Efficient Optimization in RIS-Assisted UAV System Using Deep Reinforcement Learning for mmWave-NOMA 6G Communications

Sima Sobhi-Givi, Mahdi Nouri, *Senior Member, IEEE*, Mahrokh G. Shayesteh, *Senior Member, IEEE*, Hamid Behroozi, *Member, IEEE*, Hyun Han Kwon, *Member, IEEE*, and Md. Jalil Piran, *Senior Member, IEEE*

***Abstract*—In the evolving landscape of wireless communications for 5G, 6G, and beyond, the deployment of unmanned aerial vehicles (UAVs) has emerged as a groundbreaking strategy to expand coverage areas due to their flexibility and ease of deployment. Simultaneously, reflecting intelligent surfaces (RISs) have introduced a transformative paradigm aimed at improving key performance metrics, such as average sum-rate and energy efficiency (EE). The seamless integration of advanced technologies, including UAVs, RIS, and nonorthogonal multiple access (NOMA), presents a promising avenue for significantly boosting the performance and efficiency of next-generation communication systems. This study investigates EE maximization for two scenarios in a NOMA-enabled mmWave network: 1) multi-UAV-mounted base stations (BSs) and 2) multi-UAV-mounted distributed RIS. In both cases, each UAV serves a NOMA cluster with imperfect successive interference cancellation (SIC), capturing the impact of hardware impairments in real-world NOMA systems. For each scenario, an optimization problem is formulated to maximize EE by jointly optimizing the beamforming matrix, phase shift matrix, NOMA power allocation, and UAV 3-D placement. The nonconvex problems are tackled using both model-based and model-free deep reinforcement learning (DRL) algorithms under constraints, such as minimum Quality of Service (QoS), beamforming and phase shift limits, and UAV trajectory constraints. The simulation results demonstrate that the proposed DRL algorithms significantly enhance spectral efficiency (SE) and EE, showcasing their suitability for 6G communication systems. Furthermore, a comparative analysis with orthogonal multiple access (OMA) and spatial-division multiple access (SDMA) confirms that NOMA outperforms both techniques, achieving substantial gains in efficiency and performance.**

***Index Terms*—mmWave, model-based deep reinforcement learning (DRL), model-free DRL, nonorthogonal multiple access (NOMA), reflecting intelligent surface (RIS), unmanned aerial vehicle (UAV).**

Received 15 June 2024; revised 30 December 2024, 22 January 2025, and 10 March 2025; accepted 13 March 2025. Date of publication 20 March 2025; date of current version 9 July 2025. The work of Sima Sobhi-Givi, Mahdi Nouri, and Mahrokh G. Shayesteh was supported by the Mobile Telecommunication Company of Iran, Research and Development (MCI-RD), Tehran, Iran, under Contract DR-51-0011-0030. The work of Hyun Han Kwon and Md. Jalil Piran was supported by the Korea Environment Industry and Technology Institute (KEITI) through Research and Development on the Technology for Securing the Water Resources Stability in Response to Future Change Project, funded by the Korea Ministry of Environment (MOE) under Grant RS-2024-00332300. *(Corresponding authors: Mahrokh G. Shayesteh; Md. Jalil Piran.)*

Sima Sobhi-Givi was with the Electrical and Computer Engineering Department, Urmia University, Urmia, Iran. She is now with the Electrical and Computer Engineering Department, University of Mohaghegh Ardabili, Ardabil 5619913131, Iran (e-mail: s.sobhi@uma.ac.ir).

Mahdi Nouri and Hamid Behroozi are with the Electrical Engineering Department, Sharif University of Technology, Tehran 3155267188, Iran (e-mail: mahdi.nouri@ee.sharif.edu; behroozi@sharif.edu).

Mahrokh G. Shayesteh is with the Electrical and Computer Engineering Department, Urmia University, Urmia 5756151818, Iran, and also with the Wireless Research Lab, ACRI, Electrical Engineering Department, Sharif University of Technology, Tehran 3155267188, Iran (e-mail: m.shayesteh@urmia.ac.ir).

Hyun Han Kwon is with the Department of Civil and Environmental Engineering, Sejong University, Seoul 05006, South Korea.

Md. Jalil Piran is with the Department of Computer Science and Engineering, Sejong University, Seoul 05006, South Korea (e-mail: piran@sejong.ac.kr).

Digital Object Identifier 10.1109/JIOT.2025.3553176

I. Introduction

A. Motivation

With the rapid development of unmanned aerial vehicles (UAVs) in the fifth-generation (5G) of mobile networks, they are playing a significant role in improving spectral efficiency (SE) [1]. UAVs have many advantages, such as high mobility, low cost, and Line-of-Sight (LoS) transmission, which can be utilized to improve throughput, average secrecy rate, and energy efficiency (EE) [2], [3].

Reflecting intelligent surface (RIS) is a technology that has gained interest due to its potential to steer signals to desired paths with low-cost surfaces. RIS uses a determined number of square patch antennas with digitally controllable amplitude and phase modules to achieve passive beamforming [4]. However, to match the performance of a relay, RIS requires a higher number of elements. Moreover, RIS may result in a lower signal-to-noise ratio (SNR) compared to massive multiple-input–multiple-output (MIMO) systems [5].

Nonorthogonal multiple access (NOMA) is a promising technology for enhancing the SE of wireless networks. In NOMA, multiple users share the same spectrum resources by assigning different power levels to different users. This power-domain multiplexing allows NOMA to support a greater number of users compared to traditional orthogonal multiple access (OMA) schemes, which allocate distinct resources to each user, leading to a more efficient use of the available bandwidth [6].

In NOMA systems, the success of the transmission relies heavily on successive interference cancellation (SIC), a

technique used to decode the signals of multiple users simultaneously by first canceling the strongest signal and then progressively canceling weaker signals. SIC allows users with different power levels to coexist in the same frequency band, but its performance is sensitive to the accuracy of the interference cancellation process. Perfect SIC refers to the ideal scenario where interference cancellation is flawless, and all signals are correctly decoded without residual interference. However, in practical scenarios, imperfect SIC is inevitable. Imperfect SIC occurs when the cancellation process is not perfect, leading to residual interference from incorrectly decoded signals. This imperfection can degrade the quality of the received signal, reducing the system's overall performance [7].

We model this imperfect SIC by introducing an SIC error term, which accounts for inaccuracies in the interference cancellation process. This error term becomes especially significant in highly dynamic environments where channel conditions fluctuate rapidly, such as in mobile networks or networks with varying interference. By incorporating SIC errors, we provide a more realistic representation of SIC performance in real-world systems, where hardware limitations, channel estimation errors, or environmental dynamics contribute to imperfect interference cancellation. This modeling approach highlights the challenges faced in practical NOMA deployments and emphasizes the need for robust algorithms that can mitigate the impact of imperfect SIC on system performance.

Also, NOMA is particularly well suited for UAV-assisted networks, as UAVs can be used to provide LoS links to users, which improves the performance of NOMA [8].

The integration of RIS, UAVs, and NOMA is motivated by the synergistic advantages these technologies offer. RIS enhances signal coverage and adaptability by strategically adjusting reflection coefficients, while UAVs provide dynamic network adaptability, extended coverage, and energy-efficient deployment. NOMA, through efficient resource sharing, boosts system capacity. The combined utilization of RIS, UAVs, and NOMA aims to create a versatile and resilient wireless communication system, addressing challenges related to coverage, adaptability, capacity, EE, and security, while harnessing the unique strengths of each technology for future-proofed and high-performance networks.

Deep reinforcement learning (DRL) tools are gaining traction in beamforming optimization due to several key advantages over traditional optimization algorithms, particularly in handling nonconvex beamforming problems [9], [10], [11], [12], [13].

1) *Handling Nonconvex Optimization:* Traditional methods struggle with nonconvex UAV-RIS beamforming problems, often getting trapped in suboptimal solutions. DRL algorithms excel by exploring solution spaces effectively, leading to better outcomes.
2) *Adaptability to Dynamic and Stochastic Channels:* DRL outperforms traditional optimization in dynamic environments, where wireless conditions change frequently. DRL's adaptability allows for real-time optimization, unlike stationary-based traditional methods.
3) *Dealing With High-Dimensionality:* DRL is adept at navigating high-dimensional decision spaces, common in mmWave channels. Its function approximation capabilities enable efficient exploration without being hindered by local optima.
4) *Learning From Data-Driven Experience:* Unlike traditional methods relying on explicit mathematical models, DRL learns directly from data, adapting to real-world scenarios without perfect models.
5) *Handling Uncertainty and Partial Information:* DRL effectively incorporates uncertain UAV-RIS channel conditions and limited feedback into decision-making, leading to robust strategies even with imperfect information.
6) *Joint Optimization and Exploration:* DRL excel in balancing conflicting objectives like throughput maximization, interference minimization, and latency reduction. Its ability to explore multiple strategies aids in finding appropriate beamforming solutions.

The key motivation behind DRL over alternative learning tools lies in its adaptability to dynamic environments, ability to handle uncertainty, and efficiency in exploring high-dimensional solution spaces.

B. Related Work

In this section, we categorize the related works into two distinct categories based on the proposed system models: 1) UAV-mounted BS and 2) UAV-mounted RIS. By organizing the literature in this way, we aim to highlight the unique contributions, challenges, and gaps in both categories, providing a structured foundation for our proposed unified framework.

1) UAV-Mounted BS: The authors in [14], [15], [16], [17], [18], [19], [20] considered UAV-mounted base station (BS) in their system model.

In [14], an efficient framework was suggested to optimize the active and passive beamforming at UAV and RIS as well as the UAV trajectory by maximizing the received signal power. The IoT wireless networks with the assistance of a UAV, and one RIS were studied in [15]. The authors investigated maximizing the total network sum-rate for optimizing the UAV trajectory, the energy harvesting scheduling of IoT devices, and the RIS phase shift matrix. However, user clustering in NOMA groups and EE optimization were not considered in [14] and [15]. In [16], the EE maximization in the RIS-assisted UAV-enabled mobile-edge computing (MEC) systems was considered, where an iterative algorithm was utilized for jointly optimizing the bit and power allocation, RIS phase shift design, and UAV trajectory. However, the NOMA technique was not included in this research.

In [17], an RIS assisted multi-UAV system with NOMA communication was investigated. The authors maximized the sum-rate of the network by optimizing the passive beamforming at RIS, and trajectory at UAV in the NOMA clusters. In [18], multiple RIS were utilized to establish additional and intelligent transmission links between NOMA users and UAV-mounted BS. The authors formulated a throughput maximization problem for optimizing the UAV

trajectory, passive beamforming of the RIS, and the power and time allocation. The optimization problem was solved using the Lagrange-based reward-constrained proximal policy optimization (LRCPPO) algorithm. However, the effects of imperfect SIC and EE maximization were not addressed in [17] and [18]. In another research [19], a combination of UAV-RIS was proposed for NOMA communication systems which focused on minimizing the power consumption by optimizing the phase shift of RIS, UAV trajectory, and power allocation for transmitting data between the UAV and the users. The work in [20] addressed the problem of maximizing EE in a UAV-RIS system with NOMA transmission. The authors utilized deep deterministic policy gradient (DDPG) to solve the optimization problem. However, they did not investigate the use of model-based DRL methods and did not consider the trajectory of the UAV and imperfect SIC. As highlighted in the mentioned studies, the system model was based on idealized channel conditions and assumptions.

2) UAV-Mounted RIS: In [21], [22], [23], [24], [25], [26], [27], [28], and [29], a UAV-mounted RIS model with the BS and users on the ground was developed.

The aim of [21] was to maximize the secrecy EE by jointly optimizing the UAV trajectory, phase shift at RIS, user association, and transmit power. However, NOMA transmission and active beamforming at the BS were not investigated. In [22], a system was assisted to aid the cell-edge users and enhance the quality of data transmission. The authors focused on optimizing beamforming vectors at the BS and phase shifters at the RIS by maximizing the EE metric. However, the NOMA transmission was not addressed. Moreover, [23] introduced an innovative system design incorporating ambient backscatter communication into UAV-mounted RIS networks, showcasing the potential of this integration in enhancing wireless connectivity. Building on this, [24] developed a detailed path loss model specifically for scenarios involving UAV-mounted absorbing metasurfaces, with experimental validation conducted in a controlled anechoic chamber, ensuring the reliability of their findings. Shifting focus to trajectory design, [25] concentrated on optimizing 3-D UAV trajectories in urban settings to enhance the SNR for ground users, addressing challenges posed by complex urban environments. Additionally, [26] explored the joint optimization of UAV trajectories and RIS configurations, aiming to facilitate efficient computational task offloading in IoT networks. These studies underline the importance of considering both communication strategies and trajectory planning to fully harness the capabilities of UAV-mounted RIS systems in diverse wireless network applications. Aung et al. [27] investigated an energy-efficient downlink communication system utilizing multiple UAV-mounted RISs. They formulated a problem to maximize EE by jointly considering RIS deployment, reflecting element on/off states, phase shifts, and power control. The problem was decomposed into three subproblems: 1) RIS deployment; 2) joint reflecting element state and phase shift optimization; and 3) power control. To address these, the authors employed the successive convex approximation (SCA) approach, actor–critic proximal policy optimization (AC-PPO), and the whale optimization algorithm (WOA) in an alternating manner. However, the study did not incorporate NOMA transmission, and the optimization problem was not solved in a fully joint manner.

In [28], a NOMA system was considered with the objective of maximizing the data rate for the proximate user while ensuring a satisfactory target rate for the distant user to assure the Quality of Service (QoS). The optimization of horizontal position of the UAV, BS beamforming vectors, and the RIS phase shifter was studied to achieve higher data-rates for users, but EE optimization and imperfect SIC were not considered. A similar system to that in [22], with a comparable objective, was examined in [29], incorporating NOMA for uplink data transmission. However, neither power control nor the effects of imperfect SIC were addressed in [22] or [29].

C. Contributions

Existing research on UAV-RIS systems has explored various optimization approaches, yet several critical challenges remain unaddressed. One major hurdle is the seamless integration of UAVs, RIS, and NOMA into a unified framework, given their diverse characteristics and functionalities. Additionally, optimizing EE requires comprehensive solutions that account for power consumption, active and passive beamforming, and environmental sustainability. Another key challenge is imperfect SIC in NOMA transmission, which affects interference cancellation but is often overlooked in prior studies. To address these gaps, this article investigates two deployment scenarios.

1) *UAV-Mounted BS With RIS Assistance:* A system where a UAV carries a BS to enable communication in areas where fixed infrastructure is impractical.
2) *UAV-Mounted RIS for Enhanced Communication:* A setup where RIS-equipped UAVs improve the quality and efficiency of wireless transmission.

Both scenarios focus on maximizing EE through the joint optimization of the UAV or BS beamforming matrix, RIS phase shift matrix, power gain in NOMA transmission, and UAV 3-D placement. To achieve this, we employ both model-free and model-based DRL algorithms, marking the first use of model-based DRL for EE maximization in RIS-assisted UAV networks. The key contributions of this work are as follows.

1) *Comprehensive UAV-RIS-NOMA Optimization Framework:* We propose a unified framework integrating UAV placement, RIS configuration, and NOMA. Two complementary scenarios are introduced.
   a) *RIS-Enhanced Multi-UAV-Mounted BS for NOMA Networks:* This scenario examines how RIS can enhance communication efficiency in UAV-mounted BS networks.
   b) *Multi-UAV-Mounted RIS for NOMA Transmission:* This scenario explores how multiple UAVs can optimize RIS configurations to improve NOMA performance.
2) *UAV-Mounted BS and UAV-Mounted RIS Deployments:* The study investigates both UAV-mounted BS and UAV-mounted RIS architectures, demonstrating their effectiveness in enhancing EE while maintaining acceptable SE.
3) *Joint Optimization of Key Parameters:* To tackle non-convex optimization challenges, we jointly optimize the UAV or BS beamforming matrix, RIS phase shift matrix,

power allocation for NOMA transmission, and UAV 3-D placement, ensuring efficient resource utilization.

4) *Novel Application of Model-Based DRL:* We introduce both model-free and model-based DRL to solve the optimization problem, with model-based DRL being used for the first time in EE maximization for UAV-assisted RIS systems.
5) *Realistic Consideration of Imperfect SIC in NOMA:* Unlike previous studies, we account for imperfect SIC, where interference cancellation is not ideal due to hardware limitations and signal processing inaccuracies. Our analysis compares scenarios with both perfect and imperfect SIC, offering a practical evaluation of system performance.
6) *Comparative Performance Analysis of Multiple Access Techniques:* Simulation results provide insights into NOMA, SDMA, and OMA transmission under various conditions. For NOMA, perfect SIC achieves superior SE and EE compared to imperfect SIC. For OMA, performance is evaluated under both zero and nonzero interference. For SDMA, we assess the impact of both perfect and imperfect detection. Across all cases, NOMA consistently outperforms SDMA and OMA in terms of SE and EE, reinforcing its potential in UAV-RIS networks.

Table I summarizes the characteristics of the related works and the proposed network.

The remainder of this article is organized as follows. Section II provides a detailed description of the proposed RIS-assisted UAV network. Section III formulates the optimization problem, focusing on the design of active and passive beamforming, power allocation, and UAV 3-D placement. Section IV introduces the proposed DRL-based solutions to address the formulated problems. Section V presents and analyzes the numerical results, while Sections VI and VII conclude this article and discuss future research directions.

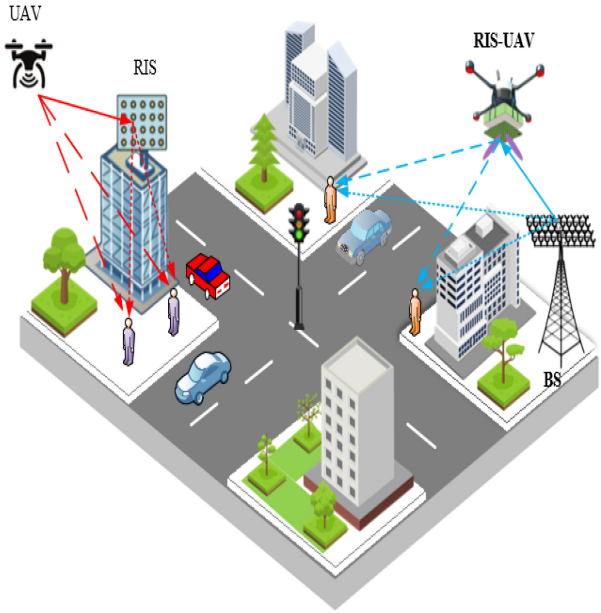

Fig. 1. Proposed system model.

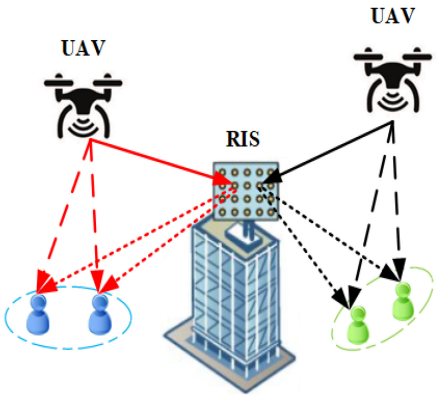

Fig. 2. Proposed multi-UAV-mounted BS NOMA network.

II. System Description

We consider a general wireless communication network consisting of UAVs, RIS, UAV-mounted RIS, and BS, as illustrated in Fig. 1. The network design aims to address various scenarios where traditional communication links face challenges.

In scenarios where the LoS between the users and the BS is obstructed by obstacles, a UAV-mounted RIS is deployed to establish a reflected communication path, thereby enhancing signal propagation and coverage. The UAV-mounted RIS actively adjusts its reflection coefficients to optimize the communication channel quality dynamically. In environments lacking a terrestrial BS, a UAV-mounted BS, in conjunction with an RIS, is employed to provide seamless connectivity to the users. The UAV-mounted BS acts as the primary signal transmitter, while the RIS assists in overcoming path loss and mitigating shadowing effects by intelligently reflecting and focusing the signal toward the users.

To simplify the analysis and facilitate a deeper understanding of the proposed network model, we categorize the system into two distinct operational scenarios.

1) *UAV-Mounted BS Scenario:* This scenario focuses on the deployment of a UAV-mounted BS for direct communication with users, complemented by the RIS to enhance coverage and signal quality.
2) *UAV-Mounted RIS Scenario:* This scenario emphasizes the role of a UAV-mounted RIS in assisting communication between users and a terrestrial BS when direct LoS is unavailable.

The technical modeling and optimization of each scenario are thoroughly analyzed in the subsequent sections, addressing the challenges and operational considerations inherent to these configurations.

A. Multi-UAV-Mounted BS

Fig. 2 illustrates the proposed narrow-band RIS-enhanced with multi-UAV-mounted BS for NOMA networks. In our proposed scenario, K rotary-wing UAVs are deployed to serve

TABLE I
SUMMARY OF THE PROPOSED METHOD AND RELATED WORKS

	Objective	UAV-mounted BS	UAV-mounted RIS	NOMA	Imperfect SIC	Energy Efficiency	Solution
Proposed	Designing active and passive beamforming matrix, power allocation, UAV 3D placement, Maximizing EE	✓	✓	✓	✓	✓	DQN, DDQN, DDPG, TD3, MBPO
[14]	Designing active and passive beamforming matrix, UAV 3D placement, Maximizing received signal power	✓	×	×	×	×	Iterative algorithm
[15]	Passive beamforming, UAV 3D placement, Maximizing sum-rate	✓	×	×	×	×	DDPG and PPO
[16]	Power allocation, passive beamforming, UAV 3D placement, Maximizing EE	✓	×	×	×	✓	Iterative algorithm
[17]	Designing passive beamforming, UAV 3D placement, Maximizing sum-rate	✓	×	✓	×	×	BCD-based iterative algorithm
[18]	Passive beamforming, power allocation, UAV 3D placement, Maximizing throughput	✓	×	✓	×	×	Lagrange multipliers and PPO
[19]	Designing passive beamforming, UAV 3D placement, power allocation, Minimizing power consumption	✓	×	✓	×	✓	DDQN
[20]	Passive beamforming, power allocation, Maximizing EE	✓	×	✓	×	✓	DDPG
[21]	Power allocation, passive beamforming, UAV 3D placement, Maximize EE	×	✓	×	×	✓	Iterative algorithm
[22]	Active and passive beamforming, Maximizing EE	×	✓	×	×	✓	Iterative algorithm
[28]	Active and passive beamforming, UAV 3D placement, Maximize data-rate	×	✓	✓	×	×	Iterative algorithm
[29]	Active and passive beamforming, Maximizing EE	×	✓	✓	×	✓	semi-definite relaxation technique

K NOMA groups, assisted by an RIS equipped with N reflecting elements. In this research, UAVs and ground user equipments (g-UEs) are considered with N_t and one antennas, respectively. It is assumed that UAVs and RIS serve g-UEs by the active and passive beamformings, respectively. The locations of the g-UEs and the RIS are fixed and represented with $\mathbf{u}_i^k = [x_i^k, y_i^k, z_i^k]^T$ for the ith g-UE of the kth UAV, and $\mathbf{r} = [x_r, y_r, z_r]^T$ for the RIS. Also, the locations of K UAVs are denoted by $\mathbf{Q} = [\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_K]^T$ where $\mathbf{q}_k = [x_k, y_k, z_k]$.

1) Channel Model: In UAV-assisted communication systems, understanding the nature of the wireless channel is crucial because it directly impacts system design and performance. To model the channel accurately, we adopt a Rician fading model, which captures the hybrid propagation environment characterized by both LoS and Non-Line-of-Sight (NLoS) components. This is particularly relevant in UAV communication systems, where the UAVs' elevated position often ensures a dominant LoS path, while multipath scattering introduces NLoS components. The channels between the kth UAV and RIS with the ith g-UE, $\mathbf{u}_i^k$, are, respectively, denoted by $\mathbf{h}_{UG}^{k,i} \in \mathbb{C}^{1\times N_t}$ and $\mathbf{h}_{RG}^{k,i} \in \mathbb{C}^{N\times 1}$, which can be expressed as [30]

$$\mathbf{h}_{UG}^{k,i} = \sqrt{\frac{\rho_0}{\left\|\mathbf{q}_j - \mathbf{u}_i^k\right\|^{\alpha_{UG}}}} \left(\sqrt{\frac{\beta_{UG}}{1+\beta_{UG}}}\bar{\mathbf{h}}_{UG}^{k,i} + \sqrt{\frac{1}{1+\beta_{UG}}}\hat{\mathbf{h}}_{UG}^{k,i}\right) \tag{1}$$

$$\mathbf{h}_{RG}^{k,i} = \sqrt{\frac{\rho_0}{\left\|\mathbf{r} - \mathbf{u}_i^k\right\|^{\alpha_{RG}}}} \left(\sqrt{\frac{\beta_{RG}}{1+\beta_{RG}}}\bar{\mathbf{h}}_{RG}^{k,i} + \sqrt{\frac{1}{1+\beta_{RG}}}\hat{\mathbf{h}}_{RG}^{k,i}\right) \tag{2}$$

where ρ_0 is the path loss at the reference distance of one meter, α_{UG} and α_{RG} represent the path loss exponents of the UAV to g-UE (U-G) and RIS to g-UE (R-G) links, respectively, β_{UG} and β_{RG} denote the Rician factors, $\bar{\mathbf{h}}_{UG}^{k,i}$, $\bar{\mathbf{h}}_{RG}^{k,i}$ show the deterministic LoS components, and $\hat{\mathbf{h}}_{UG}^{k,i}$, $\hat{\mathbf{h}}_{RG}^{k,i}$ are the stochastic nature of multipath scattering which are random Rayleigh distributed NLoS components. Therefore, the deterministic LoS term reflects the direct, unscattered path, while the random NLoS term captures the impact of multipath

propagation caused by reflections and diffraction in the environment. Both (1) and (2) include geometric dependence on the distances between the respective nodes, which accounts for the attenuation of signal power with increasing distance.

For the UAV-to-RIS channel $\mathbf{h}_{UR}^{k} \in \mathbb{C}^{N\times N_t}$ modeled in (3). According to [31], the air-to-ground communication channels are mainly dominated by the LoS links, which simplifies the representation due to the high likelihood of an unobstructed path between UAVs and RISs. Furthermore, it is assumed that the Doppler effect induced by the mobility of the UAV is perfectly compensated at the receivers [32]. The deterministic LoS component, $\bar{\mathbf{h}}_{UR}^{k}$ dominates this link, making it well suited for RIS-enabled systems where predictable and stable channels are desirable for efficient phase adjustment and beamforming. Therefore, the channel of the link from the kth UAV to RIS (U-R link), $\mathbf{h}_{UR}^{k} \in \mathbb{C}^{N\times N_t}$, can be displayed as [30]

$$\mathbf{h}_{UR}^{k} = \sqrt{\frac{\rho_0}{\|\mathbf{q}_k - \mathbf{r}\|^2}}\bar{\mathbf{h}}_{UR}^{k}. \tag{3}$$

The effective channel power gain combines the mentioned individual channels and incorporates the effect of the RIS. The RIS is modeled through the diagonal matrix Φ, which enables dynamic phase shifting to constructively combine signals from the direct and reflected paths, enhancing the received signal power. The frequency-flat RIS reflection matrix is shown by

$$\Phi = \mathrm{diag}\left(e^{j\phi_1}, e^{j\phi_2}, \ldots, e^{j\phi_N}\right) \tag{4}$$

where $\Phi \in \mathbb{C}^{N\times N}$ and $\phi_n \in [0, 2\pi)$ $\forall n$ denotes the corresponding phase shift of the nth reflecting element of the RIS [33].

This equation also includes the beamforming vector $\mathbf{w}_1^k$, which optimizes the UAV's transmission toward the intended gUE. The mathematical formulation thus encapsulates the interplay between channel characteristics, RIS functionality, and UAV beamforming, offering a complete picture of the communication link. Therefore, the effective channel power gain between the kth UAV and $\mathbf{u}_i^k$ with the aid of the RIS, $\mathbf{c}_{k,i} \in \mathbb{C}^{1\times N_t}$, is given by

$$\mathbf{c}_{k,i} = \left|(\mathbf{h}_{UG}^{k,i} + (\mathbf{h}_{RG}^{k,i})^H \Phi \mathbf{h}_{UR}^{k})\mathbf{w}_1^k\right|^2 \tag{5}$$

where $(.)^H$ is the Hermitian and $\mathbf{w}_1^k \in \mathbb{C}^{N_t\times 1}$ is the beamforming vector of the UAV.

2) Signal-to-Interference-Plus-Noise Ratio: Here, we assume that each UAV employs NOMA to provide communication services for g-UEs. Without loss of generality, we define the user ordering based on channel gains as follows:

$$\mathbf{c}_{k,N_u} \geq \cdots \geq \mathbf{c}_{k,i} \geq \mathbf{c}_{k,j} \geq \cdots \geq \mathbf{c}_{k,1} \tag{6}$$

where N_u is the number of g-UE in each cluster. In the NOMA scheme, SIC is employed at the receiver, enabling strong users (i.e., users with higher channel gains) to cancel or mitigate the interference caused by weaker users (i.e., users with lower channel gains). To ensure a more practical scenario, we consider the impact of imperfect SIC. Hence, the signal-to-interference-plus-noise ratio (SINR) of strong user $\mathbf{u}_i^k$ (denoted as $\mathrm{SINR}_{k,i}$), will be equal to

$$\mathrm{SINR}_{k,i} = \frac{\mathbf{c}_{k,i}\zeta_{k,i}P}{\sigma^2 + \xi\mathbf{c}_{k,i}P\sum_{n=1}^{i-1}\zeta_{k,n}} \tag{7}$$

where $\zeta_{k,i}$ is the transmission power gain for $\mathbf{u}_i^k$, ensuring that $\sum_{i=1}^{N_u}\zeta_{k,i} \leq 1$ and maintaining the ordering $\zeta_{k,N_u} \leq \cdots \leq \zeta_{k,i} \leq \zeta_{k,j} \leq \cdots \leq \zeta_{k,1}$. Additionally, P denotes the transmission power of the UAV and σ^2 is the noise power. The parameter ξ represents the imperfect SIC coefficient, where $\xi = 0$ corresponds to perfect SIC, $\xi = 1$ represents no SIC, and $\xi \in [0, 1)$ accounts for imperfect SIC scenarios.

Similarly, the SINR of weak user $\mathbf{u}_j^k$ (denoted as $\mathrm{SINR}_{k,j}$), can be written as

$$\mathrm{SINR}_{k,j} = \frac{\mathbf{c}_{k,j}\zeta_{k,j}P}{\sigma^2 + \mathbf{c}_{k,j}P\sum_{n=j+1}^{N_u}\zeta_{k,n}}. \tag{8}$$

This formulation captures the effect of imperfect SIC on the performance of NOMA, ensuring a more realistic representation of real-world scenarios where interference cancellation is not always ideal due to hardware limitations, channel estimation errors, and environmental dynamics.

3) Spectral Efficiency and Energy Efficiency : SE is defined as the information rate that can be transmitted over a given bandwidth. Also, EE is characterized by the number of bits that can be transmitted over a unit of power consumption. The SE of $\mathbf{u}_i^k$ ($\eta_{k,i}^{SE}$) is obtained as

$$\eta_{k,i}^{SE} = \log_2(1 + \mathrm{SINR}_{k,i}), \quad (\mathrm{bits/sec/Hz}). \tag{9}$$

In UAV-based systems, power consumption is a critical consideration due to the limited energy resources of UAVs and the need to optimize operational efficiency. The total power consumption of a UAV (P_t^k) includes contributions from the UAV's transmit power (P), the RIS power consumption (P_{RIS}), and the UAV's propulsion energy (P_{UAV}) [22]

$$P_t^k = P + P_{\mathrm{RIS}} + P_{\mathrm{UAV}} \tag{10}$$

where

$$P_{\mathrm{RIS}} = NP_n, n = 1, \ldots, N \tag{11}$$

$$P_{\mathrm{UAV}} = a_1V^3 + a_2/V. \tag{12}$$

The RIS power consumption (P_{RIS}) is detailed in (11), where it is modeled as the product of the number of RIS elements N, the per-element circuit power P_n [34]. This representation reflects the energy required to dynamically adjust the phase of incident signals, with P_n depending on the RIS hardware and resolution. The scalability of RIS power with the number of elements highlights the tradeoff between improved performance (via larger RIS arrays) and increased energy expenditure. The UAV's propulsion power (P_{UAV}) expressed in (12), is modeled as a function of its constant velocity V. The term a_1V^3 represents the power required to overcome aerodynamic drag, which increases with the cube of the velocity, while the term a_2/V accounts for the energy consumed to sustain lift, which decreases with higher velocity. The constants a_1 and a_2 depend on UAV-specific parameters,

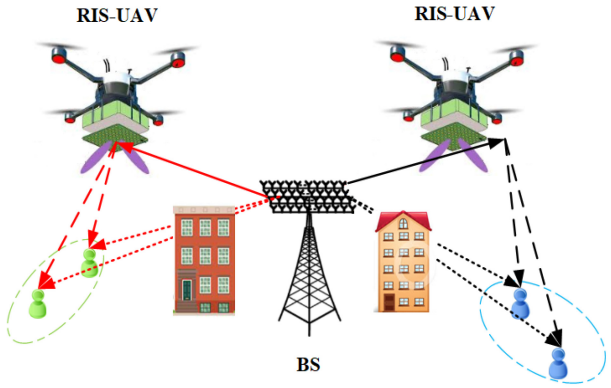

Fig. 3. Proposed multi-UAV-mounted RIS NOMA network.

such as weight, wing area, and air density [35]. This model captures the complex relationship between UAV velocity and power efficiency, offering insights into optimal speed selection for energy savings.

Together, these equations provide a comprehensive framework for analyzing power consumption in UAV-assisted systems, linking physical phenomena (e.g., drag and lift) with practical design considerations. By combining the channel and power models, the analysis becomes robust, enabling performance optimization under realistic energy constraints.

Finally, the EE of the network (η^{EE}) can be calculated as follows [36]:

$$\eta^{\text{EE}} = \frac{\sum_{k=1}^{K}\sum_{i=1}^{N_u}\eta_{k,i}^{SE}}{\sum_{k=1}^{K}P_t^k}, \quad \text{(bits/sec/Hz/J)} \tag{13}$$

where P_t^k represents the total power consumption of the kth UAV.

B. Multi-UAV-Mounted RIS

As shown in Fig. 3, a downlink multi-UAV-mounted RIS NOMA is considered with K rotary-wing UAV-RIS and one BS. Owing to the obstruction posed by tall buildings, an NLoS channel is present between the BS and each g-UE. In this particular scenario, each UAV, equipped with an RIS, functions as a passive relay to enhance communication between the g-UEs and the BS. The BS is outfitted with M antennas, and its location remains fixed and represented with $\mathbf{b} = [x_b, y_b, z_b]$.

1) Channel Model: In a similar procedure, the Rician fading channel model is assumed for all communication links. Hence, the channels between the kth UAV-RIS and $\mathbf{u}_i^k$, between the BS and $\mathbf{u}_i^k$, and between the kth UAV-RIS and BS are, respectively, denoted by $\mathbf{g}_{UG}^{k,i} \in \mathbb{C}^{N\times 1}$, $\mathbf{g}_{BG}^{k,i} \in \mathbb{C}^{1\times M}$, and $\mathbf{g}_{UB}^{k} \in \mathbb{C}^{N\times M}$ which can be expressed similar to (1)–(3). Hence, the effective channel power gain of $\mathbf{u}_i^k$, $\mathbf{d}_{k,i}$, is given by

$$\mathbf{d}_{k,i} = \left|\left(\mathbf{g}_{BG}^{k,i} + \left(\mathbf{g}_{UG}^{k,i}\right)^H \Phi \mathbf{g}_{UR}^{k}\right)\mathbf{w}_2^k\right|^2 \tag{14}$$

where $\mathbf{w}_2^k \in \mathbb{C}^{M\times 1}$ is the beamforming vector at the BS [37].

In this scenario, the BS power consumption (P_{BS}) is added to the total power consumption defined in the previous part. The calculation of SINR, SE, and EE for the second scenario is similar to the previous section.

III. Problem Formulation

A. Multi-UAV-Mounted BS

Our design focuses on maximizing the EE of the communication system by jointly optimizing the UAV beamforming matrix, RIS phase shift matrix, power gain of NOMA transmission, and UAV 3-D placement. Hence, the optimization problem is formulated as

$$\max_{\mathbf{Q},\zeta_{k,i},\Phi,\mathbf{w}_1^k} \eta^{\text{EE}} \tag{15}$$

subject to:

$$\begin{aligned}
&\mathbf{S_1}\text{: } \Phi = \text{diag}\left(e^{j\phi_1},\ldots,e^{i\phi_N}\right),\ \phi_n \in [0,2\pi) \quad \forall n\\
&\mathbf{S_2}\text{: } \|\mathbf{w}_1^k\|^2 = 1 \quad \forall k\\
&\mathbf{S_3}\text{: } \text{SINR}_{k,i} \geq \text{SINR}_{th},\ \sum_{i=1}^{N_u}\eta_{k,i}^{SE} \geq \eta_{th}^{SE} \quad \forall k\\
&\mathbf{S_4}\text{: } \sum_{i=1}^{N_u}\zeta_{k,i}P \leq P_{\max},\\
&\mathbf{S_5}\text{: } Z_{\min} \leq z_k \leq Z_{\max},\ \|\mathbf{q}_k - \mathbf{q}_j\| \geq \Delta_{\min} \quad \forall k \neq j.
\end{aligned} \tag{16}$$

Constraint $\mathbf{S_1}$ denotes the phase shift constraint of each RIS subsurface. Constraint $\mathbf{S_2}$ shows the restriction of the power for beamforming matrix at UAV. Constraint $\mathbf{S_3}$ demonstrates that the received SINR of each g-UE and sum-rate of each cluster should be higher than the threshold values (SINR_{th} and η_{th}^{SE}) to satisfy the minimum QoS requirements. According to $\mathbf{S_4}$, the maximum transmission power of UAV for each cluster is $P_{\max}$. To guarantee operational safety and prevent collisions, the flying height of the UAV and the separation distance between any two UAVs should satisfy the constraint $\mathbf{S_5}$, where $[Z_{\min}, Z_{\max}]$ defines the permissible range for UAV flying height, and $\Delta_{\min}$ represents the minimum inter-UAV distance necessary for collision avoidance.

B. Multi-UAV-Mounted RIS

Same as the previous part, our goal is to maximize the EE of the communication system by jointly optimizing the BS beamforming matrix, RIS phase shift matrix, power gain of NOMA transmission, and UAV-RIS 3-D placement. Hence, the optimization problem is reformulated as

$$\max_{\mathbf{Q},\zeta_{k,i},\Phi_k,\mathbf{w}_2^k} \eta^{\text{EE}} \tag{17}$$

subject to:

$$\begin{aligned}
&\mathbf{S_1}\text{: } \Phi = \text{diag}\left(e^{j\phi_{k,1}},\ldots,e^{i\phi_{k,N}}\right),\ \phi_n \in [0,2\pi) \quad \forall n\\
&\mathbf{S_2}\text{: } \|\mathbf{w}_2^k\|^2 = 1 \quad \forall k\\
&\mathbf{S_3} - \mathbf{S_5}.
\end{aligned} \tag{18}$$

Constraint $\mathbf{S_1}$ is related to the phase shift limitations of each UAV-RIS subsurface. Constraint $\mathbf{S_2}$ is the limitation of the power for beamforming matrix at BS. Constraints $\mathbf{S_3}$–$\mathbf{S_5}$ are the same as (16).

IV. Proposed DRL-Based Solutions

In this study, we propose a solution for optimizing the beamforming matrix, RIS phase shift matrix, power allocation, and UAV 3-D placement, as represented in (15)–(18). These objective functions are nonconvex, making them unsolvable using traditional convex optimization methods. Therefore, we turn to DRL algorithms to address these challenges. The selection of DRL models is based on their ability to effectively optimize multi-UAV-mounted BS and multi-UAV-mounted RIS systems in dynamic wireless environments, while balancing computational complexity, convergence speed, and adaptability. We employ DQN, DDQN, DDPG, TD3, and MBPO algorithms. DQN and DDQN are chosen for their simplicity and efficiency in discrete action spaces, though they are less effective in complex, high-dimensional scenarios. DDPG and TD3 are more suited for handling continuous action spaces, making them better for optimizing UAV positions, RIS phase shifts, and power allocation, with TD3 offering enhanced stability. MBPO, a model-based approach, improves learning efficiency by leveraging environmental predictions, making it particularly well suited for dynamic RIS optimization. Comparative analysis indicates that while MBPO achieves the best spectral and EE, it requires more training episodes, demonstrating its adaptability to real-world deployment challenges. The system parameters are defined as follows.

Agent: The UAV and BS are treated as agents.

State: The last received SINR of $\mathbf{u}_i^k$ at time t denotes the state ($\mathbf{s}$), hence we have

$$\mathbf{s} = \left[\text{SINR}_{k,i}\right]_{1\le i\le N_u}. \tag{19}$$

where [.] denotes vector.

Action: The action, $\mathbf{a}$, is the set of available beamforming and RIS phase shift matrix, transmission powers, and UAV 3-D placement. In deep Q-network (DQN) and double DQN (DDQN), the action space should be discrete, so we quantize the mentioned parameters to b bits.

Reward: The DRL method optimizes system performance by using a reward function that guides the agent's decisions while balancing key objectives and system constraints. In this formulation, the reward function is constructed as the sum of the EE objective function (η^{EE}) and penalty terms that enforce critical constraints. These constraints include maintaining the SINR above a required threshold, ensuring SE meets the predefined limit, adhering to the total transmission power constraint, and preserving the minimum UAV separation distance. Mathematically, the reward function is expressed as

$$\mathbf{r} = \eta^{\text{EE}} - \left|\text{SINR}_{k,i} - \text{SINR}_{th}\right| - \left|\sum_{i=1}^{N_u}\eta_{k,i}^{SE} - \eta_{th}^{SE}\right| - \left|\sum_{i=1}^{N_u}\zeta_{k,i}P - P_{\max}\right| - \left|\left\|\mathbf{q}_k - \mathbf{q}_j\right\| - \Delta_{\min}\right|. \tag{20}$$

Here, the absolute differences ensure that deviations from the required constraints are penalized, encouraging the DRL agent to find an optimal policy that satisfies all conditions while maximizing EE. This formulation allows the agent to effectively navigate the solution space, ensuring both performance and feasibility in a dynamic wireless environment.

TABLE II
Mean SE Comparison of DRL Methods (bits/s/Hz)

Proposed Scenarios	Multi-UAV mounted BS		Multi-UAV mounted RIS	
SIC Effect	Perfect	Imperfect	Perfect	Imperfect
DQN	22.6	3.8	24.6	4.4
DDQN	26.4	5.2	27.3	5.2
DDPG	28	6	30.2	7
TD3	38	8	36.9	8.6
MBPO	40	8.8	39.1	9.4

Two kinds of DRL methods defined as model-free and model-based DRL are employed to solve the optimization problems. Model-based learning uses a model of the environment to learn the optimal policy. This can be more efficient than model-free learning, but it requires a good model of the environment. Model-free method learns the optimal policy directly from experience, without using a model of the environment. This is less efficient than model-based learning, but it does not require a model of the environment. The model-free techniques are DQN [38] and DDQN [39] which are utilized to learn optimal policies in discrete action space. As well, DDPG [40] and twin-delayed deep deterministic policy gradient (TD3) [41] have continuous action space. The model-based policy optimization (MBPO) [42] is used for both discrete and continuous action spaces. In DRL algorithms, the Markov decision process (MDP) is used to update the Q-table which is given by

$$Q(\mathbf{s},\mathbf{a}) = (1-\alpha)Q(\mathbf{s},\mathbf{a}) + \alpha\left(r + \gamma\max_a Q(\mathbf{s}',\mathbf{a})\right) \tag{21}$$

where $\alpha \in (0,1]$ denotes the learning rate of the algorithm which reflects the weight of the current experience, r represents the reward function, and γ denotes the discount factor.

As will be shown in the next section, TD3 and MBPO outperform DQN, DDQN, and DDPG. Therefore, we will only explain the TD3 and MBPO algorithms for solving the optimization problem in the Appendix.

V. Performance Evaluation

This section evaluates the performance of the proposed EE maximization methods under two different scenarios: 1) multi-UAV-mounted BS and 2) multi-UAV-mounted RIS. The analysis focuses on two key performance metrics: 1) SE and 2) EE, across the proposed network architectures.

First, we compare the performance of several DRL algorithms, including DQN, DDQN, DDPG, TD3, and MBPO, to highlight their effectiveness in solving the formulated optimization problems. Additionally, we explore various multiple access techniques, such as NOMA, OMA, and SDMA, under both perfect and imperfect conditions, considering the impact of hardware impairments and SIC errors.

The influence of the number of transmit antennas on network performance is also investigated, providing insights into the scalability and adaptability of the proposed solutions

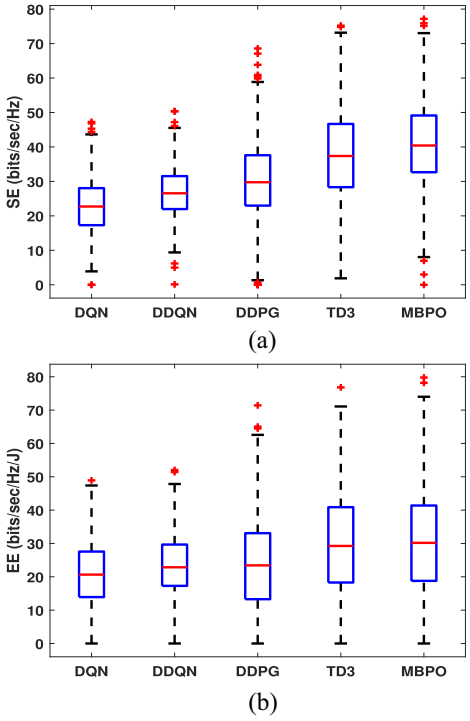

Fig. 4. Performance of the proposed multi-UAV-mounted BS network with perfect SIC effect. (a) SE. (b) EE.

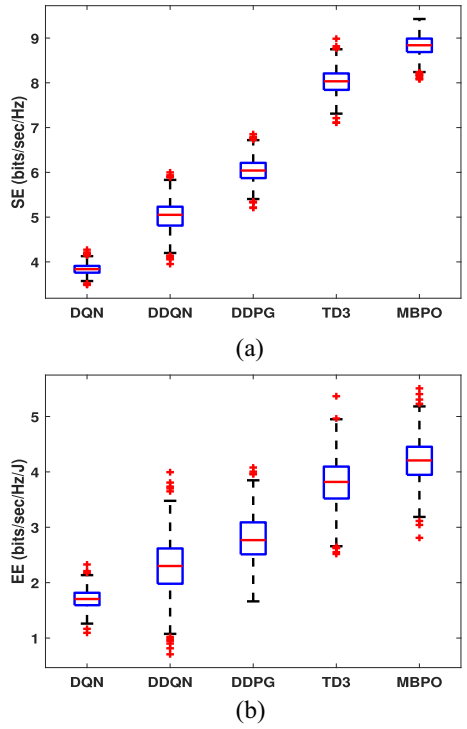

Fig. 5. Performance of the proposed multi-UAV-mounted BS network with imperfect SIC effect. (a) SE. (b) EE.

for practical deployments. Table II summarizes the simulation parameters for clarity and reproducibility. This comprehensive evaluation aims to demonstrate the robustness and applicability of the proposed approaches in addressing the challenges of modern communication systems.

A. Multi-UAV-Mounted BS

1) Comparison of DRL Algorithms: In this scenario, we analyze the performance of the proposed DRL algorithms—DQN, DDQN, DDPG, TD3, and MBPO—under both perfect and imperfect SIC conditions. The mean SE for each algorithm under perfect SIC is as follows: 22.6, 26.4, 28, 38, and 40 bits/s/Hz, respectively. Under imperfect SIC, these values decrease to 3.8, 5.2, 6, 8, and 8.8 bits/s/Hz, respectively, reflecting the impact of hardware impairments and SIC errors on system performance. Figs. 4 and 5 provide a visual comparison of SE and EE across different DRL algorithms under both perfect and imperfect SIC conditions.

The results consistently show that the MBPO algorithm outperforms other methods, achieving the highest SE and EE in both perfect and imperfect SIC scenarios. This superior performance is attributed to MBPO's ability to leverage a model-based DRL approach, which enables better prediction and optimization of outcomes in dynamic environments. This advantage is particularly evident in complex communication systems, where MBPO's model-based framework allows for more accurate and efficient solutions.

2) Comparison of Multiple Access Methods: We compare the performance of NOMA with SDMA and OMA under perfect and imperfect conditions. Under perfect conditions, NOMA achieves the highest mean SE at 28 bits/s/Hz, outperforming SDMA (22 bits/s/Hz) and OMA (13.2 bits/s/Hz). When imperfections are introduced, NOMA still performs better with 6 bits/s/Hz, compared to SDMA's 4.6 bits/s/Hz and OMA's 3.6 bits/s/Hz. The results are shown in Figs. 6 and 7, which highlight the cumulative distribution functions (CDFs) of SE and EE for NOMA, SDMA, and OMA, utilizing the DDPG algorithm for optimization.

These findings emphasize the resilience of NOMA in managing interference and maintaining higher SE and EE, even under imperfect conditions. This makes NOMA an ideal candidate for future communication networks where efficient interference management is critical.

3) Impact of Number of Transmit Antennas: The SE performance of the multi-UAV-mounted BS system, as shown in Fig. 8, improves significantly as the number of UAV transmit antennas increases. The added antennas provide more spatial degrees of freedom, improving beamforming and interference mitigation. This trend underscores the importance of optimizing antenna configurations for achieving higher SE, even under imperfect SIC conditions. The results highlight that increasing the number of UAV transmit antennas is a promising approach for improving SE in UAV-assisted communication systems.

B. Multi-UAV-Mounted RIS

1) Comparison of DRL Algorithms: In this section, we analyze the performance of various DRL algorithms, particularly focusing on the proposed multi-UAV-mounted RIS system under both perfect and imperfect SIC conditions. The results show that MBPO consistently outperforms other DRL methods

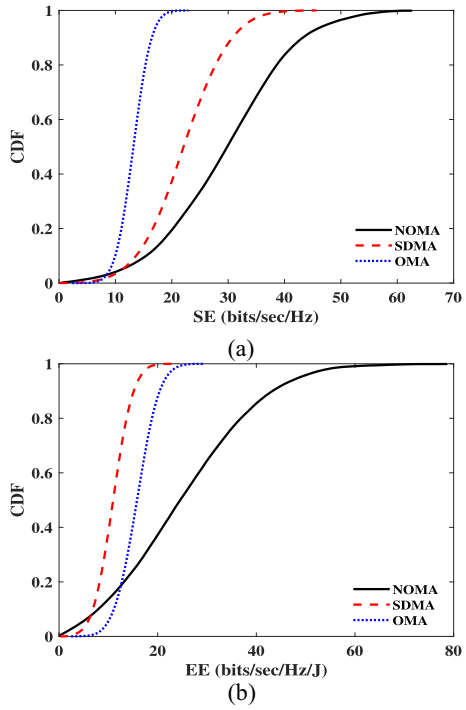

Fig. 6. Performance comparison of the proposed perfect-SIC NOMA multi-UAV-mounted BS network, perfect-detection SDMA, and zero-interference-OMA with DDPG. (a) SE. (b) EE.

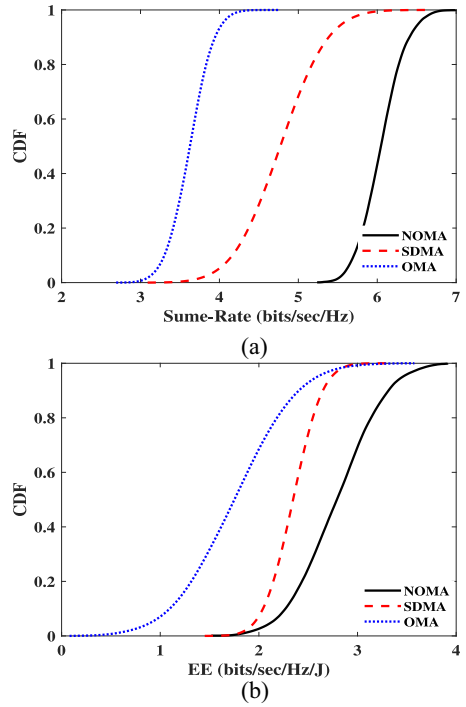

Fig. 7. Performance comparison of the proposed imperfect-SIC NOMA multi-UAV-mounted BS network, imperfect-detection SDMA, and nonzero-interference-OMA with DDPG. (a) SE. (b) EE.

in terms of SE and EE. Under perfect SIC conditions, the mean SE values for DQN, DDQN, DDPG, TD3, and MBPO are 24.6, 27.3, 30.2, 36.9, and 39.1 bits/s/Hz, respectively.

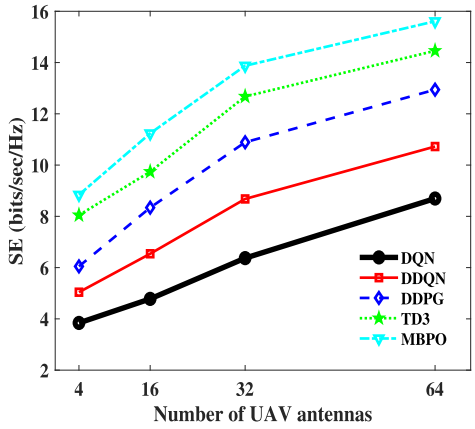

Fig. 8. Performance of the proposed imperfect-SIC NOMA multi-UAV-mounted BS network for different number of transmit antennas.

However, when SIC imperfections are introduced, the corresponding SE values drop, but MBPO maintains a notable advantage with an SE of 9.4 bits/s/Hz, compared to the others: DQN (4.4), DDQN (5.2), DDPG (7), and TD3 (8.6) bits/s/Hz. These results highlight MBPO's ability to leverage its model-based framework to tackle nonconvex optimization challenges, which is critical for adapting to dynamic communication environments.

Figs. 9 and 10 clearly illustrate the SE and EE performance of the algorithms. As seen, MBPO excels not only in achieving the highest SE but also maintains robust performance across varying SIC accuracies. This is attributed to its model-based approach, which enables more efficient decision-making by incorporating learned environmental models. The consistent superiority of MBPO across scenarios reinforces its potential as a promising solution for enhancing multi-UAV-mounted RIS systems in real-world conditions, where imperfections like SIC errors and environmental dynamics are inevitable.

2) Comparison of Multiple Access Methods: This section compares the performance of NOMA, SDMA, and OMA in the context of the multi-UAV-mounted RIS system utilizing DDPG algorithm. NOMA consistently shows superior performance in terms of SE and EE, owing to its ability to share resource blocks and leverage SIC for interference management. Under perfect conditions, the mean SE values for NOMA, SDMA, and OMA are 30.2, 23.3, and 14.1 bits/s/Hz, respectively. However, when imperfections are introduced, such as interference and detection errors, NOMA still outperforms SDMA and OMA with an SE of 7 bits/s/Hz, compared to 5.6 and 4.1 bits/s/Hz, respectively.

Figs. 11 and 12 further emphasize NOMA's effectiveness in sharing resources and managing interference. Despite imperfections, NOMA achieves higher SE and EE than SDMA and OMA, validating its robustness in dynamic network conditions. These results suggest that NOMA is a promising candidate for next-generation wireless networks, particularly when integrated with RIS technology and UAVs to enhance performance under real-world constraints.

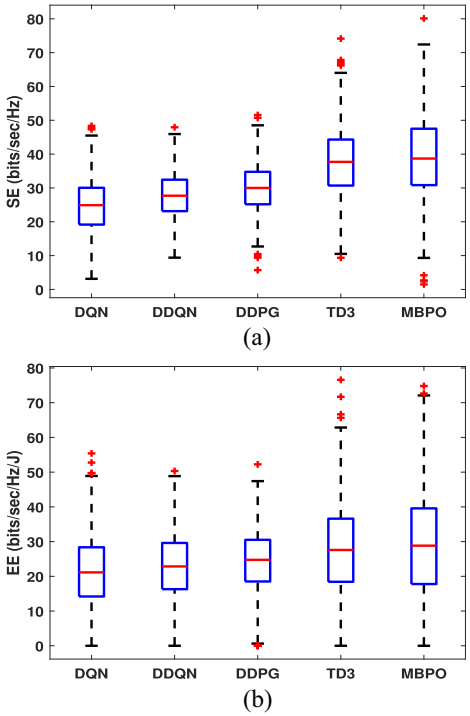

Fig. 9. Performance of the proposed multi-UAV-mounted RIS network with perfect SIC effect. (a) SE. (b) EE.

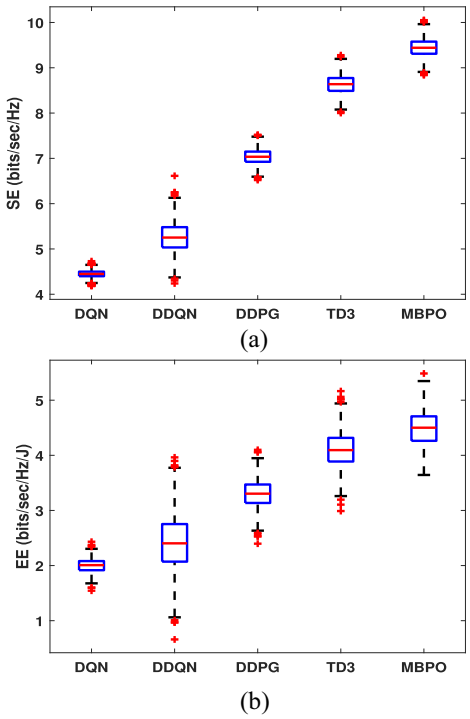

Fig. 10. Performance of the proposed multi-UAV-mounted RIS network with imperfect SIC effect. (a) SE. (b) EE.

3) Impact of Number of Transmit Antennas: Fig. 13 presents the SE performance of the system for varying numbers of BS transmit antennas under imperfect SIC conditions. The results indicate that as the number of transmit antennas increases, the beamforming accuracy improves, leading to higher SE. This trend highlights the importance of scalable antenna configurations for optimizing communication efficiency in multi-UAV-mounted RIS systems. Additionally, the MBPO algorithm remains the top performer, capitalizing on

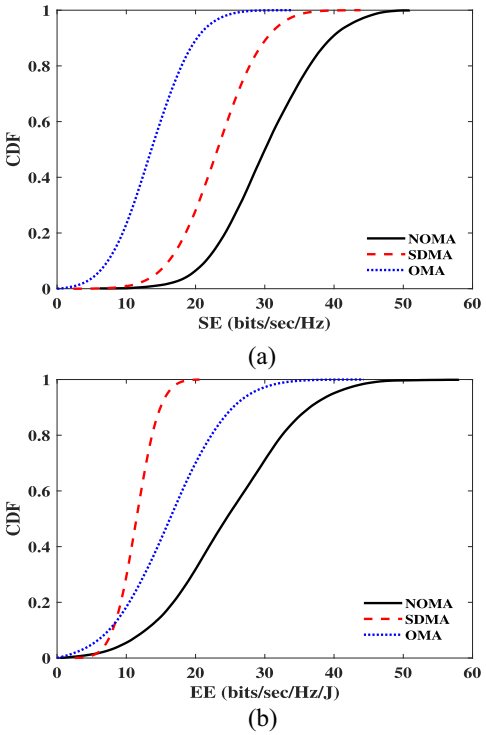

Fig. 11. Performance comparison of the proposed perfect-SIC NOMA multi-UAV-mounted RIS network, perfect-detection SDMA, and zero-interference-OMA. (a) SE. (b) EE.

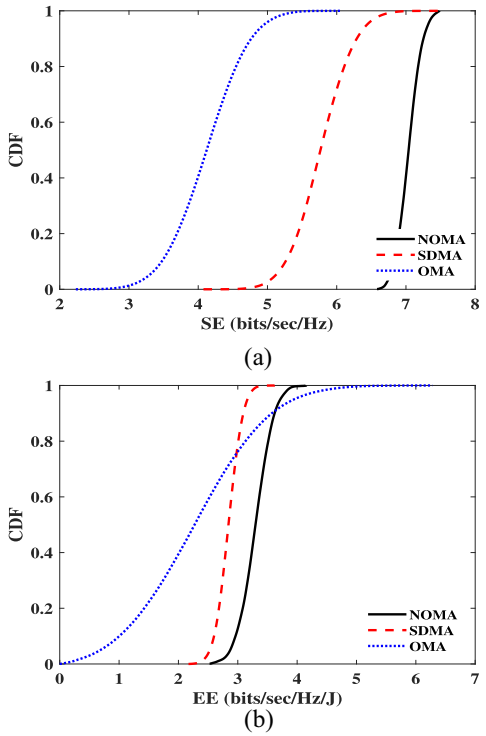

Fig. 12. Performance comparison of the proposed imperfect-SIC NOMA multi-UAV-mounted RIS network, imperfect-detection SDMA, and nonzero-interference-OMA. (a) SE. (b) EE.

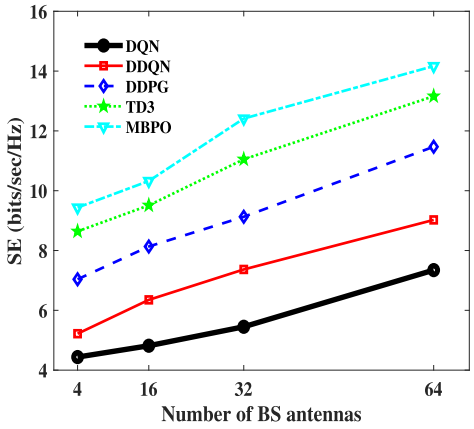

Fig. 13. Performance of the proposed imperfect-SIC NOMA multi-UAV-mounted RIS network considering different number of transmit antennas.

its model-based learning approach to effectively optimize the RIS network.

These findings demonstrate the scalability of the proposed system and underscore the critical role of advanced antenna configurations in enhancing the performance of next-generation wireless networks, particularly in dynamic and interference-prone environments.

C. Performance Comparison

Tables III and IV summarize the mean SE values for different scenarios, facilitating a clear comparison of the DRL algorithms and multiple access methods. Table III compares the performance of DQN, DDQN, DDPG, TD3, and MBPO across multi-UAV-mounted BS and RIS systems. The results show that MBPO outperforms the other methods under both perfect and imperfect SIC conditions, making it the preferred choice for maximizing SE in complex environments.

Table IV compares the SE performance of NOMA, SDMA, and OMA under both perfect and imperfect conditions. NOMA consistently achieves the highest SE, demonstrating its advantage over the other access methods, even when imperfections are introduced. This reinforces the value of NOMA in achieving efficient and robust communication in next-generation wireless systems.

Additionally, Fig. 14 compares the proposed MBPO algorithm with an iterative algorithm from [43] for maximizing EE in multi-UAV-mounted BS and RIS networks under perfect SIC conditions. The results show that MBPO significantly outperforms the iterative method, demonstrating the power of model-based DRL in optimizing complex systems and ensuring stable, reliable performance in real-world scenarios.

D. Computational Complexity Analysis

The computational complexity and convergence properties of the proposed DRL algorithms are crucial in dynamic environments. The complexity depends on factors, such as the state and action space size, training samples, and neural network architecture. Simpler algorithms like DQN and DDQN require lower computational resources and may converge faster in basic environments but struggle with stability in complex scenarios. In contrast, DDPG and TD3, designed for continuous action spaces, offer improved stability at the cost of higher computational demands, potentially slowing convergence. MBPO, with its model-based approach, introduces additional complexity by learning an environmental model, which can either enhance stability through better long-term planning or introduce instability if the model is inaccurate.

Table V presents empirical results on the number of episodes required for convergence. The results indicate that MBPO requires the highest number of episodes, followed by TD3, then DDPG, with DDQN converging faster than these, and DQN requiring the least number of episodes.

TABLE III
SIMULATION PARAMETERS [28], [34]

Parameter	Value (Unit)
Carrier frequency	28 GHz
Max transmission power	10 dBm
Number of UAV transmit antennas	16
Number of BS transmit antennas	16
Number of RIS reflecting elements	16
Noise power	-174 dBm/Hz
Minimum SE	3 bits/sec/Hz
Imperfect SIC Coefficient	0.2
Pathloss at reference distance	-30 dB
Pathloss exponents of U-G and R-G links	2.2
Rician factors of links	10 dB
UAV power consumption	10 dBm
RIS phase shifter power consumption	0.01 dBm
UAV flying height	60-100 m

TABLE IV
MEAN SE COMPARISON OF MA METHODS (BITS/S/HZ)

Proposed Scenarios	Multi-UAV mounted BS		Multi-UAV mounted RIS	
SIC Effect	Perfect	Imperfect	Perfect	Imperfect
NOMA	28	6	30.2	7
SDMA	22	4.6	23.3	5.6
OMA	13.2	3.6	14.1	4.1

TABLE V
COMPLEXITY ANALYSIS

Algorithms / Proposed	DQN	DDQN	DDPG	TD3	MBPO
Multi-UAV-mounted BS	841	1842	2758	3166	5704
Multi-UAV-mounted RIS	937	2075	2649	3129	5643

VI. CONCLUSION

In this article, we examined mmWave-NOMA transmission in two distinct scenarios: 1) multi-UAV-mounted BS and 2) multi-UAV-mounted RIS. Given the practical limitations, such as imperfect SIC decoding in NOMA systems, we formulated optimization problems to maximize the EE of the network. The optimization involved key variables, including the UAV and BS beamforming matrices, RIS phase shift matrix, UAV and BS power allocation, and UAV 3-D placement. To solve these nonconvex problems, we employed both model-based

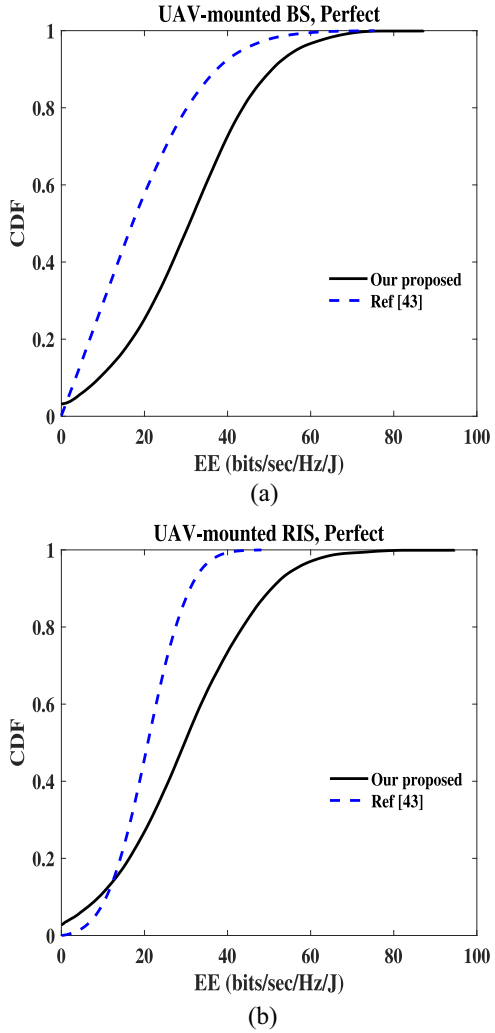

Fig. 14. Performance comparison of the proposed perfect-SIC NOMA with [43]. (a) Multi-UAV-mounted BS. (b) Multi-UAV-mounted RIS.

and model-free DRL algorithms. We considered both perfect and imperfect effects for multiple access techniques, including perfect and imperfect SIC for NOMA, zero and nonzero interference for OMA, and perfect and imperfect detection for SDMA. Our results demonstrated that the proposed methods achieved significant improvements in both SE and EE, with the model-based DRL method consistently outperforming model-free approaches. Moreover, the proposed NOMA transmission outperformed SDMA and OMA in terms of both SE and EE, making it an ideal candidate for future 6G communication networks. The practical implications of our work include improving network coverage and efficiency in UAV-assisted communication systems, with potential applications in smart cities, disaster recovery, and remote area connectivity.

VII. Future Work

This study lays the groundwork for exploring the integration of NOMA with UAV-mounted BS and RIS for energy-efficient and high-performance 6G networks. However, several research directions can be pursued to enhance and expand upon the findings of this work. One promising avenue is the development of hybrid DRL frameworks that combine the strengths of model-based and model-free approaches. Such frameworks could significantly improve optimization accuracy and adaptability while efficiently handling the computational complexity inherent in dynamic and complex wireless environments. Another critical area for future research is the incorporation of security constraints into the optimization process. UAV-based communication networks are inherently vulnerable to security threats, such as jamming and eavesdropping. Addressing these vulnerabilities would ensure more robust and secure data transmission in RIS-assisted multi-user air-to-ground communications. Additionally, the current study assumes static user locations, which may not fully capture the dynamic nature of real-world scenarios. Expanding the model to include user mobility patterns would provide more practical insights, especially in scenarios involving high-speed users or rapidly changing environments. Similarly, exploring multicell interference scenarios, where multiple UAV-mounted BSs and RIS operate in dense urban environments, would better align the framework with the complexities of real-world deployments.

Future research could also broaden the scope by evaluating additional performance metrics. While this study focuses on EE and SE, incorporating metrics, such as coverage, outage probability, and fairness, would offer a more comprehensive understanding of network performance. Finally, while this work serves as a proof of concept validated through simulations, the next logical step is to transition toward operational models and hardware implementations. Developing prototypes and conducting real-world testing will bridge the gap between theoretical optimization and practical deployment, ensuring that the proposed methods are ready for scalable and robust applications in 6G networks.

By addressing these directions, future studies can overcome the current limitations, refine the proposed framework, and advance the field of UAV and RIS-assisted communication systems.

Appendix
Description of DRL Algorithms

A. TD3 Algorithm

TD3 is a model-free DRL algorithm that improves upon the DDPG algorithm by addressing some of its drawbacks, notably overestimation bias. DDPG is an off-policy actor–critic algorithm that uses two neural networks: 1) an actor network to select actions and 2) a critic network to evaluate the value of those actions [41]. However, DDPG can over estimate the value of actions, leading to poor performance. TD3 addresses the over estimation bias problem in DDPG by using two critic networks instead of one. TD3 learns two Q-function, Q_{ϕ_1} and Q_{ϕ_2}, in almost the same way that DDPG learns its single Q-function. However, unlike DDPG, TD3 uses the minimum of the two Q-functions to form the Q-learning target. This helps to reduce the over estimation bias and improve the performance of the algorithm. Moreover, TD3 adds clipped noise to the actions generated by the target policy, $\mu_{\theta_{\text{targ}}}$, before using them to form the Q-learning target. This

Algorithm 1 Pseudocode of TD3

Initialize policy parameters θ, Q-function parameters ϕ_1 and ϕ_2, empty replay buffer $\mathcal{D}$
Set target parameters equal to main parameters $\theta_{\text{targ}} \leftarrow \theta$, $\phi_{\text{targ},1} \leftarrow \phi_1$, and $\phi_{\text{targ},2} \leftarrow \phi_2$
repeat
  Observe state $\mathbf{s}$ and select action $\mathbf{a} = \text{clip}(\mu_\theta(\mathbf{s}) + \epsilon, a_{\text{Low}}, a_{\text{High}})$, where $\epsilon \sim \mathcal{N}$
  Execute $\mathbf{a}$ in the environment
  Observe next state $\mathbf{s}'$, reward r, and done signal d to indicate whether $\mathbf{s}'$ is terminal
  Store $(\mathbf{s}, \mathbf{a}, r, \mathbf{s}', d)$ in replay buffer $\mathcal{D}$
  If $\mathbf{s}'$ is terminal, reset environment state.
  if it is time to update **then**
    for j in range **do**
      Randomly sample a batch of transitions, $B = \{(\mathbf{s}, \mathbf{a}, r, \mathbf{s}', d)\}$ from $\mathcal{D}$
      Compute target actions
      $\mathbf{a}'(\mathbf{s}') = \text{clip}\big(\mu_{\theta_{\text{targ}}}(\mathbf{s}') + \text{clip}(\epsilon, -c, c), a_{\text{Low}}, a_{\text{High}}\big)$
      Compute targets
      $y(r, \mathbf{s}', d) = r + \gamma(1-d) \min_{i=1,2} Q_{\phi_{\text{targ},i}}(\mathbf{s}', \mathbf{a}'(\mathbf{s}'))$
      Update Q-functions by one step of gradient descent using
      $\nabla_{\phi_i} \frac{1}{|B|} \sum_{(\mathbf{s},\mathbf{a},r,\mathbf{s}',d)\in B} \big(Q_{\phi_i}(\mathbf{s}, \mathbf{a}) - y(r, \mathbf{s}', d)\big)^2$
      if j mod *policy_delay* $= 0$ **then**
        Update policy by one step of gradient ascent using
        $\nabla_{\phi_i} \frac{1}{|B|} \sum_{\mathbf{s}\in B} Q_{\phi_1}(\mathbf{s}, \mu_\theta(\mathbf{s}))$
        Update target networks with
        $\phi_{\text{targ},i} \leftarrow \rho\phi_{\text{targ},i} + (1-\rho)\phi_i$
        $\theta_{\text{targ}} \leftarrow \rho\theta_{\text{targ}} + (1-\rho)\theta$
      end if
    end for
  end if
until convergence

helps to explore the action space (valid action space is $a_{\text{Low}} \le \mathbf{a} \le a_{\text{High}}$) more effectively and improve the robustness of the algorithm [41]. Thus, the target actions are obtained as

$$\mathbf{a}'(\mathbf{s}') = \text{clip}\big(\mu_{\theta_{\text{targ}}}(\mathbf{s}') + \text{clip}(\epsilon, -c, c), a_{\text{Low}}, a_{\text{High}}\big) \quad (22)$$

where clip(.) set the values in the valid range. s' denotes the next state and $\epsilon \sim \mathcal{N}(0, \sigma)$. Target policy smoothing acts as a regularizer for the algorithm, preventing it from exploiting incorrect sharp peaks in the Q-function approximator. This situation can arise in DDPG when the Q-function approximator develops an inaccurately sharp peak for certain actions. The policy may rapidly exploit that peak, leading to fragile or incorrect behavior. Target policy smoothing mitigates this issue by smoothing out the Q-function over similar actions. Both Q-functions share a single target, determined by using the smaller of the two Q-function values as

$$y(r, \mathbf{s}', d) = r + \gamma(1-d) \min_{i=1,2} Q_{\phi_{\text{targ},i}}(\mathbf{s}', \mathbf{a}'(\mathbf{s}')). \quad (23)$$

Then, both are learned by regressing to this target

$$L(\phi_1, \mathcal{D}) = \underset{(\mathbf{s},\mathbf{a},r,\mathbf{s}',d)\sim\mathcal{D}}{E}\Big[\big(Q_{\phi_1}(\mathbf{s}, \mathbf{a}) - y(r, \mathbf{s}', d)\big)^2\Big] \quad (24)$$

Algorithm 2 Pseudocode of MBPO

Initialize policy π_ϕ, predictive model p_θ, environment dataset D_{env}, model dataset D_{model}
for N epochs **do**
  Train model p_θ on D_{env} via maximum likelihood
  for E steps **do**
    Take action according to π_ϕ; add to D_{env}
    for Mmodel rollouts **do**
      Sample $\mathbf{s}$ uniformly from D_{env}
      Perform k-step model rollout starting from $\mathbf{s}$ using policy π_ϕ; add to D_{model}
      for G gradient updates **do**
        Update policy parameters on model data:
        $\phi \leftarrow \phi - \lambda_\pi \hat{\Delta}_\phi J_\pi(\phi, D_{model})$
      end for
    end for
  end for
end for

$$L(\phi_2, \mathcal{D}) = \underset{(\mathbf{s},\mathbf{a},r,\mathbf{s}',d)\sim\mathcal{D}}{E}\Big[\big(Q_{\phi_2}(\mathbf{s}, \mathbf{a}) - y(r, \mathbf{s}', d)\big)^2\Big] \quad (25)$$

where $E[.]$ denotes the expectation value. Using the smaller Q-value for the target helps to prevent overestimation in the Q-function. Finally, the policy is learned by maximizing Q_{ϕ_1} as

$$\max_\theta \underset{\mathbf{s}\sim\mathcal{D}}{\text{E}}\big[Q_{\phi_1}(s, \mu_\theta(\mathbf{s}))\big]. \quad (26)$$

The pseudocode for the TD3 is shown in Algorithm 2.

B. MBPO Algorithm

MBPO is a model-based, online, off-policy DRL algorithm that improves the performance and the efficiency of learning. The algorithm works by first interacting with the environment to collect experience data and train a model of the environment. Then, the agent updates the policy parameters using the real experience data and experience generated from the environment model. MBPO agents use the training algorithm shown in Algorithm 1, which periodically updates the environment model and the base off-policy agent [42].

MBPO uses an ensemble of Gaussian neural networks as its predictive model. Each member of the ensemble is defined as follows:

$$p_\theta(\mathbf{s}'|\mathbf{s}, \mathbf{a}) = N(\mu_\theta(\mathbf{s}, \mathbf{a}), \Sigma_\theta(\mathbf{s}, \mathbf{a})) \quad (27)$$

where $N(\mu_\theta(\mathbf{s}, \mathbf{a}), \Sigma_\theta(\mathbf{s}, \mathbf{a}))$ shows Gaussian distribution with mean $\mu_\theta(\mathbf{s}, \mathbf{a})$ and variance $\Sigma_\theta(\mathbf{s}, \mathbf{a})$. The maximum-likelihood loss used in the model training is expressed as

$$L(\theta) = \mathbb{E}\big[\log\big(p_\theta(\mathbf{s}'|\mathbf{s}, \mathbf{a})\big)\big]. \quad (28)$$

In policy optimization, the policy evaluation and policy improvement steps are, respectively, obtained as follows:

$$\text{Q}^\pi(\mathbf{s}, \mathbf{a}) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t r(\mathbf{s}, \mathbf{a})\right] \quad (29)$$

$$\min J_\pi(\phi, D) = \underset{\mathbf{s}\sim D}{\mathbb{E}}\big[D_{KL}\big(\pi \,\|\, \exp\big(\text{Q}^\pi - V^\pi\big)\big)\big]. \quad (30)$$

This update guarantees that $\text{Q}^{\pi_{new}}(\mathbf{s}, \mathbf{a}) \ge \text{Q}^{\pi_{old}}(\mathbf{s}, \mathbf{a})$ [44].

REFERENCES

[1] L. Gupta, R. Jain, and G. Vaszkun, "Survey of important issues in UAV communication networks," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 2, pp. 1123–1152, 2nd Quart., 2016.

[2] B. Duo, Q. Wu, X. Yuan, and R. Zhang, "Energy efficiency maximization for full-duplex UAV secrecy communication," *IEEE Trans. Veh. Technol.*, vol. 69, no. 4, pp. 4590–4595, Apr. 2020.

[3] G. Zhang, Q. Wu, M. Cui, and R. Zhang, "Securing UAV communications via joint trajectory and power control," *IEEE Trans. Wireless Commun.*, vol. 18, no. 2, pp. 1376–1389, Feb. 2019.

[4] Q. Wu and R. Zhang, "Towards smart and reconfigurable environment: Intelligent reflecting surface aided wireless network," *IEEE Commun. Mag.*, vol. 58, no. 1, pp. 106–112, Jan. 2020.

[5] D. Semmler, M. Joham, and W. Utschick, "High SNR analysis of RIS-aided MIMO broadcast channels," 2023, *arXiv:2210.15259*.

[6] S. Sobhi-Givi, M. G. Shayesteh, H. Kalbkhani, and N. Rajatheva, "Resource allocation and user association for load balancing in NOMA-based cellular heterogeneous networks," in *Proc. Iran Workshop Commun. Inf. Theory (IWCIT)*, 2020, pp. 1–6.

[7] S. Sobhi-Givi, M. Nouri, M. G. Shayesteh, H. Kalbkhani, and Z. Ding, "Joint power allocation and user fairness optimization for reinforcement learning over mmWave-NOMA heterogeneous networks," *IEEE Trans. Veh. Technol.*, vol. 73, no. 9, pp. 12962–12977, Sep. 2024.

[8] X.-T. Dang, H. V. Nguyen, and O.-S. Shin, "Optimization of IRS-NOMA-assisted cell-free massive MIMO systems using deep reinforcement learning," *IEEE Access*, vol. 11, pp. 94402–94414, 2023.

[9] A. Taha, Y. Zhang, F. B. Mismar, and A. Alkhateeb, "Deep reinforcement learning for intelligent reflecting surfaces: Towards standalone operation," in *Proc. IEEE 21st Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, 2020, pp. 1–5.

[10] Y. Wang, I. W.-H. Ho, S. Zhang, and Y. Wang, "Intelligent reflecting surface enabled fingerprinting-based localization with deep reinforcement learning," *IEEE Trans. Veh. Technol.*, vol. 72, no. 10, pp. 13162–13172, Oct. 2023.

[11] H. Peng and L.-C. Wang, "Energy harvesting reconfigurable intelligent surface for UAV based on robust deep reinforcement learning," *IEEE Trans. Wireless Commun.*, vol. 22, no. 10, pp. 6826–6838, Oct. 2023.

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

[13] X. Fan, M. Liu, Y. Chen, S. Sun, Z. Li, and X. Guo, "RIS-assisted UAV for fresh data collection in 3d urban environments: A deep reinforcement learning approach," *IEEE Trans. Veh. Technol.*, vol. 72, no. 1, pp. 632–647, Jan. 2023.

[14] L. Ge, P. Dong, H. Zhang, J.-B. Wang, and X. You, "Joint beamforming and trajectory optimization for intelligent reflecting surfaces-assisted UAV communications," *IEEE Access*, vol. 8, pp. 78702–78712, 2020.

[15] K. K. Nguyen, A. Masaracchia, V. Sharma, H. V. Poor, and T. Q. Duong, "RIS-assisted UAV communications for IoT with wireless power transfer using deep reinforcement learning," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 5, pp. 1086–1096, Aug. 2022.

[16] X. Qin, Z. Song, T. Hou, W. Yu, J. Wang, and X. Sun, "Joint optimization of resource allocation, phase shift and UAV trajectory for energy-efficient RIS-assisted UAV-enabled MEC systems," *IEEE Trans. Green Commun. Netw.*, vol. 7, no. 4, pp. 1778–1792, Dec. 2023.

[17] X. Mu, Y. Liu, L. Guo, J. Lin, and H. V. Poor, "Intelligent reflecting surface enhanced multi-UAV NOMA networks," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 10, pp. 3051–3066, Oct. 2021.

[18] J. Lei, T. Zhang, X. Mu, and Y. Liu, "NOMA for STAR-RIS assisted UAV networks," 2023, *arXiv:2307.14345*.

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

[20] I. Budhiraja, V. Vishnoi, N. Kumar, D. Garg, and S. Tyagi, "Energy-efficient optimization scheme for RIS-assisted communication underlaying UAV with NOMA," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2022, pp. 1–6.

[21] H. Long et al., "Joint trajectory and passive beamforming design for secure UAV networks with RIS," in *Proc. IEEE Globecom Workshops (GC Wkshps*, 2020, pp. 1–6.

[22] Z. Mohamed and S. A'ıssa, "Leveraging UAVs with intelligent reflecting surfaces for energy-efficient communications with cell-edge users," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, 2020, pp. 1–6.

[23] S. Solanki, S. Gautam, S. K. Sharma, and S. Chatzinotas, "Ambient backscatter assisted co-existence in aerial-IRS wireless networks," *IEEE Open J. Commun. Soc.*, vol. 3, pp. 608–621, 2022.

[24] A. Pitilakis et al., "On the mobility effect in UAV-mounted absorbing metasurfaces: A theoretical and experimental study," *IEEE Access*, vol. 11, pp. 79777–79792, 2023.

[25] H. Mei, K. Yang, Q. Liu, and K. Wang, "3D-trajectory and phase-shift design for RIS-assisted UAV systems using deep reinforcement learning," *IEEE Trans. Veh. Technol.*, vol. 71, no. 3, pp. 3020–3029, Mar. 2022.

[26] 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. 19584–19594, Nov. 2023.

[27] 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.

[28] S. Jiao, F. Fang, X. Zhou, and H. Zhang, "Joint beamforming and phase shift design in downlink UAV networks with IRS-assisted NOMA," *J. Commun. Inf. Netw.*, vol. 5, no. 2, pp. 138–149, Jun. 2020.

[29] Z. Mohamed and S. Aïssa, "Resource allocation for energy-efficient cellular communications via aerial IRS," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, 2021, pp. 1–6.

[30] 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.

[31] X. Lin et al., "The sky is not the limit: LTE for unmanned aerial vehicles," *IEEE Commun. Mag.*, vol. 56, no. 4, pp. 204–210, Apr. 2018.

[32] Q. Wu and R. Zhang, "Common throughput maximization in UAV-enabled OFDMA systems with delay consideration," *IEEE Trans. Commun.*, vol. 66, no. 12, pp. 6614–6627, Dec. 2018.

[33] Y. Ma, M. Li, Y. Liu, Q. Wu, and Q. Liu, "Optimization for reflection and transmission dual-functional active RIS-assisted systems," *IEEE Trans. Commun.*, vol. 71, no. 9, pp. 5534–5548, Sep. 2023.

[34] C. Huang, A. Zappone, G. C. Alexandropoulos, M. Debbah, and C. Yuen, "Reconfigurable intelligent surfaces for energy efficiency in wireless communication," *IEEE Trans. Wireless Commun.*, vol. 18, no. 8, pp. 4157–4170, Aug. 2019.

[35] Y. Zeng and R. Zhang, "Energy-efficient UAV communication with trajectory optimization," *IEEE Trans. Wireless Commun.*, vol. 16, no. 6, pp. 3747–3760, Jun. 2017.

[36] S. Sobhi-Givi, M. G. Shayesteh, and H. Kalbkhani, "Energy-efficient power allocation and user selection for mmWave-NOMA transmission in M2M communications underlaying cellular heterogeneous networks," *IEEE Trans. Veh. Technol.*, vol. 69, no. 9, pp. 9866–9881, Sep. 2020.

[37] M. Nouri et al., "Hybrid precoding based on active learning for mmWave massive MIMO communication systems," *IEEE Trans. Commun.*, vol. 71, no. 5, pp. 3043–3058, May 2023.

[38] M. Roderick, J. MacGlashan, and S. Tellex, "Implementing the deep Q-network," 2017, *arXiv:1711.07478*.

[39] M. Sewak, "Deep Q network (DQN), double DQN, and Dueling DQN: A step towards general artificial intelligence," in *Deep Reinforcement Learning: Frontiers of Artificial Intelligence*. Singapore: Springer, 2019, pp. 95–108.

[40] T. P. Lillicrap et al., "Continuous control with deep reinforcement learning," 2019, *arXiv:1509.02971*.

[41] S. Dankwa and W. Zheng, "Twin-delayed DDPG: A deep reinforcement learning technique to model a continuous movement of an intelligent robot agent," in *Proc. 3rd Int. Conf. Vis., Image Signal Process.*, 2019, pp. 1–5.

[42] M. Janner, J. Fu, M. Zhang, and S. Levine, "When to trust your model: Model-based policy optimization," in *Proc. 33rd Adv. Neural Inf. Process. Syst.*, 2019, pp. 1–12.

[43] W. Feng et al., "Resource allocation for power minimization in RIS-assisted multi-UAV networks with NOMA," *IEEE Trans. Commun.*, vol. 71, no. 11, pp. 6662–6676, Nov. 2023.

[44] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor–critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Proc. 35th Int. Conf. Mach. Learn.*, 2018, pp. 1861–1870.

Sima Sobhi-Givi received the B.Sc., M.Sc., and Ph.D. degrees in electrical engineering from Urmia University, Urmia, Iran, in 2013, 2015, and 2021, respectively.

From 2021 to 2023, she served as a Postdoctoral Research Fellow with Urmia University, supported by the Mobile Telecommunication Company of Iran (MCI), the largest mobile operator in the Middle East. She then worked as the Project Manager for 5G Communication Systems with MCI from 2023 to 2024. Since 2024, she has been an Assistant Professor with the Department of Electrical Engineering, University of Mohaghegh Ardabili, Ardabil, Iran. Her research interests include 5G and beyond-5G wireless cellular networks and the application of machine learning in communication systems.

Mahdi Nouri (Senior Member, IEEE) received the B.S. degree in telecommunication engineering from the University of Tabriz, Tabriz, Iran, in 2009, and the M.Sc. degree in telecommunication engineering from Iran University of Science and Technology, Tehran, Iran, in 2012.

From 2017 to 2019, he was an Assistant Professor (Lecturer) with the Department of Electrical Engineering, Electronics and Telecommunications, Arak University of Technology, Arak, Iran. He is currently a Research Fellow with the Sharif University of Technology (SUT), Tehran, and the Senior Project Manager of 5G Communication Systems with Mobile Telecommunication Company of Iran (MCI) who is the first and largest mobile operators in Middle East. He is also an Iran State Member and a Sector Member of MCI in the International Telecommunication Union (ITU). He has published more than 40 journal and conference papers. He was successfully selected for the Iran Postdoctoral Innovative Talent Support Program from the Iran National Science Foundation (INSF), SUT. His research interests include physical layer of 6G communication systems and B5G communication systems, deep reinforcement learning, machine learning in mobile communication systems, NFV-MANO for management, and orchestration of cloud networks.

Mr. Nouri is the winner of the 2017 Young Scientists Award from Iran's National Elites Foundation and the Outstanding Doctoral Society Award of the University of Isfahan. He served on the editorial boards of the IEEE transaction journals and conferences.

Mahrokh G. Shayesteh (Senior Member, IEEE) received the B.Sc. degree in electrical engineering from the University of Tehran, Tehran, Iran, the M.Sc. degree in electrical engineering from Khajeh Nasir University of Technology, Tehran, and the Ph.D. degree in electrical engineering from Amir Kabir University of Technology, Tehran, in 2003.

She is currently a Professor with the Department of Electrical Engineering, Urmia University, Urmia, Iran. She is also working with the Wireless Research Laboratory, Advanced Communication Research Institute, Department of Electrical Engineering, Sharif University of Technology, Tehran. Her research interests include wireless communications, signal, and image processing.

Hamid Behroozi (Member, IEEE) received the B.Sc. degree in electrical engineering from the University of Tehran, Tehran, Iran, in 2000, the M.Sc. degree in electrical engineering from the Sharif University of Technology, Tehran, in 2003, and the Ph.D. degree in electrical engineering from Concordia University, Montreal, QC, Canada, in 2007.

From 2007 to 2010, he was a Postdoctoral Fellow with the Department of Mathematics and Statistics, Queen's University, Kingston, ON, Canada. He is currently an Associate Professor with the Department of Electrical Engineering, Sharif University of Technology. His research interests include information theory, joint source–channel coding, artificial intelligence in signal processing and data science, and cooperative communications.

Dr. Behroozi was the recipient of several academic awards, including the Ontario Postdoctoral Fellowship awarded by the Ontario Ministry of Research and Innovation, the Quebec Doctoral Research Scholarship awarded by the Government of Quebec, the Hydro Quebec Graduate Award, and the Concordia University Graduate Fellowship.

Hyun Han Kwon (Member, IEEE) received the B.S. degree in civil engineering from Seoul National University of Science & Technology, Seoul, South Korea, in 1999, and the M.S. and Ph.D. degrees in water resources engineering from the University of Seoul, Seoul, in 2001 and 2004, respectively.

He is a Professor with the Department of Civil and Environmental Engineering, Sejong University, Seoul. Previously, he was a Professor with Chonbuk National University, Jeonju, South Korea, a Senior Research Scientist with the Korea Institute of Civil Engineering and Building Technology, Goyang, South Korea, and a Research Scientist with Columbia University's Water Center, New York City, NY, USA. He has authored over 300 peer-reviewed papers and has been cited more than 6 000 times. His research focuses on hydrology, risk analysis, and climate variability, with applications of machine learning, artificial intelligence, and data mining in flood forecasting, drought prediction, and water resources management.

Dr. Kwon currently serves as the Vice President of Research and President of the Industry-Academy Cooperation Foundation at Sejong University.

Md. Jalil Piran (Senior Member, IEEE) received the Ph.D. degree in electronics and information engineering from Kyung Hee University, Seoul, South Korea, in 2016.

Then, he worked as a Postdoctoral Fellow with the Networking Laboratory, Kyung Hee University. He holds a distinguished academic background and currently serves as an Associate Professor with the Department of Computer Science and Engineering, Sejong University, Seoul. He has made significant contributions to the field of artificial intelligence and data science through his extensive research publications in esteemed international journals and conferences. His areas of expertise encompass machine learning, data science, big data, Internet of Things, and cyber security.

Prof. Piran received accolades from the Iranian Ministry of Science, Technology, and Research as an Outstanding Emerging Researcher in 2017. In addition to his research endeavors, he actively engages with scholarly journals as an Editor, including the IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS and *Engineering Applications of Artificial Intelligence* (Elsevier). He has also served as a Guest Editor for the IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS and IEEE TRANSACTIONS ON CONSUMER ELECTRONICS. He is the Vice-Chair of the IEEE Consumer Technology Society on Machine Learning, Deep Learning, and AI (MDA) in Consumer Electronics. Furthermore, he assumes the role of Track Chair for Machine Learning, Deep Learning, and AI in the CE (MDA) Track for the 2024 IEEE International Conference on Consumer Electronics. In 2022, he chaired the 5G and Beyond Communications Session at the prestigious IEEE International Conference on Communications. He represents South Korea as an Active Delegate to the Moving Picture Experts Group from 2014 to 2017. His outstanding research contributions have been recognized internationally, as evidenced by the prestigious Scientist Medal of the Year 2017 awarded by IAAM in Stockholm, Sweden. His exceptional Ph.D. dissertation was honored as the Dissertation of the Year 2016 by the Iranian Academic Center for Education, Culture, and Research in the Engineering Group.