Deep Reinforcement Learning for Secrecy Energy-Efficient UAV Communication with Reconfigurable Intelligent Surface

Mau-Luen Tham*
Department of Electrical and Electronic Engineering
Universiti Tunku Abdul Rahman
Kajang, Malaysia
thamml@utar.edu.my

Yi Jie Wong*
Department of Electrical and Electronic Engineering
Universiti Tunku Abdul Rahman
Kajang, Malaysia
yjwong1999@1utar.my

Amjad Iqbal
Department of Electrical and Electronic Engineering
Universiti Tunku Abdul Rahman
Kajang, Malaysia
amjad.iqbal68@hotmail.com

Nordin Bin Ramli
MIMOS Berhad
Kuala Lumpur, Malaysia
nordin.ramli@mimos.my

Yongxu Zhu
Department of Electrical and Electronic Engineering
University of Warwick
Coventry, United Kingdom
Yongxu.Zhu@warwick.ac.uk

Tasos Dagiuklas
Cognitive Systems Research Centre
London South Bank University
London, United Kingdom
tdagiuklas@lsbu.ac.uk

***Abstract*— This paper investigates the physical layer security (PLS) issue in reconfigurable intelligent surface (RIS) aided millimeter-wave rotary-wing unmanned aerial vehicle (UAV) communications under the presence of multiple eavesdroppers and imperfect channel state information (CSI). The goal is to maximize the worst-case secrecy energy efficiency (SEE) of UAV via a joint optimization of flight trajectory, UAV active beamforming and RIS passive beamforming. By interacting with the dynamically changing UAV environment, real-time decision making per time slot is possible via deep reinforcement learning (DRL). To decouple the continuous optimization variables, we introduce a twin-twin-delayed deep deterministic policy gradient (TTD3) to maximize the expected cumulative reward, which is linked to SEE enhancement. Simulation results confirm that the proposed method achieves greater secrecy energy savings than the traditional twin-deep deterministic policy gradient DRL (TDDRL)-based method.**

***Keywords*— Secrecy energy efficiency, deep reinforcement learning, physical layer security, reconfigurable intelligent surface, unmanned aerial vehicle**

I. INTRODUCTION

Deploying unmanned aerial vehicles (UAVs) either as flying base station or relay node are expected to be an integral part of beyond 5G/6G mobile networks [1]. Together with high altitude platform systems (HAPS) and satellites, UAVs can form non-terrestrial networks (NTNs), which has been proposed in 3rd Generation Partnership Project (3GPP) TR 38.821 [2]. Compared to the conventional 2D ground space in terrestrial networks, the 3D positioning of UAV offers massive connectivity and high capacity [3]. Specifically, the high altitude of UAVs opens the possibility of forming more favorable Line-of-Sight (LoS) channels. However, the open characteristic of the wireless links renders it vulnerable to different security threats, such as eavesdropping.

Physical layer security (PLS) can secure wireless communications by improving the legitimate channel capacity while degrading the data decoding performance of the eavesdroppers. To this end, several techniques such as multiple-input multiple-output [4] and cooperative jamming [5] have been proposed. However, none of these works have considered the airborne communication platforms. Studies on UAV-based PLS have been presented in [6-7], aiming to maximize the secrecy rate via proper UAV trajectory design. Although these works have extracted substantial gains in secrecy rate, they are unable to alter the wireless propagation, which demands a new paradigm to synthetically customize the propagation environment of signal waveform. One solution is to adopt Reconfigurable Intelligent Surface (RIS), which is a programming surface structure that enables spectral- and energy-efficient wireless communications [8]. The combination of RIS and UAV has led to significant gains in coverage [9] and throughput [10]. However, their objective function is not linked with the secrecy rate.

Energy efficiency is another critical performance metric in UAV networks due to constraints of battery life. In [11], the authors have proposed an energy minimization scheme for UAV-aided IoT system via a cooperative control of UAV trajectory, communication scheduling, and transmit power allocation. However, the presence of eavesdroppers has not been considered. Recognizing the importance of both security and energy issues, the scheme in [12] has maximized the worst-case secrecy energy efficiency (SEE) by jointly optimizing transmission power allocation and trajectory construction. However, the optimal solution is attained only if the perfect state information is known. Such method is referred to as traditional model-based algorithm. In practice, finding the proper distributions to model the entire UAV-aided network is non-trivial owing to the movement of the UAV and the end users, as well as the communication and processing latency.

In the absence of complete state information, Deep Reinforcement Learning (DRL), which merges Deep Neural Networks (DNN) with Reinforcement Learning (RL), has emerged as a right candidate to tackle the real-time dynamic optimization problem [13]. Specifically, DRL can automatically extract features from various types of raw data with complex correlations via continuous interaction with the mobile environments. DRL can be classified into two main categories, first, discrete-level control such as

* The first two authors contributed equally to this work.

Deep Q Network (DQN) [14] and second, continuous-level control such as Deep Deterministic Policy Gradient (DDPG) [15] and Twin-Delayed Deep Deterministic Policy Gradient (TD3) [16]. In the context of UAV communications, the latter has been in the spotlight since UAV hovering control and transmit power allocation exist in continuous domain. In [17], the authors have proposed a twin-DDPG DRL (TDDRL) framework which maximizes the sum secrecy rate (SSR) of all authorized users by controlling the flight trajectory, RIS passive beamforming and UAV active beamforming.

Inspired by the work in [17], this paper considers a RIS aided millimeter-wave rotary-wing UAV communications under the presence of multiple eavesdroppers and imperfect channel state information (CSI). Instead of solving SSR maximization problem, our goal is to maximize the worst-case SEE via joint optimization of UAV flight trajectory, active and passive beamforming. Given the overestimation bias of DDPG, we introduce twin-TD3 (TTD3) to maximize the expected cumulative reward, which is linked to SEE enhancement. Simulation results validate that the proposed method achieves greater energy savings than the traditional TDDRL-based method.

The remainder of this paper is organized as follows. Section II presents the system model of the RIS-aided mmWave UAV communication system and formulates the SEE optimization problem. In Section III, the TTD3 based algorithm is developed and explained in detail. In Section IV, the performance of the proposed solution is compared with respect to TDDRL. Concluding remarks are provided in Section V.

II. System Model and Problem Statement

A. System Model

Fig. 1 depicts a DRL-empowered RIS-aided mmWave UAV communication framework. By creating virtual LOS, the RIS can improve the link security from UAV to K users, accompanied by P wiretappers. All users and wiretappers are assumed to have single antenna. The RIS possesses an uniform planar array (UPA) with $M = m^2$ passive reflective elements, while the UAV has an A-element uniform linear array (ULA). We divide the entire flight duration T into N time instants evenly. Let δ_t represents each individual time slot, such that $t = n\delta_t$ for $n \in N$. The RIS is statically located at coordinate $\boldsymbol{w}_R = (x_R, y_R, z_R)^T$. On the other hand, the coordinates of users and eavesdropper at time instant n are represented by $\boldsymbol{w}_i = (x_i[n], y_i[n], z_i[n])^T, \forall i \in K \cup P$. Finally, let $\boldsymbol{q}[n]$ be the coordinates of UAV at time slot n. The UAV speed can then be expressed as:

$$\|\mathbf{v}[n]\| = \sqrt{\|\boldsymbol{q}[n] - \boldsymbol{q}[n-1]\|^2}/\delta_t \tag{1}$$

Let $\boldsymbol{q}[0]$ be the initial coordinates of the UAV, B be the UAV's moving boundary, and D_{max} represents the maximum UAV maneuvering distance at time slot n. We define the UAV mobility constraints as follows:

$$\boldsymbol{q}[0] \equiv (0,0,H_U) \tag{2a}$$
$$|x[n]|, |y[n]| \leq B \tag{2b}$$
$$\sqrt{\|\boldsymbol{q}[n] - \boldsymbol{q}[n-1]\|^2} \leq D_{max} \tag{2c}$$

With flying speed $\|\mathbf{v}[n]\|$, we can derive the UAV's propulsion energy consumption for a rotary-wing UAV as follows [11]:

$$E_p[n] \approx \delta_t \left(P_0 + \frac{3P_0\|\mathbf{v}[n]\|^2}{U_{tip}^2} + \frac{1}{2} d_0 \rho s A_r \|\mathbf{v}[n]\|^3 \right) + \delta_t P_i \left(\sqrt{1 + \frac{\|\mathbf{v}[n]\|^4}{4v_0^4}} - \frac{\|\mathbf{v}[n]\|^2}{2v_0^2} \right)^{\frac{1}{2}} \tag{3}$$

Constants P_i and P_0 represent the induced power and blade profile power in hovering status, respectively. U_{tip}^2 is the rotor blade's tip speed, and v_0 denotes the average rotor induced velocity in hover. Moreover, d_0 is the fuselage drag ratio, and s is the rotor solidity. Lastly, ρ is the air density and A_r is the rotor disc area.

Our channel model follows the idea of 3D SV channel model presented in [17-18]. Let the channel gain from UAV to RIS, from UAV to p-th wiretapper, from UAV to k-th users, from RIS to k-th users and from RIS to p-th wiretappers are represented as $\boldsymbol{h}_{U\,R} \in \mathbb{C}^{M\times A}$, $\boldsymbol{H}_{U,k} \in \mathbb{C}^{A\times 1}$, $\boldsymbol{h}_{U,p} \in \mathbb{C}^{A\times 1}$, $\boldsymbol{h}_{R,k} \in \mathbb{C}^{M\times 1}$, and $\boldsymbol{h}_{R,p} \in \mathbb{C}^{M\times 1}$, respectively. The channel from UAV to users or the eavesdroppers is denoted by $\boldsymbol{H}_{C,i} = \text{diag}(\boldsymbol{h}_{R,i}^H)\boldsymbol{h}_{UR}, \forall i \in K \cup P$. Similarly, the RIS passive beamforming matrix is defined as $\boldsymbol{\theta} = \text{diag}(\beta_1 e^{j\theta_1}, \beta_2 e^{j\theta_2}, \ldots \beta_A e^{j\theta M})$, where $\beta_m \in [0,1]$, $m = \{1,2,\ldots,M\}$, $\theta_m \in [0,2\pi)$ shows the amplitude reflection and phase shift of the m-th RIS reflection elements, respectively. To maximize the power of reflecting signal and simplify the problem, we set the value of $\beta_m = 1$. The channel coefficients from UAV to all recipients can be combined as shown in Equation (4).

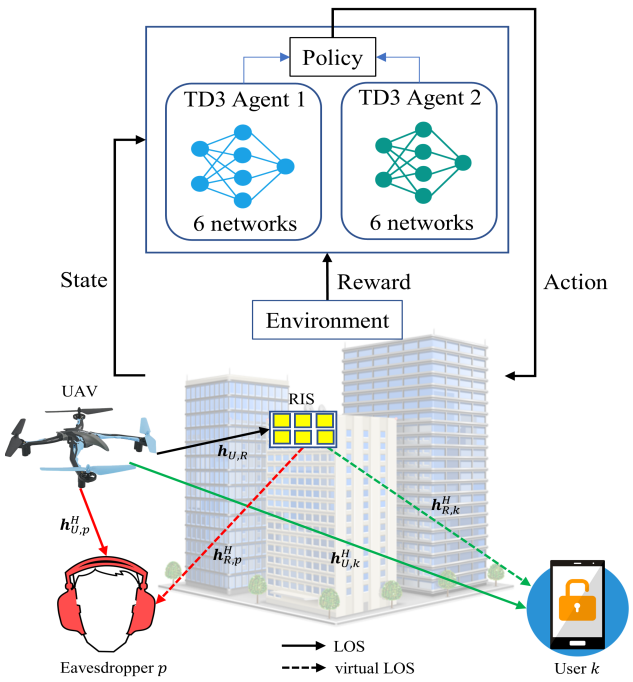

Fig. 1. DRL-empowered RIS-aided mmWave UAV communications.

$$\boldsymbol{H}_C = \left\{\boldsymbol{h}_{U,i}^H + \psi^H \boldsymbol{H}_{C,i} | \forall_i \in \kappa \cup \rho\right\} \quad (4)$$

where $\boldsymbol{\psi}$ is the RIS passive beamforming matrix that can be extracted as $\boldsymbol{\psi} = \text{vec}(\boldsymbol{\theta})$. Finally, we can express the signal received at the i-th user or wiretapper from the UAV as:

$$y_i = \left(\boldsymbol{h}_{U,i}^H + \psi^H \boldsymbol{H}_{C,i}\right)\boldsymbol{G}\boldsymbol{s} + n_i, \forall_i \in K \cup P \quad (5)$$

where $\boldsymbol{G} \in \mathbb{C}^{A \times K}$ and $\boldsymbol{s} \in \mathbb{C}^{K \times 1}$ with $E[|s_n|^2] = 1$ indicates the beamforming matrix and transmitted symbol at the UAV, respectively. n_i denotes the background noise and define as $n_i \sim \mathcal{N}(0, \sigma_n), \forall_i \in K \cup P$. Denote g_k as the k-th column of $\boldsymbol{G}$. Thus, the total attainable data rate at the k-th user can be formulated as:

$$R_k^u[n] = \log_2\left(1 + \frac{\left(\left|\boldsymbol{h}_{U.n}^H + \boldsymbol{\psi}^H \boldsymbol{H}_{C,k}\right| g_k\right)^2}{\sum_{k' \in \kappa \setminus k}\left(\left|\boldsymbol{h}_{U,k}^H + \boldsymbol{\psi}^H \boldsymbol{H}_{C,k}\right| g_{k'}\right)^2 + n_k^2}\right) \quad (6)$$

Similarly, the feasible p-eavesdropper-to- k-user rate can be represented as:

$$R_{p,k}^e[n] = \log_2\left(1 + \frac{\left(\left|\boldsymbol{h}_{U,p}^H + \boldsymbol{\psi}^H \boldsymbol{H}_{C,p}\right| g_k\right)^2}{\sum_{k' \in \kappa \setminus k}\left(\left|\boldsymbol{h}_{U,p}^H + \boldsymbol{\psi}^H \boldsymbol{H}_{C,p}\right| g_{k'}\right)^2 + n_p^2}\right) \quad (7)$$

According to [11], we can write the particular UAV-to-k-user secrecy rate as follows:

$$R_k^{\text{sec}}[n] = \left[R_k^u[n] - \max_{\forall_p} R_{p,k}^e[n]\right]^+ \quad (8)$$

where $[j]^+ = \max(0, j)$. We suggest interested readers to refer to [17] for more details of the modelling. Here, the SEE of the system model as:

$$\text{SEE}[n] = \frac{\sum_{k=1}^K R_k^{sec}[n]}{E_p[n]} = \frac{\text{SSR}[n]}{E_p[n]} \quad (9)$$

B. Problem Statement

This paper aims to maximize the long-term SEE by tuning the UAV's trajectory $\boldsymbol{Q}$, active $\boldsymbol{G}$ and passive beamforming $\boldsymbol{\theta}$. The problem can be formulated as:

$$\max_{\boldsymbol{Q},\boldsymbol{G},\boldsymbol{\theta}} \sum_{n=1}^N \text{SEE}[n] \quad (10a)$$

$$s.t. \quad (10b)$$

$$\Pr\left\{R_k^{sec}[n] \geq R_k^{sec,th}\right\} \geq 1 - \rho_k, \forall k, \forall n \quad (10c)$$

$$\theta_m \in [0, 2\pi), m = \{1, 2, \dots, M\} \quad (10d)$$

$$\text{Tr}(\boldsymbol{G}\boldsymbol{G}^H) \leq P_{max} \quad (10e)$$

Constraint (10c) is to make sure that each user k can perfectly extract its message at a data rate of $R_k^{sec,th}$, with a probability of at least $1 - \rho_k$. Owing to the non-convex characteristics of (10b), (10c), (10d) and the time-varying CSI, Problem (10) is intractable and time-correlated. Thus, we employ DRL to learn the best policy by interacting with the dynamically changing UAV environment.

III. Proposed Solution

UAV path $\boldsymbol{Q}$ is tightly connected with a great amount of CSI. Hence, optimizing all the variables simultaneously is challenging. Inspired by [17], we construct two DRL agents to decouple these variables, instead of exploiting only one DRL agent as in most DRL-based solutions. We have used TD3 [16] as DRL agents. TD3 is an improvement over the DDPG algorithm [15], in which both are used for continuous control problems. TD3 uses two critics instead of one, to reduce the over-estimation of action values. It also introduces a "delayed" update for the actor network to improve the stability of the learning process. TD3 has been shown to be more efficient and robust to hyperparameter choices than DDPG and has been used in several recent works on DRL.

At this end, we propose a Twin TD3 (TTD3) algorithm to solve problem (10). The first TD3 agent takes CSI as its input, to generate the optimal UAV active beamforming matrix $\boldsymbol{G}$, and the RIS passive beamforming matrix $\boldsymbol{\theta}$. On the other hand, the second TD3 agent is employed to obtain the best UAV trajectory $\boldsymbol{Q}$ based on the local information $\boldsymbol{W}$. The design of the TTD3 algorithm is described as follows.

A. Active and Passive Beamforming

The first TD3 agent takes the CSI as its input, to generate the optimal $\boldsymbol{G}$ and $\boldsymbol{\theta}$. The problem can be formulated as a Markov Decision Process (MDP) with the following state, action and reward.

1) ***State*** s_n^1: The state for the first TD3 agent is the predicted comprehensive CSI from the UAV to all users and eavesdropper at each time slot n.

2) ***Action*** a_n^1: The TD3 agent will generate $\boldsymbol{G}$ *and* $\boldsymbol{\theta}$ as the action. To address the complex-valued input, $\boldsymbol{G} = Re\{\boldsymbol{G}\} + Im\{\boldsymbol{G}\}$ and $\boldsymbol{\theta} = Re\{\boldsymbol{\theta}\} + Im\{\boldsymbol{\theta}\}$ are divided into real and imaginary part.

3) ***Reward*** r_n^1: Our goal is to optimize SEE as defined in (9). However, in practice, we find that directly employing (9) as reward function may lead to poor convergence and performance. This is because the DRL agent may focus on improving SEE by simply making the denominator ($E_p[n]$) as small as possible, without improving the numerator ($R_k^{sec}[n]$). This is especially true during the earlier phase of training, when $R_k^{sec}[n] \to 0$ or $R_k^{sec}[n] \leq 0$. Furthermore, the impact of $E_p[n]$ on the convergence depends on the expected value $\mathbb{E}\left(E_p[n]\right)$, which varies based on the system model setting. Without loss of generality, we reformulate the reward function as shown below:

$$r_n^1 = \tanh\left(\sum_{k=1}^K R_k^{sec}[n] - c_1 p_m - c_2 p_r - c_3 p_g - c_4 p_e\right) \quad (11)$$

Similar to [17], the p_m, p_r and p_g in (11) are the penalties when the constraints (10b), (10c), and (10e) are not fulfilled, respectively. On the other hand, p_e is the penalty for high energy consumption. We formulate p_e as shown in (12).

$$p_e = \begin{cases} 0, & \sum_{k=1}^{K} R_k^{sec}[n] < 0 \\ 0.1\left(\sum_{k=1}^{K} R_k^{sec}[n]\right)\widetilde{E_p}[n], & \sum_{k=1}^{K} R_k^{sec}[n] \geq 0 \end{cases} \quad (12)$$

For generalization purposes, we normalize $E_p[n]$ to range $0 \leq E_p[n] \leq 1$. Let normalized $E_p[n]$ denoted by $\widetilde{E_p}[n] = \frac{E_p[n]-E_{p,min}}{E_{p,max}-E_{p,min}}$. Energy consumption is the lowest when $\widetilde{E_p}[n] = 0$, and highest when $\widetilde{E_p}[n] = 1$. Instead of directly setting $p_e = \widetilde{E_p}[n]$, we set p_e to grow proportionally to $0.1(\sum_{k=1}^{K} R_k^{sec}[n])$. This can prevent the agent to blindly reduce p_e without optimizing $\sum_{k=1}^{K} R_k^{sec}[n]$. Penalty p_e is small when $\sum_{k=1}^{K} R_k^{sec}[n] \to 0$, so that the agent can first focus on improving $\sum_{k=1}^{K} R_k^{sec}[n]$. When $\sum_{k=1}^{K} R_k^{sec}[n] \leq 0$, we set $p_e = 0$. By doing so, the TD3 agent can learn sufficiently to optimize $\boldsymbol{G}$ and $\boldsymbol{\theta}$ (along with $\boldsymbol{Q}$ by the second TD3 agent), before we start penalizing the agent(s) for energy consumption.

B. UAV Trajectory

Another TD3 agent is employed to compute the optimal UAV trajectory $\boldsymbol{Q}$ by taking the local information $\boldsymbol{W}$ as input. Similarly, the problem can be designed as an MDP using the following state, action and reward.

1) ***State*** s_n^2: Since we have decoupled the UAV trajectory from CSI, the second TD3 agent only takes the local information $\boldsymbol{W}$ as input.

2) ***Action*** a_n^2: At each time slot n, the TD3 agent generates the flying direction $\boldsymbol{d}[n]$ in the 3D Cartesian dimension. Based on $\boldsymbol{d}[n]$, we can acquire the next coordinate of the UAV as $\boldsymbol{q}[n] = \boldsymbol{q}[n-1] + \boldsymbol{d}[n]$. After N time slots, the complete UAV trajectory can be expressed as $\boldsymbol{Q} = \{\boldsymbol{q}[0], \boldsymbol{q}[1], \dots, \boldsymbol{q}[n]\}$.

3) ***Reward*** r_n^2: This agent shares the same reward function as defined in (5) since both agents have the same optimization objective, which is to maximimze SEE.

IV. Simulation Results and Discussion

As mentioned earlier, a TTD3 algorithm comprising of two TD3 agents have been employed for the optimization problem (4). Both TD3 agents comprise of one actor and two critic networks, in which all networks are based on multi-layer perceptron (MLP). The hyperparameter for the TTD3 algorithm is shown in Table I.

We set the starting positions of UAV and the two users as (0 m, 25 m, 50 m), (4 m, 47 m, 0 m) and (25 m, 25 m, 0 m), respectively. The RIS and eavesdropper are fixed at (0 m, 50 m, 12.5 m) and (47 m, -4 m, 0 m), respectively. Moreover, we model the two users to move uniformly in one direction as depicted in Fig. 2. The rest of the parameters are configured as $D_{max} = 0.25$ m, $\delta_t = 0.1$ s, $T_d = 1$ s, $f_c = 28$ GHz, $C_0 = 61$ dB, $P_{max} = 30$ dBm, $\sigma_n = -114$ dBm, $\sigma_s = 3$ dB, $L = 3$, $\alpha_{ur} = 2.2$, $\alpha_u = 3.5$, $\alpha_r = 2.8$, $M = 16$, $A = 4$, $K = 2$, $P = 1$, $\Phi_l^{AoA} \in \{30,45,60\}$, $\Phi_l^{AoD} \in \{5,10,15,25\}$, $\Lambda_l^{AoD} \in \{1,3,5\}$, $\Lambda_l^{AoA} \in \{5,10,15\}$ (degrees), as defined in [17]. For energy-related parameters of the UAV, we set $P_0 = 580.65$ W, $P_i = 790.6715$ W, $U_{tip} = 200$ m/s, $d_0 = 0.3$, $\rho = 1.225$ kg/m^3, $s = 0.05$, $A_r = 0.79$ m^2 as in [11], [19].

TABLE I. Hyperparameters for TTD3

Hyperparameters	Values
TD3 Agent 1 size (actor and critics)	27 x 800 x 600 x 515 x 256 x 20
TD3 Agent 2 size (actor and critics)	3 x 400 x 300 x 256 x 128 x 2
Actor learning rate	0.0001
Critic learning rate	0.001
Number of episodes, N_{ep}	300
Time step, N	100
Batch size, N_b	64
Replay memory size	30000
Update actor interval	2

We compare our results with three benchmarks. **Benchmark 1**: We have implemented TDDRL algorithms from [17] in the system model to optimize $\boldsymbol{Q}$, $\boldsymbol{G}$ and $\boldsymbol{\theta}$ without energy constraint (7). **Benchmark 2**: We replace the TTDRL algorithms in Benchmark 1 with our proposed TTD3. **Benchmark 3**: Similar to Benchmark 1, we implement TDDRL algorithm but with energy penalty (7). Lastly, we compare our **proposed method**, which employs TTD3 with energy penalty. Each algorithm is evaluated in terms of average SSR and SEE in one complete N time steps. We run the simulation from [17] for 5 testing episodes, and averaged out the performance of each benchmark as tabulated in Table II.

A. Trajectory analysis

For each algorithm, we plot one of the trajectories from the five testing episodes, in Fig. 2. In general, all trajectories move away from the eavesdropper. Interestingly, UAVs in benchmark 2 and the proposed method (both using the proposed TTD3) move more efficiently toward the midpoint of the two user's last location. This allows the UAV to serve both users as fairly as possible, proving the superiority of TTD3 in maximizing the SSR (as shown in Table II). On the other hand, the UAVs in the other two benchmarks did not end up somewhere near the midpoint. However, they still achieve a considerably high average SSR, showing that Agent 1 successfully learns the optimal active and passive beamforming matrix.

Furthermore, it is noticed that the UAV moves faster at each time slot n, when energy penalty p_e is employed (i.e. Benchmark 2 and proposed method). Both algorithms have

TABLE II. Benchmarking of SSR and SEE

Algorithms	Average SSR (bits/s/Hz)	Total Energy Consumption (kJ)	Average SEE (bits/s/Hz/kJ)
Benchmark 1	5.03	12.4	40.8
Benchmark 2	6.05	12.7	48.2
Benchmark 3	4.68	11.2	39.4
Proposed method	5.39	11.2	48.4

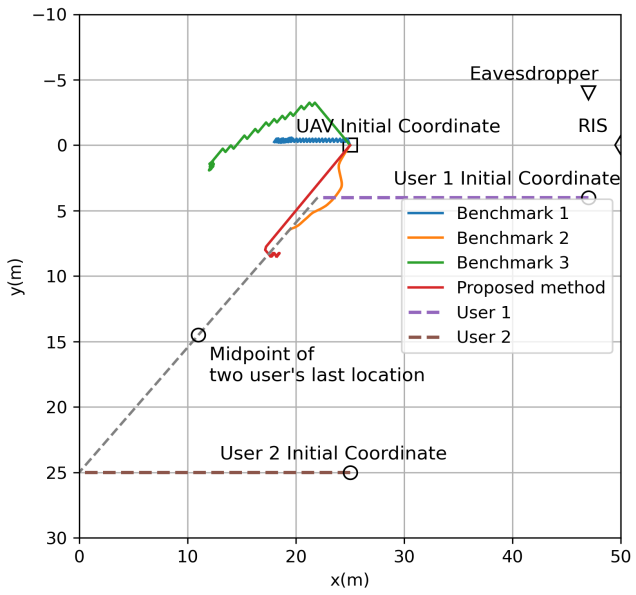

Fig. 2: The optimized trajectory by each algorithm.

a longer travelling distance compared to Benchmark 1 and 3 (without p_e). This is because for the given mobility constraint (2c), $E_p[n]$ becomes smaller when the speed of UAV increases. This is reasonable, since it takes more energy to maintain the UAV in a static position, compared to a moving UAV (where the $E_p[n]$ can be reduced by the forward momentum). In short, the $E_p[n]$ can be reduced by moving faster per time slot n, which conserves energy.

B. *Average SSR and Average SEE*

In terms of average SSR, TTD3 has outperformed TDDRL in both with and without energy penalty settings. This is because TTD3 employs TD3 agent, which is superior to the DDPG used in TDDRL. The proposed method achieved the second-highest SSR, falling behind only Benchmark 2, which also uses the TTD3 algorithm. In addition, the proposed method achieves the highest SEE and the lowest energy consumption. The reason is that the energy penalty p_e penalizes the TTD3 agents for high energy consumption. The proposed method strikes a balance between optimizing SSR and minimizing energy consumption. Generally, methods that employ energy penalty p_e (Benchmark 3 and the proposed method) consume lesser energy as compared to those without energy penalty p_e. Figures 3(a) and 3(b) show the average SSR and average SEE of each algorithm after training for 300 episodes, respectively. Based on Figure 3, it is clear that the top 2 algorithms for average SSR and average SEE are the two variations of our proposed TTD3 (with and without energy penalty).

C. *Computation complexity*

TTD3 is composed of a finite number of MLPs. Let L, n_0 and n_i denote the MLP layer numbers, the input layer size, and the number of neurons in i-th layer, respectively. During training phase, the computational complexity for an MLP to update its weights in each step can be as $O(N_b(n_0n_1 + \sum_{i=1}^{L-1} n_i n_{i+1}))$ [20]. In total, it takes N_{ep} x N steps for the TTD3 to complete its training. Hence, the total training computational complexity can be computed as $O(N_{ep}NN_b(n_0n_1 + \sum_{i=1}^{L-1} n_i n_{i+1}))$. On the other hand, the

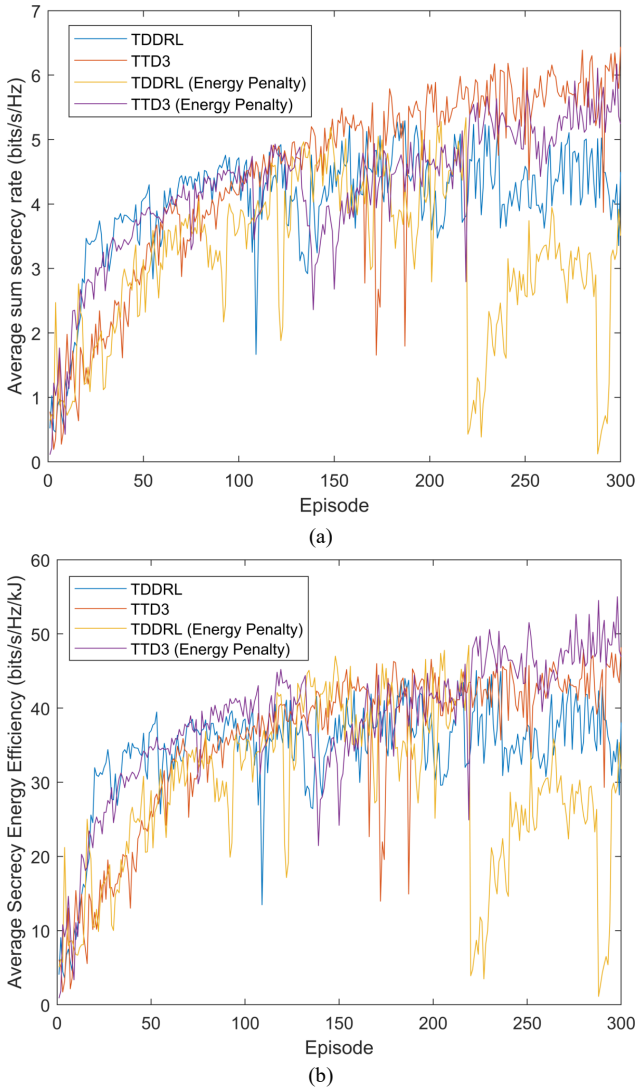

Fig. 3: Performance metric versus episodes. (a) Average SSR. (b) Average SEE.

computational complexity in online deployment mode is dramatically reduced to $O(n_0n_1 + \sum_{i=1}^{L-1} n_i n_{i+1})$. This is done by cutting off the training procedure that requires feedforward and backpropagation of N_b data points. Hence, the computational complexity can be retained at a favourable level.

V. Conclusion

In this paper, we have proposed a TTD3-based framework which maximizes the SEE of RIS aided millimeter-wave rotary-wing UAV communication system. Computational complexity of the proposed scheme has been analyzed. We showed that the reward function plays a crucial role in guiding the trajectory, active and passive beamforming towards green UAV communications. Apart from the SEE maximization, we also demonstrated the superiority of TTD3 over TDDRL in maximizing the SSR.

Acknowledgment

This work was supported by the British Council under UK-ASEAN Institutional Links Early Career Researchers Scheme with project number 913030644. Also, we express our gratitude to the authors in [17] for providing their simulation codes in https://github.com/Brook1711/WCL-pulish-code.

REFERENCES

[1] J. Lee and V. Friderikos, "Interference-aware path planning optimization for multiple UAVs in beyond 5G networks," in *Journal of Communications and Networks*, vol. 24, no. 2, pp. 125-138, April 2022, doi: 10.23919/JCN.2022.000006.

[2] I. C. Msadaa, S. Zairi and A. Dhraief, "Non-Terrestrial Networks in a Nutshell," in *IEEE Internet of Things Magazine*, vol. 5, no. 2, pp. 168-174, June 2022, doi: 10.1109/IOTM.007.2100121.

[3] Mohsan SAH, Khan MA, Alsharif MH, Uthansakul P, Solyman AAA. Intelligent Reflecting Surfaces Assisted UAV Communications for Massive Networks: Current Trends, Challenges, and Research Directions. Sensors (Basel). 2022 Jul 14;22(14):5278. doi: 10.3390/s22145278. PMID: 35890955; PMCID: PMC9322292.

[4] X. Chen, D. W. K. Ng, W. H. Gerstacker, and H.-H. Chen, "A survey on multiple-antenna techniques for physical layer security," IEEE Communications Surveys & Tutorials, vol. 19, no. 2, pp. 1027–1053, 2016.

[5] H. Hui, A. L. Swindlehurst, G. Li, and J. Liang, "Secure relay and jammer selection for physical layer security," IEEE Signal Processing Letters, vol. 22, no. 8, pp. 1147–1151, 2015.

[6] A. Li and W. Zhang, "Mobile jammer-aided secure UAV communications via trajectory design and power control," China Communications, vol. 15, no. 8, pp. 141–151, 2018.

[7] Y. Cai, F. Cui, Q. Shi, M. Zhao, and G. Y. Li, "Dual-UAV enabled secure communications: joint trajectory design and user scheduling," IEEE Journal on Selected Areas in Communications, vol. 36, no. 9, pp. 1972–1985, 2018.

[8] X. Yuan, Y.-J. A. Zhang, Y. Shi, W. Yan, and H. Liu, "Reconfigurable-intelligent-surface empowered wireless communications: Challenges and opportunities," IEEE wireless communications, vol. 28, no. 2, pp. 136–143, Feb, 2021.

[9] L. Yang, F. Meng, J. Zhang, M. O. Hasna and M. D. Renzo, "On the Performance of RIS-Assisted Dual-Hop UAV Communication Systems," in IEEE Transactions on Vehicular Technology, vol. 69, no. 9, pp. 10385-10390, Sept. 2020, doi: 10.1109/TVT.2020.3004598.

[10] X. Liu, Y. Yu, F. Li and T. S. Durrani, "Throughput Maximization for RIS-UAV Relaying Communications," in *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 10, pp. 19569-19574, Oct. 2022, doi: 10.1109/TITS.2022.3161698.

[11] C. Zhan and H. Lai, "Energy Minimization in Internet-of-Things System Based on Rotary-Wing UAV," IEEE Wireless Communications Letters, vol. 8, no. 5, pp. 1341–1344, Oct. 2019, doi: 10.1109/LWC.2019.2916549.

[12] An Li, Huohuo Han, Chuanxin Yu, "Secrecy Energy-Efficient UAV Communication via Trajectory Design and Power Control", Wireless Communications and Mobile Computing, vol. 2021, Article ID 9969311, 10 pages, 2021. https://doi.org/10.1155/2021/9969311

[13] H. Li, H. Gao, T. Lv and Y. Lu, "Deep Q-Learning Based Dynamic Resource Allocation for Self-Powered Ultra-Dense Networks," *2018 IEEE International Conference on Communications Workshops (ICC Workshops)*, Kansas City, MO, USA, 2018, pp. 1-6, doi: 10.1109/ICCW.2018.8403505.

[14] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, ``Human-level control through deep reinforcement learning," Nature, vol. 518, no. 7540, pp. 529533, Feb. 2015.

[15] T. P. Lillicrap et al., "Continuous control with deep reinforcement learning," 4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings, Sep. 2015, doi: 10.48550/arxiv.1509.02971.

[16] S. Fujimoto, H. van Hoof, and D. Meger, "Addressing Function Approximation Error in Actor-Critic Methods," 35th International Conference on Machine Learning, ICML 2018, vol. 4, pp. 2587–2601, Feb. 2018, doi: 10.48550/arxiv.1802.09477.

[17] X. Guo, Y. Chen, and Y. Wang, "Learning-Based Robust and Secure Transmission for Reconfigurable Intelligent Surface Aided Millimeter Wave UAV Communications," IEEE Wireless Communications Letters, vol. 10, no. 8, pp. 1795–1799, Aug. 2021, doi: 10.1109/LWC.2021.3081464.

[18] G. Zhou, C. Pan, H. Ren, K. Wang, M. Elkashlan, and M. D. Renzo, "Stochastic learning-based robust beamforming design for RIS-aided millimeter-wave systems in the presence of random blockages," IEEE Trans. Veh. Technol., vol. 70, no. 1, pp. 1057–1061, Jan. 2021.

[19] Y. Zeng, J. Xu, and R. Zhang, "Rotary-Wing UAV Enabled Wireless Network: Trajectory Design and Resource Allocation," 2018 IEEE Global Communications Conference, GLOBECOM 2018 - Proceedings, 2018, doi: 10.1109/GLOCOM.2018.8647595.

[20] H. Yang, Z. Xiong, J. Zhao, D. Niyato, L. Xiao, and Q. Wu, "Deep Reinforcement Learning Based Intelligent Reflecting Surface for Secure Wireless Communications," IEEE Trans Wirel Commun, vol. 20, no. 1, pp. 375–388, Feb. 2020, doi: 10.1109/twc.2020.3024860.