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

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

***Abstract*—In this letter, we study the secure transmission in millimeter-wave (mmWave) unmanned aerial vehicle (UAV) communications assisted by reconfigurable intelligent surface (RIS) under imperfect channel state information (CSI). Specifically, the active beamforming ~~policy~~ of the UAV, the ~~reflecting~~ coefficients of the RIS and the UAV trajectory are jointly ~~optimized~~ to maximize the sum secrecy rate of all legitimate users in the presence of multiple eavesdroppers. However, the formulated problem is difficult to solve due to the non-convex constrains, the intricately couple variables in the RIS-aided system and the time-related issue caused by outdated CSI. ~~And even the widely applicable AO method cannot solve the problem well. To overcome these difficulties, we provide a deep deterministic policy gradient (DDPG)-based solution. And a novel twin-DDPG deep reinforcement learing (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, secure communication, UAV communication, 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 more than 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 LoS links, thereby achieving directional signal enhancement, expanding coverage area and reducing the need for 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].

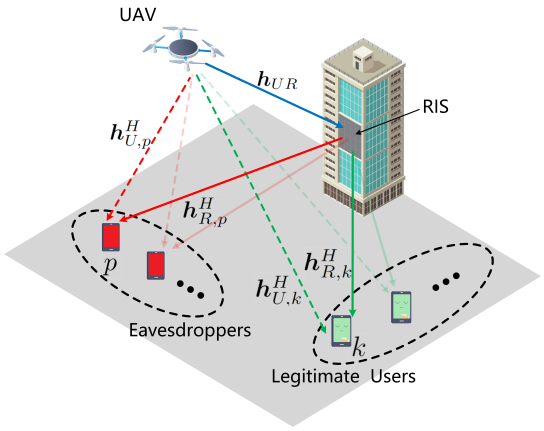

Fig. 1. RIS-aided Milimeter Wave UAV Communications.

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 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 [5]–[8] to tackle the coupled variables, and adopt the phase alignment technique [7] for the single-user system. 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. It should be pointed out that the above literature [5], [7]–[9] are based on the assumption of

Corresponding author: Ying Wang.

X. Guo, Y. Wang, and Y. Chen 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; wangying@bupt.edu.cn; chen_yuanbin@163.com).

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.

Due to good generalization, low complexity, and high accuracy, the deep reinforcement learning (DRL) is an efficient approach to jointly design the active and passive beamforming, and the UAV trajectory. 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 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 proposed a novel twin-deep deterministicpolicy gradient (DDPG) deep reinforcement learning (TDDRL) algorithm to maximize the sum secrecy rate by jointly ~~optimizing~~ the RIS ~~reflecting~~ coefficients, the active beamforming and the UAV trajectory in RIS-aided ~~and secure~~ mmWave UAV communications ~~under imperfect CSI.~~ The contributions of this work can be summarized as follows: 1) we design a robust and secure mmWave UAV communications framework, where a non-convex problem is formulated to maximize the secrecy rate; 2) we present a robust transmission design based on statistical CSI error model to address the problem brought by outdated CSI; 3) we propose a novel TDDRL algorithm by exploiting two DDPG network to solve the coupling problem between the UAV trajectory and active (passive) beamforming coefficients. Specifically, the first network provides the active (passive) beamforming policy while the second network provides the UAV trajectory policy, as such the proposed algorithm has better stability and convergence; 4) we present numerical simulations to validate the convergence and effectiveness of the proposed TDDRL algorithm.

II. System Model and Problem Formulation

A. System Model

In this letter, we consider an RIS-aided ~~milimeter wave~~ UAV ~~secure transmission~~ system where ~~a~~ 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 is with a uniform planar array (UPA) with $M{=}m^2$ passive reflecting elements (m should be 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. Let $\mathbf{w}_k{=}[x_k,y_k,0]^T,\forall k{\in}\mathcal{K}$ and $\mathbf{w}_p{=}[x_p,y_p,0]^T,\forall p{\in}\mathcal{P}$ denote the legitimate users' and the eavesdroppers' coordinates, respectively. The RIS is fixed at $\mathbf{w}_R{=}[x_R,y_R,z_R]^T$. In addition, assume that the UAV flies at a fixed altitude in a finite time span which is divided into N time slots, i.e., $T{=}N\delta_t$, where δ_t is the time slot. Then the coordinate of the UAV at the n-th time slot is denoted by $\mathbf{q}[n]{=}[x[n],y[n],H_U]^T, n{=}1,2,...,N$, 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 $\boldsymbol{h}_{U,k}{\in}\mathbb{C}^{A\times 1}$, $\boldsymbol{h}_{U,p}{\in}\mathbb{C}^{A\times 1}$, $\boldsymbol{h}_{R,k}{\in}\mathbb{C}^{M\times 1}$, $\boldsymbol{h}_{R,p}{\in}\mathbb{C}^{M\times 1}$, $\boldsymbol{H}_{UR}{\in}\mathbb{C}^{M\times A}$ be the channel gains of the UAV to k-th user, UAV to p-th eavesdropper, RIS to k-th user, RIS to p-th eavesdropper, UAV to RIS links, respectively. All the channels are modeled as milimeter wave channels [10] [11] as

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

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

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

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)$=-\mathrm{C}_0-10\alpha\log_{10}(\mathrm{D})-\mathrm{PL}_s$, C_0=61 dB 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 $\mathbf{a}_L(\theta)=\left[1,e^{j\frac{2\pi}{\lambda_c}d\sin(\theta)},\dots,e^{j\frac{2\pi}{\lambda_c}d(N-1)\sin(\theta)}\right]^{\mathrm{H}}$ [12], where θ stands for the azimuth angle-of-departure(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 $\mathbf{a}_P(\theta,\phi)=\left[1,\dots,e^{j\frac{2\pi}{\lambda_c}d(p\sin(\theta)\sin(\phi)+q\cos(\theta)\sin(\phi))},\dots\right]^{\mathrm{H}}$ [12],where $0\le p,q\le m-1$, $\theta(\phi)$ is the azimuth(elevation) AoD $\theta_{i,l}^{AoD}(\phi_{i,l}^{AoD})$ and the angle-of-arrival (AoA) $\theta_l^{AoA}(\phi_l^{AoA})$.

The cascaded channel from the UAV to the i-th user or eavesdropper can be written as $\boldsymbol{H}_{C,i}=\mathrm{diag}(\boldsymbol{h}_{R,i}^H)\boldsymbol{h}_{UR},\forall i\in\mathcal{K}\cup\mathcal{P}$. The passive beamforming matrix of the RIS [13] is ~~defined as~~ $\boldsymbol{\Theta}=\mathrm{diag}\left(\beta_1e^{j\theta_1},\beta_2e^{j\theta_2},...,\beta_Me^{j\theta_M}\right)$, where $\theta_m\in[0,2\pi)$ and $\beta_m\in[0,1]$ represent the phase shift and amplitude reflection coefficient of the m-th RIS ~~reflection~~ element, respectively. For ~~feasibility~~, the amplitude reflection coefficient subjects to unit-modulus constraints, i.e., $\beta_m{=}1$. Let $\boldsymbol{\Psi}{=}\mathrm{vec}(\boldsymbol{\Theta})^{\mathbf{T}}$ denote the vectorized passive beamforming ~~matrix~~. ~~Then~~, the received signal at the i-th user or eavesdropper from the UAV can be formulated as

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

where s_k with $E[|s_k|^2]=1$ and $\mathbf{G}\in\mathbb{C}^{\mathrm{A}\times\mathrm{K}}$ represent the transmitted symbol and the beamforming matrix at the UAV, ~~respectively,~~ 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 $\mathbf{G}$.

Then, the the achievable unsecured rate of the k-th user is given by

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

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 [14] 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 [15], the outdated CSI can be expressed as statistical CSI error model. Furthermore, 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 [16]

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

where $\boldsymbol{e}$ is independent identically distributed 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

$$\rho = J_0\left(2\pi f_D T_{\text{d}}\right), \quad (8)$$

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

$$\begin{aligned}\boldsymbol{h}_{\text{U},i} &= \rho\tilde{\boldsymbol{h}}_{\text{U},i} + \Delta\boldsymbol{h}_{\text{U},i}, \forall k\in\mathcal{K}\cup\mathcal{P},\\ \boldsymbol{h}_{\text{R},i} &= \rho\tilde{\boldsymbol{h}}_{\text{R},i} + \Delta\boldsymbol{h}_{\text{R},i}, \forall k\in\mathcal{K}\cup\mathcal{P},\\ \boldsymbol{h}_{\text{UR}} &= \rho\tilde{\boldsymbol{h}}_{\text{UR}} + \Delta\boldsymbol{h}_{\text{UR}}.\end{aligned} \quad (9)$$

Note that the system only has 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. ~~And the~~ actual CSI $\boldsymbol{h} \in \{\boldsymbol{h}_{\text{U},i}, \boldsymbol{h}_{\text{R},i}, \boldsymbol{h}_{\text{UR}}\}$ is employed to calculate achievable secrecy rate of each user which has been expressed in (4), (5), (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 $\mathbf{Q} \triangleq \{\mathbf{q}[\mathbf{n}], \mathbf{n}\in\mathcal{N}\}$ and the active (passive) beamforming matrix $\mathbf{G}(\boldsymbol{\Theta})$. The ~~optimization~~ problem is formulated as

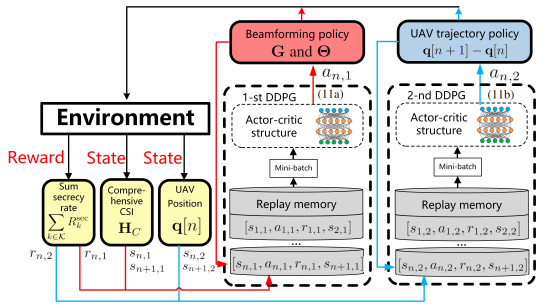

Fig. 2. Structure of the proposed TDDPG algorithm.

$$\max_{\mathbf{Q},\mathbf{G},\boldsymbol{\Theta}} \quad \sum_{k\in\mathcal{K}} R_k^{\text{sec}} \quad (10a)$$

$$s.t. \quad (1), \quad (10b)$$

$$\Pr\left\{R_k^{\text{sec}} \geq R_k^{\text{sec,th}}\right\} \geq 1-\rho_k, \forall k\in\mathcal{K}, \quad (10c)$$

$$\text{Tr}\left(\mathbf{G}\mathbf{G}^{\mathbf{H}}\right) \leq P_{\max}, \quad (10d)$$

$$\theta_m \in [0,2\pi), \forall m\in\mathcal{M}, \quad (10e)$$

where the rate outage constraint in (10c) 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 $1-\rho_k$. Constrains (10b),(10c),(10e) are non-convex. The channel gains in (10a) and (10c) are relevant to the speeds of the transceivers. Furthermore, there is no close form to express (10c). As a result, there are few traditional methods to solve this ~~non-convex~~ time-varying problem contains of constrain that has no close form to expression. In section III, a DRL-based solution is propose to overcome these challenges.

III. DRL-based Solution

To solve the non-convex problem in (10), we propose ~~the~~ TDDRL algorithm, instead of using single agent to find the optimal $\mathbf{G}$, $\boldsymbol{\Theta}$ and $\mathbf{Q}$, since $\mathbf{Q}$ would be highly coupled with large scale CSI using single agent which is actually irrelevant. As shown in Fig.2, the first network takes CSI as state to obtain the optimal $\mathbf{G}$ and $\boldsymbol{\Theta}$, while the second network takes UAV position as state to obtain the movement $\mathbf{q}[\mathbf{n}+\mathbf{1}]-\mathbf{q}[\mathbf{n}]$ of UAV at n-th time slot. Both network take the sum secrecy data rate as reward. The overall algorithm for solving problem (10) is summarized in Algorithm 1.

A. Active and Passive Beamforming

Inspired by the work of [14], the first DDPG-based network is employed to learn the optimal policy in terms of the UAV's beamforming matrix $\mathbf{G}$ and the RIS's reflecting beamforming matrix $\boldsymbol{\Theta}$ 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 $a_{n,1}$, the reward $r_{n,1}$ in n-th time slot of the first agent ~~is~~ defined as follows:

1) State $s_{n,1}$: the state of the first agent in n-th time slot contains the estimated comprehensive CSI from the UAV

Algorithm 1 TDDRL Algorithm

Initialize the 1-st DDPG network with initial Q-function $Q_1(s,a;\boldsymbol{\theta}_1)$;
Initialize the 2-nd DDPG network with initial Q-function $Q_2(s,a;\boldsymbol{\theta}_2)$;
for Episode $n_{ep}=1,2,...,N_{ep}$ **do**
  Reset the positions of the UAV and users;
  Reset the active and passive beamforming matrix $\mathbf{G}$, $\boldsymbol{\Theta}$;
  for Step $n=1,2,...,N_{step}$ **do**
    Observe all CSI as $s_{n,1}$;
    Observe the UAV position as $s_{n,2}$;
    Select actions $a_{n,1}, a_{n,2}$ with a gaussian action noise n_a with variance σ_a :

$$a_{n,1}=\arg\max_{a_{n,1}\in\mathcal{A}} Q_1\left(s_{n,1},a_{n,1};\theta_{n,1}\right)+n_a \quad (11a)$$

$$a_{n,2}=\arg\max_{a_{n,2}\in\mathcal{A}} Q_2\left(s_{n,2},a_{n,2};\theta_{n,2}\right)+n_a \quad (11b)$$

    Execute action $a_{n,1}, a_{n,2}$, receive an immediate reward $r_{n,1}$ using Eq. (12) and new states $s_{n+1,1}, s_{n+1,2}$. 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 two networks' memory queues, respectively;
    Sample a mini-batch of transitions in memory queue randomly to update the evaluation networks of both DDPG networks using proper loss function and policy gradient function [17];
    Update the target networks of both DDPG networks;
  end for
end for

to all legitimate users and eavesdroppers, i.e., $\mathbf{H}_C \triangleq \{\mathbf{h}_{U,i}^{\mathbf{H}} + \boldsymbol{\Psi}^{\mathbf{H}}\mathbf{H}_{C,i}\}, \forall \mathbf{i} \in \mathcal{K}\cup\mathcal{P}$.

2) Action $a_{n,1}$: we define the phase shift of all RIS reflecting elements $\theta_m, \forall m \in \mathcal{M}$ and the transmit beamforming matrix G as action. It is worth noting that $\mathbf{G}=\mathbf{Re}\{\mathbf{G}\}+\mathbf{Im}\{\mathbf{G}\}$ 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:

$$r_{n,1}=tanh\Big(\sum_{k=1}^{K} R_k^{\text{sec}} - p_r - p_m\Big), \quad (12)$$

where p_r is the penalty if the outage constraint (10c) is not satisfied, and p_m is the the penalty when the UAV flies out of the target area. The hyperbolic tangent function $\tanh(\cdot)$ is exploited to limite the reward in range of $(-1,1)$ for better convergence.

B. UAV Trajectory

The second DDPG is exploited to simultaneously obtain the optimal movement $\mathbf{q}[n+1]-\mathbf{q}[n]$ with $\mathbf{G}$ and $\boldsymbol{\Theta}$. It is feasible to utilize a single DDPG network to tune all parameters $\mathbf{G}, \boldsymbol{\Theta}, \mathbf{Q}$, which have been done by most works. But in this letter, ~~UAV's trajectory~~ is rarely relevant to the large amount of CSI, leading to instability and divergence by connecting irrelevant actions and feedbacks using a single network. The state $s_{n,2}$, the action $a_{n,2}$, the reward $r_{n,2}$ in n-th time slot of the second agent is defined as follows:

1) State $s_{n,2}$: as mentioned before, ~~UAV's trajectory~~ is rarely relevant to the large amount of CSI. ~~So the second network only takes the UAV's position $\mathbf{q}[n]$ as state.~~

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 can be expressed as:

$$\mathbf{q}[n+1]-\mathbf{q}[n]=\mu[n](cos\psi[n]\mathbf{e}_x+sin\psi[n]\mathbf{e}_y) \quad (13)$$

3) Reward $r_{n,2}$: the same reward function in (12) is employed, since both network have the same objective to maximize the sum secrecy rate.

C. Computational Complexity Analysis

~~Let~~ L, n_i denote the layers number of the DNN exploited in the DDPG networks and the neurons number in the i-th layer, respectively. ~~Then for~~ the training stage, the computational complexity ~~of~~ a single DNN to both deliver actions and update in a single step is $O((N_b+1)(\sum_{i=1}^{L-1} n_i n_{i+1}))$, where N_b is the size of mini-batch. Since it takes the agent $N_{ep}*N_{step}$ steps to ~~finnish~~ training, the computational complexity of the whole training procedure is $O(N_{ep}N_{step}(N_b+1)(\sum_{i=1}^{L-1} n_i n_{i+1}))$. For online working stage, the computational complexity in each step dramatically decreases to $O(\sum_{i=1}^{L-1} n_i n_{i+1})$ due to the absence of the training procedure. So the proposed TDDRL algorithm can significantly reduce the traning time cost by leveraging the off-line trained model.

IV. Simulation Results

In this section, numerical results are presented to ~~characterize~~ the performance of our proposed ~~solution~~. For the first DDPG-based network, we deploy four fully-connected hidden layers with [800, 600, 512, 256] neurons in both actor and critic networks and the AdamOptimizer is used to train the actor network with learning rate 0.0001 and critic network with learning rate 0.001. The second ~~net~~ has the same structure as the first net, but with different number of four layers [400, 300, 256, 128]. The initial coordinates of UAV and RIS are set as [0, 25 ,50], [0, 50 ,12.5]. The eavesdropper is placed at [47, -4, 0], ~~while we~~ model two legitimate users' movement as uniform motion in a straight line as shown in Fig.3. More parameters are shown in Tabel~~.~~ 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 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 ~~by~~ different benchmarks which all increase with traning episode n_{ep}. It is found that the proposed solution that jointly optimizing UAV trajectory and active (passice) beamforming achieves

TABLE I
MAIN PARAMETERS.

Parameter	Value
UAV antennas number	$\mathbf{A} = 4$
RIS reflecting elements	$\mathbf{M} = 16$
eavesdropper number	$\mathbf{P} = 1$
legitimate user number	$\mathbf{K} = 2$
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 factor [18]	$\alpha_{ur} = 2.2, \alpha_u = 3.5, \alpha_r = 2.8$
shadow fading factor	$\sigma_s = 3$ dB

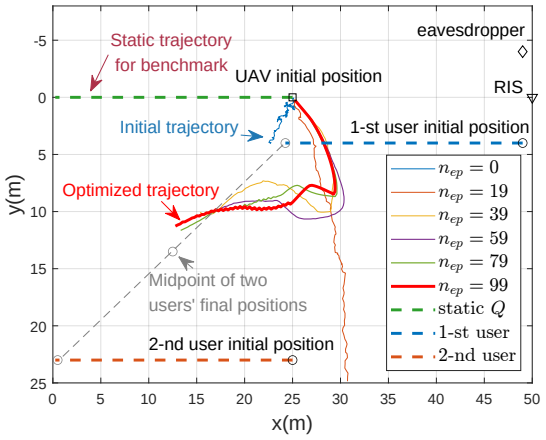

Fig. 3. Trajectory of the UAV optimized by the proposed TDDRL algorithm.

the best performance under imperfect CSI, compared with other benchmarks: fix UAV trajectory, active beamforming and passive beamforming, respectively. However, the proposed solution performs slightly better under the perfect CSI, which implies the proposed solution has good robustness. ~~Compared with the single DDPG structure, the twin DDPG structure has better convergence and performance using the same learning rate and layers number.~~ Thus, the secrecy rates of legitimate users in our proposed system under imperfect CSI can be maximized leveraging RIS and UAV by proposed DDPG-based algorithm.

图片和表格都置顶

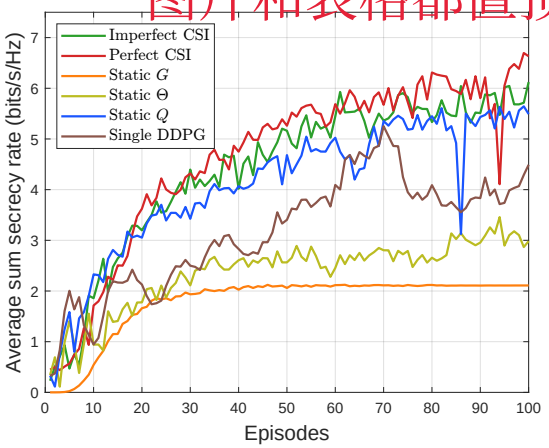

Fig. 4. Accumulated reward performance versus episodes under different RIS elements number.

V. CONCLUSION

In this letter, we investigate~~d~~ robust and secure transmission for RIS-aided mmWave UAV communications. To maximize the secrecy rates of legitimate users, we proposed a ~~DDPG-baesd optimization~~ algorithm. ~~The simulation~~ results validate~~d~~ that by jointly optimizing UAV trajectory and active (passice) beamforming, ~~the best~~ performance can be achieved under imperfect CSI compared with other benchmarks.

检查参考文献格式！

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. Sabri Abdalla, T. Faizur Rahman, and V. Marojevic, "UAVs with reconfigurable intelligent surfaces: Applications, challenges, and opportunities," *arXiv e-prints*, pp. arXiv–2012, 2020.

[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," ~~*arXiv preprint*~~ *arXiv:2011.04339*, 2020.

[6] L. Sixian, D. Bin, M. Di Renzo, T. Meixia, and Y. Xiaojun, "Robust secure uav communications with the aid of reconfigurable intelligent surfaces," *arXiv:2008.09404*, 2020.

[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 Communications Letters*, vol. 9, no. 5, pp. 716–720, 2020.

[8] M. Hua, L. Yang, Q. Wu, C. Pan, C. Li, and A. L. Swindlehurst, "Uav-assisted intelligent reflecting surface symbiotic radio system," ~~*arXiv preprint*~~ *arXiv:2007.14029*, 2020.

[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," *arXiv:2007.08380*, 2020.

[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 Transactions on Vehicular Technology*, pp. 1–1, 2021.

[11] D. Zhao, H. Lu, Y. Wang, H. Sun, Y. Gui, and J. Wu, "Joint power allocation and user association optimization for irs-assisted mmwave systems," *arXiv preprint arXiv:2010.11713*, 2020.

[12] S. K. Yong and J. S. Thompson, "A three-dimensional spatial fading correlation model for uniform rectangular arrays," *IEEE Antennas and Wireless Propagation Letters*, vol. 2, pp. 182–185, 2003.

[13] M. Zhao, Q. Wu, M. Zhao, and R. Zhang, "Two-timescale beamforming optimization for intelligent reflecting surface enhanced wireless network," pp. 1–5, 2020.

[14] 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 Transactions on Wireless Communications*, vol. 20, no. 1, pp. 375–388, 2021.

[15] 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 Transactions on Signal Processing*, vol. 68, pp. 5092–5106, 2020.

[16] 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 Transactions on Communications*, vol. 62, no. 1, pp. 59–73, 2014.

[17] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, 2015.

[18] X. Liu, Y. Liu, and Y. Chen, "Machine learning empowered trajectory and passive beamforming design in uav-ris wireless networks," *IEEE Journal on Selected Areas in Communications*, pp. 1–1, 2020.