho}$ to a continuous variable and then rewrite it into an equivalent form as follows:

$$\sum_{m\in\mathcal{M}} \sum_{k\in\mathcal{K}} (\rho_{m,k} - \rho_{m,k}^2) \leq 0, \tag{23a}$$

$$\sum_{m\in\mathcal{M}} \rho_{m,k} \leq 1 \text{ and } 0 \leq \rho_{m,k} \leq 1, \forall m\in\mathcal{M}, \forall k\in\mathcal{K}. \tag{23b}$$

We further adopt the penalty method to iteratively approximate the constraint (23a). Introducing a penalty coefficient κ and an auxiliary variable λ, we can simply optimize $\boldsymbol{\rho}$ by solving the following subproblem:

$$\max_{\boldsymbol{\rho},\lambda} \sum_{m\in\mathcal{M}} R_m^l - \kappa\lambda \tag{24a}$$

$$\text{s.t.} \quad (2),(4)-(5),(18)-(21), \text{ and } (23b), \tag{24b}$$

$$\sum_{m\in\mathcal{M}} \sum_{k\in\mathcal{K}} (\rho_{m,k} - \rho_{m,k}^2) \leq \lambda. \tag{24c}$$

Constraint (24c) can be approximated to a linear form by the first-order Taylor expansion at given point $\rho_{m,k,0}$ as follows:

$$\sum_{m\in\mathcal{M}} \sum_{k\in\mathcal{K}} \rho_{m,k} + \rho_{m,k,0}^2 - 2\rho_{m,k}\rho_{m,k,0} \leq \lambda. \tag{25}$$

By replacing constraint (24c) with the linear approximation (25), problem (24) becomes a standard convex problem. Problems (22) and (24) are alternately optimized until a stable solution is reached.

The details of the O-HDRL algorithm are summarized in Algorithm 1. Given the ARIS's passive beamforming and the GUs' association strategies, the UAVs' trajectories are updated by using the MADDPG algorithm. Then, following the trajectories, the optimization methods adapt the remaining decision variables in each time slot. To ensure the convergence of the AO algorithm, we compare the solutions of each subproblem in successive iterations and retain the solution with a higher throughput performance. Meanwhile, the O-HDRL framework significantly reduces the action space for all agents, which allows them to search for optimal actions within a lower-dimensional space. This improves not only the stability of convergence but also the overall learning performance. The SCA and the AO methods in lines 6-7 of Algorithm 1 may introduce a high computational complexity $\mathcal{O}\{(L^{3.5} + (KM)^{3.5})I_{\max}\}$ into the learning process, where $I_{\max}$ denotes the maximum number of iterations [37].

To improve the overall learning efficiency, we can design the approximate optimization methods to reduce the computation overhead of the optimization part. Note that the optimization solution within the learning process should be computation-efficient but does not have to be very precise. The GUs' association optimization relies on the GUs' channel conditions, which are dynamically controlled by the ARIS's passive beamforming. To decouple their interdependency, we can consider an ideal scenario where the ARIS has the strong capability to align each GU's reflection link with its direct link to the UAV. As such, we can first optimize the GUs'

Algorithm 1 O-HDRL for UAVs' Trajectory Planning, ARIS's Passive Beamforming, and GUs' Transmission Control

1: Initialize the DNN parameters and control variables for all UAVs agents. Set the optimization threshold $\eta \geq 0$
   % Learning for the UAVs' trajectory planning
2:    Identify the system state $\mathbf{s}_{i,t} \in \mathcal{S}$
3:    Agents adapt trajectory planning strategies $\mathbf{q}_m$ and $\mathbf{q}_r$
   % Optimizing ARIS's passive beamforming and
4:    **GUs' association**
5:    **repeat**
6:      Optimize $\boldsymbol{\theta}$ by solving subproblem (22)
7:      Optimize $\boldsymbol{\rho}$ by solving subproblem (24)
8:    **until** $\{\boldsymbol{\theta}, \rho_{m,k}\}$ reach convergence
9: **for** each UAV agent $i \in \mathcal{I}$ **do**
10:    Execute the combined action $\mathbf{a}_{c,i} = \{\mathbf{a}_{l,i}, \mathbf{a}_l\}$
11:    Observe the reward function $R(\mathbf{s}_i, \mathbf{a}_{c,i})$
12:    Record the transition to the next state $\mathbf{s}'_i$
13:    Store the transition sample $(\mathbf{s}_i, \mathbf{a}_{c,i}, R(\mathbf{s}_i, \mathbf{a}_{c,i}), \mathbf{s}'_i)$
14:    Update the parameters of actor- and critic-networks
15: **end for**

association strategy and then proceed with optimizing the ARIS's passive beamforming strategy. This transforms the AO method into a two-step process, reducing the computational complexity significantly. Furthermore, instead of using the SDR for the ARIS's passive beamforming strategy, we can heuristically align the ARIS's phase shift $\boldsymbol{\theta}$ to match the strongest UAV-GU channel within the ARIS's coverage. Since the GUs with stronger channels are decoded earlier in the SIC process, they always tolerate stronger interference from other GUs, which are decoded later. Thus, aligning the phase shift with the strongest GU-UAV channel makes it easier to meet its decoding requirement. As such, we only need to optimize the GUs' association strategy $\boldsymbol{\rho}$ with the fixed channel conditions, and thus the internal iterations between problems (22) and (24) can be avoided [11]. The computational complexity of the approximate solution is significantly reduced to $\mathcal{O}\{L + (KM)^{3.5}\}$.

V. DUAL-MODE UAVs WITH ADAPTIVE MODE SWITCHING

Deploying an ARIS can improve the UAVs' channel conditions on its trajectory to provide dynamic services for the GUs' data transmissions. Besides the fixed-mode ARIS, we extend it to the dual-mode UAVs to further improve the network flexibility and network capacity. In the dual-mode UAV system, each UAV is equipped with both a passive ARIS and an active antenna system. As such, all UAVs are capable of either passively reflecting (referred to as P-UAV) or actively collecting data (referred to as A-UAV) according to channel conditions and GUs' traffic demands, as shown in Fig. 3. It is expected that the dual-mode UAVs can offer increased flexibility for the network's interference management and capacity enhancement. In particular, when multiple A-UAVs simultaneously collect the GUs' data, each of them suffers co-channel interference from the others. The interference will become stronger as the multiple A-UAVs are
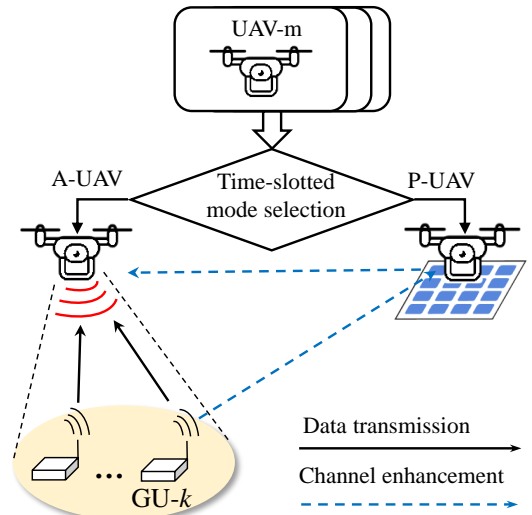

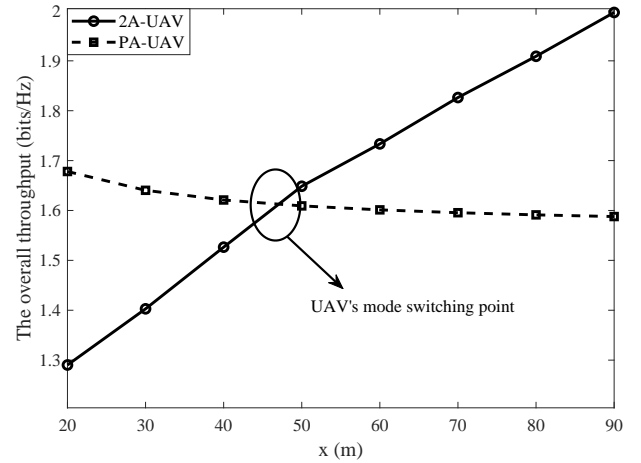Fig. 3: Dual-mode UAV-assisted wireless networks.

Fig. 4: An example for the dual-mode switching scheme.

in close proximity, which leads to a significant reduction of the overall transmission efficiency. However, if some A-UAVs are capable of switching to the P-UAVs, they can provide channel enhancement instead of the harmful interference signals to the A-UAVs. Meanwhile, multiple P-UAVs can generate collaborative beamforming to achieve higher spatial multiplexing gain for the NOMA GUs' transmissions [38]. In another aspect, when the GUs' traffic load becomes unbalanced over the service area, the P-UAVs may also switch back to the A-UAVs to help the heavy-loaded A-UAVs. The UAVs' flexible mode switching on demand can effectively enhance the overall network capacity and throughput.

To verify the effectiveness of the dual-mode switching scheme, we present an experimental confirmation to compare the overall throughput achieved by two UAVs' operation schemes, i.e, the case with two A-UAVs (denoted by 2A-UAV in Fig. 4) and the case with a hybrid of P-UAV and A-UAV (denoted by PA-UAV in Fig. 4). This experiment involves multiple GUs uniformly distributed in the simulation area. We aim to compare the overall throughput achieved by the two-UAV system as we vary the distance between the two UAVs. As depicted in Fig. 4, as two UAVs are close to each other, the PA-UAV scheme achieves a higher throughput. As the dis-

tance increases, the interference between the UAVs decreases, resulting in a higher rate of the 2A-UAV scheme. It is evident that we can switch the UAVs' operating modes on demand to achieve a better transmission performance. Note that the dual-mode UAVs may introduce additional energy consumption to operate both the active and passive systems. However, we mainly focuses on exploring the transmission improvement by the dual-mode switching scheme, and thus we omit the UAV's energy consumption in this paper. The exploration of the dual-mode UAVs' communication in energy-limited scenarios will be a promising direction for our future research.

A. Optimizing UAVs' Mode Switching

We define binary variables $\boldsymbol{\alpha} = \{\alpha_m\}_{m \in \mathcal{M}}$ as the UAVs' operation modes, and $\alpha_m = 1$ (or $\alpha_m = 0$) represents that the UAV-m works at the active (or the passive) mode in the current time slot. Given a mode selection, the equivalent channel from the GU-k to the A-UAV-m is represented as follows:

$$\widehat{\mathbf{h}}_{m,k} = \underbrace{\sum_{m' \neq m, m' \in \mathcal{M}} (1 - \alpha_{m'}) \mathbf{h}_{m,m'} \text{diag}(\tilde{\mathbf{h}}_{m',k}) \widehat{\boldsymbol{\theta}}_{m'}}_{\text{Channel enhancement by P-UAVs}} + h_{m,k}, \quad (26)$$

where $\widehat{\boldsymbol{\theta}}_m$ denote the P-UAV-m's passive beamforming. The first term of the equivalent channel (26) is enhanced by the P-UAVs' passive beamforming. The UAVs' dynamic mode switching introduces increased controllability to the wireless channel. Hence, it has the potential to improve the channel diversity for the NOMA decoding process and also create more access opportunities for GUs.

Given the mode switching strategy $\boldsymbol{\alpha}$, the data rate of the A-UAV-m from the GU-k is represented as follows:

$$\widehat{r}_{m,k} = \alpha_m \log(1 + \widehat{\text{SINR}}_{m,k}), \quad (27)$$

where $\widehat{\text{SINR}}_{m,k}$ is obtained by introducing the equivalent channel (26) into (8).

By jointly optimizing the dual-mode UAVs' trajectory planning $\mathbf{q}_m$ and mode switching $\boldsymbol{\alpha}$, passive beamforming $\widehat{\boldsymbol{\theta}}_m$ as well as the GUs' transmission control $\boldsymbol{\rho}$ strategies, the throughput maximization problem is updated as follows:

$$\max_{\boldsymbol{\alpha}, \mathbf{q}_m, \widehat{\boldsymbol{\theta}}_m, \boldsymbol{\rho}} \sum_{n \in \mathcal{N}} \sum_{m \in \mathcal{M}} \sum_{k \in \mathcal{K}} \min\{\tau \widehat{r}_{m,k}, D_k\}, \quad (28a)$$

$$\text{s.t.} \quad (1), (3) - (10), \text{ and } (26) - (27). \quad (28b)$$

The coupling of the mode switching and passive beamforming in (26) makes problem (28) challenging to solve. We first reformulate it introducing the following auxiliary matrices:

$$\mathbf{h}_m = [(1 - \alpha_1)\mathbf{h}_{m,1}, \ldots, (1 - \alpha_{m'})\mathbf{h}_{m,m'}], \quad (29a)$$

$$\overline{\mathbf{h}}_{m,k} = [\tilde{\mathbf{h}}_{1,k}, \ldots, \tilde{\mathbf{h}}_{m',k}], \quad (29b)$$

$$\tilde{\boldsymbol{\theta}}_m = [\widehat{\boldsymbol{\theta}}_1, \ldots, \widehat{\boldsymbol{\theta}}_{m'}], \quad (29c)$$

where $m' \neq m$ and $m, m' \in \mathcal{M}$. As such, the equivalent channel $\widehat{\mathbf{h}}_{m,k}$ in (26) can be reformulated as follows:

$$\widehat{\mathbf{h}}_{m,k} = \mathbf{h}_m \text{diag}(\overline{\mathbf{h}}_{m,k}) \tilde{\boldsymbol{\theta}}_m + h_{m,k}. \quad (30)$$

It is seen that the equivalent channel $\widehat{\mathbf{h}}_{m,k}$ in (30) has a similar structure to (2). Therefore, given the mode switching $\boldsymbol{\alpha}$, we
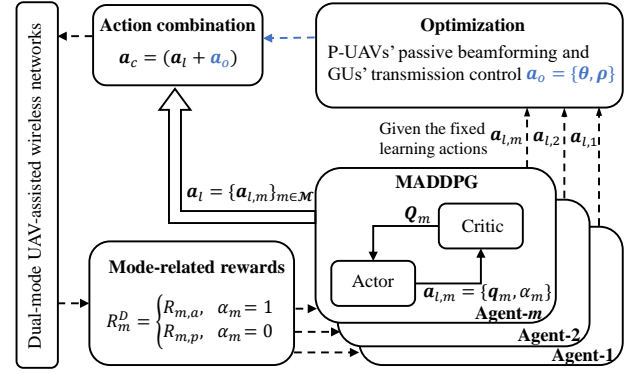

Fig. 5: The O-HDRL for dual-mode switching scheme.

can optimize the dual-mode UAVs' trajectories $\mathbf{q}_m$, the passive beamforming $\tilde{\boldsymbol{\theta}}_m$, and the GUs' transmission control $\boldsymbol{\rho}$ in problem (28) by using the similar method as proposed in the Section IV-B. Thus, the main difference between the fixed-mode ARIS and dual-mode UAV systems lies in searching for the optimal dual-mode switching strategy. From (26), we can observe that as more dual-mode UAVs switch to the P-UAVs, the number of the available reflecting links increases, which can potentially enhance the channel conditions for the A-UAVs. However, this will also reduce the number of the A-UAVs for data collection, as indicated in (27). As such, it is challenging to directly determine the UAVs' mode-switching strategy using a specific closed-form solution.

B. Learning Mode Switching between P-UAV and A-UAV

Intuitively, we can list all possible combinations of the UAVs' operating modes and formulate the optimization problem for each of them. Then, each optimization problem can be solved by Algorithm 1 proposed in the Section IV-B. However, this becomes computationally infeasible, particularly when dealing with a large number of UAVs. Hence, we also employ the MADDPG to update the mode-switching strategy. To proceed, the learning action $\mathbf{a}_{l,m}$ of each agent is modified to include the dual-mode UAVs' next position $\mathbf{q}_m$ and the mode-switching strategy $\boldsymbol{\alpha}$. The design of the reward function should reflect the dual-mode UAVs' different objectives when they dynamically select the different operation modes. Hence, the mode-related reward function R_m^D is represented as follows:

$$R_m^D = \alpha_m R_{m,a} + (1 - \alpha_m) R_{m,p}, \quad (31)$$

where $R_{m,a}$ and $R_{m,p}$ denote the rewards of the A-UAV-m and P-UAV-m, respectively. The A-UAV-m's reward $R_{m,a}$ aims to optimize the data transmission performance, while the P-UAV-m's reward $R_{m,p}$ is designed to improve the channel conditions for all A-UAVs. The expressions of $R_{m,a}$ and $R_{m,p}$ are similar to those in (13). We design a learning framework, as shown in Fig. 5. In each decision slot, the multiple agents first make decisions on the dual-mode UAVs' trajectories and mode switching strategies. Then, each agent adjusts the passive beamforming and the transmission control strategies by the optimization method. By combining the actions and interacting with the environment, each dual-mode UAV can gradually refine its decision and chooses a preferable operation mode.

TABLE I: Parameter settings in the simulations.

| Parameters	Settings
GU's transmit power p_G	30 dBm
Number of the AIRS's elements L	30
Flying altitude of the UAVs H	30 m
Background noise power σ	-90 dBm
Maximum speed of the UAVs $V_{\max}$	50 m/s
Safety distance to the UAVs $D_{\min}$	5 m
Number of time slots N	50
Actor's learning rate	10^{-4}
Critic's learning rate	10^{-3}
Reward discount factor	0.95
Size of the replay buffer	5×10^5
Size of the mini-batch	256
ϵ-greedy coefficient	0.1

VI. NUMERICAL RESULTS

In this section, numerical results are presented to evaluate the performance of the proposed ARIS-assisted and dual-mode switching schemes (denoted as Fixed-ARIS and DM-Switching, respectively) for the multi-UAV wireless networks. We consider $M = 3$ UAVs collecting data from $K = 9$ GUs randomly distributed in a (1000×1000) m^2 area, similar to that in [39]. The UAV-GU and the UAV-UAV links follow the Rican and LoS channel models, respectively. Other parameters are set as in Table I, similar to those in [40].
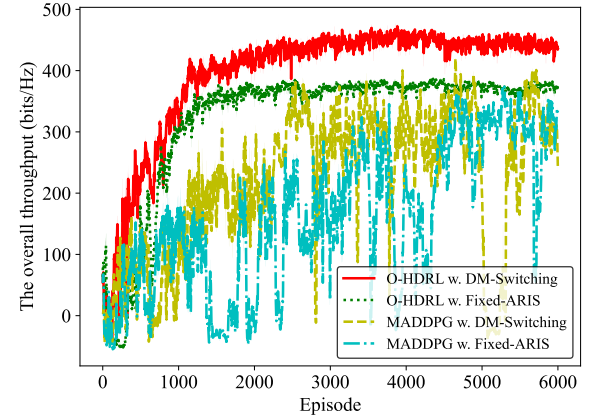
A. Convergence of the O-HDRL Framework

In Fig. 6, we validate the convergence performance of the proposed O-HDRL framework. In the multi-agent learning process, the optimization is continually introduced to refine the policies of the agents. In each time slot, the GUs' association and ARISs' passive beamforming strategies are jointly optimized by using the AO method. Fig. 6(a) illustrates the achievable throughput of different optimization schemes as the algorithm iterates. The AO method is compared with two benchmarks: (i) Fixed Phase-Shift scheme where only GUs' association strategy is optimized with the fixed ARISs' passive beamforming, and (ii) Fixed Association scheme where only the ARISs' passive beamforming strategy is optimized while keeping the GUs' association strategy fixed. The AO method achieves a higher throughput compared to the benchmarks. This validates that the joint optimization of the GUs' association and the ARISs' passive beamforming strategies improves the data transmission efficiency. Furthermore, the AO method shows a rapid convergence, which verifies that subproblems (22) and (24) are efficiently tackled.

Figure 6(b) reveals the convergence and reward performance of the O-HDRL framework with different UAVs' operation modes compared to the conventional MADDPG. In the MADDPG, all control variables are learned by the DNN at each decision-making slot. It is observed that the O-HDRL algorithms achieve superior stability and convergence performance compared to the MADDPG algorithm. Since MADDPG needs to simultaneously explore the ARISs' passive beamforming and the GUs' association strategies, it has a
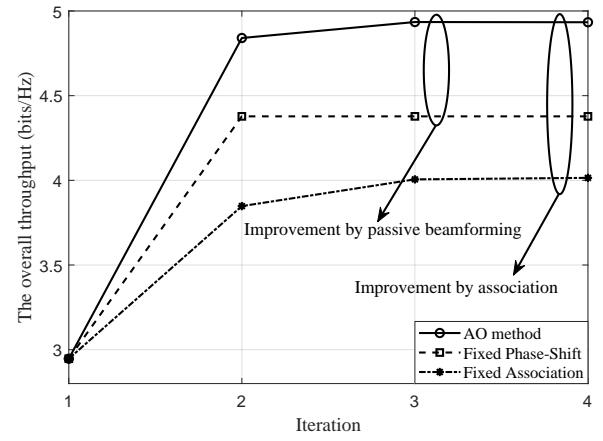
(a) Convergence performance of the optimization in one time slot.

(b) Convergence performance of the O-HDRL algorithms.

Fig. 6: Optimization improves the learning performance.

much higher dimension for the action space compared to that of the O-HDRL. The high-dimensional action space introduces instability to the multi-agent's learning process, which makes it more challenging to explore the optimal policy. For the O-HDRL algorithms, the action space is effectively reduced by introducing the optimization modules. Furthermore, by using the O-HDRL framework, the DM-Switching scheme provides 17.51% higher throughput than that of the Fixed-ARIS scheme. This validates that the UAVs' mode switching on demand can better adapt to dynamic wireless networks.

Figure 7 shows the agents' reward performance of the Fixed-ARIS and the DM-Switching schemes. The Fixed-ARIS scheme's reward performance is shown in Fig. 7(a). It is observed that the ARIS's reward increases in a stepwise manner. This is because the ARIS needs to dynamically fly to a new position to provide channel reflection for the UAVs. When the UAVs change their service places, the ARIS should also move accordingly to achieve considerable channel enhancement. Instead, the DM-Switching scheme can switch UAVs' operating modes according to the channel conditions. When P-UAVs cannot obtain good rewards by providing the channel enhancement for the other A-UAVs, they can switch to A-UAVs to continue improving their rewards by collecting data. As such, the DM-Switching scheme can make more flexible decisions and thus achieve continuous growth in reward performance, as shown in Fig. 7(b).
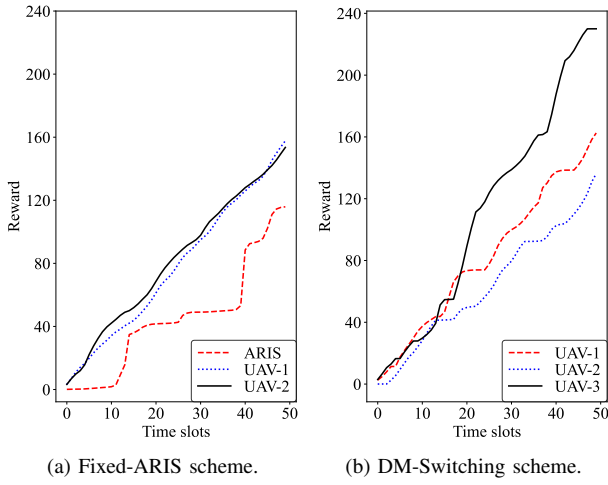
(a) Fixed-ARIS scheme.　　　　(b) DM-Switching scheme.

Fig. 7: Agents' reward improvement in different time slots.

B. Trajectory Planning with Fixed-ARIS and DM-Switching

We present the UAVs' trajectories of the Fixed-ARIS and DM-Switching schemes. We study the UAVs' trajectory planning in two different spatial distributions of the GUs, i.e., the uniform and non-uniform distributions of the GUs. In the non-uniform case, there are a few points of interests (PoIs), which are the cluster centers for the GUs' spatial distribution. The GUs' locations are generated using the Gaussian distribution centered at these PoIs as shown in Fig. 8(b) and (d).

Figure 8(a) shows the trajectory planning for the Fixed-ARIS scheme under the uniform GUs' distribution. The UAVs and the ARIS individually adapt their flying locations based on the current observation of the environment. The trajectory of the ARIS is denoted in red dashed line and the trajectories of the UAVs are marked in blue and black solid lines, respectively. The UAVs can autonomously categorize the GUs into two distinct serving groups according to the GUs' locations. Such a spatial division of the data transmission allows each UAV to focus on the GUs within a designated region, which can reduce the co-channel interference from the other UAVs. An interesting observation is that the ARIS flies between the trajectories of the two UAVs, which provides the channel enhancement and/or interference mitigation for both UAVs simultaneously. Fig. 8(b) reveals the trajectory planning under the non-uniform distribution of the GUs. The UAVs also divide the GUs into two groups, and each UAV is responsible for collecting data from its own region. Interestingly, the trajectory of the ARIS is always close to the UAV-1 assisting its data transmissions. This is because when two UAVs are far from each other, it is challenging for the ARIS to simultaneously enhance the channels of both UAVs. As such, the ARIS tends to assist the UAV-1, which has a higher data transmission load.

Figures 8(c) and (d) illustrate the trajectory planning for the DM-Switching scheme under different GUs' distributions. The dynamic changes of the UAVs' mode switching are indicated by distinct markers on the UAVs' trajectories. The solid and hollow markers represent the A-UAV and the P-UAV, respectively. In the uniform case as shown in Fig. 8(c), the dual-mode UAVs also divide the GUs into different service groups similar to that in the Fixed-ARIS scheme. The dual-mode

(a) Uniform GUs' distribution.　　(b) Non-uniform GUs' distribution.

(c) Uniform GUs' distribution.　　(d) Non-uniform GUs' distribution.

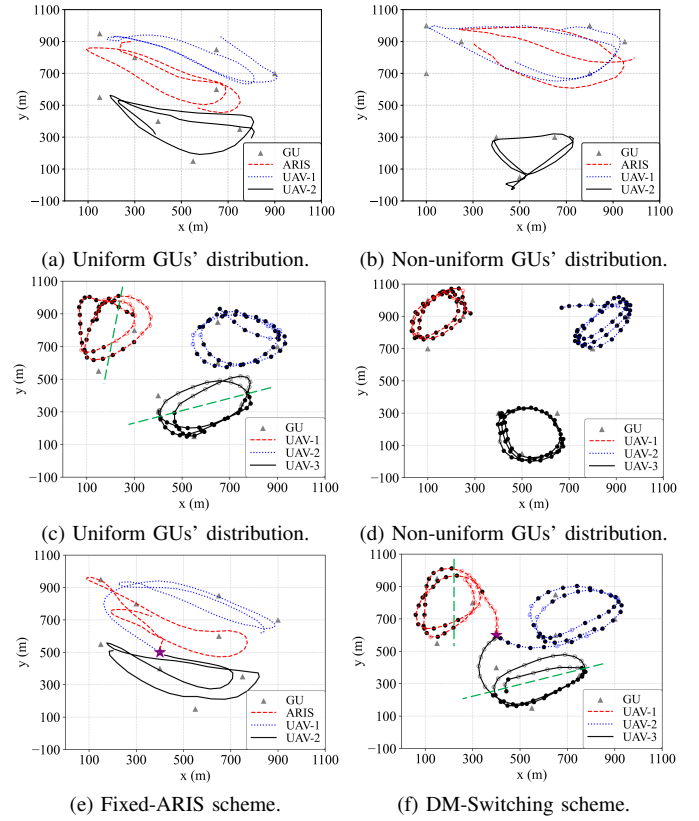(e) Fixed-ARIS scheme.　　　　(f) DM-Switching scheme.

Fig. 8: UAVs' trajectories in Fixed-ARIS and DM-Switching.

UAVs continuously switch modes along with their trajectories. When the dual-mode UAVs are in close proximity, they tend to operate as the P-UAVs. This behavior is driven by the co-channel interference experienced by the A-UAVs when they are in close proximity. By switching some A-UAVs to the P-UAVs, the interference among the A-UAVs is effectively reduced. Meanwhile, the P-UAVs also improve the channel conditions for the A-UAVs. On the contrary, as the dual-mode UAVs fly apart, they will switch back to the A-UAVs to collect data individually from the GUs. We use green dashed lines to broadly divide the mode switching between the A-UAVs and the P-UAVs. It is evident that the UAVs dynamically switch their modes based on the distances between them and the distribution of the GUs' traffic demands. These results verify that the UAVs' dual-mode switching scheme can adaptively handle the dynamic network environment.

Figure 8(d) shows the dual-mode UAVs' trajectory planning in the non-uniform case. The dual-mode UAVs mainly operate as the A-UAVs for data transmissions during the overall trajectories. This is attributed to the large distances between the UAVs, which have reduced interference to each other. Thus, there is no need for the UAVs to switch to the P-UAVs for mitigating the interference. The dual-mode UAVs prefer to operate as the A-UAVs to independently collect the data from the GUs. We further validate the trajectory planning by the Fixed-ARIS and the DM-Switching schemes starting from the same initial location, as shown in Fig. 8(e) and (f). The same initial state location implies that UAVs need greater autonomous capability to divide the GUs and plan reasonable trajectories to ensure the fair services for all GUs.
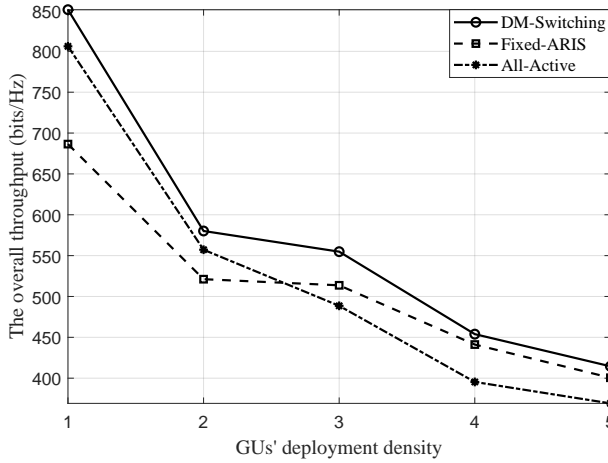
Fig. 9: The impact of GUs' deployment on the throughput.

Fig. 10: The impact of the ARIS's size on the throughput.

We use purple star to represent the UAVs' initial locations. The UAVs in the Fixed-ARIS and DM-Switching scheme can autonomously serve different GU groups and adjust their operation modes according to GUs' spatial locations and channel conditions. This validates that the O-HDRL can balance the UAVs' resource allocation across different regions.

C. Throughput with the Fixed-ARIS and DM-Switching

We delve into the throughput improvement by comparing the DM-Switching, Fixed-ARIS, and All-Active schemes under different system parameters, where the All-Active scheme represents that all dual-mode UAVs keep working as A-UAVs. We investigate the impact of different interference levels on throughput performance by varying GUs' deployment density. We first normalize the GUs' deployment density and increase the density from 1 to 5. As shown in Fig. 9, the overall throughput of all schemes decreases with the increase of the GUs' density. This is because the increased interference diminishes the NOMA transmission efficiency and enforces the competition for GUs' access. When the GUs' deployment density is lower than 2, the All-Active scheme exhibits a higher throughput than the Fixed-ARIS scheme. As the GUs' density increases, the throughput performance of the Fixed-ARIS scheme surpasses that of the All-Active scheme. This is because the All-Active scheme efficiently utilizes A-UAVs to collect data from the GUs under low interference and good channel conditions. However, as GUs' density increases, the interference becomes more severe, which intensifies channel competition. In contrast, the Fixed-ARIS scheme employs the ARIS to refine the channel conditions to lower the interference, achieving a higher throughput in a high-density scenario. Importantly, the DM-Switching scheme achieves the highest overall throughput due to the mode switching on demand. In particular, when the GUs' deployment density is 1, the DM-Switching scheme improves the throughput by 23.97% compared to the Fixed-ARIS scheme. When the GUs' deployment density is 5, the throughput enhancement of the DM-Switching scheme is 12.29% compared to the All-Active scheme.

Different access methods have unique characteristics and affect throughput performance in different ways. Besides NOMA method, other effective access methods, such as time
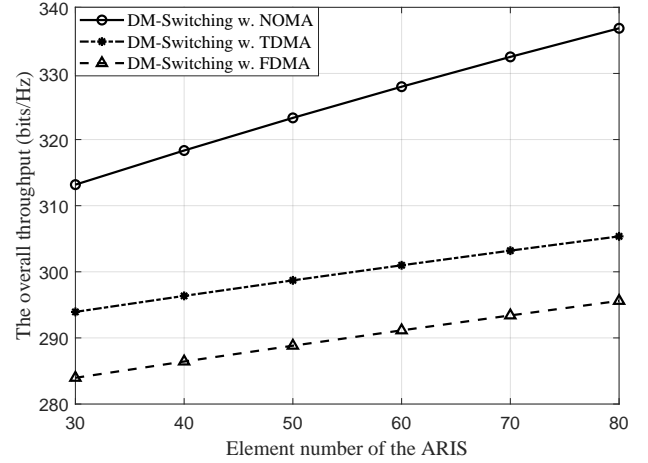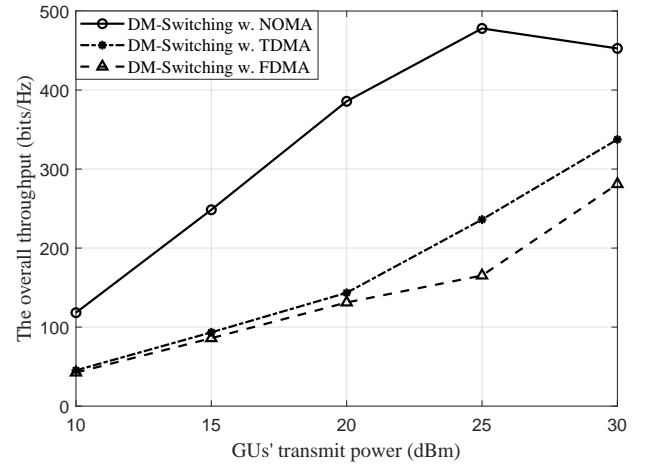

Fig. 11: The impact of GUs' transmit power on the throughput.

division multiple access (TDMA) and frequency division multiple access (FDMA), can separate GUs in time and frequency domains, respectively. To explore which access method is more suitable, we focus on the DM-Switching scheme under different GUs' access methods, i.e., NOMA, TDMA, and FDMA, which are denoted as DM-Switching w. NOMA, DM-Switching w. TDMA, and DM-Switching w. FDMA, respectively. In Fig. 10, we investigate the overall throughput as the number of the ARISs' elements L varies from 30 to 80. It is observed that all schemes' throughput performance increases with the increased ARISs' elements. This is because the large-size ARISs enhance channel conditions efficiently. The NOMA method exhibits superior performance compared to TDMA and FDMA methods. This is because the additional channel orthogonality introduced by the large-size ARISs allowing the NOMA GUs to better utilize the channel resources.

In Fig. 11, we further investigate the throughput performance of the DM-Switching scheme under different GUs' transmit power. As GUs' transmit power increases, the interference among GUs becomes more severe. It is shown that the throughput performance of the NOMA scheme surpasses that of the TDMA and the FDMA schemes. As the GUs' transmit power continuously increases, the throughput performance of the NOMA scheme initially rises and then reduces. This is because given a low transmit power of the GUs, the NOMA

method allows multiple GUs to transmit data simultaneously without much mutual interference. However, as the interference among GUs becomes stronger, it becomes hard to meet the SIC decoding requirements, resulting in a decline in the overall throughput performance.

VII. Conclusions

In this paper, we have investigated a multi-UAV-assisted NOMA transmission system. A fixed-mode ARIS scheme has been proposed to reduce the co-channel interference among the UAVs and improve the channel diversity for the GUs' NOMA transmissions. The ARIS brings greater flexibility into the channel adjustments, enabling the multi-UAV systems to better adapt to dynamic network environments. We have further proposed a dual-mode switching scheme to better deal with network dynamics. The dual-mode UAVs adaptively switch between the A-UAV mode and the P-UAV mode for either data transmissions or channel improvement according to the network dynamics. We have maximized the overall throughput in the proposed two schemes by jointly optimizing the UAVs' trajectory planning and mode switching, the ARIS's passive beamforming, and the GUs' transmission control strategies. We have proposed an O-HDRL framework that integrates the advantages of the conventional model-free DRL and the optimization methods. Simulation results have demonstrated that the O-HDRL algorithm significantly improves the learning stability and performance compared to the benchmark methods. Moreover, the dual-mode switching scheme has shown a superior throughput performance compared to the fixed-mode ARIS scheme, especially in high-density networks.

References

[1] R. Chen, M. Liu, Y. Hui, N. Cheng, and J. Li, "Reconfigurable intelligent surfaces for 6G IoT wireless positioning: A contemporary survey," *IEEE Internet Things J.*, vol. 9, no. 23, pp. 23 570–23 582, Dec. 2022.

[2] X. Liu, Y. Liu, Y. Chen, and H. V. Poor, "RIS enhanced massive non-orthogonal multiple access networks: Deployment and passive beamforming design," *IEEE J. Sel. Area. Commun.*, vol. 39, no. 4, pp. 1057–1071, Apr. 2021.

[3] Z. Ding, L. Lv, F. Fang, O. A. Dobre, G. K. Karagiannidis, N. Al-Dhahir, R. Schober, and H. V. Poor, "A state-of-the-art survey on reconfigurable intelligent surface-assisted non-orthogonal multiple access networks," *Proc. IEEE*, vol. 110, no. 9, pp. 1358–1379, Sep. 2022.

[4] X. Liu, Y. Liu, and Y. Chen, "Machine learning empowered trajectory and passive beamforming design in UAV-RIS wireless networks," *IEEE J. Sel. Area. Commun.*, vol. 39, no. 7, pp. 2042–2055, Jul. 2021.

[5] M. Vaezi, A. Azari, S. R. Khosravirad, M. Shirvanimoghaddam, M. M. Azari, D. Chasaki, and P. Popovski, "Cellular, wide-area, and non-terrestrial IoT: A survey on 5G advances and the road toward 6G," *IEEE Commun. Surveys Tuts.*, vol. 24, no. 2, pp. 1117–1174, 2nd Quart. 2022.

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

[7] B. Duo, M. He, Q. Wu, and Z. Zhang, "Joint dual-UAV trajectory and RIS design for ARIS-assisted aerial computing in IoT," *IEEE Internet Things J.*, vol. 10, no. 22, pp. 19 584–19 594, Nov. 2023.

[8] L. Ge, H. Zhang, J.-B. Wang, and G. Y. Li, "Reconfigurable wireless relaying with multi-UAV-carried intelligent reflecting surfaces," *IEEE Trans. Veh. Technol.*, vol. 72, no. 4, pp. 4932–4947, Apr. 2023.

[9] S. Zhao, Y. Long, B. Gu, N. C. Luong, B. Lyu, and S. Gong, "Exploiting mode-switching between aerial-RIS and active radio in UAV-assisted wireless networks," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Dubai, United Arab Emirates, Apr. 2024, pp. 1–6, in press.

[10] M. Fu, Y. Zhou, Y. Shi, and K. B. Letaief, "Reconfigurable intelligent surface empowered downlink non-orthogonal multiple access," *IEEE Trans. Commun.*, vol. 69, no. 6, pp. 3802–3817, Jun. 2021.

[11] S. Gong, L. Cui, B. Gu, B. Lyu, D. T. Hoang, and D. Niyato, "Hierarchical deep reinforcement learning for age-of-information minimization in IRS-aided and wireless-powered wireless networks," *IEEE Trans. Wireless Commun.*, vol. 22, no. 11, pp. 8114–8127, Nov. 2023.

[12] B. Zheng, Q. Wu, and R. Zhang, "Intelligent reflecting surface-assisted multiple access with user pairing: NOMA or OMA?" *IEEE Commun. Lett.*, vol. 24, no. 4, pp. 753–757, Apr. 2020.

[13] Z. Li, M. Chen, Z. Yang, J. Zhao, Y. Wang, J. Shi, and C. Huang, "Energy efficient reconfigurable intelligent surface enabled mobile edge computing networks with NOMA," *IEEE Trans. Cognit. Commun. Netw.*, vol. 7, no. 2, pp. 427–440, Jun. 2021.

[14] G. Yang, X. Xu, Y.-C. Liang, and M. D. Renzo, "Reconfigurable intelligent surface-assisted non-orthogonal multiple access," *IEEE Trans. Wireless Commun.*, vol. 20, no. 5, pp. 3137–3151, May 2021.

[15] Y. Liu, X. Mu, X. Liu, M. Di Renzo, Z. Ding, and R. Schober, "Reconfigurable intelligent surface-aided multi-user networks: Interplay between NOMA and RIS," *IEEE Wireless Commun.*, vol. 29, no. 2, pp. 169–176, Apr. 2022.

[16] X. Liu, Y. Liu, and Y. Chen, "Machine learning empowered trajectory and passive beamforming design in UAV-RIS wireless networks," *IEEE J. Sel. Area. Commun.*, vol. 39, no. 7, pp. 2042–2055, Jul. 2021.

[17] Y. Xiao, D. Tyrovolas, S. A. Tegos, P. D. Diamantoulakis, Z. Ma, L. Hao, and G. K. Karagiannidis, "Solar powered UAV-mounted RIS networks," *IEEE Commun. Lett.*, vol. 27, no. 6, pp. 1565–1569, Jun. 2023.

[18] H. Zhang, M. Huang, H. Zhou, X. Wang, N. Wang, and K. Long, "Capacity maximization in RIS-UAV networks: A DDQN-based trajectory and phase shift optimization approach," *IEEE Trans. Wireless Commun.*, vol. 22, no. 4, pp. 2583–2591, Apr. 2023.

[19] P. S. Aung, Y. M. Park, Y. K. Tun, Z. Han, and C. S. Hong, "Energy-efficient communication networks via multiple aerial reconfigurable intelligent surfaces: DRL and optimization approach," *IEEE Trans. Veh. Technol.*, vol. 73, no. 3, pp. 4277–4292, Mar. 2024.

[20] S. Han, J. Wang, L. Xiao, and C. Li, "Broadcast secrecy rate maximization in UAV-empowered IRS backscatter communications," *IEEE Trans. Wireless Commun.*, vol. 22, no. 10, pp. 6445–6458, Oct. 2023.

[21] M. Hua, L. Yang, C. Li, Q. Wu, and A. L. Swindlehurst, "Throughput maximization for UAV-aided backscatter communication networks," *IEEE Trans. Commun.*, vol. 68, no. 2, pp. 1254–1270, Feb. 2020.

[22] H. Hao, C. Xu, W. Zhang, S. Yang, and G.-M. Muntean, "Joint task offloading, resource allocation, and trajectory design for multi-uav cooperative edge computing with task priority," *IEEE Trans. Mobile Comput.*, pp. 1–16, 2024.

[23] S. Li, B. Duo, X. Yuan, Y.-C. Liang, and M. Di Renzo, "Reconfigurable intelligent surface assisted uav communication: Joint trajectory design and passive beamforming," *IEEE Wireless Commun. Lett.*, vol. 9, no. 5, pp. 716–720, May 2020.

[24] J. Wang, Z. Na, and X. Liu, "Collaborative design of multi-UAV trajectory and resource scheduling for 6G-enabled internet of things," *IEEE Internet Things J.*, vol. 8, no. 20, pp. 15 096–15 106, Oct. 2021.

[25] J. Xu, X. Kang, R. Zhang, Y.-C. Liang, and S. Sun, "Optimization for master-UAV-powered auxiliary-aerial-IRS-assisted IoT networks: An option-based multi-agent hierarchical deep reinforcement learning approach," *IEEE Internet Things J.*, vol. 9, no. 22, pp. 22 887–22 902, Nov. 2022.

[26] Q. Wu and R. Zhang, "Joint active and passive beamforming optimization for intelligent reflecting surface assisted SWIPT under QoS constraints," *IEEE J. Sel. Area. Commun.*, vol. 38, no. 8, pp. 1735–1748, Aug. 2020.

[27] Z. Wei, Y. Cai, Z. Sun, D. W. K. Ng, J. Yuan, M. Zhou, and L. Sun, "Sum-rate maximization for IRS-assisted UAV OFDMA communication systems," *IEEE Trans. Wireless Commun.*, vol. 20, no. 4, pp. 2530–2550, Apr. 2021.

[28] Q. Wu, X. Zhou, W. Chen, J. Li, and X. Zhang, "IRS-aided WPCNs: A new optimization framework for dynamic IRS beamforming," *IEEE Trans. Wireless Commun.*, vol. 21, no. 7, pp. 4725–4739, Jul. 2022.

[29] A. Feriani and E. Hossain, "Single and multi-agent deep reinforcement learning for AI-enabled wireless networks: A tutorial," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 2, pp. 1226–1252, 3rd Quart. 2021.

[30] L. Wang, K. Wang, C. Pan, and N. Aslam, "Joint trajectory and passive beamforming design for intelligent reflecting surface-aided UAV communications: A deep reinforcement learning approach," *IEEE Trans. Mobile Comput.*, vol. 22, no. 11, pp. 6543–6553, Nov. 2023.

[31] G. Chen, X. B. Zhai, and C. Li, "Joint optimization of trajectory and user association via reinforcement learning for UAV-aided data collection in wireless networks," *IEEE Trans. Wireless Commun.*, vol. 22, no. 5, pp. 3128–3143, May 2023.

[32] A. Khalili, E. M. Monfared, S. Zargari, M. R. Javan, N. M. Yamchi, and E. A. Jorswieck, "Resource management for transmit power minimization in UAV-assisted RIS hetnets supported by dual connectivity," *IEEE Trans. Wireless Commun.*, vol. 21, no. 3, pp. 1806–1822, Mar. 2022.

[33] Z. Ding, D. Xu, R. Schober, and H. V. Poor, "Hybrid NOMA offloading in multi-user MEC networks," *IEEE Trans. Wireless Commun.*, vol. 21, no. 7, pp. 5377–5391, Jul. 2022.

[34] S. Zhao, Y. Liu, S. Gong, B. Gu, R. Fan, and B. Lyu, "Computation offloading and beamforming optimization for energy minimization in wireless-powered IRS-assisted MEC," *IEEE Internet Things J.*, vol. 10, no. 22, pp. 19 466–19 478, Nov. 2023.

[35] N. C. Luong, D. T. Hoang, S. Gong, D. Niyato, P. Wang, Y.-C. Liang, and D. I. Kim, "Applications of deep reinforcement learning in communications and networking: A survey," *IEEE Commun. Surveys Tut.*, vol. 21, no. 4, pp. 3133–3174, Fourthquarter 2019.

[36] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, "Deep reinforcement learning: A brief survey," *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 26–38, Nov. 2017.

[37] Z.-q. Luo, W.-k. Ma, A. M.-c. So, Y. Ye, and S. Zhang, "Semidefinite relaxation of quadratic optimization problems," *IEEE Signal Process. Mag.*, vol. 27, no. 3, pp. 20–34, May 2010.

[38] B. Zheng, C. You, and R. Zhang, "Double-IRS assisted multi-user MIMO: Cooperative passive beamforming design," *IEEE Trans. Wireless Commun.*, vol. 20, no. 7, pp. 4513–4526, Jul. 2021.

[39] Y. Zeng, J. Xu, and R. Zhang, "Energy minimization for wireless communication with rotary-wing UAV," *IEEE Trans. Wireless Commun.*, vol. 18, no. 4, pp. 2329–2345, Apr. 2019.

[40] M. Hua, L. Yang, Q. Wu, C. Pan, C. Li, and A. L. Swindlehurst, "UAV-assisted intelligent reflecting surface symbiotic radio system," *IEEE Trans. Wireless Commun.*, vol. 20, no. 9, pp. 5769–5785, Sep. 2021.