





Twin delayed deep deterministic policy gradient-based physical layer security and SEE in RIS-aided UAV communication

Amjad Iqbal ^{a,*}, Ala'a Al-Habashna ^{a,b,*}, Gabriel Wainer ^a, Gary Boudreau ^c

^a Department of Systems and Computer Engineering, Carleton University, Ottawa, Canada

^b School of Computing and Informatics, Al Hussein Technical University, Amman, Jordan

^c Ericsson Canada, Kanata, Canada

ARTICLE INFO

Keywords:

Unmanned aerial vehicles
Reconfigurable intelligent surfaces
Secrecy energy efficiency
Physical layer security
Deep deterministic policy gradient

ABSTRACT

Although unmanned aerial vehicles (UAVs) have been increasingly employed, their communication systems are vulnerable to eavesdropping, primarily due to line-of-sight (LoS) channels between the air and ground. Reconfigurable intelligent surfaces (RIS) offer the capability to adapt and reshape the wireless propagation environment, making them a compelling solution for enhancing physical layer security (PLS) in UAV-integrated wireless networks. In this paper, we introduce a network paradigm that exploits RIS's reflective capabilities alongside UAVs' maneuverability to maximize the sum secrecy energy efficiency (SEE) (i.e., the ratio of the aggregated secrecy rate for all legitimate users to the total power consumption). The main objective of this research is to achieve PLS by maximizing the sum of SEE in the presence of an eavesdropper using RIS-aided UAV networks under imperfect channel state information (CSI). To achieve this, we jointly optimize the power allocation and trajectories of the UAV and beamforming (active (UAV) and passive (RIS)). To address the non-convexity arising from the outdated CSI due to UAV maneuverability, we propose an advanced deep reinforcement learning (DRL) approach, the twin delayed deep deterministic policy gradient (DDPG). The proposed twin-delayed DDPG approach effectively solves the non-convex optimization problem. Through an extensive simulation campaign, we demonstrated that the proposed approach achieves an average SEE up to 22% higher than conventional approaches (e.g., Baseline and Myopic).

1. Introduction

Wireless communication technology has undergone rapid development over the last few decades, enriching our lives in numerous ways. Unmanned aerial vehicles (UAVs) are low-cost, highly mobile, and rapidly deployable with a wide coverage range. They play a key role in real-life situations such as emergency communication, post-disaster rescue, aerial photography, and cargo transportation [1]. Furthermore, UAVs play a crucial role in wireless communication networks, contributing to a wide range of applications. For instance, UAVs can be used as aerial relays to transmit signals and compute data for ground users [2–5]. UAVs can also collect data efficiently while communicating with multiple ground nodes [6]. Despite these advantages, UAV wireless connectivity is vulnerable to various security threats, such as jamming and eavesdropping.

To defend against these threats, the physical layer security (PLS) approach has emerged to enhance secrecy performance [7]. Specifically, PLS increases channel capacity for legitimate users while reducing the ability of potential eavesdroppers to decode data. In recent

years, research has been conducted to improve the performance of the PLS in wireless networks by utilizing advanced technologies such as non-orthogonal multiple access (NOMA) and massive multiple-input multiple-output (MIMO) [8,9]. However, most of these existing PLS schemes are designed for terrestrial or static environments and do not directly address the dynamic nature of UAV-assisted communication.

Previous work in this area, such as [8] and [9], did not address the airborne communication platforms. Research on UAV-based PLS has been explored in [10,11], with the primary objective of optimizing the secrecy performance through UAV trajectory design. Such investigations have demonstrated considerable improvements in secrecy rate; however, they fall short of altering how wireless signals propagate, limiting overall performance improvements for airborne communication platforms.

To overcome these limitations, a new paradigm is required that allows the propagation environment of signal waveforms to be synthetically customized. Using reconfigurable intelligent surfaces (RISs) has proven an effective approach to addressing such problems [12]. RIS can intelligently adjust the signal environment using affordable

* Corresponding authors.

E-mail addresses: amjadiqbal3@cunet.carleton.ca (A. Iqbal), alaa.alhabashna@htu.edu.jo (A. Al-Habashna).

<https://doi.org/10.1016/j.comnet.2025.111867>

Received 23 August 2025; Received in revised form 3 November 2025; Accepted 10 November 2025

Available online 14 November 2025

1389-1286/© 2025 Elsevier B.V. All rights reserved, including those for text and data mining, AI training, and similar technologies.

passive reflecting elements on a flat surface. This enables precise 3D passive beamforming, enhancing directional signal control and improving performance [13]. Furthermore, the dynamic adaptation of the RIS to the propagation environment serves various purposes, such as boosting signal strength and preventing eavesdropping for secure communication. Consequently, incorporating RIS technology can improve the signal propagation environment, effectively directing the signal toward the intended receivers. Furthermore, data transmitted through RIS incur fewer intermediate delays than those experienced with active relays at intermediary positions. The deployment of RIS is straightforward and proves to be an effective solution to reduce overall energy consumption.

The unique and complementary characteristics of UAVs and RIS have motivated the introduction of RIS-aided UAV communications to enhance overall network performance. While UAVs' elevated altitude significantly improves their ability to communicate with legitimate users, buildings or other obstacles can occasionally obstruct this connection. Consequently, deploying an RIS on a building or in a high-altitude area provides a viable option for redirecting signals between the UAV and users [14,15].

Recent advances in machine learning (ML) have enabled the optimization of wireless networks. A growing area of ML is deep reinforcement learning (DRL), which integrates deep neural networks (DNNs) and reinforcement learning (RL) [16]. DRL algorithms can be applied for various purposes in wireless networks, such as enhancing data rate [17], minimizing energy consumption [18], and improving real-time application processing time [19]. Furthermore, DRL algorithms offer substantial advantages in wireless networks by not requiring pre-collected data for training. Instead, DRL agents actively interact with their environment and generate training samples from these interactions. Using state transitions, the NNs are trained to adjust their parameters to maximize a specific reward function. As such, trained networks can be effectively deployed for real-time predictions.

1.1. Related work

UAVs and RIS have been increasingly utilized to enhance wireless network performance by improving coverage, signal strength, and overall system efficiency. The elevated positions of UAVs enable them to overcome many terrestrial obstacles, enhancing communication capabilities. Similarly, RIS can be strategically deployed to reflect and redirect wireless signals, optimizing the propagation environment and improving network performance. However, the static nature of RIS and the dynamic nature of UAVs present unique challenges that require sophisticated optimization techniques to address effectively. In particular, the complexity of maintaining reliable and secure communications in highly dynamic network environments, characterized by mobility, channel aging, and frequent topology changes, cannot be effectively addressed by traditional static designs.

1.1.1. UAV-Based wireless networks

UAVs play a vital role in enhancing wireless network coverage and signal transmission by leveraging their elevated altitudes [20–23]. In [20], a 3D spatial arrangement of multiple UAVs is introduced to optimize the total service time and ensure users' quality-of-service (QoS) demands are met. A dynamic service area (DSA) algorithm is proposed to identify potential service areas for the placement of UAVs. Additionally, the DSA is refined within convex regions to maximize the user QoS demands and service times. In [21], a singular UAV approach is proposed for gathering data from a cluster of ground sensors in wireless networks. The primary goal is to ensure service quality, minimize energy consumption, and limit transmitted power for each sensor. In [22], the optimization of throughput and energy efficiency (EE) is investigated in UAV-aided device-to-device (D2D) networks. This work aims to optimize the sum rate and EE under power splitting. A joint optimization of the Internet of Things (IoT) transmission and UAV trajectory

is explored to maximize the system EE [23]. Despite the significant improvement in network performance achieved by UAVs, using them alone may pose challenges, such as navigating static obstacles, mitigating signal blockages caused by buildings or terrain, and efficiently adapting to dynamic changes in the communication environment. Furthermore, UAVs are vulnerable to potential interception in security-sensitive applications due to their open nature [24]. To address this issue, integrated approaches must be employed that combine mobility, adaptability, and enhanced PLS.

1.1.2. RIS-Based wireless networks

RIS technology has recently garnered considerable attention and is considered a key enabler for future cellular networks due to its unique characteristics, including low cost and low energy consumption [25–27]. The emphasis on enhancing the PLS within RIS-assisted cell-free networks is explored in [25]. This involves optimizing active beamforming at base stations (BSs) and passive beamforming at RISs to maximize the weighted sum secrecy rate. In [26], a novel alternating optimization scheme is proposed for a refracting RIS-aided hybrid satellite system to minimize the total transmit power of the satellites and the BS, resulting in significant QoS improvements. In [27], the RIS-aided cellular network is integrated with a satellite-terrestrial system to maximize the secrecy rate while satisfying system constraints. RIS has significantly improved the network performance in terms of secrecy rate. However, relying solely on RIS may pose specific communication challenges, such as overcoming mobility constraints, adapting dynamically to rapidly changing scenarios, and extending coverage to areas not easily accessible by a fixed RIS setup. Moreover, RIS is not autonomously adaptable to environmental variations and threats, such as eavesdropping, which limits its standalone effectiveness in secure and mobile environments [28].

1.1.3. RIS and UAV integrated networks

Thus, combining UAVs and RISs can significantly address these challenges, providing a comprehensive solution for enhanced wireless communication performance in complex and dynamic scenarios. In [14], an innovative methodology is presented to optimize received signal strength by integrating vector beamforming and RIS phase shift techniques. In [29], an optimization approach is employed for RIS phase shift and UAV trajectory to maximize the user's sum rate while ensuring resource allocation (RA) fairness. In [30], a RIS-aided UAV optimization model is presented to maximize achievable sum rates by optimizing resource scheduling, power allocation, and UAV trajectory. In [31], a novel energy-saving system is proposed for RIS-aided UAV networks, utilizing a game-theoretic approach. The aim is to maximize the received signal strength by UAVs and the EE experienced by each user. These studies [29–31] demonstrate that RIS-aided UAV communications have the potential to improve the quality and performance of wireless communications. However, these works exhibit two key limitations. Firstly, it is assumed that UAVs and legitimate users have perfect channel state information (CSI), implying that the communication channel is always known precisely. Such may not be the case in real-world situations. Secondly, all these works adhere to a traditional model-based approach, assuming a stationary network environment over time, which lacks practicality in scenarios featuring mobile users and highly dynamic network environments (e.g., radio). Furthermore, eavesdroppers are often overlooked, despite their critical importance in real-world deployments of aerial and intelligent reconfigurable platforms [32]. Thus, relaxing these unrealistic assumptions is imperative to exploring more practical scenarios. Ultimately, this exploration is expected to yield more versatile models that can be applied in real-time scenarios, offering substantial practical benefits by accommodating changes in the network environment and accounting for outdated CSI at each time step.

Table 1
Key differences between existing work and proposed work.

Papers	DRL Policy	Action	SEE Objective	Outdated CSI	Joint Optimization	DRL Approach
[33–36]	Value-based	Discrete	×	×	×	DQN
[37–40]	Policy-based	Continuous	×	×	×	DDPG
This work	Policy-based	Continuous	✓	✓	UAV + RIS + Power	Twin DDPG

1.1.4. ML in RIS-UAV networks

ML, particularly DRL, has emerged as a promising technique for optimizing decision-making in highly dynamic wireless networks. DRL has the potential to meet the high standards and demands of these networks. DRL trains the NNs offline, enabling them to make decisions quickly and accurately. A few recent works have leveraged these unique features to optimize RIS-aided UAV wireless networks. For instance, a deep Q network (DQN) is proposed in [33] to control RIS phase shifts and design UAV trajectories to maximize the systems' data rate and EE performance. A similar approach is proposed in [34] to jointly optimize the UAV 3D trajectory and the RIS phase shift. As a result of the proposed algorithm, the UAV's EE has been enhanced while ensuring a high level of QoS for users. In [35], the UAV trajectory and RIS phase shift are optimized to maximize system capacity under the UAV energy consumption constraint. The DRL algorithm jointly optimizes the UAV trajectory and RIS phase-shift design, thereby increasing the system capacity of RIS-UAV-assisted networks [36]. The work described in [33–36] focuses on maximizing the long-term cumulative reward using value-based DRL algorithms, which are well-suited only to discrete action spaces. However, maintaining smooth control of UAV and RIS adjustment in dynamic environments is always challenging. Policy-based (e.g., DDPG) algorithms are more suitable for dynamic environments that require continuous actions. For instance, in [37], a DDPG-based algorithm is employed to jointly optimize the RIS phase shift, the UAV's horizontal position, and the BS's power allocation, thereby maximizing the sum rate. In [38], the UAV power allocation and the RIS phase shift matrix are jointly optimized to maximize the EE performance of the considered networks. A similar approach can also be seen in [39,40]. However, these existing studies primarily focus on optimizing either the UAV trajectory, active UAV beamforming, or passive RIS beamforming alone, without exploring joint optimization. Moreover, these investigations neglect the explicit consideration of eavesdroppers, a security concern for emerging wireless networks. In addition, these studies predominantly rely on model-based environmental settings (e.g., predefined channel conditions or fixed user mobility), which may not hold in dynamic scenarios. Therefore, practical and adaptable approaches are necessary to address the complexity and dynamic nature of wireless networks in various environments. Thus, addressing dynamic control and PLS jointly remains an open research direction that our work aims to fulfill. Therefore, in this work, we employ outdated, realistic CSI and continuous control via twin DDPG,¹ and explicitly focus on SEE rather than secrecy rate alone, addressing both security and energy trade-offs in a dynamic UAV-RIS scenario compared to existing DRL-based works that either focus on rate maximization or assume perfect CSI. In Table 1, we highlight the key differences between existing works and the proposed framework. Notably, while several studies have explored DRL-based optimization for UAV-RIS systems, our work is the first to address SEE maximization under outdated CSI using a Twin Delayed DDPG framework with joint control over UAV trajectory, RIS beamforming, and transmit power.

¹ The proposed twin-delayed DDPG framework is based on the standard TD3 [41]. However, instead of designing a new learning architecture, we adapt the TD3 principles of twin critics and delayed policy updates to two coordinated DDPG agents. Specifically, DDPG1 manages active (UAV) and passive (RIS) beamforming, while DDPG2 determines the UAV trajectory. This dual-agent setup enables simultaneous yet decoupled learning of interconnected control variables in continuous action spaces and with outdated CSI

1.2. Motivations and contributions

Wireless communication has significantly advanced thanks to RIS and UAV technologies. However, there are still several challenges to overcome, and they need to be addressed efficiently.

- Many studies assume a perfect CSI, which is impractical in real-world scenarios where channel conditions are unpredictable and constantly changing [42] and [43].
- In traditional network models, the environment is stationary, lacking adaptability to dynamic changes such as user mobility and varying communication requirements [44]. However, in a practical scenario, such assumptions can cause suboptimal UAV positioning and beamforming, leading to higher transmission power and lower EE.
- Future wireless communications (i.e., emerging wireless networks) are expected to face critical security issues, such as the threat of eavesdroppers, which are often not explicitly addressed in existing works [37,45,46]. This needs to be addressed efficiently; even minor inaccuracies in CSI can lower the secrecy rate and increase the risk of information leaks for legitimate users.
- Moreover, most existing works often oversimplify real-world variability, failing to account for environmental dynamics such as UAV mobility, RIS coverage limitations, and CSI aging, all of which significantly impact performance. These works cannot guarantee security or stable communication quality in dynamic situations, limiting their practical application in emergency response, disaster monitoring, or military communications.

This means that more practical and adaptable solutions are needed to effectively handle real-time, dynamic environments and address security concerns to meet these challenges. Traditional optimization approaches, such as alternating optimization (AO), successive convex approximation (SCA), and swarm intelligence (SI) algorithms, are widely used to solve such problems. These approaches generally require accurate CSI and are suitable for small-scale networks. However, these approaches are demanding high computationally power and often unsuitable for solving large-scale network problems, especially in UAV-RIS networks, characterized by channel aging, high mobility, and non-stationary topologies. On the other hand, DRL can learn optimal control policies directly from interactions with the environment, without requiring explicit modeling of complex, time-varying wireless channels. Once trained, the DRL agent can make real-time decisions for continuous control variables such as UAV trajectory, RIS phase shifts, and transmit power, thereby achieving better adaptability and robustness to outdated CSI. Moreover, the proposed framework mitigates the overestimation bias present in the standard DRL approach and ensures more stable convergence in continuous action spaces. Therefore, a DRL-based approach is more suitable for addressing the joint UAV-RIS optimization problem under realistic, dynamic conditions compared to traditional non-DRL optimization methods. Keeping these motivations in mind, this paper aims to overcome the identified limitations by incorporating a joint optimization approach. Specifically, we address the outdated CSI at each time step t and explicitly consider the UAV trajectory, as well as active (UAV) and passive (RIS) beamforming, and UAV transmit power. In our prior work [47], we proposed an approach to maximize the secrecy rate using DRL. We extend our previous work by focusing on maximizing SEE rather than secrecy rate. In summary, the paper's main contributions are listed in the following:

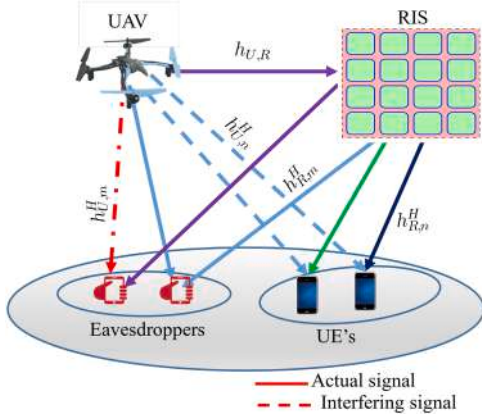


Fig. 1. System model setup.

- RIS-aided UAV Network for Secure Communication:** Existing studies primarily focus on optimizing either UAV trajectory, active UAV beamforming, or passive RIS beamforming individually, without considering joint optimization. In this paper, we jointly optimize these elements through an RIS-aided UAV network to ensure secure communication. The UAV is deployed to serve specific legitimate user equipment (UE) subject to the presence of an eavesdropper. The RIS is employed to minimize the success of eavesdropping attempts, which are posed by the severe shadowing effect on the legitimate UE, and to enhance the signal quality received from the associated UAV, thereby ensuring secure communication. Furthermore, the impact of RIS is assessed, and its performance is compared with the No-RIS case.
- Formulation of SEE Maximization Problem:** We formulate the SEE maximization problem for RIS-aided UAV networks, taking into account key constraints such as the secrecy rate requirement for each legitimate UE, the power budget across UAVs, and the RIS phase shift coefficient. The common assumption of perfect CSI is relaxed, and the impact of outdated CSI is evaluated at each time step t .
- MDP-based Solution with Twin DDPG:** For the formulated SEE problem, we define the state space, action space, and reward function using the Markov Decision Process (MDP) tool [48]. Solving such a problem using a single deep deterministic policy gradient (DDPG (*Baseline approach*)) faces an overestimation bias problem, which leads to suboptimal policy updates. To solve the formulated problem efficiently, we proposed a twin delay DDPG methodology. The first DDPG (i.e., *DDPG1*) is used to find the optimal beamforming policies for active (UAV) and passive (RIS), while the second DDPG (i.e., *DDPG2*) is used to determine the optimal trajectory for UAV.

1.3. Paper organization

Section 2 describes the system model and formulates the problem of maximizing the SEE of UAV-assisted wireless communications augmented by RIS. Section 3 provides a foundational understanding of DRL. Section 4 introduces the policy-based DDPG approach, focusing on joint optimization. Section 5 presents simulation results that demonstrate the efficiency of our proposed methods compared to conventional techniques. Finally, Section 6 encapsulates the conclusion and future direction.

2. System model and problem formulation

In this paper, we consider a wireless downlink network consisting of multiple UAVs $\mathcal{U} = \{1, 2, \dots, U\}$ supported by an RIS, as shown in Fig. 1.

The RIS plays a pivotal role in facilitating secure communication for transmitting confidential information from the UAVs to N single-

antenna UEs amidst the presence of M single-antenna eavesdroppers. In this case, a uniform linear array (ULA) with A -element is employed by UAVs, while the uniform planar array (UPA) with $A = a^2$ is utilized by RIS, such that a is an integer. Moreover, the sets of UEs and the eavesdroppers are denoted as $\mathbb{N} = \{1, 2, \dots, N\}$, $\mathcal{M} = \{1, 2, \dots, M\}$, respectively. A 3D Cartesian coordinate system is used to place all entities. The fixed coordinates are assigned to RIS at $w_R = (x_R, y_R, z_R)^T$. Without loss of generality and to avoid confusion, the ground BS interference is excluded in this work, as the focus is on the interactions within the UAV and RIS-assisted network. This approach demonstrates the effectiveness of RIS's in enhancing secure communication for the proposed system model.

In the proposed scenario, the UAV is assumed to traverse at a fixed altitude throughout K finite time slots $T = \{t_k\}_{k=1}^K$, where t_k indicates the time slot. During k -th time slot, the coordinates of the UAVs and UEs/eavesdroppers can be represented as $\mathbf{q}[k] = (x_u[k], y_u[k], H_u[k])^T$ and $\mathbf{w}_i[k] = (x_i[k], y_i[k], z_i[k])^T$, $\forall i \in \mathbb{N} \cup \mathcal{M}$, respectively. Furthermore, at k -th time slot, we represent the location information as $\mathcal{W} = \{\mathbf{q}[k]\} \cup \{\mathbf{w}_i[k] \mid \forall i \in \mathbb{N} \cup \mathcal{M}\}$. The UAVs move within a predefined area and are restricted by a maximum height H_{\max} . Moreover, we assume that UAVs can detect and avoid obstacles. Therefore, the mobility constraints subject to the UAV can be formulated in Eq. (1) as:

$$\mathbf{q}[0] = (0, 0, H_U), \quad (1a)$$

$$B \geq \max(x[k], y[k]), \quad k = 1, \dots, K, \quad (1b)$$

$$H_{\max} \geq \sqrt{\|q[k+1] - q[k]\|^2}, \quad k = 1, \dots, K-1. \quad (1c)$$

The UAV's initial coordinate is represented in Eq. (1a), while the UAV's moving boundaries B and maximum distance H_{\max} at each time step are represented by Eqs. (1b) and (1c), respectively. We define the different connection channel gains for the proposed system model. Each channel gain can be represented as channel gain from the u -th UAV to the RIS as $h_{(U,R)} \in \mathbb{C}^{(M \times A)}$, from the u -th UAV to the m -th eavesdropper as $h_{(U,m)} \in \mathbb{C}^{(A \times 1)}$, from the u -th UAV to the n -th user as $h_{(U,n)} \in \mathbb{C}^{(A \times 1)}$, from the RIS to the n -th user as $h_{(R,n)} \in \mathbb{C}^{(N \times 1)}$, and from the RIS to the m -th eavesdropper as $h_{(R,m)} \in \mathbb{C}^{(M \times 1)}$, respectively. Furthermore, all of these channels are modeled using the 3D Saleh Valenzuela (SV) channel model [49].

$$h_{U,i} = \sqrt{\frac{1}{L_{UN}}} \sum_{l=1}^{L_{UN}} g_{i,l}^u q_L(\phi_{i,l}^{\text{AoD}}), \quad \forall i \in \mathbb{N} \cup \mathcal{M}, \quad (2a)$$

$$h_{R,i} = \sqrt{\frac{1}{L_{RN}}} \sum_{l=1}^{L_{RN}} g_{i,l}^r q_M(\phi_{i,l}^{\text{AoD}}, \theta_{i,l}^{\text{AoD}}), \quad \forall i \in \mathbb{N} \cup \mathcal{M}, \quad (2b)$$

$$h_{U,R} = \sqrt{\frac{1}{L_{RN}}} \sum_{l=1}^{L_{RN}} g_l^{ur} q_M(\phi_l^{\text{AoA}}, \theta_l^{\text{AoA}}) q_L(\phi_l^{\text{AoD}})^H, \quad \forall i \in \mathbb{N} \cup \mathcal{M}. \quad (2c)$$

The large-scale fading coefficients $g \in \{g_{(i,l)}^u, g_{(i,l)}^r, g_l^{(ur)}\}$ can be formulated as $\mathcal{CN}(0, 10^{(PL/10)})$, where $PL(dB) = C_0 - 10\alpha \log_{10}(D) - PL_s$, such that α , C_0 and D indicates the path-loss exponent, the reference distance for one meter's path loss, and the link distance, respectively. Furthermore, the shadow fading coefficient is represented by $PL_s \sim \mathcal{CN}(0, \sigma_s^2)$. The steering vector q_L used in ULA can be expressed as [50].

$$q_L(\varphi) = \left[1, e^{j \frac{2\pi}{\lambda_c} d \sin(\varphi)}, \dots, e^{j \frac{2\pi}{\lambda_c} d(A-1) \sin(\varphi)} \right]^H. \quad (3)$$

such that φ denotes the azimuth angle of departure (AoD) for $\varphi_{(i,l)}^{\text{AoD}}$ and $\varphi_{(i)}^{\text{AoD}}$, while λ_c and d stand for the carrier wavelength and inter-spacing of the antenna, respectively. Similarly, the UPA steering vector $q_M(\varphi, \theta)$ can be expressed as $q_M(\varphi, \theta) = [1, \dots, e^{j \frac{2\pi}{\lambda_c} d(i \sin(\varphi) \sin(\theta) + j \cos(\varphi) \sin(\theta))}, \dots]^H$, where $\varphi(\theta)$ shows azimuth (elevation) for the angle of arrival (AoA) with $\varphi_{(i,l)}^{\text{AoA}}$ and $\theta_{(i,l)}^{\text{AoA}}$ and $0 \leq i, j \leq a-1$, respectively. The optimization variable \mathbf{Q} and the CSI coupling are determined by the line-of-sight (LoS) components $\varphi(\theta)_{l=1}^{\text{AoA(AoD)}}$ involving the trajectories of users' and UAVs for each link. According to the SV channel model,

the AoA/AoDs vary with propagation paths, challenging the idea that LoS components depend solely on the location of the UAV [14]. Therefore, we mathematically represented the LoS components for each link $\varphi(\theta)_l^{\text{AoA(AoD)}}$, $l \neq 1$ as follows [51]:

$$\varphi(\theta)_l^{\text{AoA(AoD)}} = \varphi(\theta)_{l=1}^{\text{AoA(AoD)}} + \Phi(\Lambda)_l^{\text{AoA(AoD)}}, \quad l = 2, \dots, L. \quad (4)$$

where $\Phi(\Lambda)^{\text{AoA(AoD)}}$ is known as a spreading factor [51]. We assume that the communication channel between UAVs to users, or potential eavesdroppers, is imperfect and characterized by $H_{C,i} = \text{diag}(h_{R,i}^H)h_{U,R}$, $\forall i \in \mathbb{N} \cup \mathcal{M}$. The beamforming of RIS is described as $\phi = \text{diag}(\beta_1 e^{j\theta_1}, \beta_2 e^{j\theta_2}, \dots, \beta_A e^{j\theta_A})$, where $\beta_a \in [0, 1]$, $\theta_a \in [0, 2\pi)$, $a = 1, 2, \dots, A$, represent the amplitude reflection coefficient and phase shift of the a -th RIS reflection element, respectively. To maximize the reflected signal power, a constant value for $\beta_a = 1$ is considered for all elements. The combined channel gains from a UAV to all receivers can be calculated as follows:

$$H_C = \{h_{U,i}^H + \xi^H H_{C,i} \mid \forall i \in \mathbb{N} \cup \mathcal{M}\}. \quad (5)$$

where ξ shows the RIS passive beamforming matrix and can be vectorized as $\xi = \text{vec}(\phi)$. Thus, the signal received at the i -th user or eavesdropper by each UAV can be formulated as follows:

$$y_i = (h_{U,i}^H + \xi^H H_{C,i}) \mathbf{W} \mathbf{b} + o_i, \quad \forall i \in \mathbb{N} \cup \mathcal{M}. \quad (6)$$

where $\mathbf{W} \in \mathbb{C}^{(A \times N)}$ represents the UAV beamforming matrix and $\mathbf{b} \in \mathbb{C}^{(N \times 1)}$ indicates the transmitted symbol. Thus, for the n -th user at time step t , the signal-to-interference-plus-noise ratio (SINR) can be expressed as follows:

$$\text{SINR}_n^u(t) = \frac{(|h_{U,n}^H + \xi^H H_{C,n}|w_n)^2}{\sum_{n' \in \mathbb{N} \setminus \{n\}} (|h_{U,n'}^H + \xi^H H_{C,n'}|w_{n'})^2 + \sigma_n^2}. \quad (7)$$

where $o_i \sim \mathcal{N}(0, \sigma_n)$, $\forall i \in \mathbb{N} \cup \mathcal{M}$ denotes the background noise of legitimate users. As a result, the achievable rate of the n -th user can be represented as follows:

$$\partial_n^u = \log_2(1 + \text{SINR}_n^u(t)). \quad (8)$$

Similarly, for the m -th eavesdropper's signal with respect to the n -th user at the time step t , the SINR can be expressed as follows:

$$\text{SINR}_{m,n}^e(t) = \frac{(|h_{U,m}^H + \xi^H H_{C,m}|w_n)^2}{\sum_{n' \in \mathbb{N} \setminus \{n\}} (|h_{U,m'}^H + \xi^H H_{C,m'}|w_{n'})^2 + \sigma_m^2}. \quad (9)$$

The achievable rate from the m -th eavesdropper's to the n -th user can be denoted as:

$$\partial_{m,n}^e = \log_2(1 + \text{SINR}_{m,n}^e(t)). \quad (10)$$

Finally, the individual secrecy rate from the UAV to the n -th user can be formulated as follows [52].

$$\partial_n^{\text{sec}} = \left[\partial_n^u - \partial_{m,n}^e \right]^+. \quad (11)$$

where $[j]^+ = \max(0, j)$. Note that the value of $\partial_n^u - \partial_{m,n}^e$ is always non-negative for the t_k time slot. If this value becomes negative, we can make the transmit power $P_t = 0$, making the result of Eq. (11) equal to 0. This means that adjusting the transmit power P_t ensures the secrecy rate is always non-negative.

2.1. Secrecy energy efficiency (SEE)

According to [53] and [54], we define the SEE as the ratio of the sum of the secrecy rates ∂_n^{sec} of all legitimate users/UEs in N to the total power consumption. The SEE can be achieved under the constraints of active and passive (Φ, \mathbf{W}) beamforming matrices, UAV trajectory q , and the transmit power P_t . Mathematically, the SEE can be expressed as follows:

$$\Upsilon = \text{SEE}(\{\Phi, \mathbf{W}, q\}) = \frac{\sum_{n=1}^N \partial_n^{\text{sec}}}{\mathfrak{P}P_t + P_R + P_U}, \quad (12)$$

such that \mathfrak{P} is the power amplifier efficiency and is assumed constant in this work. P_t is the UAV transmit power for users N . P_U indicates the UAV's power consumption during its flight. P_R denotes the RIS power consumption in each reflecting element. Without loss of generality, we summarized all the mathematical notation used in this work in Table 2.

2.2. Problem formulation

In this subsection, the SEE maximization problem is formulated. This can be achieved by jointly optimizing the beamforming matrices (active (UAV) and passive (RIS)), the UAV trajectory, and the UAV transmit power. The UAVs must select the most appropriate coordinates to transmit signals to RIS at each time step t . The RIS will then use the local environment information to determine the optimal phase shift for a legitimate user. Accordingly, we formulate the optimization problem based on the beamforming matrices (active (UAV) and passive (RIS)), the UAV's trajectory, and transmit power to encompass all UEs.

$$(P1) : \max_{q, \Phi, \mathbf{W}, P_t} \sum_{n \in \mathcal{N}} \Upsilon \quad (13)$$

$$\text{s.t. (1)} \quad (14a)$$

$$\partial_n^{\text{sec}} \geq \partial_n^{\text{sec}, \min}, \quad \forall n \in \mathcal{N}, \quad (14b)$$

$$0 \leq \theta_m \leq 2\pi, \quad m = 1, \dots, M, \quad (14c)$$

$$P_t(\mathbf{W}\mathbf{W}^H) \leq P_{\max} \quad (14d)$$

$$\hat{H}(t) = H(t) + \Delta H(t), \quad \|\Delta H(t)\| \leq \epsilon. \quad (14e)$$

(14a) represents the UAV mobility constraint (previously defined in (1)). Constraint (14b) states that the minimum secrecy rate from the UAV to the legitimate user n can be achieved without compromising the user's QoS requirement, thereby improving the overall SEE. Constraint (14c) indicates the RIS reflecting gain and bounded in the range of 0 and 2π , whereas constraint (14d) confines the total transmitted power of UAVs during the beamforming process to not exceed the maximum threshold P_{\max} . Constraint (14e) indicates imperfect CSI to reflect realistic UAV network conditions, accounting for feedback delays and estimation errors, where $\hat{H}(t)$, $\Delta H(t)$, and ϵ represent the estimated CSI, bounded error term, and maximum uncertainty radius, respectively. This model reflects more realistic assumptions of channel estimation in time-varying UAV scenarios.

Remark: The proposed framework uses perturbed CSI during training to minimize instability caused by outdated CSI. The agent learns to account for channel variations within the uncertainty bounds, making the learning process more robust to estimation delays and feedback errors. This helps prevent oscillations and enhances algorithmic stability in time-varying conditions. By incorporating perturbed CSI during training, the proposed framework inherently adapts to dynamic environments.

Although some elements, such as the state variables and constraints of (P1), reappear in the MDP formulation, this separation is intentional. We defined the mathematical optimization objectives and constraints in P1, providing an overview of the system design. The MDP reformulation translates these objectives into a learning-based framework, such as states, actions, and rewards, suitable for training the proposed networks. This provides readers with a clear understanding of the optimization goals and the practical approach the DRL agent uses to solve the problem in a dynamic UAV-RIS network.

Although problem (P1) is reformulated in a more tractable form, however, it is still a non-convex problem due to the constraints in (14a), (14b), (14c), and time-varying CSI at each time step t . No general optimization method currently exists to address problem (P1) effectively.

Therefore, in this paper, we present an advanced DRL approach to formulate the SEE optimization problem. The aim is to obtain optimal UAV trajectories and beamforming matrices (active (UAV) and passive (RIS)), rather than directly addressing the NP-hard optimization

Table 2
List of notations.

Symbol	Definition	Symbol	Definition
$\mathcal{U} = \{1, \dots, U\}$	Set of UAVs	$\mathbb{N} = \{1, \dots, N\}$	Set of legitimate users (UEs)
$\mathcal{M} = \{1, \dots, M\}$	Set of eavesdroppers	$A = a^2$	Number of RIS reflecting elements
$T = \{t_k\}_{k=1}^K$	Time slots	$\mathbf{q}[k] = (x_u[k], y_u[k], H_u[k])^T$	UAV position at slot k
$\mathbf{w}_i[k] = (x_i[k], y_i[k], z_i[k])^T$	Position of user/eavesdropper i	$\mathbf{w}_R = (x_R, y_R, z_R)^T$	RIS location
B	UAV horizontal boundary	H_{\max}	UAV maximum altitude
$h_{(U,R)} \in \mathbb{C}^{M \times A}$	Channel UAV \rightarrow RIS	$h_{(U,n)}, h_{(U,m)}$	Channels UAV \rightarrow user / UAV \rightarrow eavesdropper
$h_{(R,n)}, h_{(R,m)}$	Channels RIS \rightarrow user / RIS \rightarrow eavesdropper	L_{UN}, L_{RN}	Number of propagation paths
$g_{i,j}^u, g_{i,j}^e, g_{i,j}^r$	Complex fading coefficients	α	Path-loss exponent
C_0	Reference path-loss constant	D	Link distance
PL_s	Shadow fading term	$q_L(\varphi)$	ULA steering vector
$q_M(\varphi, \vartheta)$	UPA steering vector	λ_c	Carrier wavelength
d	Antenna spacing	φ, ϑ	Azimuth and elevation angles (AoA/AoD)
$\Phi(\Lambda)$	Angle spreading factor	$\phi = \text{diag}(\beta_1 e^{j\theta_1}, \dots, \beta_A e^{j\theta_A})$	RIS reflection matrix
β_a, θ_a	Amplitude and phase of RIS element a	$\xi = \text{vec}(\phi)$	Vectorized RIS beamforming matrix
$\mathbf{W} \in \mathbb{C}^{A \times N}$	UAV beamforming matrix	$\mathbf{b} \in \mathbb{C}^{N \times 1}$	Transmitted symbol vector
$H_{c,j}$	Combined UAV-RIS-receiver channel	$o_i \sim \mathcal{N}(0, \sigma_i^2)$	Receiver noise
$\text{SINR}_n^u(t)$	SINR at user n	$\text{SINR}_{m,n}^e(t)$	SINR at eavesdropper m for user n
\hat{r}_n^u, \hat{r}_m^e	Achievable rates of user / eavesdropper	\hat{r}_n^{sec}	Secrecy rate of user n
Υ	Secrecy energy efficiency (SEE)	\mathfrak{P}	Power amplifier efficiency
P_i	UAV transmit power	P_R, P_U	RIS and UAV power consumption
P_{\max}	Maximum transmit power	$\hat{H}(t), \Delta H(t), \epsilon$	Estimated CSI, error, and bound
$(P1)$	SEE maximization problem		

problem. To ensure robustness under CSI uncertainty, the DRL agent is trained using perturbed CSI samples that simulate estimation noise and feedback delays. This allows the learned policy to generalize to practical operating environments with imperfect CSI. Before diving into the details of the proposed DRL approach, we first provide an overview of the DRL approach to enrich understanding.

3. Preliminary descriptins of DRL

To implement DRL algorithms that support a system, two main approaches are used: value search and policy search. The value search approach estimates the policy parameter by examining the value of a given state. Mathematically, we define the value function V following the policy π at a given state $s \in S$ as follows:

$$V^\pi(s) = \mathbb{E}[R | s, \pi], \quad (15)$$

such that \mathbb{E} , R , and S represent the expectation function, the reward function, and the state space, respectively. The optimal value function $V^*(s)$ followed by the optimal policy π^* can be mathematically represented as:

$$V^*(s) = \max_{\pi} V^\pi(s), \quad s \in S, \quad (16)$$

Following an optimal policy π^* that satisfies the Bellman equation [55], an agent chooses the action $a \in \mathcal{A}$ that maximizes the expected cumulative reward.

$$V^*(s) = V^{\pi^*}(s) = \max_{a \in \mathcal{A}} \left\{ \mathbb{E}[r(s, a)] + \zeta \sum_{s' \in S} \mathcal{P}_{ss'}(a) V^*(s') \right\}. \quad (17)$$

where $\mathcal{P}_{ss'}(a)$ signifies the transition probability followed by the current state $s = s'$ and the future state $s' = s^{(t+1)}$ from state space $s \in S$ and action space $a \in \mathcal{A}$. ζ indicates the discount factor. The state-action value $Q(s, a)$ can be calculated when the agent observes the current state $s \in S$ executes the action $a \in \mathcal{A}$, and follows the policy π :

$$Q^\pi(s, a) = \mathbb{E}[r(s, a)] + \zeta \sum_{s' \in S} \mathcal{P}_{ss'}(a) V^*(s') \quad (18)$$

From (17) and (18), we can write the optimal value function as

$$V^*(s) = \max_{a \in \mathcal{A}} Q^*(s, a), \quad (19)$$

Eq. (19) represents the action-value function with the optimal policy π^* . In policy-based search, the policy can be found directly by adjusting the policy parameters. In policy search, one popular approach is policy

gradients, which provide efficient sampling across a wide range of parameters. Our goal is to find the optimal policy π^* that can maximize the cumulative reward function and can be described as:

$$J(\theta_\pi) = \sum_{s \in S} d^\pi(s) \sum_{a \in \mathcal{A}} \pi_\theta(a | s) r^\pi(s, a), \quad (20)$$

where θ_π and d^π are the vector policy parameter and the Markov chain stationary distribution with policy π_θ , respectively. The policy parameter θ_π is adjusted using the gradient descent, relying on $\nabla_{\theta} J(\theta_\pi)$ and can be expressed mathematically as:

$$\begin{aligned} \nabla_{\theta} J &= \sum_{s \in S} d^\pi(s) \sum_{a \in \mathcal{A}} \nabla_{\theta} \pi_\theta(a | s) Q^\pi(s, a) \\ &= \mathbb{E}_{\pi_\theta} [\nabla_{\theta} \ln \pi_\theta(a | s) Q^\pi(s, a)]. \end{aligned} \quad (21)$$

The REINFORCE algorithm is a popular policy search algorithm that utilizes Monte Carlo methods and episode samples to adjust policy parameters θ_π . Finally, the optimal policy parameters can be obtained θ_π^* as:

$$\theta_\pi^* = \arg \max_{\theta_\pi} \mathbb{E} \left[\sum_a \pi(a | s; \theta_\pi) r(s, a) \right], \quad (22)$$

The gradient is defined as follows:

$$\nabla_{\theta_\pi} J = \mathbb{E}_\pi \left[\nabla_{\theta_\pi} \ln \pi(a | s; \theta_\pi) r(s, a) \Big|_{s=s^t, a=a^t} \right]. \quad (23)$$

The parameter θ_π is updated using the gradient descent as:

$$\theta_\pi \leftarrow \theta_\pi - \tau \nabla_{\theta_\pi}, \quad (24)$$

where τ is the step size parameter, whose value ranges from $0 \leq \tau \leq 1$. The optimal action a^* can be chosen from the state s with maximum probability as:

$$a^* = \arg \max_{a \in \mathcal{A}} \pi(a | s; \theta_\pi) \quad (25)$$

3.1. The DDPG method

DDPG is an advanced version of the DRL framework that uses actor-critic algorithms to handle continuous action spaces. DDPG consists of two main networks: 1) the actor-network, which maps the state value function $\mu(s; \theta_\mu)$ to a specific action with parameters θ_μ , and 2) the critic network, which is based on the policy search and evaluates the quality of the executed action $Q(s, a)$. Furthermore, the target network and experience replay buffer techniques are employed in the DDPG algorithm to minimize computational overhead and improve learning speed.

The agent explores the environment by observing the current state s^t and executing an action a^t at each time step t . Following the execution of the action, the agent receives feedback from the environment in the form of a scalar reward r^t and moves to the next state $s^{(t+1)}$. Tuples containing the values above ($s^t, a^t, r^t, s^{(t+1)}$) are then stored as an experience replay buffer \mathbb{D} for training the actor and critic networks. The replay buffer \mathbb{D} has a finite memory size, where it updates with new samples and discards the oldest ones. Once the agent has accumulated sufficient samples, it trains the NNs using a mini-batch \mathcal{B} of transitions. Specifically, both the actor network and the critic network are trained using stochastic gradient descent (SGD) over \mathcal{B} samples. The parameters used for the critic network and the target critic network can be denoted as θ_q and θ'_q , respectively. Furthermore, we update the critic network by minimizing the loss function as

$$L = \frac{1}{B} \sum_{i=1}^B (y^i - Q(s^i, a^i; \theta_q))^2, \quad (26)$$

where

$$y^i = r^i(s^i, a^i) + \zeta Q^i(s^{(i+1)}, a^{(i+1)}; \theta'_q) \Big|_{a^{(i+1)} = \mu^i(s^{(i+1)}, \theta'_\mu)}, \quad (27)$$

such that θ'_μ indicates the target actor-network parameter. Similarly, we update the actor-network parameter as follows.

$$\nabla_{\theta_\mu} J = \frac{1}{B} \sum_{i=1}^B \left[\nabla_{a^i} Q(s^i, a^i; \theta_q) \Big|_{a^i = \mu(s^i)} \nabla_{\theta_\mu} \mu(s^i; \theta_\mu) \right] \quad (28)$$

In order to update the target actor-network θ'_q and target critic networks θ'_μ , we employed soft target updates in the following manner:

$$\theta'_q \leftarrow \Psi \theta_q + (1 - \Psi) \theta_q, \quad (29)$$

$$\theta'_\mu \leftarrow \Psi \theta_\mu + (1 - \Psi) \theta_\mu, \quad (30)$$

where $0 \leq \Psi \leq 1$ is the hyperparameter. Given that the DDPG algorithm trains the deterministic policy in an off-policy manner, a noise function $\aleph(0, 1)$ is added as a way of facilitating the exploration and exploitation process as follows [56]:

$$\mu^i(s^i; \theta'_\mu) = \mu(s^i; \theta'_\mu) + \psi \aleph(0, 1) \quad (31)$$

where $0 \leq \psi \leq 1$ is the hyperparameter.

4. The proposed twin DDPG solution

To address the problem (P1), we introduce a policy-based DRL framework that enables the agent to learn the optimal beamforming and trajectory policies without requiring prior knowledge of the system. However, the strong coupling between the outdated CSI and the UAV trajectory poses a challenge for simultaneously optimizing all variables (P1), potentially leading to suboptimal performance and convergence issues. To overcome this, we propose an advanced DRL framework, twin DDPG networks, rather than a conventional single-network approach. In the conventional DDPG framework, which uses a single actor-critic network to jointly optimize (P1), a highly coupled and highly dimensional action space is created. This coupling link causes slow convergence and a suboptimal learning policy, especially in highly dynamic environments with outdated CSI. However, our proposed framework employs a twin structure with two coordinated DDPG agents trained separately. This means two separate DDPG agents are trained to optimize (P1), each interacting with the same environment to maximize a shared objective (e.g., SEE). By decoupling action spaces, the proposed approach improves convergence and adaptability in complex environments where beamforming and trajectory control are strongly coupled. The DDPG1 determines the optimal policy for beamforming selection, encompassing both active UAV and passive RIS beamforming. Simultaneously, the DDPG2 focuses on formulating the policy for UAV trajectories. It is essential to note that both networks ultimately share the same objective function. Before delving into the details of the twin DDPG algorithm,

we first define the basic components of DDPG1 and DDPG2, namely the state space, the action space, and the reward function, for our proposed framework.

4.1. DDPG1

The DDPG1 network is utilized to acquire the optimal beamforming matrix for UAV Φ , and the RIS reflecting beamforming matrix W through interaction with the entire environment. In each episode, a time span T is defined, with each step being a time slot t_k . To maximize cumulative SEE at the k -th time slot for the DDPG1 network, we define the state $s_{(k,1)}$, action $a_{(k,1)}$, and reward $r_{(k,1)}$ as follows:

- a) **State space** ($s_{(k,1)}$): In the k -th time slot, the DDPG1 agent's state ($s_{(k,1)}$) predicts the CSI between the UAV, RIS, and legitimate users/eavesdroppers. We formally represent the state space as

$$s(k, 1) = \left\{ \hat{H}_{U,n}, \hat{H}_{U,m}, \hat{H}_{UR}, \hat{H}_{C,n}, \hat{H}_{C,m} \mid n \in \mathbb{N}, m \in \mathcal{M} \right\}, \quad (32)$$

We convert all the complex-valued matrices into a real-valued vector to feed the DDPG network as $s_{(k,1)}^{(\text{real})} = [\text{Re}(s_{(k,1)}), \text{Im}(s_{(k,1)})] \in \mathbb{R}^{2D_s}$, where $D_s = A(N + M) + MA$ indicates the dimension of the observed entries.

- b) **Action space** ($a_{(k,1)}$): At each time step, the DDPG1 agent takes actions via active and passive beamforming, denoted by Φ and W , respectively, and mathematically expressed as

$$a_{(k,1)} = \{\Phi_k, W_k\} \quad (33)$$

To deal with the complex input network, we separate these values into real and imaginary parts, represented as $a(k, 1) = \text{Re}\{\Phi\} + \text{Im}\{\Phi\}$ and $\theta = \text{Re}\{W\} + \text{Im}\{W\} \in \mathbb{R}^{2(D_a)}$, where $D_a = 2(AN + A)$ represents the dimension of the action space.

- c) **Reward** ($r_{(k,1)}$): This work aims to maximize the sum of SEE defined in Eq. (13). This is achieved when the DDPG1 agent receives optimal action values. The reward function is defined as:

$$r_{k,1} = Y - c_1 p_1 - c_2 p_2 - c_3 p_3, \quad (34)$$

such that $c_i, i \in \{1, 2, 3\}$ are the weighted coefficients that balance the penalties, where p_1, p_2 and p_3 are the penalties when the constraints (14a), (14b), and (14c) are not satisfied, respectively [57].

It is essential to note that the actor and critic networks of DDPG1 operate independently and are updated using policy gradient and Bellman residual methods, respectively, as in standard DDPG updates.

4.2. DDPG2

The DDPG2 network is used to determine the optimal trajectory, \mathbf{q}_2 , for the UAV based on local information. In the k -th time slot for the DDPG2 network, we define the state $s_{(k,2)}$, action $a_{(k,2)}$, and reward $r_{(k,2)}$ as follows:

- a) **State space** ($s_{(k,2)}$): At time slot t_k , the agent of DDPG2 state is responsible for observing the location of the UAV, and can be expressed as

$$s_{k,2} = \mathbf{q}[t_k] = [x_u, y_u, H_u]^T \in \mathbb{R}^3. \quad (35)$$

- b) **Action space** ($a_{(k,2)}$): At each time slot t_k , the DDPG2 network's action generates the 3D Cartesian flying direction $d[t]$, and can be represented as

$$a_{(k,2)} = d[t_k] = [\Delta x_k, \Delta y_k, \Delta z_k]^T, \quad (36)$$

with the next position given by $\mathbf{q}[t_{k+1}] = \mathbf{q}[t_k] + d[t_k]$.

- c) **Reward** ($r_{(k,2)}$): This network aims to find the optimal UAV trajectory that maximizes the total SEE. In the end, we train DDPG1 and DDPG2 using the same reward function to maximize the total SEE as defined in Eq. (13).

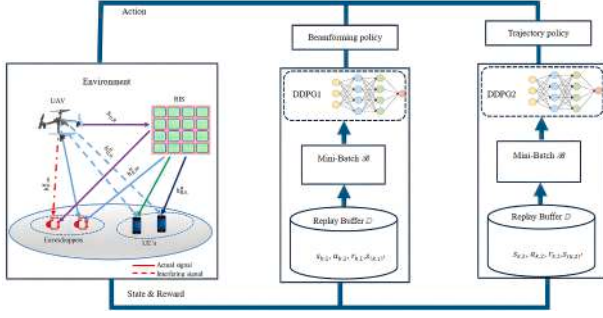


Fig. 2. Proposed twin DDPG framework.

Similar to DDPG1, the actor-critic pair for DDPG2 is trained using the same DDPG procedure. The main difference is that DDPG2's state and action spaces are defined based on the UAV's spatial movement. By learning from environmental feedback and rewards based on SEE, DDPG2 gradually learns to generate an optimal 3D trajectory policy.

Upon completing the training process, the DDPG1 network finds the optimal beamforming strategies (active (UAV) and passive (RIS)). Simultaneously, the optimal trajectory q_k for the UAV is determined by the DDPG2 network. A favorable policy is learned collaboratively by sharing reward functions with these two DDPG networks. As a result, the twin DDPG algorithm efficiently yields the optimal beamforming matrix for the active (UAV) and passive (RIS) systems, and the UAV trajectory to maximize the total SEE. The proposed twin DDPG operates in an online manner, adapting to dynamic environments, unlike the offline mode, which typically loads the policies (beamforming and trajectory) into the UAV in advance. At the same time, the system model closely follows the one introduced in [47], depicted in Fig. 2. The detailed pseudo-code of our proposed algorithm for joint optimization in RIS-aided UAV communication is provided in Algorithm 1.

Algorithm 1 Twin DDPG algorithm for joint optimization in RIS-aided UAV communication.

- 1: **Input:** Discount factor ζ , batch size B , CSI, replay buffer \mathbb{D} , learning rates for DDPG1 and DDPG2
- 2: **Output:** Optimal actions G, θ for DDPG1 and Q for DDPG2 to maximize the average SEE of RIS-aided UAV network
- 3: Initialize DDPG1 network: actor $\mu_1(\cdot)$, critic $Q_1(\cdot)$, target actor $\mu'_1(\cdot)$, target critic $Q'_1(\cdot)$
- 4: Initialize DDPG2 network: actor $\mu_2(\cdot)$, critic $Q_2(\cdot)$, target actor $\mu'_2(\cdot)$, target critic $Q'_2(\cdot)$
- 5: **for** Episode = 1, 2, ..., N^{eps} of DDPG2 **do**
- 6: $t = 0$
- 7: Reset UAV and UE positions
- 8: **for** Step $n = 1, 2, \dots, N_{step}$ **do**
- 9: Initialize DDPG1 and DDPG2 states (CSI and local info)
- 10: Select actions with Gaussian noise g_a and variance \mathcal{P}_a : $a_1 = \mu_1(\cdot) + g_a$, $a_2 = \mu_2(\cdot) + \mathcal{P}_a$
- 11: Execute $a_{k,1}, a_{k,2}$ at $s_{k,1}, s_{k,2}$; receive reward using (32)
- 12: Store transitions: DDPG1 $[s_{k,1}, a_{k,1}, r_{k,1}, s'_{k,1}]$ DDPG2 $[s_{k,2}, a_{k,2}, r_{k,2}, s'_{k,2}]$ into replay buffer \mathbb{D}
- 13: Sample mini-batch \mathcal{B} : $[s^i, a^i, r^i, s^{i+1}]$, $i \in \{1, 2\}$ from \mathbb{D}
- 14: Update critic networks (Eq. (26))
- 15: Update actor policies (Eq. (28))
- 16: Update target networks (Eqs. (29) and (30))
- 17: Update states: $s^i_{k,1} = s^{i+1}_{k,1}$, $s^i_{k,2} = s^{i+1}_{k,2}$
- 18: **end for**
- 19: **end for**

4.3. Algorithm description

At the start of the algorithm, we set up all the required input parameters, such as discount factor ζ , batch size B , CSI, replay buffer \mathbb{D} , and learning rates for both networks (i.e., DDPG1 and DDPG2). These parameters are necessary for guiding the whole training process (line 1). The goal is to find the optimal actions for DDPG1 (beamforming and phase shift (G and θ)) and trajectories (Q) for DDPG2 to maximize the average SEE of the proposed RIS-aided UAV network (line 2).

After setting the parameters, the network parameters are initialized for DDPG1, i.e., actor-network $\mu_1(\cdot)$, critic network $Q_1(\cdot)$, target actor-network $\mu'_1(\cdot)$, and target critic network $Q'_1(\cdot)$ (line 3). Similarly, the network parameters are initialized for DDPG2, i.e., actor-network $\mu'_2(\cdot)$, critic network $Q'_2(\cdot)$, target actor-network $\mu'_2(\cdot)$, and target critic network $Q'_2(\cdot)$ (line 4). The training process is structured in episodes, iterating over a predefined number (500) for DDPG2. For each episode, the time step counter is reset to zero, and the positions of the UAV and UE are also reset to ensure consistency at the start of each episode (line 7). Within each episode, the algorithm runs through multiple steps (N_{step}). During each time step, the initial state values for both DDPG1 and DDPG2 are set based on the current CSI and local information (line 9). Actions are selected for DDPG1 and DDPG2 by adding Gaussian noise to the actor networks' outputs to encourage exploration (line 10).

These actions are then executed from the environment, giving a defined reward value (Eq. 32) (line 11). All these experiences (state, action, reward, next state) for DDPG1 and DDPG2 are then stored into \mathbb{D} for future learning. From this \mathbb{D} , mini-batch \mathcal{B} samples are randomly drawn to update the networks (line 13). Finally, the critic parameters for both DDPG1 and DDPG2 are updated using (Eq. (26)) (line 14), and the actor policies are updated (Eq. (28) (line 15)). The target networks are also updated using soft update rules (Eq. (29) and (Eq. (30) (line 16)). The state values for DDPG1 and DDPG2 are then advanced to the next time step. This process repeats until all episodes and steps are completed, thereby progressively improving the policy and value estimates of the proposed twin DDPG algorithm to optimize the performance of RIS-aided UAV networks.

5. Simulation results

5.1. Network architecture and hyperparameter setup

The proposed twin DDPG network is implemented using Python 3.6 and the PyTorch 1.10.0 framework [58]. We adopted a fully customized 3D simulation framework to simulate a realistic UAV-aided RIS communication environment, where the mobility constraints, channel model, and network entities (UAV, RIS, UEs, and eavesdropper) are dynamically updated at each time slot. For DDPG1, we consider four fully connected (FC) hidden layers with neuron configurations of [512, 256, 128, 64]. Similarly, for DDPG2, we used four FC hidden layers with neuron configurations of [256, 128, 64, 64] [47]. These structures were selected after testing several layer depths (2–5) and node widths (64–512), balancing model capacity and convergence speed. The learning rates for the actor and critic networks in DDPG1 are set to 0.001 and 0.002, respectively. Similarly, for DDPG2, the learning rates for the actor and critic networks are 0.0001 and 0.0002, respectively. These values were chosen through empirical tuning to ensure stable training, with smaller learning rates for trajectory control (DDPG2) to reflect the slower dynamics of UAV mobility compared to fast-varying beamforming in DDPG1. The Adam optimizer is used for both DDPG1 and DDPG2, with a discount factor of $\zeta = 0.995$ to emphasize long-term reward accumulation. The mini-batch size B is 64, and the experience replay buffer \mathbb{D} has a capacity of 100,000 transitions per agent. Each episode consists of a 10-step training process that runs for 500 episodes. This episodic design, with each step corresponding to a time slot ($t_k = 1$ ms), strikes a balance between

Table 3
Simulation parameters.

Description	Parameter	Value
Noise power	σ_n	-90 dBm
Shadow fading component	σ_s	5 dB
Carrier frequency	f_c	28 GHz
Maximum height	H_{\max}	20 m
Path loss when $D = 1$ m	C_0	61 dB
Path loss exponent (UAV-UE)	g_u	3.5 [36]
Path loss exponent (UAV-RIS)	g_{ur}	2.2 [36]
Path loss exponent (RIS-UE)	g_r	2.8 [36]
LoS links	L	3
Time span	T_d	1 s
RIS power	P_R	0.6 dBm
UAV power	P_U	0.9 dBm

rapid policy updates and detailed environmental exploration. The Gaussian exploration noise is added to the actor output during training to promote exploration. The initial noise variance decays linearly across episodes, shifting the focus from exploration to exploitation. The spatial distributions of UAV, RIS, 2 UEs, and eavesdropper are set at (25, 0, 50)m, (0, 50, 12.5)m, (25, 25, 0) m, (4, 47, 0)m, and (47, -4, 0)m, respectively. These fixed coordinates provide a benchmark scenario for training. Additional configurations (e.g., user mobility, variable UAV altitude) are explored in the evaluation phases. Other parameters are adopted from [47] and [59], as detailed in Table 3.² We compare our proposed twin DDPG algorithm with two other schemes and discuss their limitations in the following subsection.

5.2. Benchmarking schemes

The details of the other schemes are described as follows:

Baseline: In the Baseline approach, a single DDPG agent is tasked with determining the optimal policy for the joint optimization of beamforming and UAV trajectories. It is challenging for a single agent to effectively capture the intricate dynamics and dependencies between UAV trajectories and beamforming decisions. It becomes difficult for a single DDPG agent to navigate and learn efficiently in this context, as joint optimization occurs in a complex, high-dimensional space. This complexity arises from the need to manage continuous variables for beamforming (UAV and RIS) and UAV trajectories. These variables need to be finely tuned to respond to a changing environment. Furthermore, the interdependence of these variables complicates the learning process, as changes in UAV trajectories directly influence beamforming decisions and vice versa. The DDPG agent must strike a balance between exploring new strategies and exploiting known optimal actions, a task that is challenging to maintain in such a dynamic scenario.

Myopic: In this case, we randomly select the RIS phase shift, UAV flying direction, and its distance, and optimize the UAV transmit power for each time slot. Without loss of generality, we named this procedure ‘Myopic.’ This approach is only suitable for small network scenarios, where real-time computational constraints and limited environmental variability enable effective local optimization. The Myopic approach can quickly adapt to changing environmental conditions and ensure consistent performance by focusing on immediate, local optimization without considering future states.

² This configuration is based on the controlled simulation environment to validate the proposed framework rather than to represent an exact link-budget scenario. We intentionally set the reflection coefficient of each RIS element to $|\beta_a| = 1$. This provides an ideal passive surface model, enabling the isolation of learning behavior from hardware imperfections [57,59]. Although larger RIS arrays and realistic reflection amplitudes ($|\beta_a| < 1$) can affect the absolute signal strength, they do not affect the core policy-learning or convergence properties of the proposed framework.

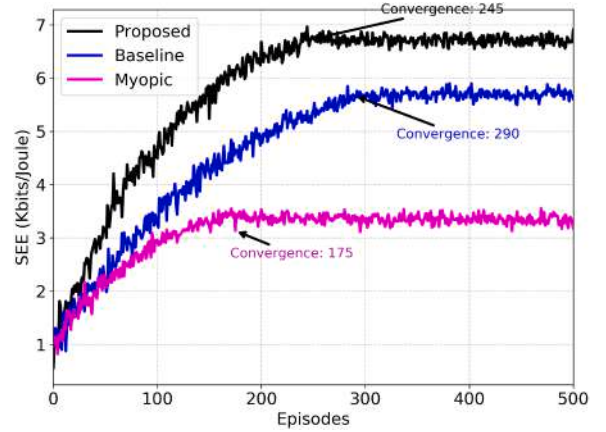


Fig. 3. Convergence analysis.

5.3. Convergence analysis

In Fig. 3, we present the SEE performance across episodes for all three approaches. The proposed twin DDPG method achieves the best results among other approaches. It can be observed from Fig. 3 that the traditional approach (i.e., Myopic) coverage is too early and requires 175 episodes. The reason is that the Myopic approach works well in a small network environment due to its greedy and short-sighted nature. However, as network complexity increases, maintaining performance becomes more difficult. On the other hand, the

Baseline approach requires more episodes than the Proposed scheme. This is because the Baseline approach requires further exploration to learn the optimal behavior from the environment to solve the joint optimization problem. The convergence is fastest for the Myopic approach, followed by the Proposed and Baseline approaches. The proposed twin DDPG method converges around 245 episodes and shows a relatively sudden increase in performance [60]. This sudden improvement is due to the interaction between the twin critic networks and the delayed policy updates. Initially, the agent requires extensive exploration of the environment, and the Gaussian noisy policy updates may lead to slow progress. However, once the critic networks stabilize and the delayed actor starts to benefit from accurate value estimates, the optimal policy can be refined quickly, leading to improved gains in SEE. Such behavior is occasionally observed in twin DDPG when the agent transitions from an exploration-dominated state to an efficient exploitation of learned dynamics. To conclude, although the DRL-based approaches (i.e., Proposed and Baseline) require more episodes to converge, their achieved SEE performance is superior to that of the Myopic approach. This demonstrates the advantage of using DRL with joint optimization in dynamic environments.

5.4. Results and analysis

Fig. 4 represents a 3D visualization illustrating the initial spatial distribution of the RIS-aided UAV network. In Fig. 4, we explicitly convey the distributions for the pivotal entities: the UAV, RIS, 2 UEs, and an eavesdropper. This figure illustrates the UAV’s responsiveness in adapting its position in response to the UEs’ changing locations. Furthermore, this 3D illustration can help to demonstrate how the UEs are trying to capture the radio signal from UAVs. This 3D illustration captures the UAV’s agile, adaptive behavior in response to the UE’s dynamic movement.

In Fig. 5, we have evaluated our proposed twin DDPG approach as well as other approaches for the average secrecy rate versus maximum transmit power P_t . We consider the following parameters for this simulation, namely $M = 1$, $N = 2$, $A = 16$, and $U = 2$. Based on our proposed twin DDPG algorithm, we obtain average secrecy rates comparable to

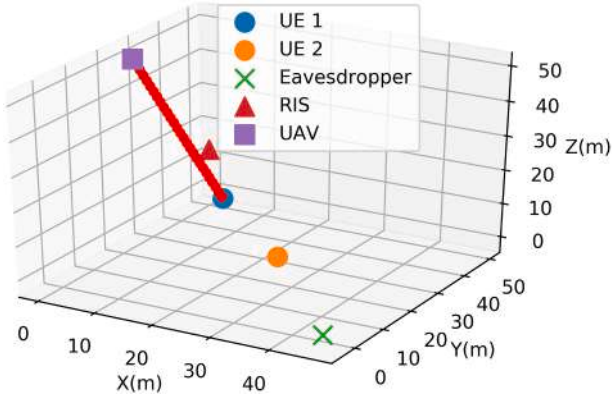
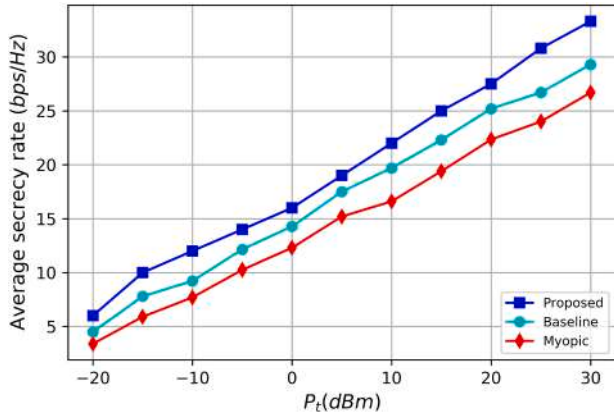


Fig. 4. 3D illustration of RIS-aided UAV network.

Fig. 5. Average secrecy rate versus different transmit power P_t level.

those of the state-of-the-art approaches (Baseline and Myopic), and we show that average secrecy rates increase with transmit power P_t across all scenarios and algorithms. We also find that the twin DDPG algorithm achieves higher average secrecy rates than the Baseline and Myopic algorithms at all power levels. This is because our proposed twin DDPG algorithm can adapt to environmental changes, making it more resilient to eavesdroppers. From Fig. 5, it can be assumed that our proposed twin DDPG emerges as a promising candidate for secure wireless communications, given its exemplary performance in attaining an above-average secrecy rate compared to other approaches.

To better understand the proposed twin DDPG algorithm, we analyze the impact of transmit power P_t , as shown in Fig. 6. Different levels of transmit power, i.e., $P_t = \{-10 \text{ dBm}, 0 \text{ dBm}, 10 \text{ dBm}, 20 \text{ dBm}, 30 \text{ dBm}\}$, are considered with average rewards plotted against time steps. Each episode consists of 10 time steps, and 500 episodes are simulated, resulting in a total of 5000 steps. The average reward at a given step Z is calculated using the following methodology:

$$\text{average_reward}(Z_i) = \frac{\sum_{z=1}^{Z_i} \text{reward}(z_i)}{Z_i}, \quad Z_i = 1, 2, \dots, Z \quad (33)$$

The average rewards generally improve as the steps progress, especially for higher values of transmit power. It can also be noticed from Fig. 6 that some curves (e.g., $P_t = 30 \text{ dBm}$) continue to rise gradually. This reflects a longer convergence period due to the greater complexity of the interaction space available to the agent at higher power levels. On the other hand, lower power levels (e.g., $P_t \geq 10 \text{ dBm}$) lead to earlier convergence but lower performance gain. The relatively smooth appearance of the reward curves is due to averaging over multiple episodes and steps, which suppresses short-term fluctuations. This behavior con-

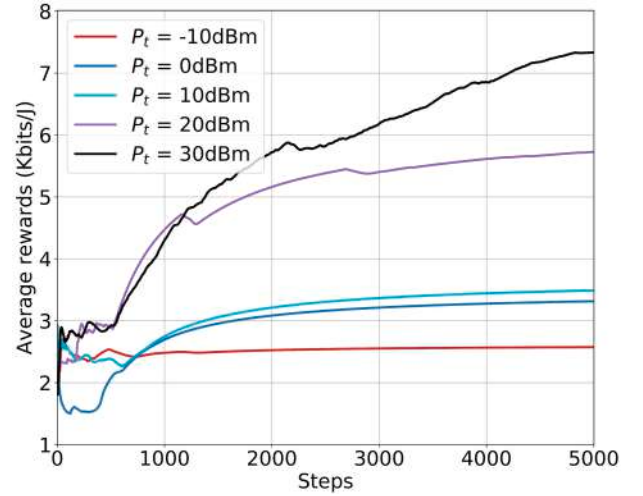
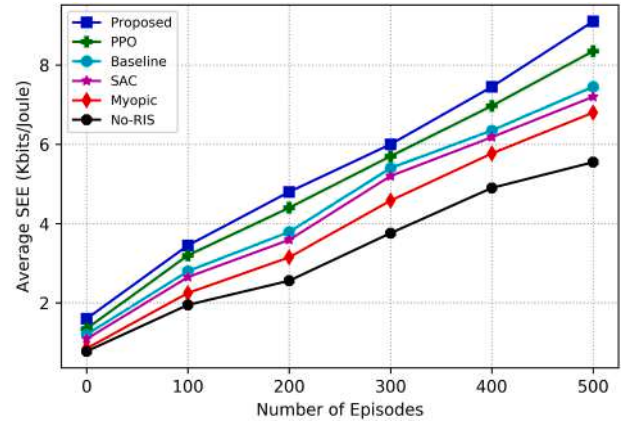
Fig. 6. Average rewards vs. time Steps under different P_t level.

Fig. 7. Average SEE vs number of episodes.

trasts with the sharper variations observed in earlier results (i.e., Fig. 3), where the SEE metric responds more directly to real-time policy shifts. The convergence of the learning process in Fig. 6 remains valid and is more accurately understood as a gradual stabilization of cumulative performance rather than abrupt changes. In this simulation, the twin DDPG algorithm demonstrates how it can adjust the UAV's trajectory and beamforming to achieve optimal performance in different environments.

The average SEE performance over the different numbers of episodes is shown in Fig. 7. As the number of episodes increases, the performance of SEE improves for all the considered approaches. This is because the agent can learn about the environment's behavior over time. However, the proposed twin DDPG algorithm outperforms the other approaches. This is due to the enhanced capability of our proposed twin DDPG algorithm in determining an optimal policy for directing radio signal strength toward UEs, thereby mitigating the impact of eavesdroppers and associated penalties. Proximal policy optimization (PPO) shows outstanding performance compared to other approaches, due to its use of an advanced clipped surrogate objective function [61]. In the case of PPO, the agent uses on-policy updates to maintain a close relationship with the old policy. This helps to remove all the unnecessary samples during the training process. Although PPO performed well compared to other approaches, it struggles with exploration in high-dimensional continuous action spaces, specifically solving the proposed joint optimization problem. In this paper, our main focus is to utilize off-policy training, and the use of PPO is beyond the scope of our work. Likewise,

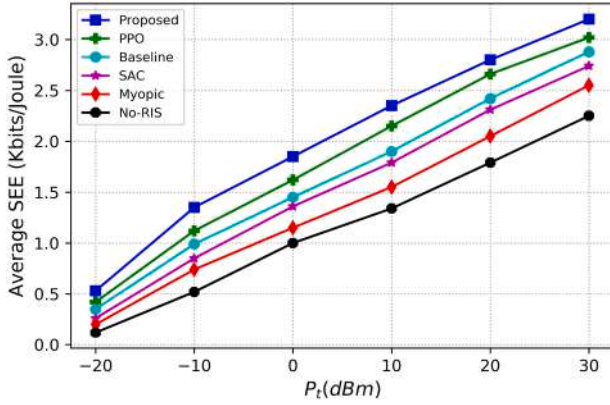


Fig. 8. Average SEE vs. the transmit power P_t budget.

the same performance of the Baseline and soft-actor critic (SAC) is observed in Fig. 7. The reason is that both methods leverage the same off-policy learning and utilize the replay buffer to maximize the SEE performance. In the case of the SAC network, an entropy function is introduced into the environment to maintain stochasticity and enhance robustness in highly dynamic environments; however, it can sometimes compromise performance. Moreover, the SAC network is temperature-sensitive; therefore, its parameters must be carefully selected to balance exploration and exploitation dynamically. If such parameters are not carefully chosen, it will result in suboptimal performance. The figures indicate that the Baseline approach outperforms the SAC approach. The Baseline approach often uses two critics (Q-functions), which helps minimize overestimation of the Q-value and improves the accuracy of policy updates. It improves training stability, resulting in better performance than SAC. The Myopic approach performs worse than the other approaches. This is because the other approaches learn from historical experience and determine the most effective RIS beamforming to direct the radio signal toward UEs. In contrast, the Myopic approach makes decisions solely based on current values without utilizing past data, resulting in lower performance. Additionally, the impact of RIS is highlighted, and its results are compared with those of the No-RIS case. This means that the transmission between the UAV and UE is solely via a direct link. From Fig. 7, it can be seen that the SEE obtained with the No-RIS is the lowest among all the other approaches. Thus, in addition to the direct link, using RIS enables users to achieve the desired QoS more efficiently. In summary, our proposed twin DDPG algorithm demonstrates a significant improvement, increasing the average SEE by up to 48% compared to the No-RIS approach.

In Fig. 8, the performance of average SEE is plotted against the different power levels achieved by different schemes. According to the results, the proposed twin DDPG has the highest SEE at various power levels compared to other methods. This demonstrates the twin DDPG algorithm's improved ability to adapt to environmental variations. The proposed twin DDPG achieves 12–19% more SEE at each power level compared to other approaches. The improvement in SEE suggests that our proposed twin DPPG approach effectively reduces power consumption and ensures secure communication.

The average SEE performance achieved by the increasing number of RIS across all approaches is shown in Fig. 9. Specifically, the SEE achieved by the proposed twin DDPG consistently increases from 2.52 Kbits/Joule to 9.5 Kbits/Joule.

Additionally, the proposed twin DDPG consistently achieves a higher average SEE than the other approaches, regardless of the number of reflecting elements. The Baseline approach achieves lower average SEE values than those obtained with the proposed twin DDPG algorithm, as a single agent explores the complete set of environmental observations. However, the Baseline approach outperforms the Myopic approach because it can select the optimal action behavior.

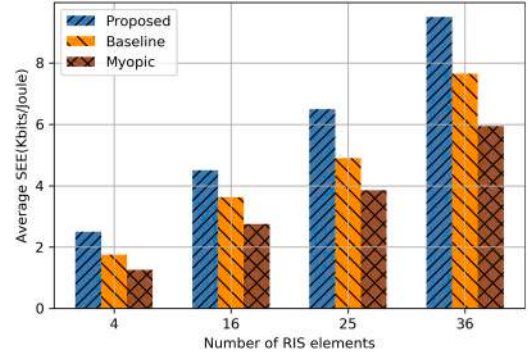


Fig. 9. Average SEE achieved with different numbers of RIS elements.

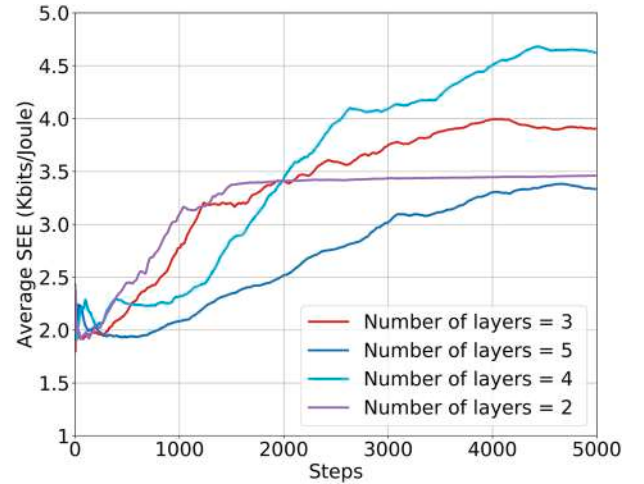


Fig. 10. Effect of hidden layers.

5.5. Effects of layers and learning rate

The performance of SEE, based on different numbers of layers with respect to steps, is shown in Fig. 10. The figure shows that the SEE performance reveals critical insights into the impact of layer depth in the proposed twin DDPG. With a few layers, i.e., 2 and 3, the SEE has not significantly improved due to its limited ability to extract all the features of the RIS-aided UAV network. Alternatively, the proposed four-layer configuration stands out as the best performer in terms of SEE. In this additional layer, the proposed DDPG can extract features, enabling the model to discover more complex patterns in the environment. With a higher level of representational capacity, the model can learn more effectively, thereby improving SEE. Despite the deeper architecture of the five-layer model, it does not show a proportional increase in performance and shows lower SEE. This is because the model overfits, and vanishing gradients impede the network's ability to generalize effectively, thereby degrading SEE performance. Based on observed performance, the selection of the layer depth is crucial for achieving optimal SEE in RL. Therefore, careful consideration and experimentation are essential to maximize efficiency. Similarly, the learning rate (lr) plays a vital role in designing DRL algorithms; therefore, it is important to select it carefully. In this simulation, we use the same learning rate (lr) for the critic and actor neural networks and investigate its impact on the performance of the proposed approach. The effect of different lr (i.e., 0.01, 0.001, 0.0001, 0.00001) for the average rewards against the number of time steps is illustrated in Fig. 11. It can be observed that lr significantly influences average reward performance. When the lr is 0.001, it achieves the best performance, although it takes longer to converge compared with 0.0001 and 0.00001, while the large lr of

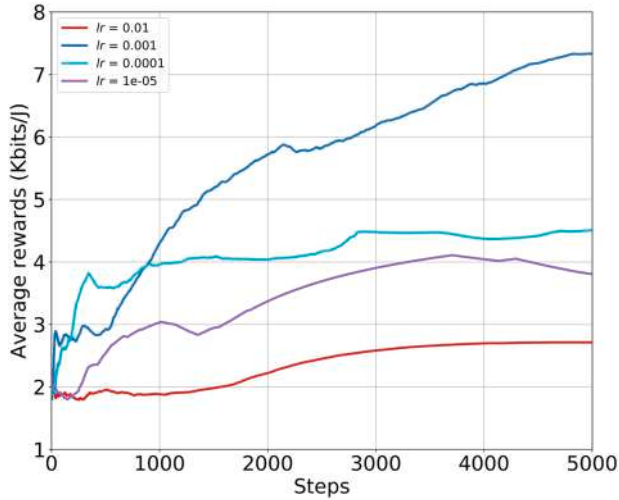


Fig. 11. Effect of learning rates.

0.01 has the worst performance. This is because too large an lr will increase the oscillation, dramatically decreasing average reward performance. To sum up, the lr should be appropriately selected, neither too large nor too small, using methods such as grid search, cross-validation, and adaptive learning rate methods. It is worth noting that the results shown in Figs. 10 and 11 are averaged over five independent training runs, which highlight the model's overall learning performance. Additionally, we applied a simple moving average filter to smooth the results, improving visual clarity. Although this presentation reduces the appearance of fluctuations, the underlying behavior of twin DDPGs (i.e., reward variance during early training) is observed during experimentation. All standard twin DDPG components (delayed policy update, target smoothing, and exploration noise) are used without modification.

5.6. UAV Trajectory

In Fig. 12, the process of the DDPG2 agent identifying the best trajectory for a UAV in a scenario with two UEs is depicted. Initially, the UAVs move towards RISs and away from eavesdroppers. As the distance between the RIS and the UEs increases, the UAV adjusts its path to follow the UEs, moving toward the midpoint between the RIS and the UEs. By increasing the distance between UAVs and RISs, the cascaded link between UAVs, RISs, and UEs is gradually weakening. Therefore, the UAV aims to serve both UEs as equitably as possible by primarily relying on direct links for transmission. As shown in Fig. 12, the proposed approach adapts better to environmental variations than other methods when considering two UEs together. This indicates that the proposed method successfully places the UAV close to the RIS and the UEs. This shows that the proposed method enables the UAV to adapt flexibly to changing environments, identify the optimal trajectory, and improve system performance.

5.7. Computational complexity analysis

A finite number of multi-layer perceptron (MLPs) are introduced in the proposed twin DDPG. Let the MLP layer numbers, the number of neurons in the j -th layer, and the input layer size be denoted by \mathbb{L} , b_j , and b_0 . To update the weights of an MLP in each step, the computational complexity for the training phase is represented as $\mathcal{O}(B(b_0 b_j + \sum_{j=1}^{L-1} n_j n_{j+1}))$. It takes $E \times N$ steps for the twin DDPG to complete its training. E and N represent the total number of episodes and the total number of steps per episode, respectively. For the training mode, the total computational complexity to evaluate and update the single network can be computed as $\mathcal{O}(ENB(b_0 b_j + \sum_{j=1}^{L-1} n_j n_{j+1}))$. In the case of online deployment mode, the computational complexity can be dramatically re-

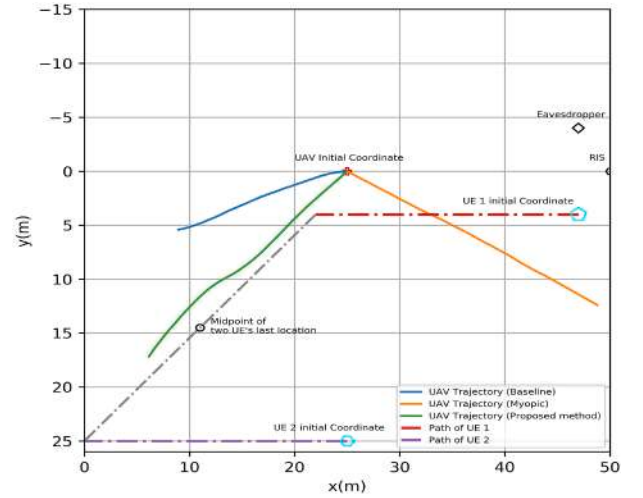


Fig. 12. Trajectory analysis.

duced to $\mathcal{O}(b_0 b_j + \sum_{j=1}^{L-1} n_j n_{j+1})$. This can be achieved by eliminating the training procedure that requires backpropagation and feedforward of B data points. Thus, a favorable level of computational complexity is maintained.

6. Conclusions

In this paper, we addressed the problem of achieving the PLS while maximizing the average SEE in RIS-aided UAV wireless networks. To manipulate the UAV's maneuverability and the RIS's reflecting capabilities, we jointly optimized the UAV's power allocation, the active (UAV) and passive (RIS) beamforming matrices, and the UAV's trajectories. Given the non-convex and intractable nature of this joint optimization problem, we proposed an advanced DRL approach, twin DDPG, for effective problem resolution. Specifically, the joint optimization problem was formulated as an MDP by defining the state space, action space, and reward functions for DDPG1 and DDPG2, with a shared reward function. According to simulation results, our proposed twin DDPG approach improved QoS satisfaction by 18% and achieved a 22% higher SEE compared to existing methods.

In the future, we plan to further explore the RIS-aided UAV network to achieve more robust and secure communication among multiple users, even in the presence of imperfect eavesdropping CSI. In addition, we plan to include interference from the ground BS.

CRedit authorship contribution statement

Amjad Iqbal: Writing – original draft, Software, Methodology, Formal analysis, Data curation, Conceptualization; **Ala'a Al-Habashna:** Writing – review & editing, Supervision, Funding acquisition; **Gabriel Wainer:** Writing – review & editing, Supervision, Resources; **Gary Boudreau:** Supervision.

Data availability

Data will be made available on request.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Amjad Iqbal reports financial support was provided by Carleton University Department of Systems and Computer Engineering. If there are other

authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is funded by Ericsson Canada and the Natural Sciences and Engineering Research Council of Canada (NSERC) .

References

- [1] F. Nait-Abdesselam, A. Alsharoa, M.Y. Selim, D. Qiao, A.E. Kamal, Towards enabling unmanned aerial vehicles as a service for heterogeneous applications, *J. Commun. Netw.* 23 (3) (2021) 212–221.
- [2] Z. Wang, F. Zhou, Y. Wang, Q. Wu, Joint 3D trajectory and resource optimization for a UAV relay-assisted cognitive radio network, *China Commun.* 18 (6) (2021) 184–200.
- [3] Z. Wang, M. Wen, S. Dang, L. Yu, Y. Wang, Trajectory design and resource allocation for UAV energy minimization in a rotary-wing UAV-enabled WPCN, *Alex. Eng. J.* 60 (1) (2021) 1787–1796.
- [4] A. Iqbal, M.L. Tham, Y.J. Wong, A. Al-Habashna, G. Wainer, Y.X. Zhu, T. Dagiuklas, Empowering non-terrestrial networks with artificial intelligence: a survey, *IEEE Access* 11 (2023) 100986–101006.
- [5] T. Zhang, Y. Xu, J. Loo, D. Yang, L. Xiao, Joint computation and communication design for UAV-assisted mobile edge computing in IoT, *IEEE Trans. Ind. Inf.* 16 (8) (2019) 5505–5516.
- [6] Z. Wang, G. Zhang, Q. Wang, K. Wang, K. Yang, Completion time minimization in wireless-powered UAV-assisted data collection system, *IEEE Commun. Lett.* 25 (6) (2021) 1954–1958.
- [7] Z. Lin, M. Guo, C. Han, Y. Sun, R. Ma, K. An, Y. He, N. Al-Dhahir, Wireless endogenous security for SAGINs: achieving ubiquitous access and secure communication in symbiosis, *IEEE Netw.* 39 (6) (2025) 155–163.
- [8] J.D.V. Sanchez, L. Urquiza-Aguiar, M.C.P. Paredes, D.P.M. Osorio, Survey on physical layer security for 5G wireless networks, *Ann. Telecommun.* 76 (3) (2021) 155–174.
- [9] S. Huang, M. Xiao, H.V. Poor, On the physical layer security of millimeter wave NOMA networks, *IEEE Trans. Veh. Technol.* 69 (10) (2020) 11697–11711.
- [10] J. Zhang, F. Wu, Y. Zhu, L. Xiao, D. Yang, Joint trajectory and power optimization for mobile jammer-aided secure UAV relay network, *Veh. Commun.* 30 (2021) 100357.
- [11] H. Han, Y. Huang, H. Hu, Y. Pan, Q. An, J. Si, Mobile jammer enabled secure UAV communication with short packet transmission, *AEU Int. J. Electr. Commun.* 157 (2022) 154434.
- [12] Y.C. Liang, J. Chen, R. Long, Z.Q. He, X. Lin, C. Huang, S. Liu, X.S. Shen, M. Di Renzo, Reconfigurable intelligent surfaces for smart wireless environments: channel estimation, system design and applications in 6G networks, *Sci. China Inf. Sci.* 64 (10) (2021) 200301.
- [13] Q. Wu, R. Zhang, Towards smart and reconfigurable environment: intelligent reflecting surface aided wireless network, *IEEE Commun. Mag.* 58 (1) (2019) 106–112.
- [14] S. Li, B. Duo, X. Yuan, Y.-C. Liang, M. Di Renzo, Reconfigurable intelligent surface assisted UAV communication: joint trajectory design and passive beamforming, *IEEE Wireless Commun. Lett.* 9 (5) (2020) 716–720.
- [15] A. Ranjha, G. Kaddoum, URLLC Facilitated by mobile UAV relay and RIS: a joint design of passive beamforming, blocklength, and UAV positioning, *IEEE Internet Things J.* 8 (6) (2020) 4618–4627.
- [16] S.E. Li, *Reinforcement Learning for Sequential Decision and Optimal Control*, Springer (2023).
- [17] A.A. Khalil, M.Y. Selim, M.A. Rahman, Deep learning-based energy harvesting with intelligent deployment of RIS-assisted UAV-CFmMIMOs, *Comput. Netw.* 229 (2023) 109784.
- [18] A. Iqbal, M.L. Tham, Y.C. Chang, Double deep Q-network-based energy-efficient resource allocation in cloud radio access network, *IEEE Access* 9 (2021) 20440–20449.
- [19] K.K. Nguyen, N.A. Vien, L.D. Nguyen, M.T. Le, L. Hanzo, T.Q. Duong, Real-time energy harvesting aided scheduling in UAV-assisted D2D networks relying on deep reinforcement learning, *IEEE Access* 9 (2020) 3638–3648.
- [20] S.F. Chou, C.Y. Yu, S.I. Sou, Efficient multi-UAV-aided communication service deployment in disaster-resilient wireless networks, in: 2023 IEEE Vehicular Networking Conference (VNC), IEEE, 2023, pp. 1–8.
- [21] K. Bani-Hani, K.F. Hayajneh, A. Jaradat, H. Shakhatreh, Energy-efficient UAV-wireless networks for data collection, *Phys. Commun.* 60 (2023) 102149.
- [22] M.A. Ouamri, G. Barb, D. Singh, A.B.M. Adam, M.S.A. Muthanna, X. Li, Nonlinear energy-harvesting for D2D networks underlying UAV with SWIPT using MADQN, *IEEE Commun. Lett.* 27 (7) (2023) 1804–1808.
- [23] T. Xiao, W. Wei, H.E. Hongliang, et al., Energy-efficient data collection for UAV-assisted IoT: joint trajectory and resource optimization, *Chin. J. Aeronaut.* 35 (9) (2022) 95–105.
- [24] A.U. Haq, S.S. Sefati, S.J. Nawaz, A. Mihovska, M.J. Beliais, Need of UAVs and physical layer security in next-generation non-terrestrial wireless networks: potential challenges and open issues, *IEEE Open J. Veh. Technol.* 6 (2025) 554–595.
- [25] W. Hao, J. Li, G. Sun, M. Zeng, O.A. Dobre, Securing reconfigurable intelligent surface-aided cell-free networks, *IEEE Trans. Inf. Forensics Secur.* 17 (2022) 3720–3733.
- [26] Z. Lin, H. Niu, K. An, Y. Wang, G. Zheng, S. Chatzinos, Y. Hu, Refracting RIS-aided hybrid satellite-terrestrial relay networks: joint beamforming design and optimization, *IEEE Trans. Aerosp. Electron. Syst.* 58 (4) (2022) 3717–3724.
- [27] L. Zhi, N. Hehao, H. Yuanzhi, A. Kang, Z. Xudong, C. Zheng, X. Pei, Self-powered absorptive reconfigurable intelligent surfaces for securing satellite-terrestrial integrated networks, *China Commun.* 21 (9) (2024) 276–291.
- [28] C. Nwufo, O. Simpson, Y. Sun, Reconfigurable intelligent surfaces (RIS) and their role in next-Generation wireless networks: an overview, *Trans. Emerg. Telecommun. Technol.* 36 (6) (2025) e70164.
- [29] L. Ge, P. Dong, H. Zhang, J.-B. Wang, X. You, Joint beamforming and trajectory optimization for intelligent reflecting surfaces-assisted UAV communications, *IEEE Access* 8 (2020) 78702–78712.
- [30] L. Wang, K. Wang, C. Pan, N. Aslam, Joint trajectory and passive beamforming design for intelligent reflecting surface-aided UAV communications: a deep reinforcement learning approach, *IEEE Trans. Mob. Comput.* 22 (11) (2022) 6543–6553.
- [31] K.K. Nguyen, A. Masaracchia, V. Sharma, H.V. Poor, T.Q. Duong, RIS-assisted UAV communications for IoT with wireless power transfer using deep reinforcement learning, *IEEE J. Sel. Top. Signal Process.* 16 (5) (2022) 1086–1096.
- [32] Y. Li, F. Khan, M. Ahmed, A.A. Soofi, W.U. Khan, C.K. Sheemar, M. Asif, Z. Han, RIS-based physical layer security for integrated sensing and communication: a comprehensive survey, *IEEE Internet Things J.* 12 (6) (2025) 32444–32468.
- [33] P. Ji, J. Jia, J. Chen, L. Guo, A. Du, X. Wang, Reinforcement learning based joint trajectory design and resource allocation for RIS-aided UAV multicast networks, *Comput. Netw.* 227 (2023) 109697.
- [34] Y. Wang, Y. Deng, L. Kang, F. Jiang, F. Jiang, Reinforcement learning-based energy efficiency optimization for RIS-assisted UAV hybrid uplink and downlink system, *Comput. Netw.* 245 (2024) 110390.
- [35] H. Mei, K. Yang, Q. Liu, K. Wang, 3D-trajectory and phase-shift design for RIS-assisted UAV systems using deep reinforcement learning, *IEEE Trans. Veh. Technol.* 71 (3) (2022) 3020–3029.
- [36] H. Zhang, M. Huang, H. Zhou, X. Wang, N. Wang, K. Long, Capacity maximization in RIS-UAV networks: a DDQN-based trajectory and phase shift optimization approach, *IEEE Trans. Wireless Commun.* 22 (4) (2022) 2583–2591.
- [37] S. Jiao, X. Xie, Z. Ding, Deep reinforcement learning-based optimization for RIS-based UAV-NOMA downlink networks, *Front. Signal Process.* 2 (2022) 915567.
- [38] K.K. Nguyen, S.R. Khosravirad, D.B. Da Costa, L.D. Nguyen, T.Q. Duong, Reconfigurable intelligent surface-assisted multi-UAV networks: efficient resource allocation with deep reinforcement learning, *IEEE J. Sel. Top. Signal Process.* 16 (3) (2021) 358–368.
- [39] X. Liu, Y. Liu, Y. Chen, Machine learning empowered trajectory and passive beamforming design in UAV-RIS wireless networks, *IEEE J. Sel. Areas Commun.* 39 (7) (2020) 2042–2055.
- [40] Y. Liu, C. Huang, G. Chen, R. Song, S. Song, P. Xiao, Deep learning empowered trajectory and passive beamforming design in UAV-RIS enabled secure cognitive non-terrestrial networks, *IEEE Wireless Commun. Lett.* 13 (1) (2023) 188–192.
- [41] S. Fujimoto, H. Hoof, D. Meger, Addressing function approximation error in actor-critic methods, in: *International Conference on Machine Learning*, PMLR, 2018, pp. 1587–1596.
- [42] Q. Zhang, Y.C. Liang, H.V. Poor, Reconfigurable intelligent surface assisted MIMO symbiotic radio networks, *IEEE Trans. Commun.* 69 (7) (2021) 4832–4846.
- [43] J. Hu, Y.-C. Liang, Y. Pei, Reconfigurable intelligent surface enhanced multi-user MISO symbiotic radio system, *IEEE Trans. Commun.* 69 (4) (2020) 2359–2371.
- [44] A. Al-Hilo, M. Samir, M. Elhattab, C. Assi, S. Sharafeddine, RIS-assisted UAV for timely data collection in IoT networks, *IEEE Syst. J.* 17 (1) (2022) 431–442.
- [45] K.K. Nguyen, S.R. Khosravirad, D.B. Da Costa, L.D. Nguyen, T.Q. Duong, Reconfigurable intelligent surface-assisted multi-UAV networks: efficient resource allocation with deep reinforcement learning, *IEEE J. Sel. Top. Signal Process.* 16 (3) (2021) 358–368.
- [46] Y. Yao, K. Lv, S. Huang, X. Li, W. Xiang, UAV Trajectory and energy efficiency optimization in RIS-assisted multi-user air-to-ground communications networks, *Drones* 7 (4) (2023) 272.
- [47] A. Iqbal, A. Al-Habashna, G. Wainer, F. Bouali, G. Boudreau, K. Wali, Deep reinforcement learning-Based resource allocation for secure RIS-aided UAV communication, in: 2023 IEEE 98th Vehicular Technology Conference (VTC2023-Fall), IEEE, 2023, pp. 1–6.
- [48] M. Wiering, M. Van Otterlo, *Conclusions, future directions and outlook*, in: *Reinforcement Learning: State-of-the-Art*, Springer, 2012, pp. 613–630.
- [49] S. Wu, C.X. Wang, M.M. Alwakeel, X. You, et al., A general 3-D non-stationary 5G wireless channel model, *IEEE Trans. Commun.* 66 (7) (2017) 3065–3078.
- [50] B. Liao, S.C. Chan, Adaptive beamforming for uniform linear arrays with unknown mutual coupling, *IEEE Antennas Wirel. Propag. Lett.* 11 (2012) 464–467.
- [51] G. Zhou, C. Pan, H. Ren, K. Wang, M. ElKashlan, M. Di Renzo, Stochastic learning-based robust beamforming design for RIS-aided millimeter-wave systems in the presence of random blockages, *IEEE Trans. Veh. Technol.* 70 (1) (2021) 1057–1061.
- [52] Y. Yao, J. Miao, T. Zhang, X. Tang, J. Kang, D. Niyato, Towards secrecy energy-efficient RIS aided UAV network: a lyapunov-Guided reinforcement learning approach, in: 2024 IEEE Wireless Communications and Networking Conference (WCNC), IEEE, 2024, pp. 1–6.
- [53] Y. Zhu, G. Zheng, M. Fitch, Secrecy rate analysis of UAV-enabled mmwave networks using matérn hardcore point processes, *IEEE J. Sel. Areas Commun.* 36 (7) (2018) 1397–1409.
- [54] Z. Lin, M. Lin, B. Champagne, W.-P. Zhu, N. Al-Dhahir, Secrecy-energy efficient hybrid beamforming for satellite-terrestrial integrated networks, *IEEE Trans. Commun.* 69 (9) (2021) 6345–6360.
- [55] C. Szepesvári, M.L. Littman, Generalized markov decision processes: dynamic programming and reinforcement-learning algorithms, in: *Proceedings of International Conference of Machine Learning*, 96, 1996.

- [56] T. Tiong, I. Saad, K.T.. Teo, H. bin Lago, Deep reinforcement learning with robust deep deterministic policy gradient, in: 2020 2nd International Conference on Electrical, Control and Instrumentation Engineering (ICECIE), IEEE, 2020, pp. 1–5.
- [57] M.I. Tham, Y.J. Wong, A. Iqbal, N.B. Ramli, Y. Zhu, T. Dagiuklas, Deep reinforcement learning for secrecy energy-efficient uav communication with reconfigurable intelligent surface, in: 2023 IEEE Wireless Communications and Networking Conference (WCNC), IEEE, 2023, pp. 1–6.
- [58] A. Yousefpour, I. Shilov, A. Sablayrolles, D. Testuggine, K. Prasad, M. Malek, J. Nguyen, S. Ghosh, A. Bharadwaj, J. Zhao, et al., Opacus: user-friendly differential privacy library in pytorch, (2021). [arXiv preprint arXiv:2109.12298](https://arxiv.org/abs/2109.12298).
- [59] X. Guo, Y. Chen, Y. Wang, Learning-based robust and secure transmission for reconfigurable intelligent surface aided millimeter wave UAV communications, *IEEE Wireless Commun. Lett.* 10 (8) (2021) 1795–1799.
- [60] S. Dankwa, W. Zheng, Twin-delayed ddpg: a deep reinforcement learning technique to model a continuous movement of an intelligent robot agent, in: Proceedings of the 3rd International Conference on Vision, Image and Signal Processing, 2019, pp. 1–5.
- [61] P. Saikia, S. Pala, K. Singh, S.K. Singh, W.-J. Huang, Proximal policy optimization for RIS-assisted full duplex 6G-V2X communications, *IEEE Trans. Intell. Veh.* 9 (7) (2023) 5134–5149.