

Joint Optimization of Trajectory and Jamming Power for Multiple UAV-Aided Proactive Eavesdropping

Delin Guo¹, Lan Tang¹, *Member, IEEE*, Xinggan Zhang², and Ying-Chang Liang³, *Fellow, IEEE*

Abstract—This paper studies a novel wireless information surveillance scenario, where the legitimate party aims to eavesdrop on multiple suspicious communication links with the help of multiple unmanned aerial vehicles (UAVs). Each suspicious link is comprised of a UAV (transmitter) and its fixed destination. To improve the eavesdropping ability, cooperative legitimate UAVs emit jamming signals to reduce the capacities of suspicious channels and plan the flight trajectory to enhance the capacity of the eavesdropping channels. Considering the system dynamics, it is natural to model this sequential decision-making problem as a Markov Decision Process (MDP), which might be solved by reinforcement learning (RL). However, it is difficult to design a policy in RL that determines jamming powers satisfying the considered eavesdropping constraints. Therefore, we decompose the optimization process into two phases, 1) obtaining the non-learning-based optimal solver for jamming power allocation under each state, and 2) optimizing the policy of moving action by RL. We will show this decoupled optimization process also holds the optimality. Considering the flying safety, we will determine the individual moving policy for each legitimate UAV rather than a centralized policy that controls all UAVs. Finally, extensive simulations are conducted to demonstrate the effectiveness of the proposed solution.

Index Terms—Proactive eavesdropping, trajectory planning, power allocation, reinforcement learning (RL), unmanned aerial vehicles (UAVs).

I. INTRODUCTION

A. Backgrounds

WIRELESS communication has been providing an efficient and convenient way for connecting people and devices and therefore playing an important role in our social life. However, the wireless communication also could be utilized by malicious users to facilitate their illegal activities. Especially in recent years, with the advancement of infrastructure-free communications, suspicious information may not go through any core infrastructure, such that it is hard to be monitored by

conventional infrastructure-based surveillance methods [1]. To improve the public safety, some authorized parties (e.g., government agencies) have strong need to monitor the suspicious communication links, e.g., the National Security Agency (NSA) of the United States has launched the Terrorist Surveillance Program to legitimately monitor wireless devices (<https://nsa.gov/1.info/surveillance/>). Therefore, the legitimate eavesdropping [2] on the physical layer has attracted widespread attention in recent years.

In wireless surveillance, a straightforward way is passive eavesdropping, which deploys a silent eavesdropper to only listen to the suspicious wireless link and then decode the received messages.¹ However, this simple method is competent only when the eavesdropping channel (i.e., the channel from the suspicious source to the eavesdropper) is better than the suspicious channel (i.e., the channel from the suspicious source to its destination) so that the information sent by the suspicious source can be reliably decoded. Yet this condition does not hold in many practical scenarios, such as the cases where the eavesdropper is deployed far away from the suspicious link, and/or the suspicious transmit beam is narrowed due to the application of high-frequency bands and massive multiple-input multiple-output (MIMO). Therefore, to improve eavesdropping efficiency, proactive eavesdropping has started to attract a lot of research interest in recent years [2]. A popular method is proactive eavesdropping via cognitive jamming, which was first studied in [1] and [3], where a full-duplex (FD) eavesdropper was deployed to degrade the channel capacity of suspicious links by transmitting jamming signals, and thus eavesdropping efficiency could be improved. After these two pioneer works, a series of jamming-assisted eavesdropping schemes were studied (e.g., [4], [5], [6], [7], [8], [9]). Another proactive eavesdropping approach is to apply spoofing relay [10]. [11] utilized an FD monitor with simultaneous eavesdropping and spoofing relaying to vary the source transmission rate in favor of the eavesdropping performance. [12] maximized the eavesdropping rate by coordinating multiple spoofing relays, the central monitor, and a jammer. [13] enhanced the eavesdropping capability by deploying a spoofing relay to realize the beam misleading in a multi-input multi-output orthogonal frequency division multiplexing (MIMO-OFDM) system. Besides, a pilot-contamination-based

Manuscript received 1 August 2022; revised 15 June 2023; accepted 29 August 2023. Date of publication 4 September 2023; date of current version 4 April 2024. This work was supported by the National Natural Science Foundation of China under Grants U1936201, 62072229, and 61976113. Recommended for acceptance by M. Li. (*Corresponding author: Lan Tang.*)

Delin Guo, Lan Tang, and Xinggan Zhang are with the School of Electronic Science and Engineering, Nanjing University, Nanjing 210093, China (e-mail: guodelin@smail.nju.edu.cn; tanglan@nju.edu.cn; zhxxg@nju.edu.cn).

Ying-Chang Liang is with the Center for Intelligent Networking and Communications (CINC), University of Electronic Science and Technology of China (UESTC), Chengdu 611731, China, and also with Yangtze Delta Region Institute (Huzhou), University of Electronic Science and Technology of China, Huzhou 313001, China (e-mail: liangyc@ieee.org).

Digital Object Identifier 10.1109/TMC.2023.3311484

¹As for how to discriminate the legitimate and suspicious users, there are some possible methods introduced in [2], such as Sensitive Contents Identification, User Mobility Profiling, and Social Network Map Analysis. Besides, some means of criminal investigation can also be applied, which is typically implemented intelligence departments.

scheme was proposed by [14], which contaminated the pilot signals during the channel estimation phase to mislead the channel estimation of the suspicious link, and thus the capacity of the suspicious channel was reduced due to the incorrect channel information so as to enhance the eavesdropping efficiency.

The legitimate eavesdropping problems considered above assumed locations of nodes, including monitor, relay, and suspicious source/destination, were fixed. Recently, unmanned aerial vehicles (UAV)-enabled communication systems have been receiving widespread research interests from industry and academia [15], [16], [17]. In legitimate eavesdropping, the UAV-enabled eavesdropper also has been considered to be a powerful tool to enhance eavesdropping efficiency since it can dynamically move to better positions for eavesdropping, e.g., flying closer to the suspicious transmitter. [18] and [19] navigated a UAV team to cooperatively eavesdrop on the suspicious ground (moving or stationary) nodes, where the trajectories of UAVs were designed to maximize the disguising performance² while guarantee the wireless eavesdropping requirements and avoid UAV collisions. [20] designed the receiving beamformers and the trajectory of the UAV eavesdropper to maximize the eavesdropping rate in a MIMO-OFDM system. [21] investigated an eavesdropping and anti-eavesdropping game between a silent UAV eavesdropper and a UAV base station, where the existence and solution of the Nash equilibrium of this game were given.

The UAV-enabled eavesdroppers considered in [18], [19], [20], [21] were silent, i.e., they can not emit any signals to help the eavesdropping. To achieve more efficient eavesdropping, many works applied FD UAVs that can simultaneously send jamming signals and receive suspicious information. In [22], a legitimate FD UAV monitored a suspicious link comprised of a suspicious UAV transmitter and a UAV receiver, where the eavesdropping rate was maximized by jointly optimizing the jamming power and trajectory of the legitimate UAV. [23] investigated a similar system model to [22], while the legitimate UAV eavesdropper patrolled in the predetermined trajectory so only the jamming power allocation was considered. [24] and [25] studied an eavesdropping-aided UAV tracking problem, where a legitimate UAV tracked a suspicious UAV pair for preventing potential crimes. The jamming-assisted proactive eavesdropping was utilized to obtain some useful information and then the tracking accuracy was improved by leveraging the eavesdropped information (such as angle-of-arrival and received signal strength). In [26], a silent eavesdropper monitored multiple ground suspicious links with the help of jamming signals sent by a mobile UAV, where the jamming power and position of the UAV were optimized. [27] optimized the jamming power over multiple frequency bands and position of the legitimate UAV to maximize the eavesdropping probability in a system with multiple suspicious sources and a suspicious destination. [28] applied a legitimate UAV eavesdropper with solar energy harvesting to monitor ground nodes, where the trajectory and jamming power were jointly designed to minimize the energy consumption while satisfying the eavesdropping performance.

²A metric proposed in [18] to measure the extent to which eavesdropping UAVs were visually noticed by suspicious nodes

B. Motivations and Contributions

As presented above, the UAV-enabled eavesdroppers have been recognized as a promising tool in legitimate eavesdropping were extensively studied [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28]. However, most of the existing works have some limitations. For example, [18], [19], [20], [21] only considered silent UAV eavesdroppers without jamming assistance. While in such cases, the eavesdropping UAV is very likely to improve the eavesdropping performance by moving close to the suspicious source or destination, which may enhance the probability of the eavesdropping intention being detected by the malicious users. [20], [21], [22], [23], [24], [25], [26], [27], [28] only adopted one legitimate UAV, which would be incompetent in eavesdropping on multiple suspicious links. Besides, except for [19] and [21], other works assumed that the suspicious source and destination were in fixed locations (or at least the distance between the suspicious source and destination was fixed, i.e., relatively stationary). Since UAVs have been extensively applied as mobile air base stations to provide services for ground users [15], [16], [17], it is also possible that UAVs are used by malicious users to send suspicious information. Therefore, it is necessary and meaningful to investigate how to monitor moving UAV suspicious links by utilizing UAVs as legitimate eavesdroppers.

In this paper, we also focus on the UAV-enabled eavesdropping system. Based on the discussions above, it is necessary to consider a more general and scalable scenario that can overcome most of the above limitations. To this end, we consider a wireless information surveillance scenario in presence of multiple legitimate UAVs and multiple suspicious links, where the legitimate party aims to eavesdrop on the information of multiple suspicious links with the help of these legitimate FD UAVs, which can simultaneously act as silent eavesdroppers and send jamming signals. On the other hand, each suspicious link is comprised of a mobile UAV (acts as the source/transmitter) and the corresponding stationary ground destination. The suspicious UAV moves around a certain area and sends the suspicious information to its destination, e.g., illegal real-time video signals. We will jointly optimize the jamming powers and trajectories of the cooperative legitimate UAVs to improve the eavesdropping performance. The main contributions of this paper are summarized as follows:

- First, we study a novel wireless information monitoring scenario with multiple legitimate UAVs versus multiple suspicious UAV-Destination pairs. We improve the eavesdropping performance by jointly optimizing the jamming power and trajectory of the cooperative legitimate UAV team. To the best of our knowledge, this jamming-assisted multiple UAV eavesdropping against multiple UAV-enabled suspicious links is first studied in this paper.
- Second, considering the unpredictable moving direction of suspicious UAVs, stochastic channel state, and the effects of the moving action on the next system state, we model this sequential decision-making problem in the dynamic system as a Markov decision process (MDP) problem, which might be solved by reinforcement learning (RL) [29]. Due to the difficulty of determining jamming power

satisfying the considered eavesdropping constraints in RL-based policy, we proposed to decompose the optimization process into two phases, 1) obtaining the non-learning-based optimal solver for jamming power under each system state, and then 2) optimizing the policy of moving action by RL. The optimality of this decoupled optimization process is proved.

- Third, to enhance the performance of the proposed method and the control safety, we separately determine an individual policy for each legitimate UAV. The reasons for applying decentralized control are as follows: 1) In practical implementations, training a policy with large action space is more difficult than training several distributed policies with small action space [30]. 2) If legitimate UAVs are controlled by a centralized policy, once a legitimate UAV can not receive control signals due to the signal being blocked or suffering jamming from malicious attackers, then it may fall from the sky.
- Finally, comprehensive numerical results are provided to verify the effectiveness of the proposed solution. It is shown that with the help of jamming signals, we can effectively guarantee the eavesdropping rate and eavesdropping success rate on multiple suspicious links by only deploying fewer eavesdropping UAVs.

C. Related Works and Differences

Due to the strong need for some authorized institutions in protecting public security, legitimate interception in the physical layer has attracted more and more attention [2]. To further improve the eavesdropping efficiency, many proactive eavesdropping approaches were proposed, by utilizing cognitive jamming [1], [3], [4], [5], [6], [7], [8], [9], spoofing relaying [11], [12], [13], [31], [32], and pilot contamination [14], [33], [34]. The above works were studied by assuming that all nodes are fixed in certain positions. While recently, due to the additional degree of freedom brought by the flexible deployment of UAVs, UAV-enabled communication systems were regarded as a prospective way to boost communication performance, and thus an increasing number of UAVs were deployed to provide communication services [15], [16], [17], [35], [36], [37]. Also, since UAVs can dynamically move to some better positions to wiretap, the UAV eavesdropper has been recognized as a powerful tool in eavesdropping, which has been extensively studied in [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28].

As we presented in Section I-B, we consider a system that adopts multiple UAV eavesdroppers to eavesdrop on multiple suspicious links, each of which is comprised of a suspicious UAV source/transmitter and its corresponding fixed destination. The most similar works to ours were given in [18], [19] and [26]. Both [18] and [19] considered using multiple UAVs to eavesdrop on multiple suspicious targets, but the adopted UAV eavesdroppers were passive so only the UAV trajectories needed to be optimized, which is different from ours. Besides, the main objectives of the two works were to optimize the disguising performance, while the eavesdropping performance was only

taken as a constraint. In [26], the eavesdropping on multiple suspicious links was implemented with the cooperation of a fixed silent eavesdropper and a mobile jamming-assisted UAV. The jamming power and position of the mobile UAV were jointly optimized to help the silent eavesdropper. The main differences from our work are that they only deployed one legitimate UAV for eavesdropping and all the suspicious nodes were in fixed locations. To be more intuitive, we have further listed the differences between our work and some relevant literature about UAV-enabled eavesdropping in Table I.

The remainder of this paper is organized as follows. We first give some necessary backgrounds of RL in Section II. The system model is elaborated in Section III. The problem interests are formulated in Section IV. In Section V, we present the proposed jamming power allocation and trajectory planning algorithm. Numerical results are conducted in Section VI. Finally, we conclude this work in Section VII.

II. PRELIMINARY

Before we give the system model, we first introduce some preliminaries of RL. RL [29] studies the sequential decision making problem of an agent in a stochastic (not always) environment, which is typically modeled as a Markov decision process (MDP). A MDP can be defined by a tuple $\langle \mathcal{S}, \mathcal{A}, p, r \rangle$. At each time step t , $s_t \in \mathcal{S}$ describes the global state of the environment (or system), and the agent executes an action $a_t \in \mathcal{A}(s_t)$, where $\mathcal{A}(s_t)$ is the feasible action space under s_t . This incurs a transition in the environment according to the state transition distribution (which is typically unknown) $p(s_{t+1}|s_t, a_t)$. Then the agent will receive the reward $r(s_t, a_t)$ (or abbreviated as r_t).

Let $\gamma \in [0, 1)$ be the discount factor, which is typically close to 1. The discount return is $R_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k}$. Therefore, γ can be interpreted as a rate which measures the weight of future rewards. The agent make decisions using a stochastic policy $\pi : \mathcal{S} \rightarrow \mathcal{P}(\mathcal{A}(\mathcal{S}))$, i.e., $a_t \sim \pi(\cdot|s_t) \in \mathcal{A}(s_t)$, where $\mathcal{P}(\mathcal{A}(\mathcal{S}))$ is the set of probability measures on $\mathcal{A}(\mathcal{S})$. Given the policy π , the definitions of the state-action value function Q^π , the state value function V^π , the advantage function A^π , and the expected discount return are as follows:

$$\begin{aligned} Q^\pi(s, a) &= \mathbb{E}[R_t | s_t = s, a_t = a; \pi], \\ V^\pi(s) &= \mathbb{E}[R_t | s_t = s; \pi], \\ A^\pi(s, a) &= Q^\pi(s, a) - V^\pi(s), \\ \eta(\pi) &= \mathbb{E}_{s_0 \sim p_0} \{V^\pi(s_0)\} \end{aligned} \quad (1)$$

where $p_0(s)$ is the given distribution of initial state s_0 . The target of RL is to find the optimal policy that maximizes the expected discount return, i.e., $\pi_* = \arg \max_{\pi} \eta(\pi)$.

Multi-Agent RL: The multi-agent RL (MARL) studies the sequential decision problem of multi-agent (e.g., multiple UAVs in this paper) in a multi-agent MDP. Let $\mathcal{N} = \{1, \dots, N\}$ denote the agent index set. We use the similar terms as those defined above. For the agent $n \in \mathcal{N}$ at the time-step t , the action is $a_t^n \in \mathcal{A}^n(s_t)$, where $\mathcal{A}^n(s_t)$ is the individual action space. The joint action is defined as $a_t = (a_t^1, \dots, a_t^N)$, and the joint action space $\mathcal{A}(s_t) = \times_{n \in \mathcal{N}} \mathcal{A}^n(s_t)$. The individual policy of

TABLE I
COMPARISONS OF UAV-ENABLED EAVESDROPPING SCHEMES IN DIFFERENT WORKS

	Ours	[18]	[19]	[20]	[21]	[22]	[23]	[24]	[25]	[26]	[27]	[28]
Jamming assistance	✓	×	×	×	×	✓	✓	✓	✓	✓	✓	✓
Multiple suspicious sources	✓	✓	✓	×	×	×	×	×	×	✓	✓	×
Multiple suspicious destinations	✓	✓	✓	×	✓	×	×	×	×	✓	×	×
Moving suspicious target	✓	×	✓	×	✓	×	×	×	×	×	×	×
Multiple UAV eavesdroppers	✓	✓	✓	×	×	×	×	×	×	×	×	×

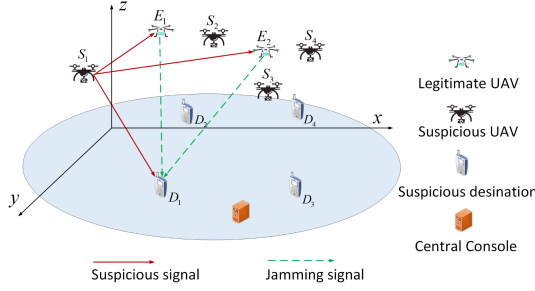


Fig. 1. Diagram of the multi-UAV eavesdropping system.

each agent is denoted as π^n , i.e., $a_t^n \sim \pi^n(\cdot|s_t)$, and the joint policy is the product of independent individual policies, i.e., $\pi(\cdot|s_t) = \prod_{n=1}^N \pi^n(\cdot|s_t)$. In cooperative tasks, all agents share the same reward function $r(s_t, a_t)$. The definitions of value functions are the same as that defined in (1). The target of MARL is to determine the individual policies that maximizes the expected discount return, i.e., $\max_{\{\pi^1, \dots, \pi^N\}} \eta(\pi)$.

III. SYSTEM MODEL

We consider a system comprised of N legitimate UAV eavesdroppers and M UAV-enabled suspicious links. The n th legitimate UAV eavesdropper is denoted by $E_n, n \in \mathcal{N} = \{1, 2, \dots, N\}$, and the m th suspicious link is comprised of a mobile suspicious UAV S_m and the corresponding stationary destination D_m , where $m \in \mathcal{M} = \{1, 2, \dots, M\}$. Each S_m flies over the area for some malicious tasks and sends suspicious information to its destinations D_m . We assume that suspicious UAVs communicate with the corresponding destinations in the frequency division multiple access (FDMA) manner, and each S_m occupies the same bandwidth. All legitimate UAVs can simultaneously receive suspicious information and send jamming signals to all suspicious destinations. They aim to cooperatively eavesdrop on these suspicious links by trajectory planning and sending jamming signals. A central console of the legitimate party could be used for decoding the information offloaded by all E_n s, and can also serve as a computer for computational tasks such as RL algorithms. The communications among the legitimate party (including UAVs and the central console) are assumed to operate in some authorized and secret channels, which will not be interfered with by suspicious UAVs. Besides, similar to [7], [11], [12], we assume the suspicious users are not aware of that they are being eavesdropped, and therefore no countermeasure are taken by them.

In Fig. 1, to be more intuitive, we provide an example of the considered system, where two legitimate UAVs are aiming

to eavesdrop on four suspicious links with given jamming and moving policies. We take the (S_1, D_1) pair for example. The suspicious UAV S_1 sends information to its destination D_1 . Due to the broadcast nature of radio signals, this suspicious signals can also be received by E_1 and E_2 . According to [1], only when the rate of joint decoding $C_{S_1}^E$ in (11) (to be defined later) is larger than the transmission rate on the suspicious channel C_1 in (11), it is possible that the information sent by the S_1 can be reliably decoded by eavesdrop party without any error. Therefore, to realize the eavesdropping under adverse monitoring condition, E_1 and E_2 send jamming signals to D_1 to force S_1 to reduce the transmission rate so that central console can decode information from received signals of E_1 and E_2 . We stress that each E_n can simultaneously send jamming signals to all D_m , but here we only plot the jamming signal to D_1 for concise. Also, the information stream among the legitimate party is not plotted for simplicity.

We assume that the system operates in an equal-length time-step fashion, where the time is slotted as $t = 1, 2, \dots$. The length of each time-step is denoted as ΔT . In each time slot, the system is assumed to be quasi-static. Next, we will elaborate on the system model in the following subsections.

A. Kinematic Model

At the time-step t , let $\mathbf{q}_{S_m}(t)$, $\mathbf{q}_{D_m}(t)$, and $\mathbf{q}_{E_n}(t)$ denote the positions of any S_m , D_m , and E_n , respectively. For any $m \in \mathcal{M}$ and $n \in \mathcal{N}$, $d_{S_m D_m}(t)$, $d_{S_m E_n}(t)$, and $d_{E_n D_m}(t)$ denote the distances from S_m to D_m , from S_m to E_n , and from E_n to D_m , respectively. Let $v_n(t) \in \mathcal{V}$ denote the velocity of E_n at the time-step t , where $\mathcal{V}$ is the feasible set of velocity. Then, the position dynamic of each E_n is given by

$$\mathbf{q}_{E_n}(t+1) = \mathbf{q}_{E_n}(t) + v_n(t)\Delta T. \quad (2)$$

The location $\mathbf{q}_{S_m}(t)$ could be obtained by overhearing the feedback location report signals sent by S_m , and also, the flying suspicious UAVs are typically visible, so the locations could directly be measured via a high resolution optical camera or synthetic aperture radar [38]. The locations of fixed destinations typically are the take-off point of suspicious UAVs, which also could be noticed. Therefore, we assume that $\mathbf{q}_{S_m}(t)$ and $\mathbf{q}_{D_m}(t)$ for all $m \in \mathcal{M}$ are available at the eavesdropping party. Besides, among legitimate UAVs, $\mathbf{q}_{E_n}(t), \forall n \in \mathcal{N}$, can easily be obtained from other teammates via information sharing in the secret channels.

B. Signal Model

In this subsection, we discuss the signal model in the time-step t . In this paper, we assume each E_n is equipped with

two sets of single-antenna, one for jamming and the other for receiving, and S_m and D_n are single-antenna. In our scheme, multiple single-antenna UAV eavesdroppers can be regarded as a distributed multi-antenna eavesdropper with a central console. Therefore, we would like to state that our model can be easily extended to the case of multi-antenna UAVs. Let x_m denote the information symbol sent by S_m and $P_0 = \mathbb{E}\{|x_m|^2\}$ be the fixed transmit power of all S_m . Since suspicious UAV communication pairs work in different frequency bands, E_n can interfere with communications of suspicious links by transmitting jamming signals with different powers in different frequency bands. Let $z_n^m(t)$ denote the jamming symbol sent by E_n to interfere with D_m . The jamming power is $p_n^m(t) = \mathbb{E}\{|z_n^m(t)|^2\}$, which satisfies $\sum_{m=1}^M p_n^m(t) \leq P_{\max}$, and $P_{\max}$ is the maximum jamming power of each E_n . For any $m \in \mathcal{M}$ and $n \in \mathcal{N}$, the channels from S_m to D_m , from S_m to E_n , and from E_n to D_m are denoted as $h_m(t)$, $h_{S_mE_n}(t)$, and $h_{E_nD_m}(t)$, respectively. The corresponding channel power gains are $\gamma_m(t) = |h_m(t)|^2$, $\gamma_{S_mE_n}(t) = |h_{S_mE_n}(t)|^2$, and $\gamma_{E_nD_m}(t) = |h_{E_nD_m}(t)|^2$, respectively.

The channel between any two UAVs is air-to-air, which is assumed to be dominated by the line-of-sight (LOS) path. Therefore, we model

$$\gamma_{S_mE_n}(t) = \alpha_0 d_{S_mD_n}^{-\beta}(t), \quad (3)$$

where α_0 is the constant path loss factor (i.e., the reference received power at 1 meter) and β is the path loss exponent. We assume that transmit signals are narrow band signals, so that all signals in different frequency bands approximately shares the same coefficient α_0 . The channels from S_m and E_n to D_m are the air-to-ground channels, which are typically comprised of the LOS and non-LOS paths. Therefore, the Rician fading channel model is adopted. The channel power gains between S_m (or E_n) and D_m are given by

$$\begin{aligned} \gamma_m(t) &= \alpha_m(t) d_{S_mD_m}^{-\beta}(t), \\ \gamma_{E_nD_m}(t) &= \alpha_{E_nD_m}(t) d_{E_nD_m}^{-\beta}(t), \end{aligned} \quad (4)$$

where $\alpha_m(t)$ and $\alpha_{E_nD_m}(t)$ are the fading factor in the Rician channel model, and they obey the probability density function (PDF) $f^R(x) = \frac{R+1}{\alpha_0} \exp(-R - (R+1)\frac{x}{\alpha_0}) I_0(2\sqrt{R(R+1)\frac{x}{\alpha_0}})$, where I_0 is the 0th order modified Bessel function of the first kind and R is the given Rician factor.

The transmitted signal x_m will be received by E_n through the channel $h_{S_mE_n}(t)$. Besides, due to the co-located transmit and receive antennas of FD eavesdropping UAVs, severe self-interference (SI) exists. Although some cancellation techniques have been proposed to relieve SI, e.g., [39] canceled the SI nearly to the receiver noise floor, we assume that the residual SI exists. The residual self-interference signal of any E_n in the m th frequency band is scaled by $\rho_n^m z_n^m(t)$, where ρ_n^m is the random residual factor with zero mean and $\mathbb{E}\{|\rho_n^m|^2\} = \sigma_\rho \ll 1$. The value of σ_ρ depends on the cancellation techniques adopted. As for the mutual-interference (MI) from other legitimate UAVs, according to [40], it can be eliminated by sharing the jamming

symbols among legitimate UAVs. Therefore, in the m th frequency band, the total signal received by E_n is

$$y_{S_mE_n}(t) = h_{S_mE_n}(t)x_m + \rho_n^m z_n^m(t) + n_{E_n}^m, \quad (5)$$

where $n_{E_n}^m$ is the Gaussian noise with zero mean and $\mathbb{E}\{|n_{E_n}^m|^2\} = N_0$.

We consider the delay-tolerant eavesdropping, where the signals received by each E_n are stored first, and then, after information collection, offloaded to the central console for centralized decoding. Therefore, the signals in the m th frequency band received by the eavesdropping party are

$$\mathbf{y}_{S_mE}(t) = \mathbf{h}_{S_mE}(t)x_m + \boldsymbol{\rho}^m \mathbf{z}^m(t) + \mathbf{n}^m, \quad (6)$$

where $\mathbf{y}_{S_mE}(t) = [y_{S_mE_1}(t), \dots, y_{S_mE_N}(t)]^T$, $\mathbf{h}_{S_mE}(t) = [h_{S_mE_1}(t), \dots, h_{S_mE_N}(t)]^T$, $\mathbf{z}^m(t) = [z_1^m(t), \dots, z_N^m(t)]^T$, $\boldsymbol{\rho}^m = \text{diag}(\rho_1^m, \dots, \rho_N^m)$, and $\mathbf{n}^m = [n_{E_1}^m, \dots, n_{E_N}^m]^T$.

By adopting Maximum ratio combining (MRC)³ to combine the received signals of N legitimate UAVs, the signal-to-interference-and-noise ratio (SINR) from S_m to the eavesdropping party is

$$\text{SINR}_{S_mE}^E(t) = \frac{\|\mathbf{h}_{S_mE}(t)\|^2 P_0}{\mathbf{w}^m(t) \mathbf{Q}_m(t) \mathbf{w}^m(t) + N_0}, \quad (7)$$

where $\mathbf{w}^m(t) = \frac{\mathbf{h}_{S_mE}^H(t)}{\|\mathbf{h}_{S_mE}(t)\|}$ is the MRC combiner and $\mathbf{Q}_m(t) = \mathbb{E}\{\boldsymbol{\rho}^m \mathbf{z}^m(t) \mathbf{z}^m(t)^H (\boldsymbol{\rho}^m)^H\}$. Equation (6) can be further transformed as the following compact form

$$\text{SINR}_{S_mE}^E(t) = \frac{\sum_{n=1}^N \gamma_{S_mE_n}(t) P_0}{N_0 + \boldsymbol{\Lambda}^m(t) \mathbf{p}^m(t)}, \quad (8)$$

where $\boldsymbol{\Lambda}^m(t) = \frac{\sigma_\rho [\gamma_{S_mE_1}(t), \dots, \gamma_{S_mE_N}(t)]}{\sum_{n=1}^N \gamma_{S_mE_n}(t)}$ and $\mathbf{p}^m(t) = [p_1^m(t), \dots, p_N^m(t)]^T$.

The signal received by D_m is

$$y_m(t) = h_m(t)x_m + \mathbf{h}_{ED_m}(t) \mathbf{z}^m(t) + n_{D_m}, \quad (9)$$

where $\mathbf{h}_{ED_m}(t) = [h_{E_1D_m}(t), \dots, h_{E_ND_m}(t)]$. The SINR at D_m is given by

$$\text{SINR}_m(t) = \frac{\gamma_m(t) P_0}{N_0 + \gamma_{ED_m}(t) \mathbf{p}^m(t)}, \quad (10)$$

where $\gamma_{ED_m}(t) = [\gamma_{E_1D_m}(t), \dots, \gamma_{E_ND_m}(t)]$.

Therefore, the channel capacity from S_m to D_m and from S_m to the eavesdropping party are

$$\begin{aligned} C_m(t) &= \log_2(1 + \text{SINR}_m(t)) \\ C_{S_mE}^E(t) &= \log_2(1 + \text{SINR}_{S_mE}^E(t)), \end{aligned} \quad (11)$$

respectively.

Channel Side Information Assumption: We assume that the suspicious UAV pairs operate in the time division duplex (TDD) mode. Due to the reciprocity of the channel in TDD systems, according to [11], we can obtain $h_{E_nD_m}(t)$ and $h_{S_mE_n}(t)$ by utilizing pilots sent by D_m and S_m during the channel training phase, respectively, and therefore $\gamma_{E_nD_m}(t)$ and $\gamma_{S_mE_n}(t)$ can

³ Although MRC is not optimal in terms of signal-to-noise ratio (SNR) when SI is present, it is a low complexity scheme.

be known. Also, the free-space path-loss $\gamma_{S_m E_n}(t)$ can be easily obtained since we can know the locations of all UAVs. $\gamma_m(t)$ can be obtained by overhearing the feedback information sent from D_m to S_m . Although $\gamma_m(t)$ is assumed to be available, we will discuss both the solutions when it is known and unknown. We assume that all node in the legitimate party (including UAVs and the central console) can share their individual information with others via the authorized and secret channels.

IV. PROBLEM FORMULATION

According to [1], if $C_{S_m}^E(t) \geq C_m(t)$, then the suspicious information can be decoded perfectly by the legitimate eavesdropping party, otherwise it can not be decoded without any error. Therefore, at each time-step t , the eavesdropping rate of all suspicious links is denoted by

$$r_t^J = \sum_{m=1}^M C_m(t) \cdot \mathbb{1}\{C_{S_m}^E(t) \geq C_m(t)\}, \quad (12)$$

where the indicator function $\mathbb{1}\{x\} = 1$ if the boolean value x is True, and 0 otherwise.

Besides, the collision avoidance of UAVs is also necessary. To this end, we design the following penalty function,

$$r_t^K = -\nu \sum_{n=1}^N \exp\left(-\left(\frac{d_{\min}^n(t)}{\sigma_d}\right)^2\right), \quad (13)$$

where $\nu \geq 0$ is the penalty factor, $\sigma_d > 0$ is a given parameter, and $d_{\min}^n(t)$ is the minimum distance between E_n and all other UAVs, i.e., $d_{\min}^n(t) = \min_{X \in \{S_m\}_{m=1}^M, \{E_i\}_{i \neq n}^N} \{d_{X E_n}(t)\}$. If $d_{\min}^n(t)$ is small, there will be a penalty, and if $\frac{d_{\min}^n(t)}{\sigma_d} \geq 2$, the n th penalty component basically diminishes. Therefore, the total reward is $r_t = r_t^J + r_t^K$.

The problem is to maximize the long run reward, which is given by

$$\max_{\{v(t), p(t)\}_{t=1}^T} \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T r_t \quad (14)$$

$$\text{s.t. : } v_n(t) \in \mathcal{V}, \forall n \in \mathcal{N}, \forall t \quad (14a)$$

$$p_n^m(t) \geq 0, \forall n \in \mathcal{N}, \forall m \in \mathcal{M}, \forall t, \quad (14b)$$

$$\sum_{m=1}^M p_n^m(t) \leq P_{\max}, \forall n \in \mathcal{N}, \forall t, \quad (14c)$$

$$C_{S_m}^E(t) \geq C_m(t), \forall m \in \mathcal{M}, \forall t, \quad (14d)$$

where $v(t) = (v_1(t), \dots, v_N(t))$, and $p(t) = (p^1(t), \dots, p^M(t))$. Alternatively, we can denote $p(t) = (p_1(t), \dots, p_N(t))$, where $p_n(t) = [p_n^1(t), \dots, p_n^M(t)]$.

The constraint (14d) is to guarantee the successful eavesdropping on all suspicious communication links. Since r_t^J is the summation of the eavesdropping rate of all suspicious links, when the objective in (14) is maximized, eavesdropping on some suspicious links may be failed, i.e., $C_m(t) > C_{S_m}^E(t)$ for some $m \in \mathcal{M}$. However, the suspicious UAVs are most likely to send the video signals to their destinations, which could be regarded

as the event-based scenario [1]. In such a case, the delivered suspicious messages (e.g., the video clips) in different fading states have the same significance to report and infer such series of events, although they may be with different data rates, e.g., different resolutions.

A. MDP Representation

Due to the Markov and stochastic properties of the system, including the unpredictable flying of suspicious UAVs, stochastic channel state, and the effects of the moving action on the next system state, we model this sequential decision-making problem (14) as an MDP introduced in Section II. Next, We give the definitions of state, action, and reward in detail.

State: The global state of the system should include sufficient information for making jamming and moving decisions. At the time step t , the global system state is denoted as

$$s_t = \left(\{p_{S_m}(t), p_{D_m}(t)\}_{m=1}^M, \{p_{E_n}(t)\}_{n=1}^N, \{\gamma_m(t)\}_{m=1}^M, \{\{\gamma_{E_n D_m}(t)\}_{n=1}^N\}_{m=1}^M \right). \quad (15)$$

Let $\mathcal{S}$ denote the state space, and therefore $s_t \in \mathcal{S}$. If the global state is available, we call the MDP fully observable, otherwise partial observable. Let o_t denote the observation at the time-step t . If it is fully observable, $o_t = s_t$, otherwise they are not equal.

Action: At the time-step t , the joint action is

$$a_t = (v(t), p(t)). \quad (16)$$

In each time-step t , the action a_t should satisfy the constraints (14a)–(14d). Especially, the intersection of constraints (14b)–(14d) can be regarded as the action space of jamming power $p(t)$. However, this intersection may be an empty set for some s_t , which implies that the current locations are not proper eavesdropping places, so that UAVs cannot realize perfect monitoring even with the aid of jamming signals. In this case, we assume that eavesdropping UAVs will keep silence, i.e., $p_n^m(t) = 0, \forall m, n$.

Denote the intersection of (14b)–(14d) at the time-step t as $\tilde{\mathcal{A}}^J(s_t)$. Then the action space of $p(t)$ under s_t is defined as

$$\mathcal{A}^J(s_t) = \begin{cases} \tilde{\mathcal{A}}^J(s_t), & \text{if } \tilde{\mathcal{A}}^J(s_t) \neq \emptyset \\ 0, & \text{otherwise.} \end{cases} \quad (17)$$

Besides, the space of moving action, constrained by (14a), is denoted as $\mathcal{A}^K = \times_{n \in \mathcal{N}} \mathcal{V}$, which has no relevant with the current state s_t .

Therefore, the joint action space under the state s_t is $\mathcal{A}(s_t) = \mathcal{A}^J(s_t) \times \mathcal{A}^K$, and $a_t \in \mathcal{A}(s_t)$. We denote $a_t^K = v(t) \in \mathcal{A}^K$ as the kinematic action, and $a_t^J = p(t) \in \mathcal{A}^J(s_t)$ as the jamming action. Based on the action space formulation, the constraints (14a)–(14d) are represented by the action space constraints in the later analysis.

Reward: The reward function has been defined as $r_t = r_t^J + r_t^K$. With the definitions of state and action in this section, and noting that the eavesdropping reward is only affected by the jamming action, and the distance penalty is only decided by the current positions of UAVs (the position information is included

in s_t), we can also express the reward as

$$r(s_t, a_t) = r^J(s_t, a_t^J) + r^K(s_t). \quad (18)$$

These representations might be used interchangeably in the following contents.

B. Problem Reformulation

Define the policy $\pi : \mathcal{S} \rightarrow \mathcal{P}(\mathcal{A}(\mathcal{S}))$. In each time-step t , the eavesdropping party choose actions according to $a_t \sim \pi(\cdot|s_t) \in \mathcal{A}(s_t)$. With the given π , the problem (14) is represented by

$$\max_{\pi} J(\pi) = \lim_{T \rightarrow \infty} \mathbb{E} \left\{ \frac{1}{T} \sum_{t=1}^T r_t | \pi \right\} \quad (19)$$

Instead, we will maximize the expected discounted return,

$$\max_{\pi} \eta(\pi) \quad (20)$$

Next, we will explain why we do not directly maximize the original objective $J(\pi)$, but $\eta(\pi)$. First, $J(\pi)$ is well proportionally approximated by $\eta(\pi)$. According to the chapter 10.4 in [29], we have that $\eta(\pi) \approx \frac{1}{1-\gamma} J(\pi)$, where when γ approaches to 1, the approximation can asymptotically be replaced by the equal sign. Therefore, in the practice, we can select a γ which is close to 1 (such as 0.999) to make (20) a good approximation. Specially, if $p_0(s) = \mu_{\pi}(s)$, where $\mu_{\pi}(s) = \lim_{t \rightarrow \infty} \Pr(s_t = s | \pi)$ is the steady-state distribution under π , then $\eta(\pi) = \frac{1}{1-\gamma} J(\pi)$ [29]. Second, although there exist some RL algorithms which can directly solve the averaged reward, the convergence guarantees of these algorithms are hard to establish. For example, the convergence of R-learning [41], which is probably the most well-known RL algorithm based on the averaged objective, is not proved. Therefore, almost all popular RL algorithms and their convergences are established on the discounted objective that has become the standard objective in RL nowadays.

V. SOLUTION TO PROBLEM (20)

Based on the discussions in the previous section, the original problem (14) is reformulated into the problem (20). However, due to the existence of the constraint (14d), it is difficult to design a policy π (typically parameterized by some type of neural network) in RL that can choose the action a_t from the action space $\mathcal{A}(s_t)$ defined in the last section. So the problem can hardly be solved by only applying RL algorithms.

For the problem in (20), we have the following observations: 1) under the quasi-static assumption of the system, the moving actions a_t^K do not affect the current eavesdropping reward r_t^J but the next state and the next distance penalty r_{t+1}^K , 2) the jamming action a_t^J will not affect the future state, 3) the action space $\mathcal{A}^K$ is fixed. Inspired by the above properties, we design a framework for solving the problem (20), which decomposes the solution into the two following steps.

First, we are going to find an optimal solver which can maximize the eavesdropping reward r^J under any state $s \in \mathcal{S}$, i.e., solving the following problem for all $s \in \mathcal{S}$,

$$\max_{a^J \in \mathcal{A}^J(s)} r^J(s, a^J). \quad (21)$$

To be consistent with the representation of policy in (20), we denote the optimal solver of problem (21) as $\hat{\pi}^J$, which satisfies $\sum_{a^J \in \mathcal{A}^J(s)} \hat{\pi}^J(a^J|s) r^J(s, a^J) \geq \sum_{a^J \in \mathcal{A}^J(s)} \pi^J(a^J|s) r^J(s, a^J)$ for all $s \in \mathcal{S}$ and $\forall \pi^J : \mathcal{S} \rightarrow \mathcal{P}(\mathcal{A}^J(\mathcal{S}))$.

Second, given the optimal solver $\hat{\pi}^J$, we optimize $\pi^K : \mathcal{S} \rightarrow \mathcal{P}(\mathcal{A}^K)$. Let $\hat{\pi}^K$ denote the optimal policy of moving actions with the given $\hat{\pi}^J$, i.e., $\hat{\pi}^K = \arg \max_{\pi^K} \eta(\hat{\pi}^J, \pi^K)$.

Proposition 1. $\hat{\pi} = \hat{\pi}^J \times \hat{\pi}^K$ is the optimal policy. That is, $\eta(\hat{\pi}) \geq \eta(\pi), \forall \pi$.

Proof. Let π be an arbitrary policy. $\pi^J(a^J|s) = \sum_{a^K \in \mathcal{A}^K} \pi(a|s)$ and $\pi^K(a^K|s) = \sum_{a^J \in \mathcal{A}^J(s)} \pi(a|s)$ are the marginal distributions of π . Note that we do not assume $\pi(a|s) = \pi^J(a^J|s) \cdot \pi^K(a^K|s)$.

According to Bellman's equation, we have

$$\begin{aligned} \eta(\hat{\pi}) &= \mathbb{E}_{s \sim p_0} \left\{ \sum_a \hat{\pi}(a|s) [r(s, a) + \gamma \sum_{s'} p(s'|s, a) V^{\hat{\pi}}(s')] \right\} \\ &\stackrel{(a)}{=} \mathbb{E}_{s \sim p_0} \left\{ \sum_{a^J} \hat{\pi}^J(a^J|s) r^J(s, a^J) + r^K(s) \right. \\ &\quad \left. + \gamma \sum_{a^K} \hat{\pi}^K(a^K|s) \sum_{s'} p(s'|s, a^K) V^{\hat{\pi}}(s') \right\} \\ &\stackrel{(b)}{\geq} \mathbb{E}_{s \sim p_0} \left\{ \sum_{a^J} \hat{\pi}^J(a^J|s) r^J(s, a^J) + r^K(s) \right. \\ &\quad \left. + \gamma \sum_{a^K} \pi^K(a^K|s) \sum_{s'} p(s'|s, a^K) V^{\hat{\pi}^J, \pi^K}(s') \right\} \\ &\stackrel{(c)}{\geq} \mathbb{E}_{s \sim p_0} \left\{ \sum_{a^J} \pi^J(a^J|s) r^J(s, a^J) + r^K(s) \right. \\ &\quad \left. + \gamma \sum_{a^K} \pi^K(a^K|s) \sum_{s'} p(s'|s, a^K) V^{\hat{\pi}^J, \pi^K}(s') \right\} \\ &\stackrel{(d)}{\geq} \eta(\pi), \end{aligned} \quad (22)$$

The equality (a) holds due to the previous observations 1) and 2). The inequality (b) holds because according to the definition of $\hat{\pi}^K$, for all π^K , $\eta(\hat{\pi}^J, \hat{\pi}^K) \geq \eta(\hat{\pi}^J, \pi^K)$. The inequality (c) holds because $\hat{\pi}$ is optimal for the problem (21). Finally, by expanding $V^{\hat{\pi}^J, \pi^K}$ in the right side of inequality (c), we have

$$\begin{aligned} V^{\hat{\pi}^J, \pi^K}(s) &= \sum_{a^J} \hat{\pi}^J(a^J|s) r^J(s, a^J) + r^K(s) \\ &\quad + \gamma \sum_{a^K} \pi^K(a^K|s) \sum_{s'} p(s'|s, a^K) V^{\hat{\pi}^J, \pi^K}(s') \\ &\geq \sum_{a^J} \pi^J(a^J|s) r^J(s, a^J) + r^K(s) \\ &\quad + \gamma \sum_{a^K} \pi^K(a^K|s) \sum_{s'} p(s'|s, a^K) V^{\hat{\pi}^J, \pi^K}(s') \\ &\geq \text{unrolling} \cdots \geq V^{\pi^J, \pi^K}(s) = V^{\pi}(s), \end{aligned} \quad (23)$$

Algorithm 1: General Framework for Solving (20).

```

1 Initialize  $\pi^K$ ;
2 for iteration = 1, 2, ... do
3    $\mathcal{D} \leftarrow \text{GenerateData}(\pi^K)$ ;
4    $\pi^K \leftarrow \text{RLProcess}(\mathcal{D})$ ;
5 end
6 function GenerateData( $\pi^K$ ):
7   Initialize a memory buffer  $\mathcal{D}$ ;
8   for  $t = 1, 2, \dots, T_l$  do
9     Solve Problem (21) to obtain  $a_t^J$ ;
10    Execute  $a_t^J$  and then receive  $r_t$ ;
11    Choose action  $a_t^K$  according to  $\pi^K(\cdot|s_t)$ ;
12    System moves to the next state  $s_{t+1}$ ;
13    Store the transition  $(s_t, a_t^K, r_t)$  into  $\mathcal{D}$ ;
14   end
15 return  $\mathcal{D}$ 

```

where “unrolling” means to iteratively expand the value function $V^{\hat{\pi}^J, \pi^K}$ using Bellman’s equation. Therefore, the last inequality (d) holds. This completes the proof.

Proposition 1 indicates the optimality of the proposed solution that sequentially optimizes the jamming action policy π^J and moving action policy π^K . Based on this proposition, we design a general framework for solving the problem (20), which is given in Algorithm 1. In each time step t of the GenerateData, we solve the problem (21) to get the jamming action a_t^J and execute it. Then the data $\{s_t, a_t^K, r_t\}_{t=1}^{T_l}$ is used to update π^K using RL (i.e., RLProcess function), where T_l is the length of each sample episode. In the next two subsections, we will first introduce how to solve the sub-problem (21), and then provide the practical RL process (i.e., the function RLProcess in Algorithm 1) and the whole algorithm in detail.

A. Solution to Subproblem (21)

In this subsection, we provide the solution to the problem (21), i.e., finding the optimal solver $\hat{\pi}^J$. As discussed in (17), when $\tilde{\mathcal{A}}^J(s_t)$ is an empty set, $\mathcal{A}^J(s_t)$ is a singleton. Therefore, we only need to optimize this problem when $\tilde{\mathcal{A}}^J(s_t)$ is not empty. In such a case, the problem (21) is re-written as

$$\max_{\mathbf{p}(t)} \sum_{m=1}^M C_m(t) \quad (24)$$

$$\text{s.t.} : p_n^m(t) \geq 0, \forall n \in \mathcal{N}, \forall m \in \mathcal{M}, \quad (24a)$$

$$\sum_{m=1}^M p_n^m(t) \leq P_{\max}, \forall n \in \mathcal{N}, \quad (24b)$$

$$C_{S_m}^E(t) \geq C_m(t), \forall m \in \mathcal{M}, \quad (24c)$$

where $\mathbb{1}\{C_{S_m}^E(t) \geq C_m(t)\}$ in the objective is omitted since it is 1 when the problem (24) is feasible.

With some rearranging, the constraint (24c) is transformed to

$$(\gamma_{ED_m}(t) \sum_{n=1}^N \gamma_{S_m E_n}(t) - \gamma_m(t) \Lambda^m(t)) \mathbf{p}^m(t)$$

$$\geq (\gamma_m(t) - \sum_{n=1}^N \gamma_{S_m E_n}(t)) N_0, \forall m \in \mathcal{M}. \quad (25)$$

Therefore, all constraints in (24) are linear inequality constraints.

Denote the objective function in (24) as $\Phi(\mathbf{p}(t)) = \sum_{m=1}^M C_m(t)$. By applying the *restriction to a line rule* (see 8.2.2.3 in [42]), it can be easily proved that $\Phi(\mathbf{p}(t))$ is convex. Due to the convexity of $\Phi(\mathbf{p}(t))$, we can not solve this problem by standard convex optimization methods. Using the fact that the first-order Taylor approximation of a convex function is a global under-estimator, $\Phi(\mathbf{p}(t))$ at the given initial point $\mathbf{p}^{(i)}(t) \geq 0$ is lower-bounded by

$$\begin{aligned} \Phi(\mathbf{p}(t)) &\geq \Phi^{lb}(\mathbf{p}(t); \mathbf{p}^{(i)}(t)) \\ &= \Phi(\mathbf{p}^{(i)}(t)) + \Psi(\mathbf{p}^{(i)}(t)) \cdot (\mathbf{p}(t) - \mathbf{p}^{(i)}(t)) \end{aligned} \quad (26)$$

where $\Psi(\mathbf{p}^{(i)}(t)) = \frac{\partial \Phi(\mathbf{p}(t))}{\partial \mathbf{p}(t)}|_{\mathbf{p}(t)=\mathbf{p}^{(i)}(t)}$.

Based on the above analysis, the problem in (24) can be efficiently approximated as the following linear programming problem:

$$\begin{aligned} \max_{\mathbf{p}(t)} \quad & \Phi^{lb}(\mathbf{p}(t); \mathbf{p}^{(i)}(t)) \\ \text{s.t.} : \quad & (24a) - (24c), \end{aligned} \quad (27)$$

which can be easily solved.

However, the linear function $\Phi^{lb}(\mathbf{p}(t); \mathbf{p}^{(i)}(t))$ is just a local approximation to $\Phi(\mathbf{p}(t))$. We need to approach the real optimal value by successively solving the problem (27). The flows are as follows: 1) Given a point $\mathbf{p}^{(i)}(t)$ and then solving the problem (27) to get the solution $\mathbf{p}^*(t)$. 2) If $|\Phi(\mathbf{p}^*(t)) - \Phi(\mathbf{p}^{(i)}(t))| < \sigma$, where σ is a small threshold for stop criterion, then $\mathbf{p}^*(t)$ is the optimal solution. Otherwise we set $\mathbf{p}^{(i)}(t) = \alpha \mathbf{p}^*(t) + (1 - \alpha) \mathbf{p}^{(i-1)}(t)$, $\alpha \in (0, 1)$ and go back to the step 1) for a new round optimizing. According to Theorem 1 in [43], this successive convex approximation (SCA) approach will converge to a Karush–Kuhn–Tucker (KKT) point.

Partial Observable: In some situations, the eavesdropping party may not be able to obtain the instantaneous channel power gain $\gamma_m(t)$, but only its statistical distribution. In this case, we aim to obtain the solution that is robust against the uncertainty.

Let us start from the constraint (24c). According to Chapter 10.3.4 in [42], we can seek the probabilistically robust solution which satisfies $\Pr(C_{S_m}^E(t) \geq C_m(t)) \geq \epsilon_c$, where ϵ_c is a desired level of probabilistic robustness of the constraint (24c). Let $\gamma_m^{\epsilon_c}$ satisfy $\Pr(\gamma_m^{\epsilon_c} \geq \gamma_m(t)) = \epsilon_c$. Then we define

$$C_m^{\epsilon_c}(t) = \log_2 \left(1 + \frac{\gamma_m^{\epsilon_c} P_0}{N_0 + \gamma_{ED_m}(t) \mathbf{p}^m(t)} \right). \quad (28)$$

Replacing $C_m(t)$ with $C_m^{\epsilon_c}(t)$ in the constraint (24c), if the modified constraint $C_{S_m}^E(t) \geq C_m^{\epsilon_c}(t)$ holds, then $\Pr(C_{S_m}^E(t) \geq C_m(t)) \geq \epsilon_c$, such that the ϵ_c -level probabilistic robustness is obtained.

As for the objective (24), we could maximize its lower bound. However, the lower bound of $C_m(t)$ is zero since $\gamma_m(t) \geq 0$. Therefore, similar to the probabilistic robustness in the constraint (24c), we turn to maximizing the following probabilistic

lower bound

$$C_m^{\epsilon_o}(t) = \log_2 \left(1 + \frac{\gamma_m^{\epsilon_o} P_0}{N_0 + \gamma_{ED_m}(t) p^m(t)} \right), \quad (29)$$

where $\Pr(\gamma_m(t) \geq \gamma_m^{\epsilon_o}) = \epsilon_o$, and thus $\Pr(C_m(t) \geq C_m^{\epsilon_o}(t)) = \epsilon_o$.

By replacing the objective in (24) with $\sum_{m=1}^M C_m^{\epsilon_o}(t)$ and the constraint (24c) with $C_{S_m}^E(t) \geq C_m^{\epsilon_o}(t)$, we can solve the modified problem using the same method as when $\{\gamma_m(t)\}_{m=1}^M$ are known.

B. RL Process

In this subsection, we will detail the update rules in RLProcess in Algorithm 1. That is, with the given $\hat{\pi}^J$, we are going to solve $\max_{\pi^K} \eta(\hat{\pi}^J, \pi^K)$. We rewrite $\eta(\hat{\pi}^J, \pi^K)$ as $\eta(\pi^K)$ in this subsection since $\hat{\pi}^J$ is given.

The most straightforward way is to apply a common single-agent RL algorithm, such as proximal policy optimization (PPO) [44] or trust region policy optimization (TRPO) [45], to learn a centralized policy which makes decisions for all agents (legitimate eavesdropping UAVs). However, we assume that agents make moving decisions in a decentralized manner, i.e., making decisions according to their individual policy. More clearly, in each time-step t , the individual moving action of the agent E_n is $a_t^{K,n} = v_n(t)$, and the joint action $a_t^K = (a_t^{K,1}, \dots, a_t^{K,N})$. The joint policy π^K is decomposed to $\pi^K(a_t^K | s_t) = \prod_{n=1}^N \pi^{K,n}(a_t^{K,n} | s_t)$.

The reasons for applying the decentralized control are as follows: 1) In the practical implementations, training a policy with large action space is probably more difficult than training several decomposed policies with small action space [30]. 2) If legitimate UAVs are controlled by the command signal from a central console (centralized policy), once some legitimate UAV can not receive control signals due to the signal blocked or suffering jamming from malicious attackers, they may fall from the sky.

In this paper, the multi-agent PPO (MAPPO) [46] will be adopted to learn the decentralized moving policies, which is an advanced MARL algorithm developed from the single-agent PPO and verified to have excellent learning performances. MAPPO restrains the KL divergence of policy by clipping the policy ratio, so that some aggressive update steps would be removed, and thus the learning performance and stability could be improved. Next, we will introduce MAPPO in detail to keep the self-contained of this paper. According to [45], let π^K be the current policy, and $\tilde{\pi}^K$ denotes an any candidate policy,

$$\begin{aligned} \eta(\tilde{\pi}^K) &= \eta(\pi^K) + \sum_s \rho_{\tilde{\pi}^K}(s) \sum_{a^K} \tilde{\pi}^K(a^K | s) A^{\pi^K}(s, a^K) \\ &= \eta(\pi^K) + \mathbb{E}_{\tilde{\pi}} \{A^{\pi^K}(s, a)\}, \end{aligned} \quad (30)$$

where $\rho_{\tilde{\pi}^K}(s) = \sum_{t=0}^{\infty} \gamma^t \Pr(s_t = s | \tilde{\pi}^K)$ is the (unnormalized) discounted state distribution and $\mathbb{E}_{\tilde{\pi}} \{\dots\} = \sum_s \rho_{\tilde{\pi}^K}(s) \sum_a \tilde{\pi}^K(a^K | s) [\dots]$. If we aim to improve $\eta(\pi^K)$ by making $\eta(\tilde{\pi}^K) \geq \eta(\pi^K)$, it should be satisfied that $\sum_s \rho_{\tilde{\pi}^K}(s) \sum_a \tilde{\pi}^K(a^K | s) A^{\pi^K}(s, a) \geq 0$.

However, due to the implicit form of $\rho_{\tilde{\pi}^K}(s)$, it is difficult to optimize the (30) directly. Besides, we want to train decentralized policies for each legitimate UAV by themselves. Therefore, we construct the following approximation to $\eta(\tilde{\pi}^K)$,

$$L_{\pi^K}(\tilde{\pi}^K) = \eta(\pi^K) + \sum_{n=1}^N \mathbb{E}_{\pi^K} \left\{ \frac{\tilde{\pi}^{K,n}(a^{K,n} | s)}{\pi^{K,n}(a^{K,n} | s)} A^{\pi^K}(s, a^K) \right\}, \quad (31)$$

where the joint policy is decoupled to summation of individual policies and the expectation is taken over the given current policy π^K . The following theorem reveals the relationship between the approximation $L_{\pi^K}(\tilde{\pi}^K)$ and the origin $\eta(\tilde{\pi}^K)$.

Theorem 1 ([46]). : Let $D_{\text{KL}}(\cdot || \cdot)$ denote the Kullback-Leibler (KL) divergence, and $D_{\text{KL}}^{\max}(\pi^{K,n} || \tilde{\pi}^{K,n}) = \max_{s \in \mathcal{S}} D_{\text{KL}}(\pi^{K,n}(\cdot | s) || \tilde{\pi}^{K,n}(\cdot | s))$. It holds that

$$\eta(\tilde{\pi}^K) \geq L_{\pi^K}(\tilde{\pi}^K) - C \max_n \sqrt{D_{\text{KL}}^{\max}(\pi^{K,n} || \tilde{\pi}^{K,n})}, \quad (32)$$

where $C = \frac{4N\epsilon}{(1-\gamma)}$ and $\epsilon = \max_{s,a} |A^{\pi^K}(s, a^K)|$.

We define that

$$M_{\pi^K}(\tilde{\pi}^K) = L_{\pi^K}(\tilde{\pi}^K) - C \max_n \sqrt{D_{\text{KL}}^{\max}(\pi^{K,n} || \tilde{\pi}^{K,n})}. \quad (33)$$

Since $\eta(\tilde{\pi}^K) \geq M_{\pi^K}(\tilde{\pi}^K)$ and $\eta(\pi^K) = M_{\pi^K}(\pi^K)$, if we can guarantee that $M_{\pi^K}(\tilde{\pi}^K) \geq M_{\pi^K}(\pi^K)$, then $\eta(\tilde{\pi}^K)$ is also improved. However, maximizing $M_{\pi^K}(\tilde{\pi}^K)$ is also difficult since the penalty coefficient C is large (γ is typically close to 1), which leads to prohibitively update small steps. Therefore, in order to take larger update steps in a robust way, we can convert the penalty term to the constraints. By expanding $L_{\pi^K}(\tilde{\pi}^K)$, the problem of maximizing $M_{\pi^K}(\tilde{\pi}^K)$ is approximated as

$$\max_{\{\pi^{K,n}\}_{n=1}^N} \mathbb{E}_{\pi^K} \left\{ \sum_{n=1}^N \frac{\tilde{\pi}^{K,n}(a^{K,n} | s)}{\pi^{K,n}(a^{K,n} | s)} A^{\pi^K}(s, a^K) \right\} \quad (34a)$$

$$\text{subject to : } D_{\text{KL}}^{\max}(\pi^{K,n} || \tilde{\pi}^{K,n}) \leq \delta, \forall n \in \mathcal{N}. \quad (34b)$$

where δ is called the trust region.

However, the KL divergence in (34b) is bounded at every point in the state space, which makes the problem difficult to be solved due to the large number of constraints. In order to make the optimization process more simple, the clipped objective is adopted in MAPPO, which removes the KL-penalty by clipping the policy ratio,

$$L_{\pi^K}^{\text{clip}}(\tilde{\pi}^K) = \sum_{n=1}^N \mathbb{E}_{\pi^K} \{f(r(\tilde{\pi}^{K,n}), A^{\pi^K}(s, a^K))\}, \quad (35)$$

where

$$\begin{aligned} f(r(\tilde{\pi}^{K,n}), A^{\pi^K}(s, a^K)) &= \\ \min\{r(\tilde{\pi}^{K,n}) A^{\pi^K}(s, a^K), c^\varepsilon(r(\tilde{\pi}^{K,n}) A^{\pi^K}(s, a^K))\} \end{aligned} \quad (36)$$

where $c^\varepsilon(x)$ is a function that clips x into the interval $[1 - \varepsilon, 1 + \varepsilon]$, and $r(\tilde{\pi}^{K,n}) = \frac{\tilde{\pi}^{K,n}(a^{K,n} | s)}{\pi^{K,n}(a^{K,n} | s)}$. As the (35) implies, when the policy ratio $r(\tilde{\pi}^{K,n})$ is out of the clip range, the gradient moving the policy outside of the clip interval will be zero. Therefore, the KL divergence could be restrained.

Based on the surrogate objective (35), we will give the practical algorithm for updating decentralized policies $\tilde{\pi}^{K,n}, \forall n$. In the deep reinforcement learning, the policy and value function are typically parameterized to tractable functions using deep neural networks (DNN). In this work, we parameterize $\tilde{\pi}^{K,n}$ and $\pi^{K,n}$ with parameters $\tilde{\theta}^{K,n}$ and $\theta^{K,n}$, respectively, for all $n \in \mathcal{N}$. The state-value function V^{π^K} is parameterized as V_ϕ with parameters ϕ .

For a sampled data segment $\tau = \{s_t, a_t^K, r_t\}_{t=1}^{T_l}$ where $a_t^K = (a_t^{K,1}, \dots, a_t^{K,N})$, $A^{\pi^K}(s_t, a_t^K)$ is estimated by the following generalized advantage estimator (GAE) in [47], which is given by $\hat{A}_t^K = \delta_t + (\gamma\lambda)\delta_{t+1} + \dots + (\gamma\lambda)^{T_l-t}\delta_{T_l}$, where $\delta_t = r_t + \gamma V_\phi(s_{t+1}) - V_\phi(s_t)$, and $\lambda \in [0, 1]$ is a hyperparameter that balances the bias and variance. For example, when $\lambda = 1$, it is the Monte-Carlo estimation, and if $\lambda = 0$, GAE degrades to the one-step temporal difference (TD) error. Then the gradient for updating $\tilde{\pi}^{K,n}, \forall n \in \mathcal{N}$ is estimated by

$$\Delta \tilde{\theta}^{K,n} = \nabla_{\tilde{\theta}^{K,n}} \hat{\mathbb{E}}_t \{f(r_t(\tilde{\theta}^{K,n}), \hat{A}_t^K)\}, \quad (37)$$

where $\hat{\mathbb{E}}_t\{\cdot\}$ is the sample average over time-steps t and $r_t(\tilde{\theta}^{K,n}) = \frac{\tilde{\pi}^{K,n}(a_t|s_t)}{\pi^{K,n}(a_t|s_t)}$.

Since the parameterized function V_ϕ is arbitrary given at the beginning of the learning process, it should be updated to approach to the real value function by minimizing the following loss function

$$L(\phi) = \frac{1}{2} \hat{\mathbb{E}}_t [(V_\phi(s_t) - y_t)^2], \quad (38)$$

where $y_t = \hat{A}_t^K + V_\phi(s_t)$. The gradient of (38) is

$$\Delta \phi = \hat{\mathbb{E}}_t [(V_\phi(s_t) - y_t) \nabla_\phi V_\phi(s_t)]. \quad (39)$$

Practical Algorithm: Based on the gradients (37) and (39) for updating $\tilde{\theta}^n$ and ϕ , respectively, and the solver for jamming action given in the last subsection, we provide the detailed procedure of the MAPPO-based trajectory planning and power allocation (TPPA) method in Algorithm 2, which is an instantiation of Algorithm 1. The algorithm iterates between the data collection phase and the training phase, where N_{iter} is the number of total iteration rounds.

In the data collection phase (i.e., lines 2-7 in Algorithm 2), at each time-step t , the *central console* receives s_t from E_n s and then solves the problem (21) using the method in Section V-A and then sends the solution $a_t^J = (p_1(t), \dots, p_N(t))$ to E_n s. Each E_n transmits jamming power $a_t^{J,n} = p_n(t)$ and executes the moving action $a_t^{K,n}$ according to $\pi^{K,n}(\cdot|s_t)$. The suspicious information received by E_n s will be fed back to the central console for estimating the reward r_t . After collecting a data segment $\tau = \{s_t, a_t^K, r_t\}_{t=1}^{T_l}$, the central console calculates $\{\hat{A}_t\}_{t=1}^{T_l}$ and $\{y_t\}_{t=1}^{T_l}$ for training. We assume that the training of each $\pi^{K,n}$ will be executed on the *on-board computer* of each E_n , while the training of V_ϕ is executed on the *central console*. Therefore, $\{\hat{A}_t\}_{t=1}^{T_l}$ is sent to all E_n s and $\{y_t\}_{t=1}^{T_l}$ is stored in the console.

In the training phase (i.e., lines 12-20 in Algorithm 2), we update the parameters in U rounds with collected data to improve the data efficiency. In each round, parameters are updated in the mini-batch style, where the j th mini-batch data $\mathcal{D}_j$ has B groups

Algorithm 2: MAPPO-Based TPPA.

Input: Initialized policies π^n with θ^n and $\tilde{\pi}^n$ with $\tilde{\theta}^n \leftarrow \theta^n$ for all n ; the value function V_ϕ ; A memory buffer $\mathcal{D}$.

Output: Well-trained policies $\pi^n, \forall n$

```

1 for iteration = 1, 2, ...,  $N_{\text{iter}}$  do
2   for  $t = 1, 2, \dots, T_l$  do
3     Solve the problem (21) to get  $a_t^J$ ;
4     Each  $E_n$  selects  $a_t^{K,n} \sim \pi^{K,n}(\cdot|s_t)$ ;
5     Execute  $a_t^J$  and then receive the reward  $r_t$ ;
6     Execute  $a_t^K$  and then transfer to  $s_{t+1}$ ;
7   end
8   Get a data segment  $\tau = \{s_t, a_t^K, r_t\}_{t=1}^{T_l}$ ;
9   Compute  $\{\hat{A}_t\}_{t=1}^{T_l}$  using GAE;
10  Compute  $\{y_t = \hat{A}_t + V_\phi(s_t)\}_{t=1}^{T_l}$ ;
11  Store data  $\{s_t, a_t^K, \hat{A}_t, y_t\}_{t=1}^{T_l}$  into  $\mathcal{D}$ ;
12  for  $u = 1, 2, \dots, U$  do
13    for  $j = 0, 1, \dots, \frac{T_l}{B} - 1$  do
14      Select a mini-batch data
15       $\mathcal{D}_j = \{s_t, a_t^K, \hat{A}_t, y_t\}_{t=1+Bj}^{B(j+1)}$ ;
16      Compute  $\Delta \tilde{\theta}^n, \forall n$  according to (37);
17      Compute  $\Delta \phi$  according to (39);
18      Apply gradient descent on  $\phi$  using  $\Delta \phi$  by
19      Adam optimizer;
20      Apply gradient ascent on  $\tilde{\theta}^n, \forall n$  using
21       $\Delta \tilde{\theta}^n$  by Adam optimizer;
22    end
23  end
  
```

of data. Each agent E_n updates $\tilde{\pi}^{K,n}$ using the gradient (37) with gradient ascent, and the central console updates the global value function V_ϕ by gradient descent using the gradient (39), where Adam optimizer in [48] is applied.

Optimality and Convergence: In this section, we have designed a general framework for solving the problem (20). That is, we first determine the optimal jamming policy $\hat{\pi}^J$, and then find the optimal moving policy $\hat{\pi}^K$ under the given $\hat{\pi}^J$, i.e., $\hat{\pi}^K = \arg \max_{\pi^K} \eta(\hat{\pi}^J, \pi^K)$. In proposition 1, we have proved that the proposed procedure holds the *optimality*: Based on Proposition 1, a general framework for optimally solving the problem (20) is given in Algorithm 1. Then in Section V-A, we solve the non-convex problem (21) using SCA. Based on this near-optimal solver, in Section V-B, we adopt MAPPO, a policy-based MARL algorithm, to learn $\hat{\pi}^K$. In theory, the policy-based RL algorithms can be proved to converge to a local optimum under some conditions [49]. Therefore, the proposed method can converge to at least a local optimum in theory. However, in deep RL, due to the application of DNN, the compatible representation condition (see [49] for details) required for convergence is not satisfied anymore, and thus the convergence of deep RL algorithms is still an open question. In this work, we empirically evaluate the convergence performance of the proposed algorithm in Fig. 4.

Complexity Analysis and Scalability: The complexity of our method mainly concludes two parts, one is the complexity of SCA for solving the problem (24), and the other part is the complexity of MAPPO. For the SCA part, we solve a linear programming (LP) with NM variables in each iteration. According to [50], the complexity of SCA part is about $\mathcal{O}_1 = \mathcal{O}(K_{\text{iter}}(NM)^{2.37} \log(NM/\sigma))$, where K_{iter} is the number of iterations in SCA and σ is the relative accuracy. The complexity of RL part depends on the size of the adopted DNNs. Let I denote the input size and O denote the output size of a DNN. Assume there are L hidden layers and each layer has $F_l, \forall l \in \{1, \dots, L\}$ neurons. If we consider the standard matrix multiplication algorithm with cubic complexity, then the complexity is $\mathcal{O}(IF_1 + OF_L + \sum_{l=1}^{L-1} F_l F_{l+1})$. In the fully observation case, the input size $I = |s_t| = MN + 7M + 3^N$. The output size O for each policy $\pi^{n,K}$ is $|\mathcal{V}|$, and for V_ϕ is 1. We assume these N policies and the value function have the same DNN structure. Therefore, the overall complexity of all DNNs is $\mathcal{O}_2 = \mathcal{O}(N(IF_1 + |\mathcal{V}|F_L + \sum_{l=1}^{L-1} F_l F_{l+1})) + \mathcal{O}(IF_1 + F_L + \sum_{l=1}^{L-1} F_l F_{l+1}) = \mathcal{O}((N+1)(IF_1 + \sum_{l=1}^{L-1} F_l F_{l+1}) + (N|\mathcal{V}| + 1)F_L)$. As shown in Algorithm 2, each iteration round consists of T_l data sample steps and UT_l training steps. Note that the SCA solver is only executed during the data collection phase and does not need to be trained. Therefore, the complexity of each data collection phase (i.e., lines 2 to 7 in Algorithm 2) is $T_l(\mathcal{O}_1 + \mathcal{O}_2)$, and the complexity of each training phase (i.e., lines 12 to 20 in Algorithm 2) is $UT_l \mathcal{O}_2$. So the overall complexity is $N_{\text{iter}}(T_l(\mathcal{O}_1 + \mathcal{O}_2) + UT_l \mathcal{O}_2)$.

As for scalability of the proposed TPPA algorithm, the SCA part is only to iteratively solve a LP problem, which has a very low complexity $\mathcal{O}_1$. Therefore, it can easily be applied to problems with large NM . As for the MARL part, its scalability cannot simply be measured by the complexity and is still an open question and highly empirical-based.

VI. NUMERICAL RESULTS

In this section, we first give the simulation setups. Then we present the simulation results with discussion and analysis.

A. Simulation Setups

System Settings: The following settings are given in default, and these parameters may be changed to show impacts on the eavesdropping performance. The number of legitimate UAVs and suspicious links are $N = 2$ and $M = 4$, respectively. When $N \geq M$, the eavesdropping task will be very easy, so we default $N < M$. According to [51], the empirical Rician factor of air-to-ground channel is about $10 \sim 60$, so we set $R = 20$ here. The other parameters are mainly referred to [19] and [26]. The reference path loss factor $\alpha_0 = -40$ dB, the path loss exponent $\beta = 2.4$, the background noise $N_0 = 10^{-9}$ mW, $P_0 = 10$ mW, $P_{\text{max}} = 40$ mW, and $\Delta T = 1$ s. The SI factor σ_ρ is defaulted as 0, and the penalty factor of collision $\nu = 0$.

We will describe locations by Cartesian coordinates in meters. The four default $\{D_m\}_{m=1}^4$ are located at $(200, 200, 0)$,

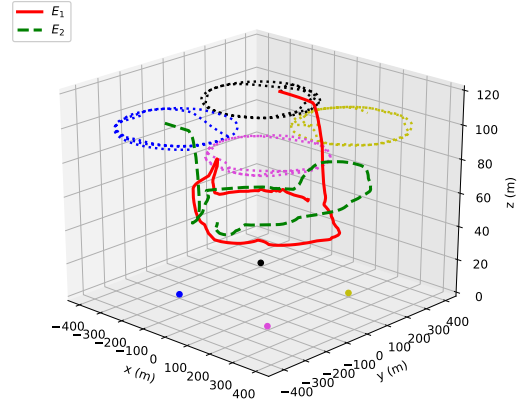


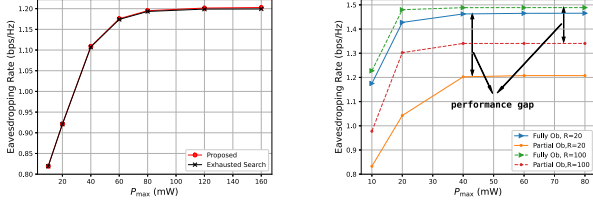
Fig. 2. 3D diagram of flight trajectories in the default setting.

$(-200, 200, 0)$, $(-200, -200, 0)$, and $(200, -200, 0)$, respectively. Each S_m flies in 3 m/s, at an altitude of 100 m, with D_m as the center and 200 m as the radius. The initial locations of $\{E_n\}_{n=1}^2$ are randomly scattered within $[-400, 400] \times [-400, 400]$ with a height of 100 m. We consider the discrete velocity set $\mathcal{V}$ for simplicity. We assume $\mathcal{V}$ consists of 11 discrete actions, which is given by $v_0 \times \{[1, 0, 0], [-1, 0, 0], [0, 1, 0], [0, -1, 0], [0, 0, 1], [0, 0, -1], [0, 0, 0], [\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0], [\frac{-1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0], [\frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}}, 0], [\frac{-1}{\sqrt{2}}, \frac{-1}{\sqrt{2}}, 0]\}$, where $v_0 = 3$ m/s is the constant magnitude of speed. The minimum flying altitude of each E_n is 50 m. To be more intuitive, we provide a three dimensional (3D) diagram of the default system in the following Fig. 2, where the trajectories of UAVs are learned by the proposed algorithm. Fig. 2 is also the 3D version of Fig. 6(b).

Algorithm Settings: We parameterize the policies and value function using the two-layer fully-connected neural network with rectified linear unit (ReLU) activation, and each hidden layer has 64 neurons. The episode length $T_l = 256$, the data reuse frequency $U = 4$, and the number of minibatch $B = 4$. The discount factor $\gamma = 0.99$, GAE factor $\lambda = 0.95$, and the clip range $\varepsilon = 0.2$ [44]. The learning rate of Adam optimizer is 5×10^{-4} , which decays to zero with the progress of training. To accelerate the learning speed, similar to [52], the parameters are shared among all agents, where the one-hot agent indices are appended to the input s_t to enable the neural networks to distinguish different agents. Besides, the parallel learning framework in [53] is adopted, where four parallel sample processes are conducted. Therefore, in each training phase, there are $T_l \times 4 = 1024$ group of data. The number iteration rounds $N_{\text{iter}} = 7000$, so there is total of $7000 \times 1024 = 7.168 \times 10^6$ steps.

B. Simulation Results

We first introduce metrics which may appear in this subsection. *Eavesdropping rate* is the sample average of eavesdropping rate per suspicious pair, i.e., $\frac{1}{M} \mathbb{E}_t \{r_t^J\}$, and (*eavesdropping*) *Success rate* is the probability of successful eavesdropping, that is estimated by $\frac{1}{M} \mathbb{E}_t \{\sum_{m=1}^M \mathbb{1}\{\mathcal{C}_{S_m}^E(t) \geq C_m(t)\}\}$. All simulation results provided in the following contents are averaged over 10^5 Monte-Carlo steps.



(a) The comparisons between the proposed solution and exhausted cases with partial and fully observations (b) The comparisons between proposed solution and exhausted cases with partial and fully observations

Fig. 3. Verification for the feasibility of the proposed solution to the subproblem (25).

In Fig. 3, we verify the effectiveness of the SCA solver proposed in Section V-A to the subproblem (24). Here, we mainly focus on jamming power allocation, so the positions of E_1 and E_2 are fixed at (200,0,50) and (-200,0,50), respectively. In Fig. 3(a), we compare the proposed solution with the exhausted search under full observations. Since the dimension of $p(t)$ is $N \times M$, which is impractical to apply the exhausted search, we therefore consider the comparison in a simplified case in which $p_n^m(t)$ for all m are the same. In this case, we only need to optimize $N = 2$ jamming powers in the default setting. In the exhaustive method, $p_n^m(t)$ is quantized to 400 quantization intervals, such that the quantization accuracy is $\frac{P_{\max}}{400 \cdot M}$. It is shown that the proposed optimization scheme achieves the eavesdropping rate close to the exhaustive method with very low computational complexity.⁴ In the exhaustive method, when the quantization level is fixed, the quantization accuracy decreases with the increase of $P_{\max}$, so that when $P_{\max}$ becomes larger, the proposed solution even slightly outperforms the exhausted search. The comparison in this simplified case verifies the effectiveness of the proposed solution. In Fig. 3(b), we demonstrate the performance of the probabilistically robust solution in the partially observable scenario, where the confidence levels are $\epsilon_o = \epsilon_c = 0.95$. The curve “Fully Ob” indicates the fully observation, and “Partial Ob” means that $\gamma_m(t)$ s are unavailable but their statistical distributions are known. When $P_{\max} \geq 40$ mW, the gaps in eavesdropping rate between “Fully Ob” and “Partial Ob” become stable. The gaps reveal that the performance degradation brought by the lack of channel information $\gamma_m(t)$. It is shown that the performance gap with $R = 100$ is obviously smaller than that with $R = 20$, because the uncertainty (or variance) of $\gamma_m(t)$ decreases with a bigger R .

In Fig. 4, we verify the performance of the MARL algorithm introduced in Section V-B, MAPPO, by comparing it with its fully centralized version, CPPO, which utilizes the single-agent PPO in [44] to learn a centralized policy for controlling all legitimate UAVs. Each episode has $T_l \times 4 = 1024$ samples. When $N = 2$ and $M = 4$, the eavesdropping rate obtained by MAPPO is extremely close to that obtained by CPPO, which is fully centralized and thus should achieve the best performance in theory. This result shows that training distributed policies using

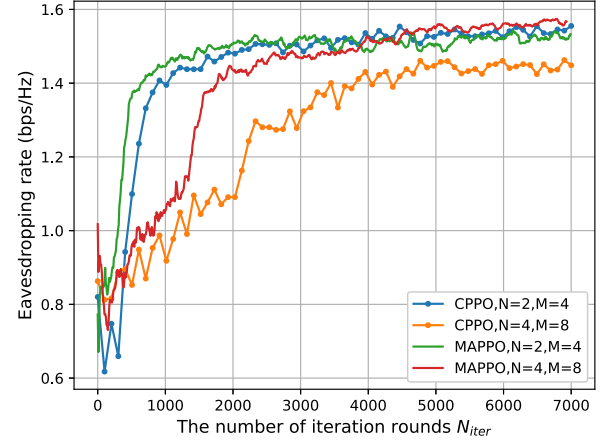


Fig. 4. Learning processes of different RL algorithms.

MAPPO will not cause extra performance loss. In the setting with $N = 4$ and $M = 8$, the four added D_m s are located at (600, 200,0), (600, -200,0), (-600, -200,0), and (-600, 200,0), respectively. In this case, MAPPO outperforms CPPO with a non-negligible gap in terms of eavesdropping rate. Besides, the convergence speed of CPPO is much slower than that of MAPPO. Specifically, CPPO takes 4000 episodes to surpass 1.4 bps/Hz, but MAPPO takes less than 2000 episodes. These results verify our previous declaration in Section V-B, that is, in the practical implementations, training a policy with large action space is more difficult than training several decomposed policies with small action space. In summary, when N is small, MAPPO achieves almost the same performance as CPPO and outperforms CPPO when N gets larger.

In Fig. 5, we compare the proposed TPPA algorithm with the following benchmarks. 1) MPC [19]: The model predictive control (MPC) method was adopted to control the mobility of eavesdropping UAVs in [19]. Since they did not consider jamming power allocation, in our comparison, we will adopt MPC for trajectory planning and the jamming power is still allocated by the SCA solver in Section V-A. MPC needs to know the states of several future time-steps. In this comparison, we assume the information on the next two steps is available. 2) AS-HD [26]: In [26], alternative optimization (AO) and SCA (AS) were applied to optimize the jamming power and position of a UAV. Since [26] considered a half-duplex (HD) scenario, for a fair comparison, we compare AS with our method in an HD variant. Specifically, in AS-HD and TPPA-HD, we consider 4 HD UAVs, two of which are moving eavesdroppers and the other two are jammers with fixed locations (200,0,50) and (-200,0,50), respectively. For AS-HD, in each time-step t , the jamming powers and positions of UAVs are optimized by AS to obtain the optimal position $\mathbf{q}_{E_n}^*$. Then, the moving action is determined by $v_n(t) = v_0 \frac{\mathbf{q}_{E_n}^* - \mathbf{q}_{E_n}(t)}{\|\mathbf{q}_{E_n}^* - \mathbf{q}_{E_n}(t)\|}$. The jamming power $p(t)$ is optimized by the method in Section V-A. For TPPA-HD, the curve is plotted by applying the proposed TPPA algorithm to the HD-variant system. 3) OJFP (Optimal Jamming with Fixed Position): In this scheme, two legitimate UAVs implement

⁴We have provided the execution time in Table II later.

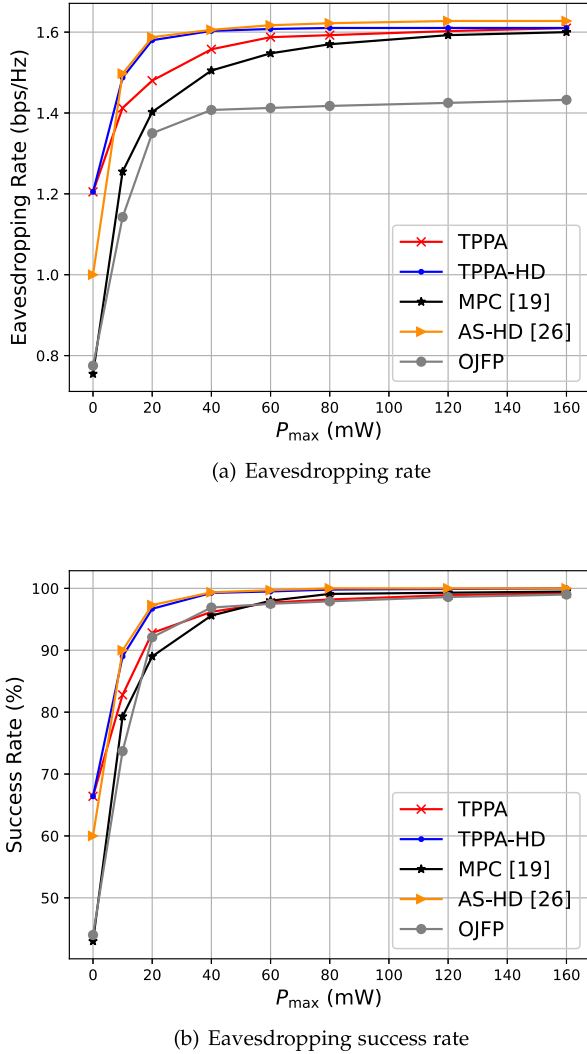


Fig. 5. Comparison of eavesdropping rate and success rate between different schemes.

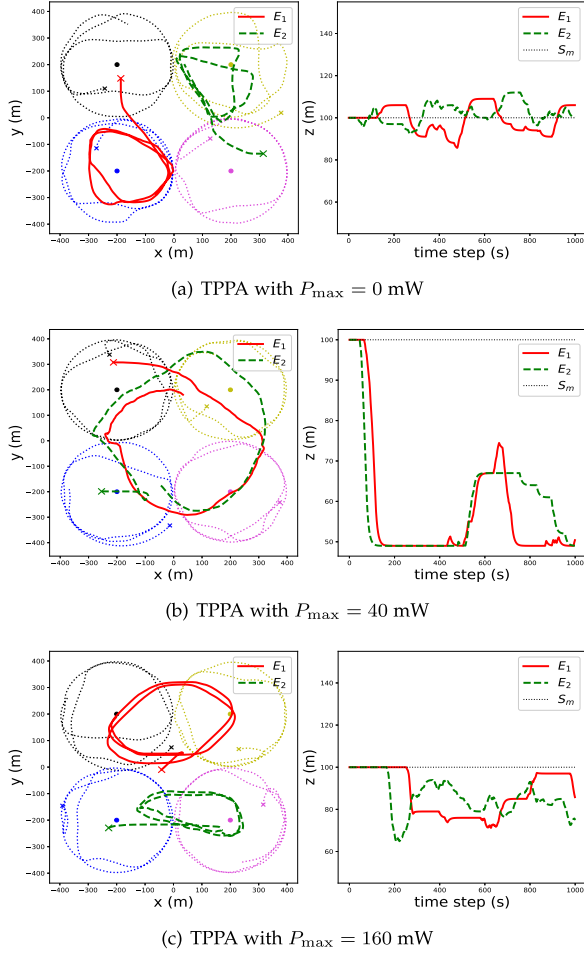
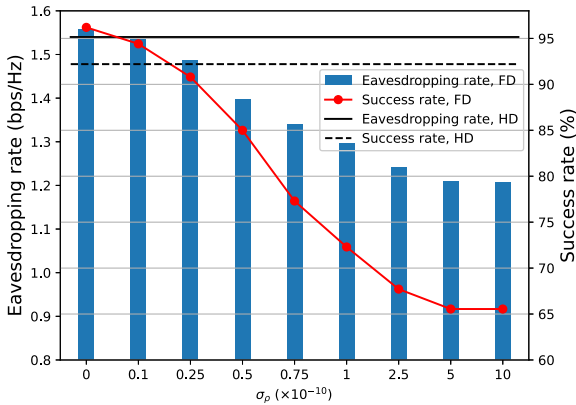
optimal jamming (using the method in Section V-A) in fixed positions (200,0,50) and (−200,0,50), respectively. We stress that results in the default FD setting and the HD variant (i.e., curves with or without the suffix “-HD”) are not comparable in general since the performances in the HD variant are also related to the fixed positions of two HD jammers, that are manually selected. Fig. 5(a) shows that, in the default system with two FD UAVs, the proposed TPPA outperforms MPC and OJFP in terms of eavesdropping rate. The gap between TPPA and MPC enlarges with the decrease of $P_{\max}$, which means TPPA has a greater advantage over MPC in the low jamming power region. Besides, the MPC method requires to predict future system states accurately, which is typically hard in practice. While for the proposed MAPPO-based TPPA algorithm, it learns from the data by interacting with the environment to maximize the expected long-run return, so the policy can consider future states from the perspective of expectations when making decisions, and thereby the uncertainty of future states can be well tackled. Furthermore, the complexity of MPC in [19] is high, where

TABLE II
EXECUTION TIME AND FPOS WITH DIFFERENT N AND M

Scenario	Execution Time (ms)			FPOs($\times 10^6$)	
	SCA	FF	Training	FF	Training
$N = 2, M = 4$	4.24	0.52	45.3	0.57	936.8
$N = 4, M = 4$	4.35	0.56	73.8	1.54	2812.64
$N = 4, M = 8$	4.5	0.62	82.8	2.06	3752.64

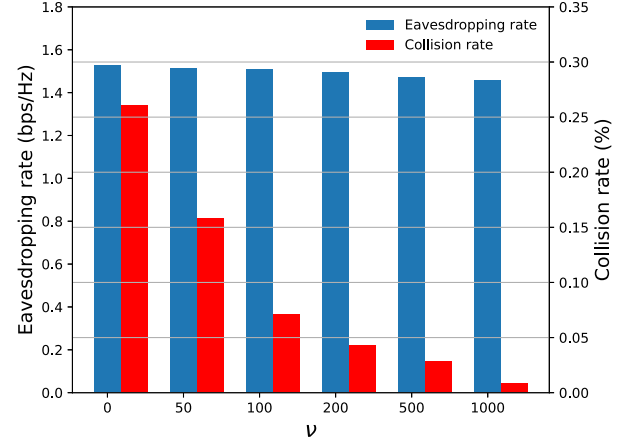
each decision-making of moving action takes about 2.8 seconds (while $\Delta T = 1$ s in our system) so that is not suitable to make decisions online. Although the MAPPO part of the proposed TPPA needs time to train, it can make decisions online after the training process (see Table II for details). By comparing TPPA-HD with AS-HD, it is shown that the AS-HD slightly outperforms the proposed TPPA-HD since AS-HD tried to find the best monitoring locations for UAVs at each time step. However, similar to MPC, the AS-HD method is time-consuming (each decision-making costs 1.6 seconds) since it needs to iterate two-tier alternative optimization processes in each time-step. Therefore, the AS method is also not able to make decisions in real time. Fig. 5(b) shows the eavesdropping success rate of different schemes. When $P_{\max} \geq 40$ mW, these schemes can achieve success rates of more than 95%. It is shown that further increasing $P_{\max}$ will not significantly improve both the eavesdropping rate and success rate, which is why we default $P_{\max} = 40$ mW.

In Fig. 6, we show the trajectories of UAVs under different $P_{\max}$, which may provide some intuitions and insights. All these trajectories are obtained by using the policies learned by the proposed TPPA algorithm. Each left subfigure depicts the movements in the x-y (horizon) plane in 1000 seconds. We have not plotted the legends for S_m s and D_m s for picture clarity. The points located at $\{(200,200), (-200,200), (-200,-200), (200,-200)\}$ denote the suspicious destinations D_1 to D_4 , respectively. The trajectory of each suspicious UAV is denoted by a dotted line which is the same color as its destination point. The marker 'x' denotes the corresponding start points of each UAV. Each right subfigure shows the height of UAVs versus time. Since all S_m fly at the fixed height of 100 m, we denote them with one dotted line. Fig. 6(a) shows that, when $P_{\max} = 0$ mW (i.e., passive eavesdropping), E_1 and E_2 fly around two symmetrical suspicious links. This might be because they cannot jam, such that they have to partly give up eavesdropping on the other two suspicious links and only focus on two links. The flight height of passive UAVs is around 100 m since they need to get as close as possible to S_m to increase the capacity of wiretap channels. When $P_{\max} = 40$ mW, as shown in Fig. 6(b), the flying coverage of legitimate UAVs becomes larger since they can eavesdrop on more suspicious links with the help of jamming. Besides, the flight heights of UAVs with $P_{\max} = 40$ mW are lower than that of passive UAVs (close to the minimum height 50 m). This is because E_n s need to seek a balance between jamming and eavesdropping. Fig. 6(c) shows that, when $P_{\max}$ further increases to 160 mW, the flight heights of legitimate UAVs are closer to 100 m as that of passive UAVs. This is because when the jamming power is sufficient, legitimate UAVs


 Fig. 6. Trajectories of UAVs with different $P_{\max}$.

 Fig. 7. Eavesdropping rate and success rate versus different SI factor σ_ρ .

can move closer to S_m s to increase $C_{S_m}^E$, without worrying about insufficient jamming on D_m s.

In Fig. 7, we investigate the impacts of SI on the eavesdropping rate and success rate. Obviously, both the eavesdropping rate and success rate decrease rapidly with the increase of σ_ρ . When $\sigma_\rho \geq 0.5 \times 10^{-10}$, the eavesdropping success rate is lower than 85%, which is bad for eavesdropping on possible video signals. When $\sigma_\rho \geq 5 \times 10^{-10}$, the eavesdropping


 Fig. 8. Eavesdropping rate and collision rate versus different penalty factor ν .

performances are essentially the same as the passive scheme (i.e., TPPA with $P_{\max} = 0$) since when the SI factor is high, the jamming signals would cause greater damage to the eavesdropping UAVs themselves than the suspicious links. Although some techniques were proposed to reduce SI, there are still some residuals in practice, which is one of the reasons why FD radios are not currently being deployed on a large scale. Therefore, we provide a HD scheme for situations where SI is not well cancelled. Specifically, we employ 3 HD moving UAVs, one of which only transmits jamming signals and the other two only passively listen, so that SI will not exist while MI can be easily eliminated. The eavesdropping rate and success rate of the HD scheme are depicted by the solid and dashed horizontal lines, respectively. It shows that the eavesdropping rate of HD is very close to that of the FD model with perfect SI cancellation and the success rate surpasses 92%, at the cost of one more UAV.

In Fig. 8, we investigate the effectiveness of the penalty term in collision avoidance. Since in 3D space, the collision hardly happens even without the collision penalty. To highlight the performance of collision avoidance, we here enhance the probability of collisions by setting E_n s to fly on the same 2D horizontal plane with S_m s. The collision rate is defined as $C(d_{\text{safe}}) = \mathbb{E}_t\{\frac{1}{N} \sum_{n=1}^N \mathbb{1}\{d_{\min}^n(t) \leq d_{\text{safe}}\}\}$. Here, we set $\sigma_d = 10$ (a parameter in (13)) and the collision rate is calculated by $C(10)$. It is shown that as the penalty factor ν increases, the collision rate significantly decreases (from 0.26% to 0.009%) while there are only slight sacrifices of the eavesdropping rates. Note that the collision rate is calculated with the safe distance $d_{\text{safe}} = 10\text{m}$. While in real-world scenarios, the radius of a UAV is typically less than 1 m. When $\nu \geq 200$, $C(2)$ is 0%, which indicates that the physical collisions can be absolutely eliminated even in the 2D plane.

In Table II, we provide the empirical complexity in terms of the (averaged) execution time and floating point operations (FPOs) in our practical implementations.⁵ The column ‘SCA’ indicates the cost of solving the problem (24) (line 3 in Algorithm

⁵Our simulations are conducted on a computer with CPU Intel Core i7-8086 k. Codes are written in Python 3.6, where CVXPY is adopted to solve the problem (24), and Tensorflow 1.12 is adopted for the MARL part.

2), ‘FF’ is the feedforward cost of DNNs in selecting moving actions (line 4 in Algorithm 2), and ‘Training’ means the total cost in each training phase (lines 12-20 in Algorithm 2). The second big column shows the averaged execution time (in ms), and the third big column shows FPOs, which is a deterministic value and is reported by the built-in function provided by Tensorflow. The FPOs of SCA is not provided since the detailed solving process of CVXPY is a black box and they do not provide a corresponding tool for obtaining FPOs. Taking the learning curve ‘MAPPO, $N = 2, M = 4$ ’ in Fig. 4 for example, the data collection phase costs $7000 \times 1024 \times (4.24 + 0.52)ms/4 = 8529.2s$ (the reason of ‘/4’ above is because we adopt 4 parallel processes to collect the data), and the time consumed on training is only $7000 \times 45.3ms = 317.1s$. Although the whole learning process is time-consuming, after the learning, we can make decisions using the proposed SCA solver and the well-trained policies in real-time, where the total decision time in each time-step t is less than 6 ms. It could be noticed that the complexity in FPOs in $N = 4, M = 8$ is about 4 times that of $N = 2, M = 4$, while the training time is only about 1.8 times. This is because the real execution time is influenced not only by FPOs but also by some other factors, e.g., the compilation time of problems in CVXPY and some optimizations of the Matrix multiplication algorithm in Tensorflow.

VII. CONCLUSION

In this paper, we employed multiple jamming-enabled UAVs to collaboratively eavesdrop on the communication information of multiple suspicious links, where each link is comprised of a movable UAV sender and its ground destination. Because it is difficult to use RL to optimize the jamming power that meets the eavesdropping constraints, we propose to determine the jamming power with SCA, and then learns the decentralized policies of moving action by applying MAPPO. We prove that this decomposed optimization also holds the optimality. Simulation results verify the effectiveness of both the solver for jamming power and the decentralized policies for moving action learned by MARL, and also showed that with the help of jamming signals, fewer UAVs can effectively eavesdrop on a larger number of suspicious links.

REFERENCES

- [1] J. Xu, L. Duan, and R. Zhang, “Proactive eavesdropping via cognitive jamming in fading channels,” *IEEE Trans. Wireless Commun.*, vol. 16, no. 5, pp. 2790–2806, May 2017.
- [2] J. Xu, L. Duan, and R. Zhang, “Surveillance and intervention of infrastructure-free mobile communications: A new wireless security paradigm,” *IEEE Wireless Commun.*, vol. 24, no. 4, pp. 152–159, Aug. 2017.
- [3] J. Xu, L. Duan, and R. Zhang, “Proactive eavesdropping via jamming for rate maximization over Rayleigh fading channels,” *IEEE Wireless Commun. Lett.*, vol. 5, no. 1, pp. 80–83, Feb. 2016.
- [4] J. Moon, H. Lee, C. Song, S. Lee, and I. Lee, “Proactive eavesdropping with full-duplex relay and cooperative jamming,” *IEEE Trans. Wireless Commun.*, vol. 17, no. 10, pp. 6707–6719, Oct. 2018.
- [5] B. Li, Y. Yao, H. Chen, Y. Li, and S. Huang, “Wireless information surveillance and intervention over multiple suspicious links,” *IEEE Signal Process. Lett.*, vol. 25, no. 8, pp. 1131–1135, Aug. 2018.
- [6] X. Jiang, H. Lin, C. Zhong, X. Chen, and Z. Zhang, “Proactive eavesdropping in relaying systems,” *IEEE Signal Process. Lett.*, vol. 24, no. 6, pp. 917–921, Jun. 2017.
- [7] C. Zhong, X. Jiang, F. Qu, and Z. Zhang, “Multi-antenna wireless legitimate surveillance systems: Design and performance analysis,” *IEEE Trans. Wireless Commun.*, vol. 16, no. 7, pp. 4585–4599, Jul. 2017.
- [8] D. Guo, H. Ding, L. Tang, X. Zhang, L. Yang, and Y.-C. Liang, “A proactive eavesdropping game in MIMO systems based on multi-agent deep reinforcement learning,” *IEEE Trans. Wireless Commun.*, vol. 21, no. 11, pp. 8889–8904, Nov. 2022.
- [9] D. Hu, Q. Zhang, P. Yang, and J. Qin, “Proactive monitoring via jamming in amplify-and-forward relay networks,” *IEEE Signal Process. Lett.*, vol. 24, no. 11, pp. 1714–1718, Nov. 2017.
- [10] J. Xu, L. Duan, and R. Zhang, “Fundamental rate limits of physical layer spoofing,” *IEEE Wireless Commun. Lett.*, vol. 6, no. 2, pp. 154–157, Apr. 2017.
- [11] Y. Zeng and R. Zhang, “Wireless information surveillance via proactive eavesdropping with spoofing relay,” *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 8, pp. 1449–1461, Dec. 2016.
- [12] J. Moon, H. Lee, C. Song, S. Kang, and I. Lee, “Relay-assisted proactive eavesdropping with cooperative jamming and spoofing,” *IEEE Trans. Wireless Commun.*, vol. 17, no. 10, pp. 6958–6971, Oct. 2018.
- [13] J. Chen, L. Tang, D. Guo, Y. Bai, L. Yang, and Y.-C. Liang, “Proactive eavesdropping in massive MIMO-OFDM systems via deep reinforcement learning,” *IEEE Trans. Veh. Technol.*, vol. 71, no. 11, pp. 12315–12320, Nov. 2022.
- [14] X. Zhou, B. Maham, and A. Hjørungnes, “Pilot contamination for active eavesdropping,” *IEEE Trans. Wireless Commun.*, vol. 11, no. 3, pp. 903–907, Mar. 2012.
- [15] J. Cui, Y. Liu, and A. Nallanathan, “Multi-agent reinforcement learning-based resource allocation for UAV networks,” *IEEE Trans. Wireless Commun.*, vol. 19, no. 2, pp. 729–743, Feb. 2020.
- [16] Y. Sun, D. Xu, D. W. K. Ng, L. Dai, and R. Schober, “Optimal 3D-trajectory design and resource allocation for solar-powered UAV communication systems,” *IEEE Trans. Commun.*, vol. 67, no. 6, pp. 4281–4298, Jun. 2019.
- [17] A. V. Savkin, H. Huang, and W. Ni, “Securing UAV communication in the presence of stationary or mobile eavesdroppers via online 3D trajectory planning,” *IEEE Wireless Commun. Lett.*, vol. 9, no. 8, pp. 1211–1215, Aug. 2020.
- [18] H. Huang, A. V. Savkin, and W. Ni, “Navigation of a UAV team for collaborative eavesdropping on multiple ground transmitters,” *IEEE Trans. Veh. Technol.*, vol. 70, no. 10, pp. 10450–10460, Oct. 2021.
- [19] H. Huang, A. V. Savkin, and W. Ni, “Decentralized navigation of a UAV team for collaborative covert eavesdropping on a group of mobile ground nodes,” *IEEE Trans. Automat. Sci. Eng.*, vol. 19, no. 4, pp. 3932–3941, Oct. 2022.
- [20] J. Wang, P. Zhang, L. Tang, Y. Bai, and L. Yang, “Intelligent passive eavesdropping in massive MIMO-OFDM systems via reinforcement learning,” *IEEE Wireless Commun. Lett.*, vol. 11, no. 6, pp. 1248–1252, Jun. 2022.
- [21] H. Wu, M. Li, Q. Gao, Z. Wei, N. Zhang, and X. Tao, “Eavesdropping and anti-eavesdropping game in UAV wiretap system: A differential game approach,” *IEEE Trans. Wireless Commun.*, vol. 21, no. 11, pp. 9906–9920, Nov. 2022.
- [22] M. Huang, Y. Chen, and X. Tao, “Proactive eavesdropping in UAV systems via trajectory planning and power optimization,” in *Proc. IEEE Wireless Commun. Netw. Conf. Workshops*, 2021, pp. 1–6.
- [23] X. Wang, K. Li, S. S. Kanhere, D. Li, X. Zhang, and E. Tovar, “PELE: Power efficient legitimate eavesdropping via jamming in UAV communications,” in *Proc. 13th Int. Wireless Commun. Mobile Comput. Conf.*, 2017, pp. 402–408.
- [24] K. Li, S. S. Kanhere, W. Ni, E. Tovar, and M. Guizani, “Proactive eavesdropping via jamming for trajectory tracking of UAVs,” in *Proc. 15th Int. Wireless Commun. Mobile Comput. Conf.*, 2019, pp. 477–482.
- [25] K. Li, R. C. Voicu, S. S. Kanhere, W. Ni, and E. Tovar, “Energy efficient legitimate wireless surveillance of UAV communications,” *IEEE Trans. Veh. Technol.*, vol. 68, no. 3, pp. 2283–2293, Mar. 2019.
- [26] G. Hu, Y. Cai, and Y. Cai, “Joint optimization of position and jamming power for UAV-aided proactive eavesdropping over multiple suspicious communication links,” *IEEE Wireless Commun. Lett.*, vol. 9, no. 12, pp. 2093–2097, Dec. 2020.
- [27] G. Hu et al., “Maxmin fairness for UAV-enabled proactive eavesdropping with jamming over distributed transmit beamforming-based suspicious communications,” *IEEE Trans. Commun.*, vol. 71, no. 3, pp. 1595–1614, Mar. 2023.

- [28] S. Hu, Q. Wu, and X. Wang, "Energy management and trajectory optimization for UAV-enabled legitimate monitoring systems," *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, pp. 142–155, Jan. 2021.
- [29] R. Sutton and A. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. Cambridge, MA, USA: MIT Press, 2018.
- [30] G. Weisz, P. Budzianowski, P.-H. Su, and M. Gašić, "Sample efficient deep reinforcement learning for dialogue systems with large action spaces," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 11, pp. 2083–2097, Nov. 2018.
- [31] M. Zhang, Y. Chen, X. Tao, and I. Darwazeh, "Power allocation for proactive eavesdropping with spoofing relay in UAV systems," in *Proc. 26th Int. Conf. Telecommun.*, 2019, pp. 102–107.
- [32] D. Xu, "Legitimate surveillance of suspicious multichannel DF relay networks with monitor mode selection," *IEEE Wireless Commun. Lett.*, vol. 10, no. 2, pp. 401–405, Feb. 2021.
- [33] C. Kai, X. Zhang, X. Hu, and W. Huang, "Joint pilot design and beamforming optimization in massive MIMO surveillance systems," *China Commun.*, vol. 19, no. 4, pp. 83–97, 2022.
- [34] Z. Wang and Y. Wang, "A pilot contamination attacker-defender model for wireless networks under Stackelberg game," in *Proc. IEEE 95th Veh. Technol. Conf.*, 2022, pp. 1–5.
- [35] M. Alzenad, A. El-Keyi, F. Lagum, and H. Yanikomeroglu, "3-D placement of an unmanned aerial vehicle base station (UAV-BS) for energy-efficient maximal coverage," *IEEE Wireless Commun. Lett.*, vol. 6, no. 4, pp. 434–437, Aug. 2017.
- [36] J. Lyu, Y. Zeng, R. Zhang, and T. J. Lim, "Placement optimization of UAV-mounted mobile base stations," *IEEE Commun. Lett.*, vol. 21, no. 3, pp. 604–607, Mar. 2017.
- [37] M. Mozaffari, W. Saad, M. Bennis, Y.-H. Nam, and M. Debbah, "A tutorial on UAVs for wireless networks: Applications, challenges, and open problems," *IEEE Commun. Surv. Tut.*, vol. 21, no. 3, pp. 2334–2360, Third Quarter 2019.
- [38] G. Zhang, Q. Wu, M. Cui, and R. Zhang, "Securing UAV communications via joint trajectory and power control," *IEEE Trans. Wireless Commun.*, vol. 18, no. 2, pp. 1376–1389, Feb. 2019.
- [39] D. Bharadia, E. McMillin, and S. Katti, "Full duplex radios," in *Proc. ACM SIGCOMM Conf. SIGCOMM*, 2013, pp. 375–386.
- [40] H. Xing, L. Liu, and R. Zhang, "Secrecy wireless information and power transfer in fading wiretap channel," *IEEE Trans. Veh. Technol.*, vol. 65, no. 1, pp. 180–190, Jan. 2016.
- [41] A. Schwartz, "A reinforcement learning method for maximizing undiscounted rewards," in *Proc. 10th Int. Conf. Mach. Learn.*, 1993, pp. 298–305.
- [42] G. C. Calafiore and L. El Ghaoui, *Optimization Models*. Cambridge, U.K.: Cambridge Univ. Press, 2014.
- [43] M. Razaviyayn, "Successive convex approximation: Analysis and applications," PhD dissertation, Faculty Grad. School, Univ. Minnesota, Minneapolis, MN, USA, 2014.
- [44] J. Schulman et al., "Proximal policy optimization algorithms," 2017, *arXiv:1707.06347*.
- [45] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1889–1897.
- [46] D. Guo, L. Tang, X. Zhang, and Y.-C. Liang, "Joint optimization of handover control and power allocation based on multi-agent deep reinforcement learning," *IEEE Trans. Veh. Technol.*, vol. 69, no. 11, pp. 13124–13138, Nov. 2020.
- [47] J. Schulman et al., "High-dimensional continuous control using generalized advantage estimation," in *Proc. Int. Conf. Learn. Representations*, 2016, pp. 1–14.
- [48] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
- [49] V. Konda and J. Tsitsiklis, "Actor-critic algorithms," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2000, pp. 1008–1014.
- [50] M. B. Cohen, Y. T. Lee, and Z. Song, "Solving linear programs in the current matrix multiplication time," *J. ACM*, vol. 68, no. 1, pp. 1–39, 2021.
- [51] S. M. Andreu, "The Rician channel model for LAP-ground communication," BS thesis, Universitat Politècnica de Catalunya, 2016. [Online]. Available: <https://upcommons.upc.edu/bitstream/handle/2117/96926/memoria.pdf?sequence=1&isAllowed=y>
- [52] J. Foerster et al., "Counterfactual multi-agent policy gradients," 2017, *arXiv:1705.08926*.
- [53] V. Mnih et al., "Asynchronous methods for deep reinforcement learning," 2016, *arXiv:1602.01783*.



Delin Guo received the BS degree in electronic information science and technology from the Jilin University, Jilin, China, in 2017. He is currently working toward the PhD degree with the Department of Electronic Science and Engineering, Nanjing University, China. His current research interests include physical-layer security, reinforcement learning, and multi-agent reinforcement learning.



Lan Tang (Member, IEEE) received the BS and MS degrees in communication engineering from the Jilin University, Jilin, China, in 2002 and 2005, respectively, and the PhD degree in comm and information science from Southeast University, Nanjing, China, in 2009. Since 2009, she has joined the School of Electronic Science and Engineering, Nanjing University, China, where she is now an associate professor. Her current research interests include key technologies in 5G/6G, machine learning and optimization theory in wireless networks, and integrated communication-radar systems.



Xinggan Zhang received the BS degree in electrical engineering from the Nanjing University of Aeronautics and Astronautics, Nanjing, P.R. China, in 1982, and the MS and PhD degrees from the Nanjing University of Aeronautics and Astronautics, in 1988 and 2001, respectively. In 1992, he joined the Department of Electronic Engineering, NUAU, where he was an associate professor. In 1999, he joined the Department of Electronic Science and Engineering, Nanjing University, where he is currently a professor. His research interests include target recognition and image processing.



Ying-Chang Liang (Fellow, IEEE) is currently a professor with the University of Electronic Science and Technology of China, China, where he leads the Center for Intelligent Networking and Communications (CINC). He was a professor with the University of Sydney, Australia, a principal scientist and technical advisor with the Institute for Infocomm Research, Singapore, and a visiting scholar with Stanford University, USA. His research interests include wireless networking and communications, cognitive radio, symbiotic communications, dynamic spectrum

access, the Internet-of-Things, artificial intelligence, and machine learning techniques. He has been recognized by Thomson Reuters (now Clarivate Analytics) as a highly cited researcher since 2014. He received the Prestigious Engineering Achievement Award from the Institution of Engineers, Singapore, in 2007, the Outstanding Contribution Appreciation Award from the IEEE Standards Association in 2011, and the Recognition Award from the IEEE Communications Society Technical Committee on Cognitive Networks in 2018. He is the recipient of numerous paper awards, including the IEEE Communications Society Stephen O. Rice Prize in 2021, the IEEE Vehicular Technology Society Jack Neubauer Memorial Award in 2014, and the IEEE Communications Society APB Outstanding Paper Award in 2012. He is a foreign member of Academia Europaea. He is the founding editor-in-chief of the *IEEE Journal on Selected Areas in Communications: Cognitive Radio Series*, and the key founder and now the editor-in-chief of *IEEE Transactions on Cognitive Communications and Networking*. He is also serving as an associate editor-in-chief of the *China Communications*. He was a guest/associate editor of the *IEEE Transactions on Wireless Communications*, *IEEE Journal of Selected Areas in Communications*, *IEEE Signal Processing Magazine*, *IEEE Transactions on Vehicular Technology*, and *IEEE Transactions on Signal and Information Processing Over Network*. He was also an associate editor-in-chief of the *World Scientific Journal on Random Matrices: Theory and Applications*. He was a distinguished lecturer of the IEEE Communications Society and the IEEE Vehicular Technology Society. He was the chair of the IEEE Communications Society Technical Committee on Cognitive Networks, and served as the TPC chair and executive co-chair of the IEEE Globecom'17.