This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LWC.2021.3081464, IEEE Wireless Communications Letters

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 robust and secure transmission in the millimeter-wave (mmWave) unmanned aerial vehicle (UAV) communication assisted by a 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 CSI is coupled with the UAV trajectory, which results in complex constraints. Furthermore, the time-related issue caused by the outdated CSI also makes the formulated problem intractable to solve. To tackle these challenges, 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*—Reconfigurable intelligent surface, physical layer security, unmanned aerial vehicle, millimeter-wave communications, deep reinforcement learning.**

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. In addition, broadcasting and superposition, as two basic properties of the wireless communication, make wireless transmissions inherently susceptible to security breaches [2]. Hence, secure transmission is a pivotal issue in UAV communication systems which attracted extensive interest of researches [3].

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 [4]. By appropriately tuning the reflection coefficients, the reflected signal can be enhanced at legitimate users and weakened at the eavesdroppers to ensure the secure communications [2], [5]. 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 [1].

A crucial issue in the RIS-aided mmWave UAV communication system is to jointly design the active and passive beamforming and the UAV trajectory to guarantee the secure transmission. However, unlike the general RIS-aided wireless communication model, the UAV mobility-induced variation of the angles of arrival/departure (AoA/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 [1], [3], [6], [7], some of which, in particular, leverage alternating optimization (AO) method to tackle the coupled variables [3], [6]. In [7], 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 in [1]–[4], [6], [7] reply upon the assumption of the perfect channel state information (CSI), which weakens the versatility and practicality of the model. Moreover, the UAV mobility-induced outdated CSI should be taken into account.

The deep reinforcement learning (DRL) is an efficient approach to jointly design the active and passive beamforming, and the UAV trajectory, due to its good generalization, low complexity, and high accuracy characteristics. The motivation of utilizing DRL approach is mainly for two reasons: i) it is fairly difficult to tackle the intricately coupled 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 outdated, 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. First, 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. Moreover, to solve this problem, a novel twin-deep deterministic policy gradient (DDPG)

This paper is supported by the Natural Science Foundation of Beijing, China (GrantNo. L192003). (*Corresponding author: Ying Wang.*)

The authors 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).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LWC.2021.3081464, IEEE Wireless Communications Letters

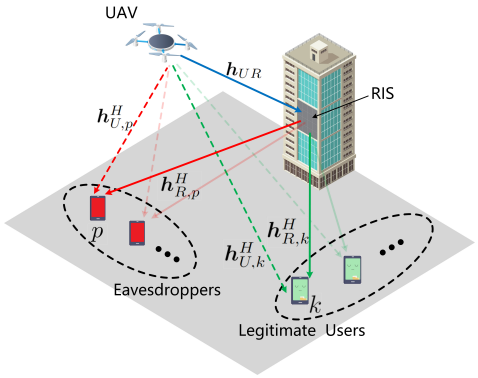

Fig. 1. RIS-aided mmWave UAV communications.

deep reinforcement learning (TDDRL) algorithm is proposed. Specifically, the first DDPG is utilized to provide policy for the active and passive beamforming while the second DDPG provides policy for the UAV trajectory, which, as compared with the conventional single DDPG structure, has the potential to effectively decouple the input information and manifests better robustness. Finally, the simulation results are provided to validate the effectiveness 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:

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

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

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

where D_{max} is the maximum moving distance at each time slot, and B is the moving boundary of the UAV.

Let $\boldsymbol{h}_{U,k}{\in}\mathbb{C}^{A\times 1}$, $\boldsymbol{h}_{U,p}{\in}\mathbb{C}^{A\times 1}$, $\boldsymbol{h}_{R,k}{\in}\mathbb{C}^{M\times 1}$, $\boldsymbol{h}_{R,p}{\in}\mathbb{C}^{M\times 1}$, $\boldsymbol{H}_{UR}{\in}\mathbb{C}^{M\times A}$ be the channel gains from the UAV to the k-th user, from the UAV to the p-th eavesdropper, from the RIS to the k-th user, from the RIS to the p-th eavesdropper, from the UAV to the RIS, respectively. All the channels are modeled according to 3D SV channel model [8] that has been widely used to characterize the mmWave channels:

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

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

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

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)$=-C_0-10\alpha\log_{10}(D)-\mathrm{PL}_s$, C_0 is the path loss at a reference distance of one meter, D (in 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 $\mathbf{a}_L(\varphi){=}\left[1,e^{j\frac{2\pi}{\lambda_c}d\sin(\varphi)},\dots,e^{j\frac{2\pi}{\lambda_c}d(A-1)\sin(\varphi)}\right]^H$, where φ stands for the azimuth AoD $\vartheta_{i,l}^{AoD}$ 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 $\mathbf{a}_P(\varphi,\vartheta){=}\left[1,\dots,e^{j\frac{2\pi}{\lambda_c}d(p\sin(\varphi)\sin(\vartheta)+q\cos(\varphi)\sin(\vartheta))},\dots\right]^H$, where $0\le p,q\le m-1$, and $\varphi(\vartheta)$ is the azimuth (elevation) AoD $\varphi_{i,l}^{AoD}(\vartheta_{i,l}^{AoD})$ and the AoA $\varphi_l^{AoA}(\vartheta_l^{AoA})$. The LoS component of each link, i.e., $\varphi(\vartheta)_{l=1}^{AoA(AoD)}$, is determined by the trajectories of both the users and the UAV, which results in the coupling of the optimization variable $\boldsymbol{Q}$ and the CSI. However, since AoA/AoDs vary in different propagation paths in the SV channel model, the assumption that the LoS components only depends on the locations of the UAV [6], [9] may be ideal in the practical senario. Thus, following the case shown in [10], the angles $\varphi(\vartheta)_l^{AoA(AoD)}, l\neq 1$ can be further expressed by:

$$\varphi(\vartheta)_l^{AoA(AoD)} = \varphi(\vartheta)_{l=1}^{AoA(AoD)} + \Phi(\Lambda)_l^{AoA(AoD)},\ l=2,\cdots,L \tag{3}$$

where $\Phi(\Lambda)_l^{AoA(AoD)}$ are the spreading factors [10].

The cascaded channel from the UAV to the i-th user or the eavesdropper can be denoted by $\boldsymbol{H}_{C,i}{=}\mathrm{diag}(\boldsymbol{h}_{R,i}^H)\boldsymbol{h}_{UR}, \forall i \in \mathcal{K}\cup\mathcal{P}$. The passive beamforming matrix of the RIS is defined as $\boldsymbol{\Theta}{=}\mathrm{diag}\left(\beta_1 e^{j\theta_1},\beta_2 e^{j\theta_2},...,\beta_M e^{j\theta_M}\right)$, 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 [11]. 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 can be vectorized as $\boldsymbol{\Psi}{=}\mathrm{vec}(\boldsymbol{\Theta})$. Thus, the received signal at the i-th user or eavesdropper from the UAV can be formulated as

$$y_i = (\boldsymbol{h}_{U,i}^H + \boldsymbol{\Psi}^H\boldsymbol{H}_{C,i})\boldsymbol{G}\boldsymbol{s} + n_i, \forall i \in \mathcal{K}\cup\mathcal{P}, \tag{4}$$

where $\boldsymbol{s} \in \mathbb{C}^{K\times 1}$ 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

$$R_k^u = \log_2\left(1+\frac{|(\boldsymbol{h}_{U,k}^H+\boldsymbol{\Psi}^H\boldsymbol{H}_{C,k})\boldsymbol{g}_k|^2}{\sum_{k'\in\mathcal{K}\setminus k}|\boldsymbol{h}_{U,k}^H+\boldsymbol{\Psi}^H\boldsymbol{H}_{C,k})\boldsymbol{g}_{k'}|^2+n_k^2}\right). \tag{5}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LWC.2021.3081464, IEEE Wireless Communications Letters

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

$$R_{p,k}^{e}=\log_2\left(1+\frac{|(\boldsymbol{h}_{U,p}^{H}+\boldsymbol{\Psi}^{H}\boldsymbol{H}_{C,p})\boldsymbol{g}_k|^2}{\sum_{k'\in\mathcal{K}\setminus k}|\boldsymbol{h}_{U,p}^{H}+\boldsymbol{\Psi}^{H}\boldsymbol{H}_{C,p})\boldsymbol{g}_{k'}|^2+n_p^2}\right). \tag{6}$$

The achievable individual secrecy rate from the UAV to the k-th user [11] can be expressed by

$$R_k^{\mathrm{sec}}=\left[R_k^{\mathrm{u}}-\max_{\forall p}R_{p,k}^{\mathrm{e}}\right]^{+}, \tag{7}$$

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

In the practical system, the perfect CSI is not available at the UAV due to the transmission delay and processing delay, as well as the mobility of the UAV and the users. The CSI may be stale at the time when the UAV transmits the data stream to the RIS and the users, which results in an inevitable performance loss once this outdated CSI is employed for transmission. Thus, the outdated CSI should be explicitly considered in the system design.

Let T_d be the delay between the outdated CSI 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 [12]

$$\boldsymbol{h}\left(t+T_{\mathrm{d}}\right)=\varrho\tilde{\boldsymbol{h}}(t)+\sqrt{1-\varrho^2}\Delta\boldsymbol{h}, \tag{8}$$

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

The autocorrelation factor ϱ bridges the real-time $\boldsymbol{h}\left(t+T_{\mathrm{d}}\right)$ with the estimated $\tilde{\boldsymbol{h}}(t)$ that is easily outdated. The distributions of the actual CSI $\boldsymbol{h}\triangleq\{\boldsymbol{h}_{\mathrm{U},i},\boldsymbol{h}_{\mathrm{R},i},\boldsymbol{h}_{\mathrm{UR}},\forall i\in\mathcal{K}\cup\mathcal{P}\}$ can be expressed as the form in (8), where ϱ is the function of speed which is determined by the trajectories of both the users and the UAV. Therefore, the trajectories have the influence on the imperfectness of the outdated CSI. Furthermore, the system can only access to the estimated CSI $\tilde{\boldsymbol{h}}\triangleq\{\tilde{\boldsymbol{h}}_{\mathrm{U},i},\tilde{\boldsymbol{h}}_{\mathrm{R},i},\tilde{\boldsymbol{h}}_{\mathrm{UR}},\forall i\in\mathcal{K}\cup\mathcal{P}\}$[1] that are outdated, and the actual CSI $\boldsymbol{h}$ given by (8) is employed to calculate achievable secrecy rate in (5)-(7).

B. Problem Formulation

In this letter, we aim to maximize the sum secrecy rate $\sum_{k=1}^{K}R_k^{\mathrm{sec}}$ by jointly optimizing the UAV's trajectory $\boldsymbol{Q}\triangleq\{\boldsymbol{q}[n],n=1,2,\cdots,N\}$ and the active (passive) beamforming matrix $\boldsymbol{G}(\boldsymbol{\Theta})$, which yields the following problem

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

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

$$\Pr\left\{R_k^{\mathrm{sec}}\geq R_k^{\mathrm{sec,th}}\right\}\geq 1-\rho_k,\forall k\in\mathcal{K}, \tag{9c}$$

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

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

[1]We assume that the CSI can be obtained by adopting the channel estimation method in the RIS aided system, such as [13].

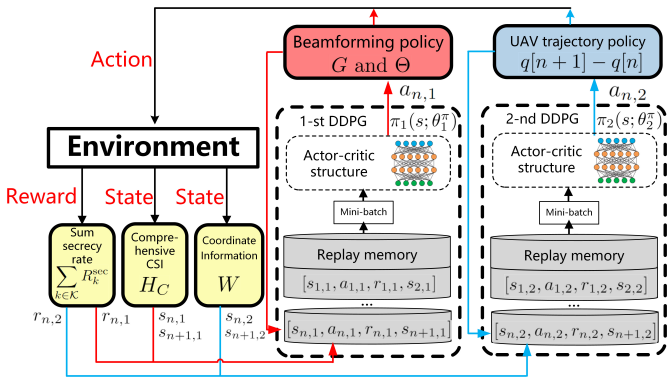

Fig. 2. Structure of the proposed TDDRL algorithm.

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 $R_k^{\mathrm{sec,th}}$ 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 closed-form expression, and the time varying CSI described in (8). 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.

III. DRL-BASED SOLUTION

To solve problem (9), we propose a TDDRL algorithm, which allows the agent to learn the policies of the beamforming and the trajectory without any prior knowledge of the system. 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, i.e., $\boldsymbol{W}$, as the state 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 is elaborated as follows.

A. Active and Passive Beamforming

Inspired by the work of [11], the first DDPG network is employed to learn the optimal policy of the beamforming matrix $\boldsymbol{G}$ at the UAV and the reflecting beamforming matrix $\boldsymbol{\Theta}$ at the RIS by interacting with the whole environment. 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., $\boldsymbol{H}_C$. Furthermore, the small-scale component in $\boldsymbol{h}$ may not be known at the UAV in advance. The small-scale information can be collected in real-time as the status of

Algorithm 1 TDDRL Algorithm

Initialize the actor and critic networks $\pi_1(s;\boldsymbol{\phi}_1^{\pi})$, $Q_1(s,a;\boldsymbol{\phi}_1^{Q})$, target actor and critic networks $\pi_1^{'}(s;\boldsymbol{\phi}_1^{\pi^{'}})$, $Q_1^{'}(s,a;\boldsymbol{\phi}_1^{Q^{'}})$ of the first DDPG network;
Similarly, Initialize $\pi_2(s;\boldsymbol{\phi}_2^{\pi})$, $Q_2(s,a;\boldsymbol{\phi}_2^{Q})$, $\pi_2^{'}(s;\boldsymbol{\phi}_2^{\pi^{'}})$, $Q_2^{'}(s,a;\boldsymbol{\phi}_2^{Q^{'}})$ 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 :
    $a_{n,1}=\pi_1(s;\boldsymbol{\phi}_1^{\pi})+n_a$, $a_{n,2}=\pi_2(s;\boldsymbol{\phi}_2^{\pi})+n_a$
    Execute actions $a_{n,1}, a_{n,2}$, receive an immediate reward $r_{n,1}$=$r_{n,2}$ According to Eq. (10) and receive new states $s_{n+1,1}, s_{n+1,2}$ from the environment;
    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 $\boldsymbol{\phi}_i^{\pi}$, $\boldsymbol{\phi}_i^{Q}$, $i\in\{1,2\}$;
    Update $\boldsymbol{\phi}_i^{\pi^{'}}$, $\boldsymbol{\phi}_i^{Q^{'}}$, $i\in\{1,2\}$;
  end for
end for

the network, and thus the proposed algorithm has the capability of adapting the variations of the environment in an online manner.

2) **Action** $a_{n,1}$: we define the the passive beamforming matrix $\boldsymbol{\Theta}$ and the active beamforming matrix $\boldsymbol{G}$ as action. It is worth noting that $\boldsymbol{G}=Re\{\boldsymbol{G}\}+Im\{\boldsymbol{G}\}$ and $\boldsymbol{\Theta}=Re\{\boldsymbol{\Theta}\}+Im\{\boldsymbol{\Theta}\}$ 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:

$$r_{n,1}=\tanh(\sum_{k=1}^{K}R_k^{\mathrm{sec}}-c_1p_m-c_2p_r-c_3p_g), \quad (10)$$

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. As for the probabilistic constraint in (9c), more particularly, we can get the outage probabilities by sampling the CSI to calculate the distribution of each R_k^{sec}, i.e., $1-\Pr\left\{R_k^{\mathrm{sec}}\geq R_k^{\mathrm{sec,th}}\right\}\approx N_{outage}/N_{sample}$, where N_{outage} and N_{sample}, respectively, represent the number of the samples the secrecy rate R_k^{sec} being below the threshold $R_k^{\mathrm{sec,th}}$ and all CSI samples generated.

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 the UAV at the n-th time slot can be expressed as: $\boldsymbol{q}[n+1]-\boldsymbol{q}[n]=\mu[n](\cos\psi[n]\boldsymbol{e}_x+\sin\psi[n]\boldsymbol{e}_y)$, where $\boldsymbol{e}_x$, $\boldsymbol{e}_y$ are the unit vector on the X-axis and the Y-axis.
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 delivers 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**. Compared with the offline mode that usually loads the policy of the beamforming and the trajectory to the UAV in advance and is insensitive to the dynamic environment, our proposed TDDRL works in an online manner and has the ability to capture the instantaneous CSI at each time slot which contains the fast-varying factors such as $\boldsymbol{g}$.

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 $\mathcal{O}(N_b(\sum_{i=1}^{L-1}n_in_{i+1}))$ [11], 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 $\mathcal{O}(N_{ep}N_{step}N_b(\sum_{i=1}^{L-1}n_in_{i+1}))$. In the online working mode, the computational complexity in each step can be dramatically decreased to $\mathcal{O}(\sum_{i=1}^{L-1}n_in_{i+1})$ by cutting off the training procedure once the performance of the network finally converges, which, thus, retains the computational complexity at a favorable level.

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. 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 TDDRL is trained in 100 episodes and each episode has 100 steps where each step is considered as a time slot. The initial coordinates of the UAV and the fixed RIS are set as (0 m, 25m, 50m) and (0m, 50m, 12.5m), respectively. The eavesdropper is placed at (47m, -4m, 0m). Furthermore, we model the movements of two legitimate users as uniform motion in a straight line as shown in Fig. 3. Other system parameters are set as δ_t=0.1 ms, T_d=1 s, f_c=28 GHz, C_0=61 dB, $P_{\max}$=30 dBm, σ_n=$-$114 dBm, L=3, α_{ur}=2.2, α_u=3.5, α_r=2.8 [7], σ_s=3 dB, A=4, M=16, P=1, K=2, $\Phi_l^{AoD}\in\{5,10,15,25\}$, $\Phi_l^{AoA}\in\{30,45,60\}$, $\Lambda_l^{AoD}\in\{1,3,5\}$, $\Lambda_l^{AoA}\in\{5,10,15\}$(degrees) [10].

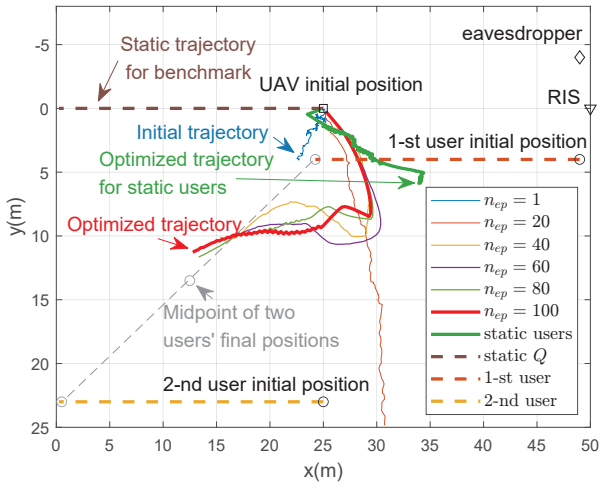

Fig. 3. The optimized UAV trajectory by exploiting TDDRL.

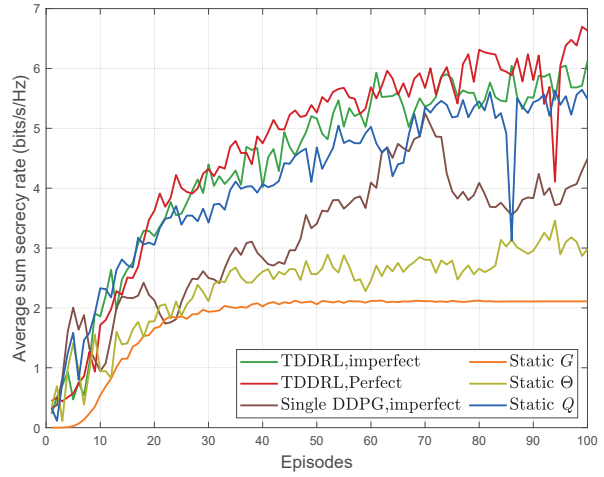

Fig. 4. Accumulated reward performance versus episodes.

Fig. 3 illustrates the process of exploring the optimal UAV trajectory under different simulation configurations. Intuitively, the UAV tends to approach the RIS at first and moves far away from the eavesdropper. As the distance between the RIS and two legitimated users grows larger, what is interesting is that the UAV follows them instead and moves towards the midpoint of the two users. Such a phenomenon can be attributed to the fact that as the cascaded UAV-RIS-users links gradually turn to be weak due to the increase of distance, the direct links dominate the transmission and thus the UAV tries to serve two legitimate uses as fairly as possible. Furthermore, to verify the effectiveness of our proposed TDDRL, i.e., TDDRL has the capability of adapting the variations in the environment, the case of two users without mobility is configured as a benchmark. In this case, the UAV finally hovers close to the RIS and users. Thus, this demonstrates that under TDDRL, the UAV can flexibly adapt to the dynamic environment and the system performance can be improved by jointly designing the UAV trajectory and beamforming.

In Fig. 4, we investigate the average sum secrecy rate versus the episodes as the TDDRL explores the optimal strategy. Several benchmarks are considered for comparison under the imperfect CSI: **benchmark I:** jointly optimize $\boldsymbol{G}$, $\boldsymbol{\Theta}$ using TDDRL and use static $\boldsymbol{Q}$ depicted in Fig. 3; **benchmark II:** use static $\boldsymbol{\Theta}$; **benchmark III:** use static $\boldsymbol{G}$. With respect to the average sum secrecy rate, the proposed TDDRL shows better performance under perfect CSI, as compared with that under imperfect CSI, which implies that the acquisition of CSI is the main bottleneck to hinder the improvement of system performance. In addition, there exists a performance gap between the TDDRL and the single DDPG. The reason is that our proposed TDDRL has the potential to decouple the intricate variables and the network status (such as CSI and the UAV position). The extensive interconnection and information redundancy between the CSI $\boldsymbol{H}_C$ and the position information $\boldsymbol{W}$ lead to non-convergence of the conventional single DDPG-based algorithm. Thus, TDDRL enables the beamforming to be more effective, which underscores the importance of jointly designing the UAV trajectory and the beamforming.

V. Conclusion

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

[2] S. Hong, C. Pan, H. Ren, K. Wang, and A. Nallanathan, "Artificial-noise-aided secure MIMO wireless communications via intelligent reflecting surface," *IEEE Trans. Commun.*, vol. 68, no. 12, pp. 7851–7866, Dec. 2020.

[3] S. Li, B. Duo, M. Di Renzo, M. Tao, and X. Yuan, "Robust secure UAV communications with the aid of reconfigurable intelligent surfaces," *IEEE Trans. Wireless Commun.*, to appear, 2021.

[4] Q. Wu and R. Zhang, "Intelligent reflecting surface enhanced wireless network via joint active and passive beamforming," *IEEE Trans. Wireless Commun.*, vol. 18, no. 11, pp. 5394–5409, 2019.

[5] H. Yang, Z. Xiong, J. Zhao, D. Niyato, Q. Wu, H. V. Poor, and M. Tornatore, "Intelligent reflecting surface assisted anti-jamming communications: A fast reinforcement learning approach," *IEEE Trans. Wireless Commun.*, vol. 20, no. 3, pp. 1963–1974, 2021.

[6] 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.

[7] 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.*, to appear, 2020.

[8] 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.

[9] Y. Chen, Y. Wang, J. Zhang, and M. Di Renzo, "QoS-driven spectrum sharing for reconfigurable intelligent surfaces (RISs) aided vehicular networks," *IEEE Trans. Wireless Commun.*, to appear, 2021.

[10] 3GPP, "Study on channel model for frequencies from 0.5 to 100 GHz," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 38.901, 01 2020, version 16.1.0.

[11] 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.

[12] 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.

[13] Z. Wang, L. Liu, and S. Cui, "Channel estimation for intelligent reflecting surface assisted multiuser communications: Framework, algorithms, and analysis," *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6607–6620, Oct. 2020.