Learning-based Robust and Secure Transmission for Reconfigurable Intelligent Surface Aided Millimeter Wave UAV Communications

Xufeng Guo, Yuanbin Chen and Ying Wang, *Member, IEEE*

***Abstract*—In this letter, we study the secure transmission in the millimeter-wave (mmWave) unmanned aerial vehicle (UAV) communications assisted by the reconfigurable intelligent surface (RIS) under imperfect channel state information (CSI). Specifically, the active beamforming of the UAV, the coefficients of the RIS elements and the UAV trajectory are jointly designed to maximize the sum secrecy rate of all legitimate users in the presence of multiple eavesdroppers. However, the formulated problem is intractable mainly due to the complex constraints result from the intricate coupled variables and the time-related issue caused by outdated CSI. To tackle these difficulties, by leveraging the deep deterministic policy gradient (DDPG) framework, a novel and effective twin-DDPG deep reinforcement learning (TDDRL) algorithm is proposed. Simulation results demonstrate the effectiveness and robustness of the proposed algorithm, and the RIS can significantly improve the sum secrecy rate.**

***Index Terms*—Deep reinforcement learnig, reconfigurable intelligent surface, physical layer security, unmanned aerial vehicle, millimeter-wave communications.**

I. Introduction

Millimeter-wave (mmWave) communications with multi-gigahertz bandwidth availability boost much higher capacity and transmission rate than conventional sub-6GHz communications. Unmanned aerial vehicles (UAVs), which are featured by their high mobility and flexible deployment, are promising candidates to compensate most of the deficiencies of mmWave signals, preserve its advantages, and provide more opportunities [1]. However, the mmWave signals transmitted by UAVs are prone to deteriorate due to their high sensitivity to the presence of spatial blockages, especially in the complex propagation environment (such as in urban areas), which thus degrades the reliability of the communication links. As a result, a more powerful and novel solution is essential.

Recently, the reconfigurable intelligent surface (RIS) composed of a large number of passive reflecting elements has become a revolutionary technology to achieve high spectral and energy efficiency in a cost-effective way [2]. By appropriately tuning the reflection coefficients, the reflected signal can be enhanced or weakened at different receivers. Since the RIS has significant passive beamforming gain, it can be incorporated into the mmWave UAV communication system to generate virtual line-of-sight (LoS) links, thereby achieving directional signal enhancement, expanding coverage area and reducing radio frequency (RF) chains [3]. In addition, broadcasting and superposition, as two basic properties of the wireless communication, make wireless transmissions inherently susceptible to security breaches [4]. Hence, secure transmission is also a pivotal issue in UAV communication systems which attracted extensive interest of researches [5], [6].

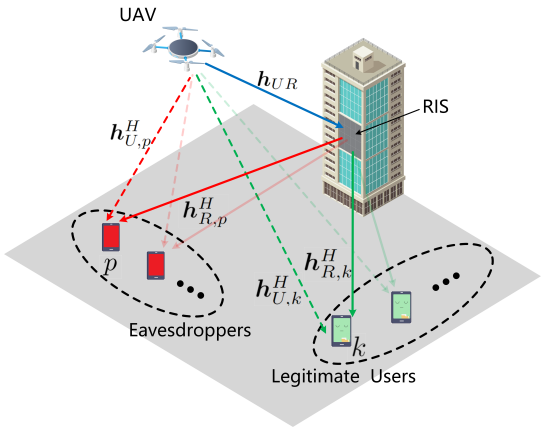

Fig. 1. RIS-aided Milimeter Wave UAV Communications.

A crucial issue in the RIS-aided mmWave UAV communication system is to jointly design the active and passive beamforming and the UAV trajectory. However, unlike the general RIS-aided wireless communication model, the UAV mobility-induced variation of the angles of arrival/departure (AoAs/AoDs) render the channel gains of all links (including direct links and cascaded links) to be optimization variables that need to be well-designed. Such variables are intricately coupled together with the active and passive beamforming matrix, which greatly increases the difficulty of the design. To circumvent this issue, several researches have been investigated in [5]–[9], some of which, in particular, leverage alternating optimization (AO) method to tackle the coupled variables [5]–[8]. In [9], a deep reinforcement learning approach is utilized to jointly optimize the passive beamforming and the UAV trajectory, in which, however, the active beamforming is not considered in this approach. Nevertheless, all these existing works [5], [7]–[9] are based on the assumption of the perfect channel state information (CSI), which weakens the versatility and practicality of the model. Furthermore, the UAV mobility-induced outdated CSI should also be taken into account.

The deep reinforcement learning (DRL) is an efficient

Corresponding author: Ying Wang.

X. Guo, Y. Chen, and Y. Wang are with the State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing, China 100876 (e-mail:brook1711@bupt.edu.cn; chen_yuanbin@163.com; wangying@bupt.edu.cn).

approach to jointly design the active and passive beamforming, and the UAV trajectory, due to its good generalization, low complexity, and high accuracy. The motivation of utilizing DRL approach is mainly for two reasons: i) it is fairly difficult to tackle the intricately couple variables in the RIS-aided UAV system, and even the widely applicable AO method cannot solve this problem well, especially for the multi-user system. ii) the UAV mobility-induced CSI is easily outdate, and there is in general no effective method to solve such a time-related issue.

In this letter, motivated by these considerations, we investigate the secure transmission problem in the RIS-aided mmWave UAV communication system. The active beamforming at the UAV, the passive beamforming at the RIS and the UAV trajectory are jointly designed by explicitly taking into account imperfect CSI. To enhance the robustness of the considered system, we study a secrecy rate maximization problem subject to the secrecy outage probability resulted from the statistical CSI error model. To solve this problem, a novel twin-deep deterministic policy gradien (TDDRL) deep reinforcement learnin algorithm is proposed. More specifically, the first network is utilized to provide policy for the active and passive beamforming while the UAV trajectory is coordinated by the second network. The obtained simulation results demonstrate the effectiveness and the performance benefits of the proposed TDDRL algorithm.

II. System Model and Problem Formulation

A. System Model

In this letter, we consider an RIS-aided mmWave UAV communication system where an RIS is exploited to assist the secure downlinks from the UAV to K single-antenna legitimate users in the presence of P single-antenna eavesdroppers. Specifically, the UAV is equipped with an A-element uniform linear array (ULA), and the RIS has a uniform planar array (UPA) with $M{=}m^2$ passive reflecting elements (m is an integer). The set of the legitimate users and the eavesdroppers are denoted by $\mathcal{K}{=}\{1,2,...,K\}$, $\mathcal{P}{=}\{1,2,...,P\}$, respectively. As shown in Fig.1, all entities are placed in the three dimensional (3D) Cartesian coordinate system. The RIS is fixed at $\boldsymbol{w}_R{=}(x_R,y_R,z_R)^T$. We assume that the UAV flies at a fixed altitude in a finite time span which is divided into N time slots, i.e., $T{=}N\delta_n$, where δ_n is the time slot. Then the coordinate of the UAV and the coordinates of the legitimate users and eavesdroppers at the n-th time slot are denoted by $\boldsymbol{q}[n]{=}(x_U[n],y_U[n],H_U)^T$ and $\boldsymbol{w}_i{=}(x_i[n],y_i[n],z_i[n])^T,\forall i\in\mathcal{K}\cup\mathcal{P}$, respectively. The location information at the n-th time slot is defined as $\boldsymbol{W}{\triangleq}\{\boldsymbol{q}[n]\}\cup\{\boldsymbol{w}_i[n]|\forall i\in\mathcal{K}\cup\mathcal{P}\}$. The UAV is subject to the following mobility constraints:

$$\|\boldsymbol{q}[n+1]-\boldsymbol{q}[n]\|^2\le D^2, n=1,\dots,N-1, \tag{1a}$$

$$|x[n]|,|y[n]|\le B, n=1,\dots,N, \tag{1b}$$

$$\boldsymbol{q}[0]\equiv(0,0,H_U), n=1,\dots,N-1. \tag{1c}$$

Let $\boldsymbol{h}_{U,k}{\in}\mathbb{C}^{A\times1}$, $\boldsymbol{h}_{U,p}{\in}\mathbb{C}^{A\times1}$, $\boldsymbol{h}_{R,k}{\in}\mathbb{C}^{M\times1}$, $\boldsymbol{h}_{R,p}{\in}\mathbb{C}^{M\times1}$, $\boldsymbol{H}_{UR}{\in}\mathbb{C}^{M\times A}$ be the channel gains from the UAV to k-th user, the UAV to the p-th eavesdropper, the RIS to the k-th user, the RIS to the p-th eavesdropper, the UAV to the RIS links, respectively. All the channels are modeled according to 3D SalehValenzuela channel model [10] which has been widely used to characterize the mmWave channels:

$$\boldsymbol{h}_{U,i}=\sqrt{\frac{1}{L_{UK}}}\sum_{l=1}^{L_{UK}}g_{i,l}^{u}\mathbf{a}_L\left(\theta_{i,l}^{AoD}\right),\forall i\in\mathcal{K}\cup\mathcal{P}, \tag{2a}$$

$$\boldsymbol{h}_{R,i}=\sqrt{\frac{1}{L_{RK}}}\sum_{l=1}^{L_{RK}}g_{i,l}^{r}\mathbf{a}_P\left(\theta_{i,l}^{AoD},\phi_{i,l}^{AoD}\right),\forall i\in\mathcal{K}\cup\mathcal{P}, \tag{2b}$$

$$\boldsymbol{h}_{UR}=\sqrt{\frac{1}{L_{RK}}}\sum_{l=1}^{L_{RK}}g_{l}^{ur}\mathbf{a}_P\left(\theta_{l}^{AoA},\phi_{l}^{AoA}\right)\mathbf{a}_L\left(\theta_{l}^{AoD}\right)^H. \tag{2c}$$

In (2), the large-scale fading coefficients defined by $g\in\{g_{i,l}^{u},g_{i,l}^{r},g_{l}^{ur}\}$ follow a complex Gaussian distribution as $\mathcal{CN}(0,10^{\frac{PL}{10}})$, where PL(dB) $=-C_0{-}10\alpha\log_{10}(\mathrm{D})-\mathrm{PL_s}$, C_0 is the path loss at a reference distance of one meter, D (meters) is the link distance, α denotes the path-loss exponent, and $\mathrm{PL}_s{\sim}\mathcal{CN}(0,\sigma_s^2)$ is the shadow fading component. The steering vector of the ULA is denoted by $\mathbf{a}_L(\theta){=}\left[1,e^{j\frac{2\pi}{\lambda_c}d\sin(\theta)},\dots,e^{j\frac{2\pi}{\lambda_c}d(A-1)\sin(\theta)}\right]^H$, where θ stands for the azimuth AoD $\theta_{i,l}^{AoD}$ and θ_{l}^{AoD}, d is the antenna inter-spacing, and λ_c is the carrier wavelength. The steering vector of the UPA is denoted by $\mathbf{a}_P(\theta,\phi){=}\left[1,\dots,e^{j\frac{2\pi}{\lambda_c}d(p\sin(\theta)\sin(\phi)+q\cos(\theta)\sin(\phi))},\dots\right]^H$, where $0\le p,q\le m-1$, and $\theta(\phi)$ is the azimuth(elevation) AoD $\theta_{i,l}^{AoD}(\phi_{i,l}^{AoD})$ and the AoA $\theta_{l}^{AoA}(\phi_{l}^{AoA})$.

The cascaded channel from the UAV to the i-th user or the eavesdropper can be written as $\boldsymbol{H}_{C,i}{=}\mathrm{diag}(\boldsymbol{h}_{R,i}^H)\boldsymbol{h}_{UR},\forall i\in\mathcal{K}\cup\mathcal{P}$. Let $\boldsymbol{H}_C\triangleq\{\boldsymbol{h}_{U,i}^H+\boldsymbol{\Psi}^H\boldsymbol{H}_{C,i}|\forall i\in\mathcal{K}\cup\mathcal{P}\}$ denote the combined channel gains between the UAV and all receivers. The passive beamforming matrix [11] of the RIS is denoted by $\boldsymbol{\Theta}{=}\mathrm{diag}\left(\beta_1e^{j\theta_1},\beta_2e^{j\theta_2},...,\beta_Me^{j\theta_M}\right)$, where $\theta_m\in[0,2\pi)$, $\beta_m\in[0,1],m{=}\{1,2,\cdots,M\}$ represent the phase shift and amplitude reflection coefficients of the m-th RIS reflection element, respectively. The amplitude reflection coefficients are set to one, i.e., $\beta_m{=}1$ to simplify the problem and maximize the power of the reflecting signal [12]. Let $\boldsymbol{\Psi}{=}\mathrm{vec}(\boldsymbol{\Theta})$ denote the vectorized passive beamforming vector. Thus, the received signal at the i-th user or eavesdropper from the UAV can be formulated as

$$y_i=(\boldsymbol{h}_{U,i}^H+\boldsymbol{\Psi}^H\boldsymbol{H}_{C,i})\boldsymbol{G}\boldsymbol{s}+n_i,\forall i\in\mathcal{K}\cup\mathcal{P}, \tag{3}$$

where $\boldsymbol{s}\in\mathbb{C}^{K\times1}$ with $E[|s_k|^2]{=}1$ and $\boldsymbol{G}\in\mathbb{C}^{A\times K}$ represents the transmitted symbol and the beamforming matrix at the UAV, and it is assumed that $n_i\sim\mathcal{N}(0,\sigma_n),\forall i\in\mathcal{K}\cup\mathcal{P}$. Let $\boldsymbol{g}_k$ be the k-th column of the beamforming matrix $\boldsymbol{G}$. Then, the achievable rate of the k-th user is given by

$$R_k^u=\log_2\left(1+\frac{|(\boldsymbol{h}_{U,k}^H+\Psi^H\boldsymbol{H}_{C,k})\boldsymbol{g}_k|^2}{\sum_{k'\in\mathcal{K}\setminus k}|\boldsymbol{h}_{U,k}^H+\Psi^H\boldsymbol{H}_{C,k})\boldsymbol{g}_{k'}|^2+n_k^2}\right). \tag{4}$$

If the p-th eavesdropper aims to eavesdrop the signal of the k-th user, its achievable rate can be denoted by

$$R_{p,k}^{e} = \log_2\left(1+\frac{|(\boldsymbol{h}_{U,p}^H+\Psi^H\boldsymbol{H}_{C,p})\boldsymbol{g}_k|^2}{\sum_{k'\in\mathcal{K}\setminus k}|\boldsymbol{h}_{U,p}^H+\Psi^H\boldsymbol{H}_{C,p})\boldsymbol{g}_{k'}|^2+n_p^2}\right). \quad (5)$$

The achievable individual secrecy rate from the UAV to the k-th user [12] can be expressed by

$$R_k^{\text{sec}} = \left[R_k^{\text{u}} - \max_{\forall p} R_{p,k}^{\text{e}}\right]^+, \quad (6)$$

where $[z]^+ = \max(0, z)$.

It is worth noting that the outdated CSI will lead to substantial performance loss in practical systems. According to [13], the outdated CSI can be expressed as the statistical CSI error model. Furthermore, let T_d be the delay between the outdated and the real-time CSI. The relation between the outdated channel vector $\boldsymbol{h}(t)$ and the real-time channel vector $\boldsymbol{h}(t+T_d)$ can be expressed as [14]

$$\boldsymbol{h}\left(t+T_{\text{d}}\right) = \rho\boldsymbol{h}(t) + \sqrt{1-\rho^2}\boldsymbol{e}, \quad (7)$$

where $\boldsymbol{e}$ is independent identically distributed (i.i.d) with $\boldsymbol{h}\left(t+T_{\text{d}}\right)$ and $\boldsymbol{h}(t)$, ρ is the autocorrelation function of the channel gain $\boldsymbol{h}(t)$, given by the zeroth-order Bessel function of the first kind as $\rho = J_0\left(2\pi f_D T_{\text{d}}\right)$, where f_D is the Doppler spread which is expressed as $f_D = vf_c/c$, where v, f_c, c represent the velocity of the transceivers, the carrier frequency and the speed of light, respectively.

Then, the actual channel coefficients can be rewritten as

$$\begin{aligned}
\boldsymbol{h}_{\text{U},i} &= \rho\tilde{\boldsymbol{h}}_{\text{U},i} + \Delta\boldsymbol{h}_{\text{U},i}, \forall i \in \mathcal{K}\cup\mathcal{P},\\
\boldsymbol{h}_{\text{R},i} &= \rho\tilde{\boldsymbol{h}}_{\text{R},i} + \Delta\boldsymbol{h}_{\text{R},i}, \forall i \in \mathcal{K}\cup\mathcal{P},\\
\boldsymbol{h}_{\text{UR}} &= \rho\tilde{\boldsymbol{h}}_{\text{UR}} + \Delta\boldsymbol{h}_{\text{UR}}.
\end{aligned} \quad (8)$$

Note that the system only has the access to the estimated CSI $\tilde{\boldsymbol{h}} \in \{\tilde{\boldsymbol{h}}_{\text{U},i}, \tilde{\boldsymbol{h}}_{\text{R},i}, \tilde{\boldsymbol{h}}_{\text{UR}}\}$, which are outdated, to generate active and passive beamforming and UAV trajectory. The actual CSI $\boldsymbol{h} \in \{\boldsymbol{h}_{\text{U},i}, \boldsymbol{h}_{\text{R},i}, \boldsymbol{h}_{\text{UR}}\}$ given by (8) is employed to calculate achievable secrecy rate expressed in (4)-(6).

B. Problem Formulation

In this letter, we aim to maximize the sum secrecy rate $\sum_{k=1}^{K} R_k^{\text{sec}}$ by jointly optimizing the UAV's trajectory $\boldsymbol{Q} \triangleq \{\boldsymbol{q}[n], n \in \mathcal{N}\}$ and the active (passive) beamforming matrix $\boldsymbol{G}(\boldsymbol{\Theta})$, which yields the following problem

$$\max_{\boldsymbol{Q},\boldsymbol{G},\boldsymbol{\Theta}} \quad \sum_{k\in\mathcal{K}} R_k^{\text{sec}} \quad (9a)$$

$$s.t. \quad (1), \quad (9b)$$

$$\Pr\left\{R_k^{\text{sec}} \geq R_k^{\text{sec,th}}\right\} \geq 1-\rho_k, \forall k \in \mathcal{K}, \quad (9c)$$

$$\text{Tr}\left(\boldsymbol{G}\boldsymbol{G}^H\right) \leq P_{\max}, \quad (9d)$$

$$\theta_m \in [0, 2\pi), m = \{1, 2, \cdots, M\}, \quad (9e)$$

where the secrecy rate outage constraint in (9c) guarantees that the probability that each legitimate user can successfully decode its message at a data rate of $R_k^{\text{sec,th}}$ is no less than $1-\rho_k$. Problem (9) is intractable mainly for the non-convex constraints in (9b), (9c), and (9e), the secrecy outage constraint without close-form expression, and the time varying CSI described in (7). There is in general no standard method to solve such a probability-constrained non-convex optimization. Next, a DRL-based approach is proposed to overcome these challenges effectively.

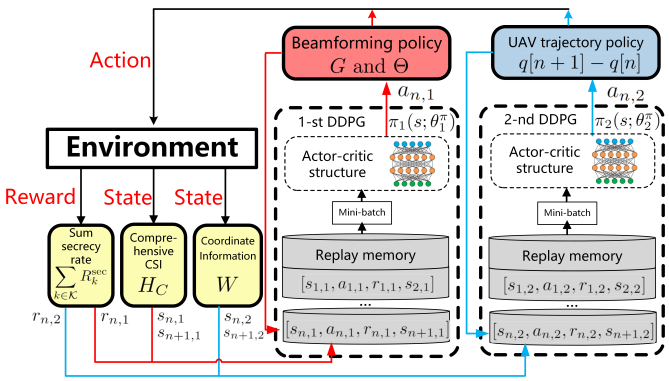

Fig. 2. Structure of the proposed TDDPG algorithm.

III. DRL-based Solution

To solve problem (9), we propose a TDDRL algorithm, which is able to allow the agent to learn the policies of the beaforming and the trajectory without any prior knowledge of the system [15]. Since the UAV trajectory $\boldsymbol{Q}$ is highly coupled with large amounts of CSI, it is difficult to optimize all the variables simultaneously, which may incur a poor convergence and performance. To tackle this issue, two DDPG networks are constructed to decouple these variables instead of employing a single agent in the conventional DRL-based network. In particular, as illustrated in Fig.2, the first network takes the CSI, i.e., $\boldsymbol{H}_C$ as the state to obtain the optimal $\boldsymbol{G}$ and $\boldsymbol{\Theta}$, the second network takes the coordinates of the UAV and all legitimate users and eavesdroppers as the state, i.e., $\boldsymbol{W}$ to obtain the UAV movement which consists of the flying distance $\mu[n]$ and direction $\psi[n]$ at the n-th time slot. Both networks share the same reward function. The design of the DRL-based solution are elabarated as follows.

A. Active and Passive Beamforming

Inspired by the work of [12], the first DDPG network is employed to learn the optimal policy in terms of the beamforming matrix $\boldsymbol{G}$ of the UAV and reflecting beamforming matrix $\boldsymbol{\Theta}$ of the RIS by interacting with the whole system. Each episode is defined as a time span T, where each step is defined as a time slot δ_n. In order to maximize the sum secrecy rate, the state $s_{n,1}$, the action $a_{n,1}$ and the reward $r_{n,1}$ of the first network at the n-th time slot are defined as follows:

1) **State** $s_{n,1}$: the state of the first agent in the n-th time slot contains the estimated comprehensive CSI from the UAV to all legitimate users and eavesdroppers, i.e., $\boldsymbol{H}_C$.
2) **Action** $a_{n,1}$: we define the the RIS reflecting matrix $\boldsymbol{\Theta}$ and the transmit beamforming matrix $\boldsymbol{G}$ as action. It is worth noting that $\boldsymbol{G} = Re\{\boldsymbol{G}\} + Im\{\boldsymbol{G}\}$ and

Algorithm 1 TDDRL Algorithm

Initialize the actor and critic networks $\pi_1(s;\boldsymbol{\theta}_1^{\pi})$, $Q_1(s,a;\boldsymbol{\theta}_1^{Q})$, target actor and critic networks $\pi_1^{'}(s;\boldsymbol{\theta}_1^{\pi^{'}})$, $Q_1^{'}(s,a;\boldsymbol{\theta}_1^{Q^{'}})$ of the first DDPG network;
Similarly, Initialize $\pi_2(s;\boldsymbol{\theta}_2^{\pi})$, $Q_2(s,a;\boldsymbol{\theta}_2^{Q})$, $\pi_2^{'}(s;\boldsymbol{\theta}_2^{\pi^{'}})$, $Q_2^{'}(s,a;\boldsymbol{\theta}_2^{Q^{'}})$ for the second DDPG network;
for Episode $n_{ep} = 1, 2, ..., N_{ep}$ of the second DDPG network **do**
  Reset the positions of the UAV and all users;
  for Step $n = 1, 2, ..., N_{step}$ **do**
    Observe $\boldsymbol{H}_C$ as $s_{n,1}$, and $\boldsymbol{W}$ as $s_{n,2}$;
    Select actions $a_{n,1}, a_{n,2}$ with a gaussian action noise n_a with variance σ_a :
    $a_{n,1} = \pi_1(s;\boldsymbol{\theta}_1^{\pi}) + n_a$, $a_{n,2} = \pi_2(s;\boldsymbol{\theta}_2^{\pi}) + n_a$
    Execute actions $a_{n,1}, a_{n,2}$, receive an immediate reward $r_{n,1}$ According to Eq. (10) and receive new states $s_{n+1,1}, s_{n+1,2}$ from the environment. Note that $r_{n,1} = r_{n,2}$;
    Store the transitions $[s_{n,1}, a_{n,1}, r_{n,1}, s_{n+1,1}]$ and $[s_{n,2}, a_{n,2}, r_{n,2}, s_{n+1,2}]$ into the memory queues;
    Sample mini batchs to update $\boldsymbol{\theta}_i^{\pi}$, $\boldsymbol{\theta}_i^{Q}$, $i\in\{1,2\}$;
    Update $\boldsymbol{\theta}_i^{\pi^{'}}$, $\boldsymbol{\theta}_i^{Q^{'}}$, $i\in\{1,2\}$;
  end for
end for

$\boldsymbol{\Theta} = Re\{\boldsymbol{\Theta}\} + Im\{\boldsymbol{\Theta}\}$ are separated as real part and imaginary part to tackle with the real input problem.

3) **Reward** $r_{n,1}$: the reward function is defined as:

$$r_{n,1} = \tanh(\sum_{k=1}^{K} R_k^{\text{sec}} - c_1 p_m - c_2 p_r - c_3 p_g), \quad (10)$$

where p_m, p_r and p_g are the penalties when the constraints (9b), (9c) and (9d) are not satisfied, respectively. The coefficients $c_i, i\in\{1,2,3\}$ are the weights for balancing the penalties and the sum secrecy rate. The value of the outage probabilities at each time step are estimated by 500 $\boldsymbol{H}_C$ samples generated according to the statistical CSI error model in (8). The hyperbolic tangent function $\tanh(\cdot)$ is exploited to limit the reward in the range of $(-1, 1)$ for a better convergence.

B. UAV Trajectory

The second DDPG network is exploited to simultaneously obtain the optimal movement $\mu[n]$ and $\psi[n]$ with $\boldsymbol{G}$ and $\boldsymbol{\Theta}$. The state $s_{n,2}$, the action $a_{n,2}$ and the reward $r_{n,2}$ of the second network at the n-th time slot are defined as follows:

1) **State** $s_{n,2}$: as mentioned before, the UAV trajectory is highly coupled with the large amounts of CSI. Thus, we take only the location information $\boldsymbol{W}$ as the state of the second network to decouple the variables.
2) **Action** $a_{n,2}$: the action contains the UAV's flying distance $\mu[n]$ and the direction $\psi[n]$. Then, the movement of UAV at the n-th time slot can be expressed as:

$$\boldsymbol{q}[n+1] - \boldsymbol{q}[n] = \mu[n](\cos\psi[n]\boldsymbol{e}_x + \sin\psi[n]\boldsymbol{e}_y), \quad (11)$$

TABLE I
MAIN PARAMETERS.

Parameter	Value
UAV antennas number	$A = 4$
RIS reflecting elements	$M = 16$
eavesdropper number	$P = 1$
legitimate user number	$K = 2$
step number	$N_{step} = 100$
episode number	$N_{ep} = 100$
carrier frequency	$f_c = 28$ GHz
max transmission power	$P_{max} = 30$ dBm
noise power	$\sigma_n = -114$ dBmW
path loss at one meter	$C_0 = 61$ dB
path loss factor [16]	$\alpha_{ur} = 2.2, \alpha_u = 3.5, \alpha_r = 2.8$
shadow fading factor	$\sigma_s = 3$ dB

3) **Reward** $r_{n,2}$: the same reward function in (10) is employed, since both networks have the same objective to maximize the sum secrecy rate.

As the training process turns to converge, the first network derives the optimal active and passive beamforming strategy, and the second network imparts the optimal trajectory. The shared reward function and environment information allow these two networks to coordinate with each other to learn a favorable policy. Thus, the beamforming matrix ($\boldsymbol{G}$, $\boldsymbol{\Theta}$), and the UAV trajectory $\boldsymbol{Q}$ are achieved according to the proposed TDDRL algorithm. The overall algorithm for solving problem (9) is summarized in Algorithm 1.

C. Computational Complexity Analysis

This subsection mainly discusses the computational complexity of the proposed TDDRL algorithm. In particular, let L and n_i denote the layers number of the deep neural network (DNN) exploited in the DDPG networks and the neurons number in the i-th layer, respectively. For the training mode, the computational complexity for a single DNN to both evaluate and update in a single step is $\mathcal{O}(N_b(\sum_{i=1}^{L-1} n_i n_{i+1}))$ [12], where N_b is the size of the mini-batch. Since the TDDRL algorithm is composed of finite number of DNNs, and it takes $N_{ep} * N_{step}$ steps to finish training, the total training computational complexity of the TDDRL algorithm is $\mathcal{O}(N_{ep} N_{step} N_b(\sum_{i=1}^{L-1} n_i n_{i+1}))$. For the working mode, the computational complexity in each step dramatically decreases to $\mathcal{O}(\sum_{i=1}^{L-1} n_i n_{i+1})$ due to the absence of the training procedure.

IV. SIMULATION RESULTS

In this section, numerical results are presented to evaluate the performance of the proposed TDDRL algorithm. For the first DDPG network, we deploy four fully-connected hidden layers with [800, 600, 512, 256] neurons in both actor and critic networks and the adaptive moment estimation optimizer is used to train the actor network with learning rate 0.0001 and the critic network with learning rate 0.001. The second network has the same structure as the first network, but with different number of four layers [400, 300, 256, 128]. The initial coordinates of the UAV and the fixed RIS are set as (0 m, 25 m, 50m) and (0 m, 50 m, 12.5 m), respectively. The

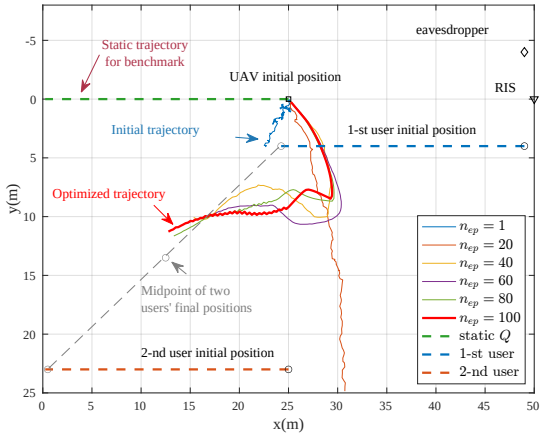

Fig. 3. Trajectory of the UAV optimized by the proposed TDDRL algorithm.

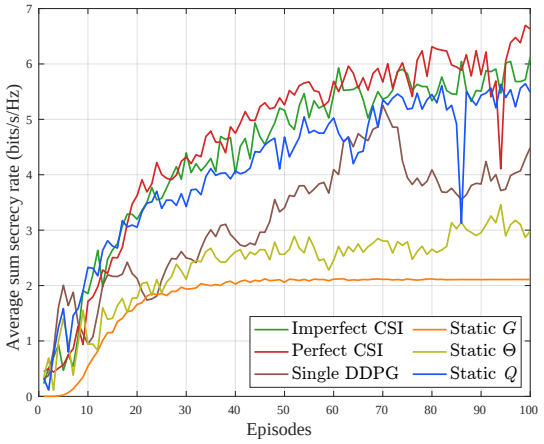

Fig. 4. Accumulated reward performance versus episodes under different RIS elements number.

eavesdropper is placed at (47 m, -4 m, 0 m). Furthermore, we model the movements of two legitimate users as uniform motion in a straight line as shown in Fig.3. More detailed parameters are shown in Table I.

Fig.3 illustrates the optimized trajectory, which eventually converges as the learning procedure is over. It can be observed that the UAV tends to move away from the eavesdropper. Moreover, the UAV is inclined to chase and follow the midpoint of two legitimate users' position, while keeping relatively close distance to the RIS. This implies the UAV trajectory is jointly optimized with the active and passive beamforming, and the proposed algorithm can adapt to dynamic conditions brought by the users' mobility.

Fig.4 plots the average sum secrecy rate versus training episodes, where the following benchmarks are used for comparison: 1) the proposed TDDRL algorithm under imperfect CSI; 2) the proposed TDDRL algorithm under perfect CSI; 3) jointly design $\boldsymbol{G}$, $\boldsymbol{\Theta}$ and $\boldsymbol{Q}$ with a single DDPG network; 4) jointly design $\boldsymbol{\Theta}$ and $\boldsymbol{Q}$ while using static $\boldsymbol{G}$; 5) using static $\boldsymbol{\Theta}$; 6) using static $\boldsymbol{Q}$. It is found that the TDDRL algorithm which jointly optimizes $\boldsymbol{G}$, $\boldsymbol{\Theta}$ and $\boldsymbol{Q}$ achieves the best performance under imperfect CSI. Compared with the single DDPG scheme, the TDDRL algorithm has a better convergence and performance by configuring the same learning rate and layer number. However, there exists the performance gap in terms of the secrecy rate obtained by TDDRL between perfect and imperfect CSI. This is expected and substantiates the importance of the robust design in the actual system.

V. Conclusion

In this letter, we investigate robust and secure transmission for RIS-aided mmWave UAV communications. To maximize the sum secrecy rate of all legitimate users, we propose a TDDRL algorithm to effectively tackle the concerned issues. Simulation results validate that by jointly optimizing UAV trajectory and active (passive) beamforming, a better performance can be achieved compared with several benchmarks.

References

[1] C. Zhang, W. Zhang, W. Wang, L. Yang, and W. Zhang, "Research challenges and opportunities of UAV millimeter-wave communications," *IEEE Wireless Commun.*, vol. 26, no. 1, pp. 58–62, Feb. 2019.

[2] M. Di Renzo, A. Zappone, M. Debbah, M. S. Alouini, C. Yuen, J. de Rosny, and S. Tretyakov, "Smart radio environments empowered by reconfigurable intelligent surfaces: How it works, state of research, and the road ahead," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 11, pp. 2450–2525, Nov. 2020.

[3] A. S. Abdalla, T. F. Rahman, and V. Marojevic, "UAVs with reconfigurable intelligent surfaces: Applications, challenges, and opportunities," Dec. 2020. [Online]. Available: https://arxiv.org/pdf/2012.04775.pdf

[4] X. Yu, D. Xu, Y. Sun, D. W. K. Ng, and R. Schober, "Robust and secure wireless communications via intelligent reflecting surfaces," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 11, pp. 2637–2652, Nov. 2020.

[5] W. Wang, H. Tian, W. Ni, and M. Hua, "Intelligent reflecting surface aided secure UAV communications," Nov. 2020. [Online]. Available: https://arxiv.org/pdf/2011.04339.pdf

[6] S. Li, B. Duo, M. Di Renzo, M. Tao, and X. Yuan, "Robust secure UAV communications with the aid of reconfigurable intelligent surfaces," Dec. 2020. [Online]. Available: https://arxiv.org/pdf/2008.09404.pdf

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

[8] M. Hua, L. Yang, Q. Wu, C. Pan, C. Li, and A. L. Swindlehurst, "UAV-assisted intelligent reflecting surface symbiotic radio system," Jan. 2021. [Online]. Available: https://arxiv.org/pdf/2007.14029

[9] L. Wang, K. Wang, C. Pan, W. Xu, and N. Aslam, "Joint trajectory and passive beamforming design for intelligent reflecting surface-aided UAV communications: A deep reinforcement learning approach," Jul. 2020. [Online]. Available: https://arxiv.org/pdf/2007.08380

[10] 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.*, Jan. 2021, accepted to appear.

[11] M. M. Zhao, Q. Wu, M. J. Zhao, and R. Zhang, "Intelligent reflecting surface enhanced wireless networks: Two-timescale beamforming optimization," *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, pp. 2–17, Jan. 2021.

[12] 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. Wireless Commun.*, vol. 20, no. 1, pp. 375–388, Jan. 2021.

[13] G. Zhou, C. Pan, H. Ren, K. Wang, and A. Nallanathan, "A framework of robust transmission design for IRS-aided MISO communications with imperfect cascaded channels," *IEEE Trans. Signal Process.*, vol. 68, pp. 5092–5106, Aug. 2020.

[14] Y. Huang, F. Al-Qahtani, C. Zhong, Q. Wu, J. Wang, and H. Alnuweiri, "Performance analysis of multiuser multiple antenna relaying networks with co-channel interference and feedback delay," *IEEE Trans. Commun.*, vol. 62, no. 1, pp. 59–73, Jan. 2014.

[15] C. Huang, R. Mo, and C. Yuen, "Reconfigurable intelligent surface assisted multiuser MISO systems exploiting deep reinforcement learning," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1839–1850, Aug. 2020.

[16] X. Liu, Y. Liu, and Y. Chen, "Machine learning empowered trajectory and passive beamforming design in UAV-RIS wireless networks," *IEEE J. Sel. Areas Commun.*, Dec. 2020, accepted to appear.