

Collaborative Reinforcement Learning Based Unmanned Aerial Vehicle (UAV) Trajectory Design for 3D UAV Tracking

Yujiao Zhu, *Student Member, IEEE*, Mingzhe Chen, *Member, IEEE*,
Sihua Wang, *Student Member, IEEE*, Ye Hu, *Member, IEEE*,
Yuchen Liu, *Member, IEEE*, and Changchuan Yin, *Senior Member, IEEE*

Abstract—In this paper, the problem of using one active unmanned aerial vehicle (UAV) and four passive UAVs to localize a 3D target UAV in real time is investigated. In the considered model, each passive UAV receives reflection signals from the target UAV, which are initially transmitted by the active UAV. The received reflection signals allow each passive UAV to estimate the signal transmission distance which will be transmitted to a base station (BS) for the estimation of the position of the target UAV. Due to the movement of the target UAV, each active/passive UAV must optimize its trajectory to continuously localize the target UAV. Meanwhile, since the accuracy of the distance estimation depends on the signal-to-noise ratio of the transmission signals, the active UAV must optimize its transmit power. This problem is formulated as an optimization problem whose goal is to jointly optimize the transmit power of the active UAV and trajectories of both active and passive UAVs so as to maximize the target UAV positioning accuracy. To solve this problem, a Z function decomposition based reinforcement learning (ZD-RL) method is proposed. Compared to value function decomposition based RL (VD-RL), the proposed method can find the probability distribution of the sum of future rewards to accurately estimate the expected value of the sum of future rewards thus finding better transmit power of the active UAV and trajectories for both active and passive UAVs and improving target UAV positioning accuracy. Simulation results show that the proposed ZD-RL method can reduce the positioning errors by up to 39.4% and 48.1%, compared to VD-RL and independent deep RL methods, respectively.

Index Terms—Unmanned aerial vehicles, localization, trajectory design, Z function decomposition based reinforcement learning.

I. INTRODUCTION

Unmanned aerial vehicle (UAV) localization has gained significant attention from academic and commercial fields

Y. Zhu, S. Wang, and C. Yin are with the Beijing Laboratory of Advanced Information Network, Beijing University of Posts and Telecommunications, Beijing, 100876, China (E-mail: yjzhu@bupt.edu.cn; sihuawang@bupt.edu.cn; ccyin@bupt.edu.cn).

M. Chen is with the Department of Electrical and Computer Engineering and Institute for Data Science and Computing, University of Miami, Coral Gables, FL, 33146, USA (Email: mingzhe.chen@miami.edu).

Y. Hu is with the Department of Industrial and System Engineering, University of Miami, Coral Gables, FL, 33146, USA (Email: yehu@miami.edu).

Y. Liu is with the Department of Computer Science, North Carolina State University, Raleigh, NC, 27695, USA (Email: yuchen.liu@ncsu.edu).

A preliminary version of this work [1] is submitted to the Proceedings of the 2023 IEEE International Global Communications Conference (GLOBECOM).

since it supports a wide range of applications in military, assistance and industrial scenarios [2]–[4]. For example, when UAVs perform attack missions in the military field, it is necessary to locate and track unauthorized UAVs in real time [5]. However, achieving accurate UAV positioning faces several challenges. First, UAVs are moving at a high speed, and thus estimating the real-time positions of UAVs is challenging. Second, since the coordinates of UAVs are three-dimensional (3D), estimating 3D coordinates of UAVs requires more sensors (at least four sensors [6]) and complex positioning algorithms. Third, dynamic wireless environments such as electromagnetic interference, transmit power allocation, and available communication resources will affect the transmission of pilot signals used for UAV localization thus affecting UAV localization accuracy [7].

A. Related Works

Recently, several existing works such as [8]–[11] have studied problems for UAV localization. The authors in [8] and [9] considered the use of a single camera sensor to track movement of UAVs. However, the positioning algorithms used in [8] and [9] must be implemented based on unique hardware and high computational resource. The authors in [10], [11] used ratio-frequency (RF) signals to estimate the positions of UAVs. In particular, in [10], the authors obtained the arrival time of transmitted signals from several sensors and determined the 3D positions of UAVs. The authors in [11] jointly used angle and arrival time of transmitted signals to estimate the positions of UAVs thus reducing the number of sensors used for UAV localization. However, the authors in [8]–[11] did not consider how the positions of sensors affect the UAV localization accuracy and they also did not consider the optimization of the deployment of sensors. In fact, the positions of sensors will significantly affect the UAV positioning accuracy [12]. Meanwhile, most of these works [8]–[11] assumed that the values of signal-to-noise ratio (SNR) of transmitted signals are constant, which is impractical in actual wireless networks. In addition, most of these works [8]–[11] assumed that a central controller knows the positions of all sensors and channel state information (CSI) in advance such that the central controller will directly use this information for

UAV positioning. Therefore, these works [8]–[11] cannot be used for scenarios where the central controller cannot obtain the positions of sensors or CSI.

Recently, a number of existing works [13]–[17] have studied the use of reinforcement learning (RL) [18] for UAV localization in the networks where the central controller cannot obtain all the information needed for UAV localization. In particular, the authors in [13] selected different ground sensors to optimize the UAV localization performance using a double deep Q-network based RL method. The authors in [14] developed a domain randomization based RL algorithm and estimated the real-time position of a UAV using a monocular camera while considering environmental impacts such as wind gusts. The authors in [15] used time difference of signal arrival information measured by ground sensors to estimate 3D coordinates of UAVs and applied deep deterministic policy gradient (DDPG) and soft actor-critic methods to optimize Taylor series linearized localization approach. The authors in [16] analyzed the effects of measurement uncertainty on the performance of UAV localization based on a proximal policy optimization (PPO) algorithm in an environment with dynamic noise. In [17], the authors mapped UAVs' initial sensory measurements into control signals for localization and navigation by an actor-critic based deep reinforcement learning (DRL) algorithm. However, the central controller in these works [13]–[17] must collect sensing data from all sensors to determine the UAV movement, which will increase the communication overhead and the time used for UAV localization. Meanwhile, most of these works [14]–[17] considered the use of statically installed sensors for UAV localization, which may not be used for localizing a UAV with a high movement speed.

B. Contributions

The main contribution of this work is to design a novel framework that can real-time monitor the position of a target UAV by controlled UAVs including four passive UAVs and one active UAV. The main contributions include:

- We propose a UAV-based localization system to estimate the positions of the target UAV in which the active UAV transmits signals to the target UAV, while four passive UAVs collect the arrival time of signals transmitted from the active UAV to the target UAV, and then from the target UAV to passive UAVs. Next, each passive UAV estimates the distance from the active UAV to the target UAV, and then to the passive UAV. Such distance information is transmitted to the BS, which calculates the position of the target UAV.
- In the considered UAV localization system, since the target UAV will change its position according to its performed task, each controlled UAV must optimize its trajectory to accurately localize the target UAV. Meanwhile, the accuracy of the distance information estimated by passive UAVs depends on the SNR of the signals transmitted from the active UAV and hence the active UAV must optimize its transmit power according to the movements of the target UAV and passive UAVs. This

problem is formulated as an optimization problem that aims to maximize the localization accuracy of the target UAV via optimizing the transmit power of the active UAV and the trajectories of the active and passive UAVs.

- To solve this problem, we propose a Z function decomposition based reinforcement learning (ZD-RL) method that enables each controlled UAV to determine its trajectory and the active UAV to determine its transmit power via its individual observation. Compared to value function decomposition methods [19], the Z function decomposition can find the probability distribution of the sum of future rewards such that each controlled UAV can accurately estimate the expected value of the sum of future rewards to update the parameters of its deep neural networks (DNNs). Hence, the proposed ZD-RL method can improve the efficiency and stability of optimizing the transmit power of the active UAV and the trajectories of controlled UAVs to minimize the positioning error of the target UAV.
- To further minimize the positioning error of the target UAV, we analyze how the positions of the controlled UAVs affect the positioning error of the target UAV. Our analytical results show that the minimum positioning error of the target UAV can be achieved when the distance between each controlled UAV and the target UAV is minimized.

Simulation results show that the proposed ZD-RL method can achieve up to 39.4% and 48.1% reduction in the positioning error of the positions of the target UAV compared to traditional value function decomposition based RL (VD-RL) and independent DRL methods, respectively. *To the best of our knowledge, this is the first work that presents a UAV localization framework that utilizes one active UAV and four passive UAVs for 3D UAV positioning.*

The rest of this paper is organized as follows. The system model and problem formulation are described in Section II. The Z function decomposition based power allocation and trajectory design method is discussed in Section III. The optimal deployment of controlled UAVs for target UAV localization are analyzed in Section IV. In Section V, numerical simulation results are presented and analyzed. Finally, conclusions are drawn in Section VI.

II. SYSTEM MODEL AND PROBLEM FORMULATION

Consider a UAV-assisted positioning network in which a ground BS and a set \mathcal{M} of five controlled UAVs jointly monitor the position of the target UAV in real time, as shown in Fig. 1. The controlled UAVs consist of an active UAV and four passive UAVs. In our model, the active UAV transmits signals to the target UAV which will reflect the signals to passive UAVs. Then, passive UAVs estimate the signal transmission distance from the active UAV to the target UAV, and then to passive UAVs. The estimated signal transmission distance will be transmitted to the BS to calculate the position of the target UAV. Finally, the BS sends the estimated position of the target UAV to the target UAV. We assume that the real-time

TABLE I
LIST OF NOTATIONS

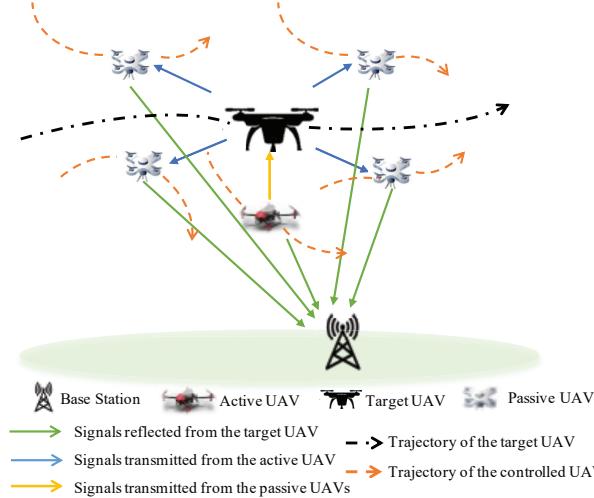


Fig. 1. Illustration of the considered UAV localization network.

3D coordinates of the controlled UAVs are known to the BS. Next, we first introduce the movement model of the active and passive UAVs. Then, the transmission links among the active UAV, target UAV, passive UAVs, and the BS are introduced. Finally, the positioning model and the optimization problem is formulated.

Let $\mathbf{u}_{m,t} = [x_{m,t}, y_{m,t}, z_{m,t}]^T$ be the 3D coordinates of UAV m at time slot t . Hereinafter, we use a sequence number 0 to represent the active UAV and a sequence number from 1 to 4 to represent a passive UAV. For example, $\mathbf{u}_{0,t}$ represents the coordinate of the active UAV and $\mathbf{u}_{m,t}$ with $1 \leq m \leq 4$ is the coordinate of a passive UAV. Then, the coordinates of UAV m is

$$\mathbf{u}_{m,t+1}(\phi_{m,t}, \varphi_{m,t}) = \mathbf{u}_{m,t} + v_{m,t} \Delta_t \begin{bmatrix} \cos \varphi_{m,t} \cos \phi_{m,t} \\ \sin \varphi_{m,t} \cos \phi_{m,t} \\ \sin \phi_{m,t} \end{bmatrix}, \quad (1)$$

where $\varphi_{m,t}$ is the yaw angle, $\phi_{m,t}$ is the pitch angle, $v_{m,t}$ is the flight speed, and Δ_t is the time duration of a time slot.

A. Transmission Model

Here, we introduce the models for transmission links a) from the active UAV to the target UAV and then reflected to passive UAVs, b) from passive UAVs to the ground BS.

1) *Active UAV-Target UAV-Passive UAV Links:* In our model, the active UAV transmits a signal a_t to the target UAV. We assume that there is no occlusion in the path from the active UAV to the target UAV, and paths from the target UAV to passive UAVs. Let $\tau_{m,t}$ denote the time of transmitting signal a_t from the active UAV to passive UAV m via the target UAV. Then, $\tau_{m,t}$ can be given by

$$\tau_{m,t} = \frac{r_{m,t}(\mathbf{u}_{0,t}, \mathbf{s}_t, \mathbf{u}_{m,t})}{c}, \quad (2)$$

where c is the speed of light and $r_{m,t}(\mathbf{u}_{0,t}, \mathbf{s}_t, \mathbf{u}_{m,t}) = d_{0,t}(\mathbf{u}_{0,t}, \mathbf{s}_t) + d_{m,t}(\mathbf{s}_t, \mathbf{u}_{m,t})$ is the distance from the active

Notation	Description
M	Number of controlled UAVs
$\mathbf{u}_{m,t}$	Position of controlled UAV m
$v_{m,t}$	Flight speed of controlled UAV m
Δ_t	Time duration of a time slot
$\varphi_{m,t}$	Yaw angle of controlled UAV m
$\phi_{m,t}$	Pitch angle of controlled UAV m
$\tau_{m,t}$	Transmit time of signals
c	Speed of light
\mathbf{s}_t	Position of the target UAV
$d_{m,t}$	Distance from the target UAV to controlled UAV m
$p_{m,t}$	Transmit power of controlled UAV m
$\omega_{m,t}$	Random Gaussian noise
a_t	Transmitting signal
$y_{m,t}$	Received signals at passive UAV m
$x_{m,t}$	Scattering coefficient of the target UAV
$h_{m,t}$	Path loss between UAVs
β_0	LoS path loss at a reference distance
$\gamma_{m,t}^A$	SNR of signals received by passive UAV m
σ^2	Variance of measurement error
$E_{m,t}$	Energy consumption of the active UAV
$k_{m,t}$	Distance between the BS and passive UAV m
\mathbf{s}_B	Position of the BS
$\chi_{m,t}$	Elevation angle of passive UAV m
L_{FS}	Free-space path loss
$l_{m,t}^{LoS}$	LoS path loss from UAV m to the BS
$\Pr(l_{m,t}^{LoS})$	Probability of LoS
$l_{m,t}^{NLoS}$	NLoS path loss from UAV m to the BS
D	Data size of the distance information
$\gamma_{m,t}^B$	SNR of signals received at the BS
W	Bandwidth
ϵ^2	Variance of Gaussian noise
$T_{m,t}^A$	Transmission delay between UAVs
\hat{r}_t	Distance measurement information
r_t	Actual distance
$n_{m,t}$	Measurement information error
$T_{m,t}^B$	Transmission delay from passive UAV m to the BS
V	Number of time slots
$\hat{\mathbf{s}}_t$	Estimated position of the target UAV

UAV to the target UAV and then from the target UAV to passive UAV m with $d_{0,t}(\mathbf{u}_{0,t}, \mathbf{s}_t) = \|\mathbf{u}_{0,t} - \mathbf{s}_t\|$ being the distance between the active UAV and the target UAV located at $\mathbf{s}_t = [x_t, y_t, z_t]^T$ and $d_{m,t}(\mathbf{s}_t, \mathbf{u}_{m,t}) = \|\mathbf{s}_t - \mathbf{u}_{m,t}\|$ being the distance between the target UAV and passive UAV m .

The signals transmitted from the active UAV, reflected by the target UAV, and received by passive UAV m at time slot t is given by

$$y_{m,t} = \sqrt{p_{0,t}} h_{m,t} x_{m,t} h_{0,t} a_t e^{-j2\pi f_t t} + w_{m,t}, \quad (3)$$

where $p_{0,t}$ is the transmit power of the active UAV at time slot t , $x_{m,t}$ represents the scattering coefficient of the target UAV [20], and $w_{m,t}$ is Gaussian noise with zero mean and ϵ^2 variance. $h_{0,t} = \sqrt{\beta_0} d_{0,t}^{-1}(\mathbf{u}_{0,t}, \mathbf{s}_t)$ represents the path loss from the active UAV to the target UAV, and $h_{m,t} = \sqrt{\beta_0} d_{m,t}^{-1}(\mathbf{u}_{m,t}, \mathbf{s}_t)$ represents the path loss from the target UAV to passive UAV m with $\sqrt{\beta_0}$ being the line-of-sight (LoS) path loss at a reference distance [21]. We use LoS links to model the link between the active UAV and the target UAV and the links between the target UAV and passive UAVs.

At passive UAV m , the signal-to-noise ratio (SNR) of the signal transmitted by the active UAV and reflected by the target UAV is given by [22]

$$\gamma_{m,t}^A(\mathbf{u}_{0,t}, \mathbf{u}_{m,t}, p_{0,t}) = \frac{p_{0,t}|h_{m,t}x_{m,t}h_{0,t}|^2}{\epsilon^2}. \quad (4)$$

The transmission delay from the active UAV to the target UAV and from the target UAV to passive UAV m is given by

$$T_{m,t}^A(\mathbf{u}_{0,t}, \mathbf{u}_{m,t}, p_{0,t}) = \frac{D_A}{W \log_2(1 + \gamma_{m,t}^A(\mathbf{u}_{m,t}))}, \quad (5)$$

where D_A is the size of the transmitting signals and W is the bandwidth. The energy consumption of the active UAV is given by

$$E_{m,t}(\mathbf{u}_{0,t}, \mathbf{u}_{m,t}, p_{0,t}) = p_{0,t}T_{m,t}^A(\mathbf{u}_{0,t}, \mathbf{u}_{m,t}, p_{0,t}). \quad (6)$$

Due to the limited energy of the active UAV, the transmit power of the active UAV must be optimized to minimize the positioning error of the target UAV while satisfying the energy consumption requirements of the active UAV.

2) *Passive UAV-BS Links:* Passive UAVs require to use their received signals to calculate the distance $\hat{r}_{m,t}$ from the active UAV to the target UAV and then from the target UAV to the passive UAV. Then, each passive UAV will transmit its calculated distance $\hat{r}_{m,t}$ to the BS. We use probabilistic LoS and non-line-of sight (NLoS) links to model the links between passive UAVs and the BS. The LoS and NLoS path loss of passive UAV m transmitting signals to the BS located at s_B at time slot t is given by

$$\begin{aligned} l_{m,t}^{\text{LoS}}(\mathbf{u}_{m,t}) \\ = L_{\text{FS}}(k_0) + 10\mu_{\text{LoS}} \log(k_{m,t}(\mathbf{u}_{m,t}, s_B)) + \lambda_{\sigma_{\text{LoS}}}, \end{aligned} \quad (7)$$

$$\begin{aligned} l_{m,t}^{\text{NLoS}}(\mathbf{u}_{m,t}) \\ = L_{\text{FS}}(k_0) + 10\mu_{\text{NLoS}} \log(k_{m,t}(\mathbf{u}_{m,t}, s_B)) + \lambda_{\sigma_{\text{NLoS}}}, \end{aligned} \quad (8)$$

where $L_{\text{FS}}(k_0) = 20 \log(k_0 f_0^B 4\pi/c)$ is the free-space path loss with k_0 being the free-space reference distance and f_0^B being the carrier frequency. $k_{m,t}(\mathbf{u}_{m,t}, s_B)$ is the distance between passive UAV m and the BS at time slot t . $\lambda_{\sigma_{\text{LoS}}}$ and $\lambda_{\sigma_{\text{NLoS}}}$ are the shadowing random variables, which are Gaussian variables in dB with zero mean and $(\sigma_{\text{LoS}}^B)^2$, $(\sigma_{\text{NLoS}})^2$ dB variances. The probability of LoS is given by

$$\Pr(l_{m,t}^{\text{LoS}}(\mathbf{u}_{m,t})) = (1 + X \exp(-Y[\chi_{m,t} - X]))^{-1}, \quad (9)$$

where X and Y are constants which are related to the environment factors, and $\chi_{m,t}$ is the elevation angle of passive UAV m at time slot t , which satisfies $\sin(\chi_{m,t}) = \frac{z_{m,t}}{k_{m,t}(\mathbf{u}_{m,t}, s_B)}$. Therefore, the path loss from passive UAV m to the BS at time slot t is given by

$$\begin{aligned} \bar{l}_{m,t}(\mathbf{u}_{m,t}) &= \Pr(l_{m,t}^{\text{LoS}}(\mathbf{u}_{m,t})) \times l_{m,t}^{\text{LoS}}(\mathbf{u}_{m,t}) \\ &\quad + (1 - \Pr(l_{m,t}^{\text{LoS}}(\mathbf{u}_{m,t}))) \times l_{m,t}^{\text{NLoS}}(\mathbf{u}_{m,t}). \end{aligned} \quad (10)$$

The SNR of the signal transmitted from passive UAV m to the BS at time slot t is given by

$$\gamma_{m,t}^B(\mathbf{u}_{m,t}) = \frac{p_{m,t}}{\epsilon^2} 10^{-\bar{l}_{m,t}(\mathbf{u}_{m,t})/10}, \quad (11)$$

where $p_{m,t}$ is the transmit power of passive UAV m at time slot t . The transmission delay from passive UAV m to the BS at time slot t is given by

$$T_{m,t}^B(\mathbf{u}_{m,t}) = \frac{D_B}{W \log_2(1 + \gamma_{m,t}^B(\mathbf{u}_{m,t}))}, \quad (12)$$

where D_B is the data size of the distance information transmitted from passive UAVs to the BS.

B. Model for Positioning

Let $\hat{r}_t = [\hat{r}_{1,t}, \dots, \hat{r}_{4,t}]^T$ be the distance measurement information received by the BS from passive UAVs. Then, the BS uses \hat{r}_t to estimate the position of the target UAV. A two-stage weighted least-squares (TSWLS) method [23] is exploited to determine the position of the target UAV. Hence, we assume that the distance measurements \hat{r}_t from the active UAV to passive UAV m via the target UAV involves an error, and can be expressed by $\hat{r}_{m,t} = r_{m,t} + n_{m,t}(p_{0,t}, \mathbf{u}_{0,t}, \mathbf{u}_{m,t})$, where $n_{m,t}(p_{0,t}, \mathbf{u}_{0,t}, \mathbf{u}_{m,t})$ is the independent Gaussian measurement error with zero mean and variance $\sigma_{m,t}^2(p_{0,t}, \mathbf{u}_{0,t}, \mathbf{u}_{m,t})$ [24]. Based on the distance measurement information \hat{r}_t , 3D position of the controlled UAVs $\mathbf{U}_t = [\mathbf{u}_{0,t}, \dots, \mathbf{u}_{4,t}]^T$ and the transmit power $p_{0,t}$ of the active UAV at time slot t , the estimated position of the target UAV $\hat{s}_t(\mathbf{U}_t, p_{0,t})$ can be obtained via the TSWLS method in [23].

C. Problem Formulation

Given the defined system model, our goal is to minimize the mean squared error (MSE) between the estimated position and the actual position of the target UAV over a time period T that consists of V time slots under the delay and movement constraints of UAVs. This minimization problem includes optimizing the transmit power of the active UAV and the trajectories of passive and active UAVs. The optimization problem is given by

$$\min_{p_{0,t}, \varphi_t, \phi_t} \sum_{t=1}^V \sqrt{(\hat{s}_t(\mathbf{U}_t, p_{0,t}) - s_t)^2}, \quad (13)$$

$$\text{s.t. } E_{m,t} \leq E_{\max}, \quad (13a)$$

$$T_{m,t}^B(\mathbf{u}_{m,t}) \leq \xi, \quad \forall m \in \mathcal{M}, \quad (13b)$$

$$\varphi^{\min} \leq \varphi_{m,t} \leq \varphi^{\max}, \quad \forall m \in \mathcal{M}, \quad (13c)$$

$$\phi^{\min} \leq \phi_{m,t} \leq \phi^{\max}, \quad \forall m \in \mathcal{M}, \quad (13d)$$

$$L_{\min} \leq \|\mathbf{u}_{m,t+1} - \mathbf{s}_{t+1}\| \leq L_{\max}, \quad \forall m \in \mathcal{M}, \quad (13e)$$

$$L_{\min} \leq \|\mathbf{u}_{m,t+1} - \mathbf{u}_{m',t+1}\| \leq L_{\max}, \quad \forall m, m' \in \mathcal{M}, \quad (13f)$$

where $p_{0,t}$ is the transmit power of the active UAV, $\varphi_t = [\varphi_{0,t}, \dots, \varphi_{4,t}]^T$ and $\phi_t = [\phi_{0,t}, \dots, \phi_{4,t}]^T$ are the yaw angle vector and the pitch angle vector for the active UAV

and passive UAVs, respectively. (13a) is a maximum energy consumption constraint for the active UAV. (13b) is the delay needed to transmit distance information from each passive UAV to the BS. (13c) and (13d) are the yaw angle and the pitch angle constraints for the controlled UAVs. (13e) is the constraint of the distance between a controlled UAV and the target UAV, and (13f) is the constraint of the distance between any two controlled UAVs.

The problem (13) is challenging to solve by conventional optimization algorithms due to the following reasons. First, since the Hessian matrix of objective function in (13) is not a positive semi-definite matrix, the problem (13) is non-convex. Second, the BS must know the coordinates of the target UAV to optimize the transmit power of the active UAV and trajectories of controlled UAVs using optimization methods. However, the target UAV is moving and hence the BS may not be able to obtain the real-time position of the target UAV. To solve the optimization problem (13), we use a distributed RL algorithm which finds the probability distribution of the sum of future rewards to estimate the expected value of the sum of future rewards accurately. The proposed method enables the active UAV to determine its transmit power and each controlled UAV to determine its trajectory using its individual observation. Hence, using distributed RL, the BS and controlled UAVs can minimize the positioning error of the target UAV.

III. PROPOSED Z FUNCTION DECOMPOSITION BASED RL

In this section, we introduce a ZD-RL method to solve the optimization problem in (13). Compared to standard RL algorithms [19] such as deep Q-network (DQN), the ZD-RL method aims to find the probability distribution of the sum of future rewards and uses this probability distribution to accurately estimate the expected value of the sum of future rewards. Hence, the ZD-RL method can improve the efficiency of optimizing the transmit power of the active UAV and trajectories of controlled UAVs. Next, we first introduce the components of the ZD-RL method. Then, the process of using the ZD-RL method to find the global optimal transmit power for the active UAV and trajectories for controlled UAVs is explained.

A. Components of the ZD-RL method

The ZD-RL method consists of six components: a) agents, b) actions, c) states, d) rewards, e) individual Z function, f) global Z function, which are specified as follows:

- *Agents*: The agents that perform the ZD-RL method are the controlled UAVs. Each passive UAV must decide its yaw angle and pitch angle and the active UAV must decide its transmit power, yaw angle, and pitch angle at each time slot.
- *State space*: A state of each agent is used to describe the local environment of each controlled UAV. In particular, a state of each passive UAV consists of its 3D coordinates and the distance measurements from the active UAV to the target UAV, and then from the target UAV to the

passive UAV. Hence, a state of a passive UAV m at time slot t is $\mathbf{o}_{m,t} = [x_{m,t}, y_{m,t}, z_{m,t}, \hat{r}_{m,t}]$. Since the active UAV cannot obtain the distance measurement, and the BS does not need the distance measurement of the active UAV to estimate the position of the target UAV, the state of the active UAV is $\mathbf{o}_{0,t} = [x_{0,t}, y_{0,t}, z_{0,t}]$. The states of all agents at time slot t can be represented by a vector $\mathbf{o}_t = [\mathbf{o}_{0,t}, \dots, \mathbf{o}_{4,t}]$.

- *Actions*: The action of each passive UAV is the yaw angle and the pitch angle and the action of the active UAV is the transmit power, the yaw angle and the pitch angle. Hence, an action of passive UAV m at time slot t can be expressed as $\mathbf{a}_{m,t} = [\varphi_{m,t}, \phi_{m,t}]$, and an action of the active UAV at time slot t is $\mathbf{a}_{0,t} = [p_{0,t}, \varphi_{0,t}, \phi_{0,t}]$. The actions of all controlled UAVs at time slot t is $\mathbf{a}_t = [\mathbf{a}_{0,t}, \dots, \mathbf{a}_{4,t}]$.
- *Reward*: The reward of each controlled UAV captures the positioning accuracy of the target UAV resulting from a selected action. Given the global state \mathbf{o}_t and the selected action \mathbf{a}_t , the reward of each controlled UAV at time slot t is $R_t(\mathbf{o}_t, \mathbf{a}_t) = -\sqrt{(\hat{s}_t(\mathbf{U}_t, p_{0,t}) - s_t)^2}$. Note that, $R_t(\mathbf{o}_t, \mathbf{a}_t)$ increases as the positioning error in (13) decreases, which implies that maximizing the reward of each controlled UAV can minimize the positioning error.
- *Individual Z function*: Z function is defined as the sum of future reward under a given state $\mathbf{o}_{m,t}$, a selection action $\mathbf{a}_{m,t}$, and a policy π , which can be expressed as $Z(\mathbf{o}_{m,t}, \mathbf{a}_{m,t}) = \sum_{t=0}^{\infty} \gamma^t R(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})$, where γ is a discounted factor. Given the definition, our purpose is to estimate the probability distribution of $Z(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})$. This is different from DQN [19] that uses a neural network to estimate the sum of expected future reward. In particular, the relationship between Q function and our defined Z function is expressed as

$$\begin{aligned} Q(\mathbf{o}_{m,t}, \mathbf{a}_{m,t}) &= \mathbb{E}_{\pi}[Z(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})] \\ &= \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t R(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})\right]. \end{aligned} \quad (14)$$

The advantage of estimating Z function instead of Q function is that Q function values estimated using the probability distribution of Z function are more accurate compared to Q function values directly estimated by DQN [25]. Hence, the ZD-RL method ensures the stability and effectiveness of model convergence [26]. Next, we introduce the process of estimating the probability distribution of Z function. First, we introduce the cumulative distribution function (CDF) of $Z(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})$, which is given by

$$F(z) = \mathbb{P}(Z(\mathbf{o}_{m,t}, \mathbf{a}_{m,t}) \leq z), \quad (15)$$

where $F(z)$ represents the probability that $Z(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})$ is smaller than a value z . To estimate the probability distribution of $Z(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})$, we use a DNN. The input of the DNN is the individual state $\mathbf{o}_{m,t}$, individual action $\mathbf{a}_{m,t}$ and a probability value ς_i , and the output is a value

of Z function, such as $\hat{Z}_{\omega_m}(\mathbf{o}_{m,t}, \mathbf{a}_{m,t}, \varsigma_i)$, where ω_m is the parameters of the DNN. The relationship between the input of DNN and its output can be expressed as

$$\varsigma_i = \mathbb{P}\left(Z(\mathbf{o}_{m,t}, \mathbf{a}_{m,t}) \leq \hat{Z}_{\omega_m}(\mathbf{o}_{m,t}, \mathbf{a}_{m,t}, \varsigma_i)\right). \quad (16)$$

From (16), we can see that Z function is to find a value of $\hat{Z}_{\omega_m}(\mathbf{o}_{m,t}, \mathbf{a}_{m,t}, \varsigma_i)$ such that $\mathbb{P}(Z(\mathbf{o}_{m,t}, \mathbf{a}_{m,t}) \leq \hat{Z}_{\omega_m}(\mathbf{o}_{m,t}, \mathbf{a}_{m,t}, \varsigma_i)) = \varsigma_i$. Given the relationship between ς_i and $\hat{Z}_{\omega_m}(\mathbf{o}_{m,t}, \mathbf{a}_{m,t}, \varsigma_i)$, the next step is to determine the value of ς_i such that we can use less DNN outputs to estimate the entire probability distribution of $Z(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})$. To this end, we use a quantile vector $\varsigma = [\varsigma_1, \dots, \varsigma_N]$ with $\varsigma_i = \frac{i}{N}, i = 1, \dots, N$.

- *Global Z function:* The global Z function $Z_T(\mathbf{o}_t, \mathbf{a}_t)$ is used to estimate the probability distribution of all controlled UAVs' achievable future rewards at each global state \mathbf{o}_t and action \mathbf{a}_t . Similarly to individual Z functions, the probability distribution of the global Z function is approximated by a set of global Z function values with a quantile vector ς , and the approximated global Z function is represented by $\hat{Z}_T(\mathbf{o}_t, \mathbf{a}_t, \varsigma)$. Based on the distributional individual-global-max principle [27], the relationship between $\hat{Z}_T(\mathbf{o}_t, \mathbf{a}_t, \varsigma)$ and $\hat{Z}_{\omega_m}(\mathbf{o}_{m,t}, \mathbf{a}_{m,t}, \varsigma)$ is given by

$$\begin{aligned} \hat{Z}_T(\mathbf{o}_t, \mathbf{a}_t, \varsigma) &= \sum_{m=0}^4 M(\mathbf{o}_{m,t}, \mathbf{a}_{m,t}, \varsigma) \\ &+ \sum_{m=0}^4 \left(\hat{Z}_{\omega_m}(\mathbf{o}_{m,t}, \mathbf{a}_{m,t}, \varsigma) - M(\mathbf{o}_{m,t}, \mathbf{a}_{m,t}, \varsigma) \right), \end{aligned} \quad (17)$$

where $M(\mathbf{o}_{m,t}, \mathbf{a}_{m,t}, \varsigma)$ is the approximated expected value of $\hat{Z}_{\omega_m}(\mathbf{o}_{m,t}, \mathbf{a}_{m,t}, \varsigma)$ and can be written as $M(\mathbf{o}_{m,t}, \mathbf{a}_{m,t}, \varsigma) = \frac{1}{N} \sum_{i=1}^N \hat{Z}_{\omega_m}(\mathbf{o}_{m,t}, \mathbf{a}_{m,t}, \varsigma_i)$.

B. Training of the ZD-RL Method

Here, we describe the entire training process of the ZD-RL method for optimizing the transmit power of the active UAV and trajectories of all controlled UAVs. In particular, we will first introduce the loss function of the ZD-RL method. Then, we introduce the training procedures. The total loss of the ZD-RL method is defined as the sum of the pair-wise loss for two values ς_i, ς_j based on quantile Huber loss [28], where $\varsigma_i, \varsigma_j \in \varsigma$. Compared to mean-square-error (MSE) loss and mean absolute error (MAE) used in traditional RL, the quantile Huber loss can reduce the sensitivity to abnormal samples that deviate from the normal range. The total loss is

$$\begin{aligned} \mathcal{L}_T(\omega_0, \dots, \omega_4) &= \frac{1}{N} \sum_{t=1}^V \sum_{i=1}^N \sum_{j=1}^N |\varsigma_i - \mathbb{1}_{\{u(\mathbf{o}_t, \mathbf{a}_t, \varsigma_i, \varsigma_j) < 0\}}| \frac{G(u(\mathbf{o}_t, \mathbf{a}_t, \varsigma_i, \varsigma_j))}{\eta}, \end{aligned} \quad (18)$$

where $\mathbb{1}_{\{x\}} = 1$ when $x < 0$ and $\mathbb{1}_{\{x\}} = 0$, otherwise. $u(\mathbf{o}_t, \mathbf{a}_t, \varsigma_i, \varsigma_j) = R_t(\mathbf{o}_t, \mathbf{a}_t) + \gamma \hat{Z}_T(\mathbf{o}_{t+1}, \mathbf{a}_{t+1}, \varsigma_j) -$

Algorithm 1 ZD-RL method for Solving Problem (13)

```

1: Initialize the DNN parameters  $\omega_m$  of each controlled UAV, a quantile vector  $\varsigma$ .
2: for each iteration do
3:   for each controlled UAV  $m$  do
4:     for each time slot  $t$  do
5:       Observe the observation  $\mathbf{o}_{m,t}$ .
6:       Select an action according to a  $\epsilon$ -greedy scheme.
7:       Calculate individual Z function values  $\hat{Z}_{\omega_m}(\mathbf{o}_{m,t}, \mathbf{a}_{m,t}, \varsigma)$  and  $\hat{Z}_{\omega_m}(\mathbf{o}_{m,t+1}, \mathbf{a}_{m,t+1}, \varsigma)$ .
8:     end for
9:     Controlled UAVs transmit  $\mathbf{o}_{m,t}, \hat{Z}_{\omega_m}(\mathbf{o}_{m,t}, \mathbf{a}_{m,t}, \varsigma)$ , and  $\hat{Z}_{\omega_m}(\mathbf{o}_{m,t+1}, \mathbf{a}_{m,t+1}, \varsigma)$  to the BS.
10:    end for
11:    The BS calculates the reward and global Z function, and transmits to controlled UAVs.
12:    for each controlled UAV  $m$  do
13:      Update  $\omega_m$  using  $R(\mathbf{o}_t, \mathbf{a}_t), \hat{Z}_T(\mathbf{o}_t, \mathbf{a}_t, \varsigma)$  and  $\hat{Z}_T(\mathbf{o}_{t+1}, \mathbf{a}_{t+1}, \varsigma)$  based on (19).
14:    end for
15:  end for

```

$\hat{Z}_T(\mathbf{o}_t, \mathbf{a}_t, \varsigma_i)$ with $\mathbf{a}_{m,t+1} = \arg \max_{\mathbf{a}'_m} M(\mathbf{o}_{m,t+1}, \mathbf{a}'_m, \varsigma)$ [29]. $G(u(\mathbf{o}_t, \mathbf{a}_t, \varsigma_i, \varsigma_j))$ is given by

$$G(u(\mathbf{o}_t, \mathbf{a}_t, \varsigma_i, \varsigma_j)) = \begin{cases} \frac{1}{2} (u(\mathbf{o}_t, \mathbf{a}_t, \varsigma_i, \varsigma_j))^2, & \text{if } |u(\mathbf{o}_t, \mathbf{a}_t, \varsigma_i, \varsigma_j)| \leq \eta, \\ \eta (|u(\mathbf{o}_t, \mathbf{a}_t, \varsigma_i, \varsigma_j)| - \frac{1}{2} \eta), & \text{otherwise,} \end{cases}$$

where η is a hyper-parameter that determines the emphasis of Huber loss on MSE or MAE.

The training process consists of the following three steps:

- *Step 1 (training at controlled UAVs):* Given a quantile vector $\varsigma = [\varsigma_1, \dots, \varsigma_N]$, each controlled UAV observes its local state $\mathbf{o}_{m,t}$, takes an action $\mathbf{a}_{m,t}$ according to a ϵ -greedy algorithm, and calculates its individual Z function values $\hat{Z}_{\omega_m}(\mathbf{o}_{m,t}, \mathbf{a}_{m,t}, \varsigma)$, $\hat{Z}_{\omega_m}(\mathbf{o}_{m,t+1}, \mathbf{a}_{m,t+1}, \varsigma)$. Then, each UAV transmits its state $\mathbf{o}_{m,t}$, individual Z function values $\hat{Z}_{\omega_m}(\mathbf{o}_{m,t}, \mathbf{a}_{m,t}, \varsigma)$ and $\hat{Z}_{\omega_m}(\mathbf{o}_{m,t+1}, \mathbf{a}_{m,t+1}, \varsigma)$ to the BS.
- *Step 2 (training at the BS):* After collecting individual state and individual Z function values from all controlled UAVs, the BS calculates the reward $R_t(\mathbf{o}_t, \mathbf{a}_t)$ and the global Z function values $\hat{Z}_T(\mathbf{o}_t, \mathbf{a}_t, \varsigma)$, $\hat{Z}_T(\mathbf{o}_{t+1}, \mathbf{a}_{t+1}, \varsigma)$ based on (17), and transmits $R_t(\mathbf{o}_t, \mathbf{a}_t)$, $\hat{Z}_T(\mathbf{o}_t, \mathbf{a}_t, \varsigma)$, and $\hat{Z}_T(\mathbf{o}_{t+1}, \mathbf{a}_{t+1}, \varsigma)$ to controlled UAVs. Here, the BS does not need to implement and update any neural networks.
- *Step 3 (updating at controlled UAVs):* Each UAV updates DNN parameters to approximate the probability distribution of its individual Z function using its collected global reward and global Z function values. The update of each controlled UAV m is

$$\omega_m = \omega_m + \alpha_m \nabla_{\omega_m} \mathcal{L}_T(\omega_0, \dots, \omega_4), \quad (19)$$

where α_m is the step size. The entire training process of the ZD-RL method is summarized in Algorithm 1.

C. Convergence, Implementation, and Complexity Analysis

Next, we analyze the convergence, implementation and complexity of training the proposed ZD-RL method.

1) *Convergence Analysis:* Here, we analyze the convergence of the proposed ZD-RL algorithm. We first analyze the gap between the optimal expected value of the individual Z function of controlled UAV m and the expected value of individual Z function of controlled UAV m obtained by the proposed ZD-RL method. Then, we show that this gap will converge to zero. In particular, the gap between the optimal expected value of individual Z function of controlled UAV m and the expected value of individual Z function of controlled UAV m obtained by the proposed ZD-RL method is

$$e(\mathbf{o}_{m,t}, \mathbf{a}_{m,t}) = M(\mathbf{o}_{m,t}, \mathbf{a}_{m,t}) - M^*(\mathbf{o}_{m,t}, \mathbf{a}_{m,t}), \quad (20)$$

where $M^*(\mathbf{o}_{m,t}, \mathbf{a}_{m,t}) = \mathbb{E}[Z^*(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})]$ is the expected value of the optimal individual Z function of controlled UAV m with respect to future Z functions (i.e., $Z^*(\mathbf{o}_{m,t+1}, \mathbf{a}_{m,t+1}), Z^*(\mathbf{o}_{m,t+2}, \mathbf{a}_{m,t+2}), \dots$). From (20), we can see that if the gap $e(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})$ converges to zero, the proposed ZD-RL method converges [30]. To prove that the gap $e(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})$ will finally converge to zero, we need to analyze how the gap changes as the number of training iterations increases. In particular, we define a distributional Bellman operator to find a relationship between the individual Z function of controlled UAV m at two continuous time slots. In particular, the distributional Bellman operator of the individual Z function is defined as

$$\mathcal{T}(Z(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})) \stackrel{D}{=} R(\mathbf{o}_{m,t}, \mathbf{a}_{m,t}) + \gamma Z(\mathbf{o}_{m,t+1}, \mathbf{a}_{m,t+1}), \quad (21)$$

where $\mathbf{a}_{m,t+1} = \arg \max_{\mathbf{a}'_m} M(\mathbf{o}_{m,t+1}, \mathbf{a}'_m)$. Based on the above definition, the convergence of the proposed ZD-RL algorithm is shown in the following lemma.

Lemma 1. The proposed ZD-RL method is guaranteed to converge to zero, if the following conditions are satisfied [31]:

- 1) The gap $e(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})$ satisfies

$$\begin{aligned} e_{k+1}(\mathbf{o}_{m,t}, \mathbf{a}_{m,t}) \\ = (1 - \alpha_m) e_k(\mathbf{o}_{m,t}, \mathbf{a}_{m,t}) + \alpha_m F(\mathbf{o}_{m,t}, \mathbf{a}_{m,t}), \end{aligned} \quad (22)$$

where $F(\mathbf{o}_{m,t}, \mathbf{a}_{m,t}) = R(\mathbf{o}_{m,t}, \mathbf{a}_{m,t}) + \gamma M(\mathbf{o}_{m,t+1}, \mathbf{a}_{m,t+1}) - M^*(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})$.

- 2) $\|\mathbb{E}[F(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})]\|_\infty \leq \gamma \|e(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})\|_\infty, \forall \gamma \in (0, 1)$, where $\|\cdot\|_\infty$ represents the infinite norm taking the maximum value of the absolute value of the elements, $\mathbb{E}[F(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})]$ is the expected value of $F(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})$ with respect to the state transition probability distribution.
- 3) $\text{Var}(\mathbb{E}[F(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})]) \leq C_F(1 + \|e(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})\|_\infty^2)$, where $\text{Var}(\mathbb{E}[F(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})])$ is the variance of $\mathbb{E}[F(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})]$, and C_F is a constant with $C_F \geq 0$.

Proof: See Appendix A. \square

2) *Implementation Analysis:* Next, we explain the implementation of the proposed ZD-RL method for UAV localization. The proposed ZD-RL method includes an offline

training stage and an online decision-making stage. In the offline training phase, each controlled UAV requires 1) the positioning error between the estimated position and the actual position of the target UAV and 2) the global Z function value to update its DNN parameters based on (18) and (19). To calculate the positioning error, the BS needs to collect the distance measurement information $\hat{r}_{m,t}$, the transmit power of the active UAV, and the positions of controlled UAVs. The distance information is estimated by the signals transmitted from the active UAV to the passive UAV and reflected by the target UAV. The transmit power of the active UAV is notified by the active UAV, and the positions of controlled UAVs are transmitted by controlled UAVs. To calculate the global Z functions, the BS needs to collect individual Z functions as shown in (17) in our training stage. In the online decision-making stage, the well trained DNN can be directly used to determine the transmit power, yaw angle, and pitch angle of controlled UAVs.

3) *Complexity Analysis:* The complexity of the proposed algorithm lies in training the DNN of each controlled UAV. To analyze the complexity of training the designed ZD-RL method, we first assume that the value of the transmit power $p_{m,t}$ of controlled UAV m at time slot t is selected from a set of $\{p_{m,t}^1, \dots, p_{m,t}^{N_p}\}$, the yaw angle $\varphi_{m,t}$ of controlled UAV m is selected from a set $\{\varphi_{m,t}^1, \dots, \varphi_{m,t}^{N_1}\}$, and the pitch angle $\phi_{m,t}$ is selected from a set $\{\phi_{m,t}^1, \dots, \phi_{m,t}^{N_2}\}$ with N_p , N_1 , and N_2 being the number of elements in their corresponding sets. Since we only consider optimizing the transmit power of the active UAV and the transmit power of passive UAVs are constant, we have $N_p = 1$, when $m = 1, \dots, 4$. The interval of two yaw angles $\Delta\varphi_m$ is defined as $\Delta\varphi_m = \varphi_{m,t}^{i+1} - \varphi_{m,t}^i, i = 1, \dots, N_1 - 1$ and the interval of two pitch angles $\Delta\phi_m$ is defined as $\Delta\phi_m = \phi_{m,t}^{i+1} - \phi_{m,t}^i, i = 1, \dots, N_2 - 1$. Hence, the relationship between N_1 , N_2 and the interval of angles $\Delta\varphi_m$ and $\Delta\phi_m$ is $N_1 = \frac{\varphi_{m,t}^{N_1} - \varphi_{m,t}^1}{\Delta\varphi_m} + 1$, and $N_2 = \frac{\phi_{m,t}^{N_2} - \phi_{m,t}^1}{\Delta\phi_m} + 1$. Then, the complexity of training the designed ZD-RL method is shown in the following proposition.

Proposition 1. The time complexity of training the proposed ZD-RL method is

$$\begin{aligned} \mathcal{O} \left(\sum_{l=1}^{L-1} l_i l_{i+1} + |\mathbf{o}_{m,t}| l_1 + N l_L \right. \\ \left. + l_L \left(N_p \left(\frac{\varphi_{m,t}^{N_1} - \varphi_{m,t}^1}{\Delta\varphi_m} + 1 \right) \left(\frac{\phi_{m,t}^{N_2} - \phi_{m,t}^1}{\Delta\phi_m} + 1 \right) \right) \right), \end{aligned} \quad (23)$$

where $|\mathbf{o}_{m,t}|$ is the size of state space, l_i is the number of neurons in hidden layer i , L is the number of hidden layers, N is the number of elements in the quantile vector.

Proof: Based on [32], at each iteration, the time-complexity of training ZD-RL method is $\mathcal{O} \left(\sum_{l=1}^{L-1} l_i l_{i+1} + |\mathbf{o}_{m,t}| l_1 + N l_L + |\mathbf{a}_{m,t}| l_L \right)$, where $|\mathbf{a}_{m,t}|$ is the size of action space. Since $|\mathbf{a}_{m,t}|$ depends on

the interval $\Delta\varphi_m$ of two adjacent yaw angles and the interval $\Delta\phi_m$ of two adjacent pitch angles, $|\mathbf{a}_{m,t}|$ can be given by

$$|\mathbf{a}_{m,t}| = N_p \times \left(\frac{\varphi_{m,t}^{N_1} - \varphi_{m,t}^1}{\Delta\varphi_m} + 1 \right) \times \left(\frac{\phi_{m,t}^{N_2} - \phi_{m,t}^1}{\Delta\phi_m} + 1 \right), \quad (24)$$

where $N_p = 1$ when $m = 1, \dots, 4$. This is because we only consider optimizing the transmit power of the active UAV and the transmit power of passive UAVs are constant. Based on (24), the time-complexity of training the proposed ZD-RL method is

$$\begin{aligned} \mathcal{O} & \left(\sum_{l=1}^{L-1} l_l l_{l+1} + |\mathbf{o}_{m,t}| l_1 + N l_L \right. \\ & \left. + l_L \left(N_p \left(\frac{\varphi_{m,t}^{N_1} - \varphi_{m,t}^1}{\Delta\varphi_m} + 1 \right) \left(\frac{\phi_{m,t}^{N_2} - \phi_{m,t}^1}{\Delta\phi_m} + 1 \right) \right) \right). \end{aligned} \quad (25)$$

This completes the proof. \square

From proposition 1, we see that as the interval $\Delta\varphi_m$ and $\Delta\phi_m$ of two adjacent angles decreases, the time-complexity of training the proposed ZD-RL method at each iteration increases and hence the number of iterations that the ZD-RL method required to converge increases. However, when the intervals $\Delta\varphi_m$ and $\Delta\phi_m$ increases, the controlled UAVs may find better yaw angles and pitch angles for the target UAV localization thus improving localization performance.

IV. CONTROLLED UAV DEPLOYMENT FOR TARGET UAV LOCALIZATION

In this section, we aim to find the positions of controlled UAVs that can minimum the positioning error of the target UAV. At each time slot, the relationship between the positions of controlled UAVs and the distance $r_{m,t}$ from the active UAV to the target UAV and then from the target UAV to passive UAV m is given by

$$r_{m,t} = d_{m,t}(\mathbf{u}_{m,t}, \mathbf{s}_t) + d_{0,t}(\mathbf{u}_{0,t}, \mathbf{s}_t), \quad (26)$$

Taking differentiation at both sides of (26), we have

$$\begin{aligned} dr_{m,t} &= \left(\frac{x_t - x_{m,t}}{d_{m,t}} + \frac{x_t - x_{0,t}}{d_{0,t}} \right) dx_t \\ &+ \left(\frac{y_t - y_{m,t}}{d_{m,t}} + \frac{y_t - y_{0,t}}{d_{0,t}} \right) dy_t \\ &+ \left(\frac{z_t - z_{m,t}}{d_{m,t}} + \frac{z_t - z_{0,t}}{d_{0,t}} \right) dz_t, \quad m = 1, 2, 3, 4. \end{aligned} \quad (27)$$

Then, we can rewrite (27) as

$$dr_t = M ds_t \quad (28)$$

where $dr_t = [dr_{1,t}, dr_{2,t}, dr_{3,t}, dr_{4,t}]^T$, $ds_t = [dx_t, dy_t, dz_t]^T$, and

$$M = \begin{bmatrix} \frac{x_t - x_{1,t}}{d_{1,t}} + \frac{x_t - x_{0,t}}{d_{0,t}} & \frac{y_t - y_{1,t}}{d_{1,t}} + \frac{y_t - y_{0,t}}{d_{0,t}} & \frac{z_t - z_{1,t}}{d_{1,t}} + \frac{z_t - z_{0,t}}{d_{0,t}} \\ \frac{x_t - x_{2,t}}{d_{2,t}} + \frac{x_t - x_{0,t}}{d_{0,t}} & \frac{y_t - y_{2,t}}{d_{2,t}} + \frac{y_t - y_{0,t}}{d_{0,t}} & \frac{z_t - z_{2,t}}{d_{2,t}} + \frac{z_t - z_{0,t}}{d_{0,t}} \\ \frac{x_t - x_{3,t}}{d_{3,t}} + \frac{x_t - x_{0,t}}{d_{0,t}} & \frac{y_t - y_{3,t}}{d_{3,t}} + \frac{y_t - y_{0,t}}{d_{0,t}} & \frac{z_t - z_{3,t}}{d_{3,t}} + \frac{z_t - z_{0,t}}{d_{0,t}} \\ \frac{x_t - x_{4,t}}{d_{4,t}} + \frac{x_t - x_{0,t}}{d_{0,t}} & \frac{y_t - y_{4,t}}{d_{4,t}} + \frac{y_t - y_{0,t}}{d_{0,t}} & \frac{z_t - z_{4,t}}{d_{4,t}} + \frac{z_t - z_{0,t}}{d_{0,t}} \end{bmatrix}. \quad (29)$$

Based on (28), the MSE between the estimated position $\hat{\mathbf{s}}_t$ and the actual position \mathbf{s}_t of the target UAV in (13) at time slot t can be expressed as $e_t = \sqrt{(dx_t)^2 + (dy_t)^2 + (dz_t)^2}$ [33]. Hence, we have $e_t = \sqrt{\text{tr}(\mathbb{E}[ds_t ds_t^T])}$, where $\text{tr}(\cdot)$ is the trace of the matrix. Then, the minimum value of the positioning error e_t of the target UAV is shown in the following proposition.

Theorem 2. If the distances between passive UAVs and the target UAV satisfy $d_{1,t} = d_{2,t} = d_{3,t} = d_{4,t}$, the minimum positioning error of the target UAV e_t is

$$e_t = \sqrt{4k(L_{\min})^2 \text{tr}((M^T M)^{-1})}. \quad (30)$$

Proof: See Appendix B. \square

From Theorem 2, we can see that the minimum positioning error of the target UAV depends on the safety distance L_{\min} between any two UAVs in constraint (13e), and the value of $\text{tr}((M^T M)^{-1})$ which relies on the positions of controlled UAVs. Theorem 2 also shows that as the distance between each controlled UAV and the target UAV is minimum (i.e., $d_{1,t} = d_{2,t} = d_{3,t} = d_{4,t} = L_{\min}$), the positioning error can be minimized.

Based on Theorem 2, next, we can also derive the minimum positioning error of the target UAV when the position of the active UAV is given, which is shown in the following proposition.

Lemma 2. Given the positions of the target UAV \mathbf{s}_t and the active UAV $\mathbf{u}_{0,t}$, if the distances from passive UAVs to the target UAV satisfy $d_{1,t} = d_{2,t} = d_{3,t} = d_{4,t}$, the minimum positioning error of the target UAV is

$$e_t = \frac{3}{2}(L_{\min} + d_{0,t})\sqrt{k}, \quad (31)$$

where k is a coefficient [24].

Proof: See Appendix C. \square

From Lemma 2, we see that when the positions of the active UAV and the target UAV are given, the minimum positioning error only depends on the distance L_{\min} between each passive UAV and the target UAV.

V. SIMULATION RESULTS AND ANALYSIS

For our simulations, five controlled UAVs and a BS jointly localize a target UAV. The moving speed of each controlled UAV is $v_{m,t} = 10$ m/s and the time duration of a time slot is

TABLE II
PARAMETERS

Parameters	Values	Parameters	Values
c	$3e^8$ m/s	$p_{m,t}$	5 W
ϵ^2	-95 dBm	W	1 MHz
$(\sigma_{\text{LoS}}^B)^2$	8.41	$(\sigma_{\text{NLoS}}^B)^2$	33.78
E_{\max}	100 kJ	ξ	1 s
L_{\min}	100 m	L_{\max}	10 km
ϕ_{\min}	-15°	ϕ_{\max}	15°
φ_{\min}	-15°	φ_{\max}	15°
D_B	5 bit	V	30
μ_{LoS}^B	2	μ_{NLoS}^B	2.4
Y	0.13	X	11.9

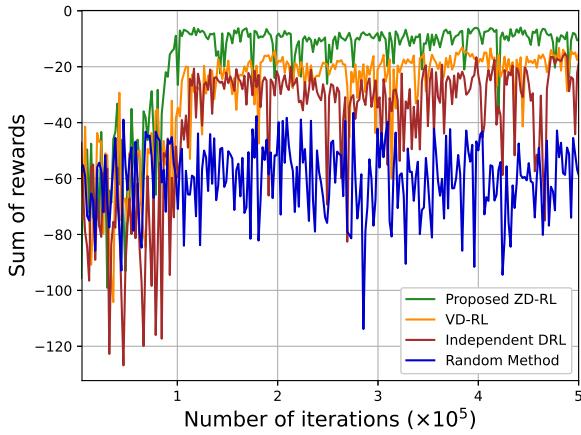


Fig. 2. Value of the sum of rewards as the total number of iterations.

$\Delta_t = 1$ s. We use the TSWLS method to estimate the position of the target UAV at each time slot [23]. Other parameters are listed in Table II. For comparison, we consider three baselines: a) independent DRL method in which each controlled UAV uses a DQN to optimize its trajectory without considering other controlled UAVs' movements, b) VD-RL method in which controlled UAVs collaboratively determine their trajectories to minimize positioning errors by decomposing the global Q function value into several individual Q function values [19], c) a random method in which each controlled UAV determines its yaw angle and pitch angle randomly.

Fig. 2 shows how the positioning accuracy changes as the number of iterations varies. From Fig. 2, we can see that the proposed ZD-RL method can improve the sum of rewards by up to 39.4% and 48.1% compared to the VD-RL and the independent DRL. The 39.4% gain stems from the fact that the proposed ZD-RL method finds the probability distribution of the sum of future rewards to estimate the expected value of the sum of future rewards using Z function values. Thus, the expected value of the sum of future rewards obtained by the proposed ZD-RL method is more accurate than that directly estimated by the VD-RL. The 48.1% gain stems from the fact that the proposed ZD-RL method enables controlled UAVs to cooperatively optimize their trajectories to minimize positioning errors while in independent DRL, each controlled UAV determines its trajectory according to its individual Q function without considering other controlled UAVs' movements. Fig. 2 also shows that the proposed ZD-

RL method can reduce the number of iterations required to converge by up to 9.6% and 15.2% compared to VD-RL and independent DRL. The reason is that the proposed method cooperatively train the trajectories of controlled UAVs and the transmit power of the active UAV using the probability distribution of the sum of future rewards. Compared to the VD-RL and independent DRL that estimate the expected value of the sum of future rewards, the proposed method are more stable and accurate thus reducing the number of iterations required to convergence.

Fig. 3 shows the actual and the estimated trajectories of the target UAV obtained by the considered algorithms. In Figs. 3(a), 3(b), and 3(c), the target UAV moves in a straight line from the stating position (500 m, 500 m, 100 m) to (789 m, 500 m, 116 m) and five controlled UAVs are randomly distributed in a sphere of radius 1000 m centered on the target UAV. In Figs. 3(d), 3(e), and 3(f), the target UAV moves in the curve of "C". In Figs. 3(g), 3(h), and 3(i), the target UAV follows the curve of "S". From Fig. 3, we can also see that the gaps between the real trajectories and estimated trajectories obtained by the proposed ZD-RL increase as the trajectories of the target UAV become more complex. This is because as the trajectories of the target UAV becomes more complex, it becomes more difficult for the proposed ZD-RL method to control the trajectories of controlled UAVs to keep small distances with the target UAV in real time. From Fig. 3, we can also see that the proposed method can estimate the target UAV position more accurately compared to the VD-RL, and independent DRL method. As the target UAV moves from the initial position to the end position, the gap between the actual positions and the positions estimated by the proposed ZD-RL method is small while the gap resulting from each baseline increases. This is due to the fact that, the proposed ZD-RL method enables controlled UAVs to cooperatively select the pitch angle and yaw angle based on the global Z function, which is generated by the BS using a set of individual Z functions thus the proposed ZD-RL method can accurately optimize the trajectories of controlled UAVs in time to track the target UAV as the target UAV moves in different trajectories.

Fig. 4 shows how the positioning error changes as the speed of the target UAV varies when the target UAV moves in the curve of "S". In Fig. 4, we can see that as the speed of the target UAV increases, the positioning errors of the considered algorithms increase. This is due to the fact that as the speed of the target UAV increases, controlled UAVs cannot follow the target UAV and the distances between the target UAV and controlled UAVs increase. Fig. 4 also shows that the proposed ZD-RL method can achieve up to 28.9% and 39.6% gains in terms of the positioning accuracy compared to the VD-RL method and independent DRL method, respectively, in the case that the target UAV moving at the speed of 22 m/s. The 28.9% gain stems from the fact that the VD-RL method obtains the global value function by linearly calculating the sum of the expected value of future rewards at each controlled UAV. However, the proposed ZD-RL method calculates the global Z

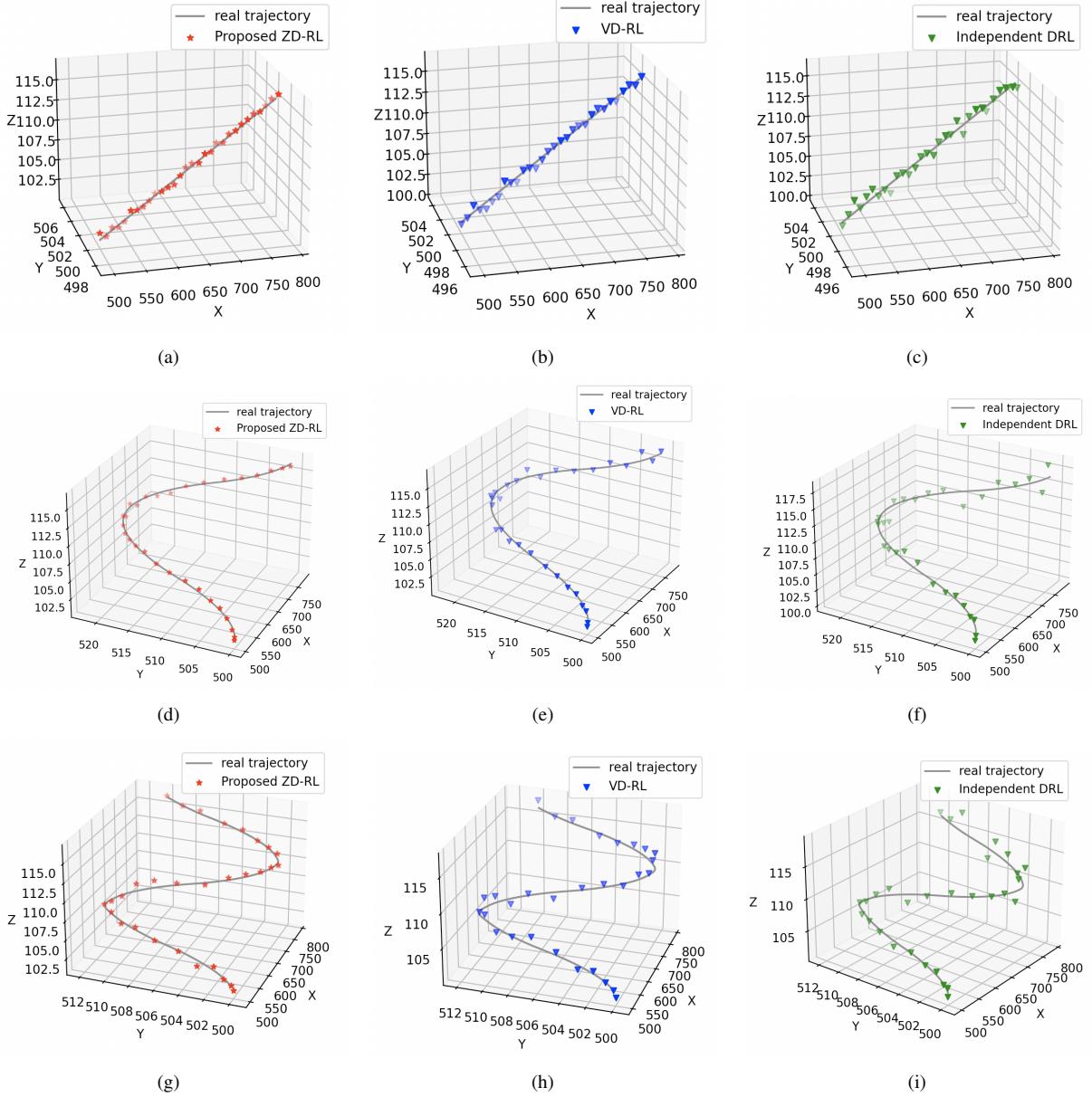


Fig. 3. The actual trajectories of the target UAV and the estimated trajectories obtained by different methods.

function using a set of global Z functions, which contains more interaction information with the environment thus being able to select pitch angle and yaw angle for controlled UAVs and optimize the transmit power for the target UAV to localize the target UAV accurately. The 39.6% gain is because the proposed ZD-RL uses the global observation information and global reward generated by the BS to train DNN parameters of each controlled UAV and enables controlled UAVs to select accurate actions by learning the movements from each other thus improving the localization accuracy cooperatively.

Fig. 5 shows how the average positioning errors change as the distance between each controlled UAV and the target UAV varies. In this simulation, the target UAV moves in the curve of “S” and the distances between each controlled UAV and the target UAV satisfy $d_{1,t} = d_{2,t} = d_{3,t} = d_{4,t}$. The yellow line in Fig. 5 represents the theoretically analytical result of

the minimum positioning error obtained by Lemma 2. In Fig. 5, we can see that the minimum positioning error obtained by the proposed ZD-RL method is 1.61 m while the theoretical positioning error is 1.18 m when $d_{m,t} = 100$ m. Hence, there is a gap between the theoretical and the simulation results. This is because the measurement information estimated by passive UAVs may have errors and the controlled UAVs may not be able to keep the minimum safety distance with the target UAV in real time. From Fig. 5, we can also see that the positioning errors of considered algorithms increase as the distance between each controlled UAV and the target UAV increases. This stems from the fact that the SNR of signals transmitted from the active UAV to each passive UAV via the target UAV decreases as the distance between each controlled UAV and the target UAV increase. Fig. 5 also shows that the proposed ZD-RL method can reduce the

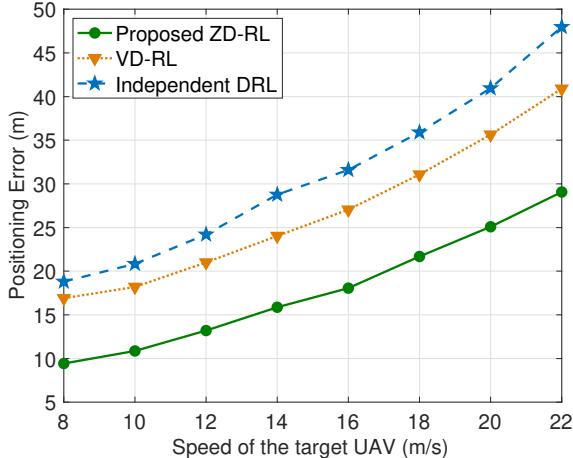


Fig. 4. Value of the positioning error as the speed of the target UAV varies.

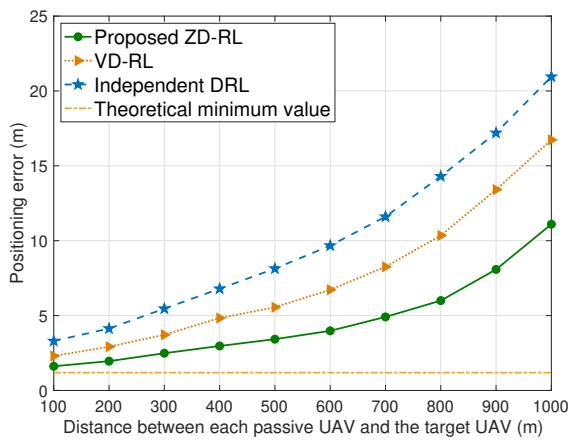


Fig. 5. Value of the positioning error as the distance between each controlled UAV and the target UAV.

positioning error by up to 33.6% and 46.7% compared to the VD-RL and independent DRL methods when $d_{m,t} = 1000$ m. This is because the proposed ZD-RL algorithm enables each controlled UAV to update its DNN parameters based on the approximated probability distribution of individual Z function and adjust its trajectory to minimize the positioning error of the target UAV cooperatively.

Fig. 6 shows how the positioning errors change as the SNR of signals transmitted from the active UAV to each passive UAV varies. From Fig. 6, we can see that as SNR increases, the positioning errors obtained by considered algorithms decrease. This stems from the fact that the variance of measurement errors of each passive UAV increases as SNR decreases. Fig. 6 also shows that the proposed algorithm can reduce positioning errors by up to 24.3% and 37.1% compared to VD-RL method and independent DRL method, respectively, when the SNR is 0 dB. This is because the proposed ZD-RL can approximate the expected value of the sum of future rewards using a non-linear weight function thus improve approximation accuracy.

Fig. 7 shows how the positioning errors obtained by the proposed ZD-RL method change as the number of elements N in the quantile vector varies. From Fig. 7, we can see that

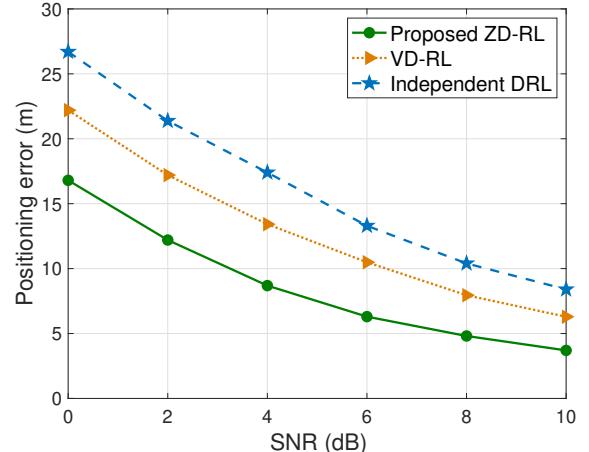


Fig. 6. Value of the positioning error as the SNR of signals transmitted from the target UAV to passive UAVs varies. ($d_{1,t} = d_{2,t} = d_{3,t} = d_{4,t} = 900$ m)

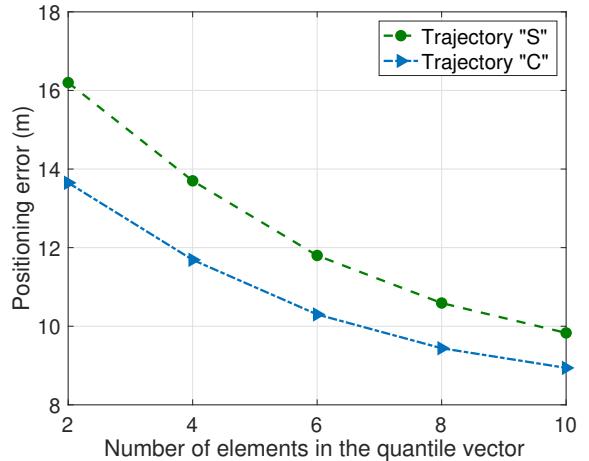


Fig. 7. Value of the positioning error as the number of elements N in the quantile vector varies when the target UAV moves in the curve of "S" and "C".

as the value of N increases, the positioning errors obtained by the proposed ZD-RL method decrease. This stems from the fact that when the number of elements in the quantile vector increases, each agent can obtain more values of the sum of future rewards with different quantiles thus approximating the probability distribution of individual Z functions more accurately. Fig. 7 also shows that the positioning error first drops rapidly when the number of quantiles is small and then decreases more slowly as the number of quantiles increases sufficiently. This is because as the number of quantiles is quite small, the localization performance is mainly limited by the fact that the proposed algorithm cannot accurately approximate the probability distribution of individual Z functions. When N gradually increases, the main limitation shifts from the number of quantiles to the trajectory of the target UAV.

VI. CONCLUSION

In this paper, a novel localization framework that uses several controlled UAVs to localize a target UAV has been proposed. We have modeled this localization problem as an

optimization problem that aims to optimize the positioning accuracy by jointly optimizing the transmit power of the active UAV and trajectories of all controlled UAVs. To solve this problem, we have proposed a ZD-RL method, which uses the probability distribution of the sum of future rewards to estimate the expected values of the sum of future rewards instead of directly estimating the expected values of the sum of future rewards as done in Deep Q. Hence, the proposed method enables each controlled UAV to find its optimal transmit power and trajectory to minimize the positioning errors efficiently. To further reduce the positioning error of the target UAV, we have derived the relationship between the positions of controlled UAVs and the positioning error of the target UAV. Based on the derived expression of the positioning error, we can obtain the minimum positioning error of the target UAV. Simulation results have shown that the proposed method yielded significant improvements in terms of the positioning accuracy compared to baselines.

APPENDIX

A. Proof of Lemma 1

We first explain why the proposed ZD-RL method satisfies condition 1). From (18), the update rule of individual Z function of controlled UAV m can be given by

$$\begin{aligned} Z_{k+1}(\mathbf{o}_{m,t}, \mathbf{a}_{m,t}) &= Z_k(\mathbf{o}_{m,t}, \mathbf{a}_{m,t}) + \alpha_m (R(\mathbf{o}_{m,t}, \mathbf{a}_{m,t}) \\ &\quad + Z(\mathbf{o}_{m,t+1}, \mathbf{a}_{m,t+1}) - Z(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})). \end{aligned} \quad (32)$$

Taking the expectation of individual Z function with respect to transition probability distribution $\mathbb{P}(\mathbf{o}'_m | \mathbf{o}_{m,t}, \mathbf{a}_{m,t})$ and subtracting $M^*(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})$ at both sides, we have

$$\begin{aligned} \mathbb{E}[Z_{k+1}(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})] - M^*(\mathbf{o}_{m,t}, \mathbf{a}_{m,t}) &= \\ (1 - \alpha_m) (\mathbb{E}[Z(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})] - M^*(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})) &+ \alpha_m (R(\mathbf{o}_{m,t}, \mathbf{a}_{m,t}) + \gamma \mathbb{E}[Z(\mathbf{o}_{m,t+1}, \mathbf{a}_{m,t+1})] \\ &\quad - M^*(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})). \end{aligned} \quad (33)$$

Since $e(\mathbf{o}_{m,t}, \mathbf{a}_{m,t}) = M(\mathbf{o}_{m,t}, \mathbf{a}_{m,t}) - M^*(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})$ and $F(\mathbf{o}_{m,t}, \mathbf{a}_{m,t}) = R(\mathbf{o}_{m,t}, \mathbf{a}_{m,t}) + \gamma M(\mathbf{o}_{m,t+1}, \mathbf{a}_{m,t+1}) - M^*(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})$, we have

$$\begin{aligned} e_{k+1}(\mathbf{o}_{m,t}, \mathbf{a}_{m,t}) &= \\ (1 - \alpha_m) e_k(\mathbf{o}_{m,t}, \mathbf{a}_{m,t}) + \alpha_m F(\mathbf{o}_{m,t}, \mathbf{a}_{m,t}). & \end{aligned} \quad (34)$$

Hence, the proposed ZD-RL method satisfies condition 1). Next, we explain why the proposed ZD-RL method satisfies condition 2). To prove condition 2), we first find the expected value of $F(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})$, which is given by

$$\begin{aligned} \mathbb{E}[F(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})] &= \\ \mathbb{E}[R(\mathbf{o}_{m,t}, \mathbf{a}_{m,t}) + \gamma M(\mathbf{o}_{m,t+1}, \mathbf{a}_{m,t+1}) &\quad - M^*(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})] \\ &= \mathbb{E}[R(\mathbf{o}_{m,t}, \mathbf{a}_{m,t}) + \gamma \mathbb{E}[Z(\mathbf{o}_{m,t+1}, \mathbf{a}_{m,t+1})]] \\ &\quad - \mathbb{E}[Z^*(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})] \end{aligned} \quad (35)$$

$$\stackrel{(a)}{=} \mathbb{E}[\mathcal{T}(Z(\mathbf{o}_{m,t}, \mathbf{a}_{m,t}))] - \mathbb{E}[\mathcal{T}(Z^*(\mathbf{o}_{m,t}, \mathbf{a}_{m,t}))]$$

$$\stackrel{(b)}{=} \mathcal{T}(\mathbb{E}[Z(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})]) - \mathcal{T}(\mathbb{E}[Z^*(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})]),$$

where equation (a) and equation (b) follow from the results in [30, Lemma 4]. According to the results in [30, Lemma 3], we have

$$\begin{aligned} &||\mathcal{T}(\mathbb{E}[Z(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})]) - \mathcal{T}(\mathbb{E}[Z^*(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})])||_\infty \\ &\leq \gamma ||\mathbb{E}[Z(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})] - \mathbb{E}[Z^*(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})]||_\infty. \end{aligned} \quad (36)$$

Based on (36), (35) can be written as

$$\begin{aligned} &||\mathbb{E}[F(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})]||_\infty \\ &= ||\mathcal{T}(\mathbb{E}[Z(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})]) - \mathcal{T}(\mathbb{E}[Z^*(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})])||_\infty \\ &\leq \gamma ||\mathbb{E}[Z(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})] - \mathbb{E}[Z^*(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})]||_\infty \\ &= \gamma ||M(\mathbf{o}_{m,t}, \mathbf{a}_{m,t}) - M^*(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})||_\infty \\ &= \gamma ||e(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})||_\infty. \end{aligned} \quad (37)$$

Hence, condition 2) is satisfied. For condition 3), using (35), $\text{Var}(\mathbb{E}[F(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})])$ can be rewritten as

$$\begin{aligned} &\text{Var}(\mathbb{E}[F(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})]) \\ &= \mathbb{E}[(F(\mathbf{o}_{m,t}, \mathbf{a}_{m,t}) - \mathbb{E}[F(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})])^2] \\ &= \mathbb{E}[F(\mathbf{o}_{m,t}, \mathbf{a}_{m,t}) \\ &\quad - (\mathcal{T}(M(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})) - \mathcal{T}(M^*(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})))^2] \\ &= \mathbb{E}[R(\mathbf{o}_{m,t}, \mathbf{a}_{m,t}) + \gamma M(\mathbf{o}_{m,t+1}, \mathbf{a}_{m,t+1}) \\ &\quad - (R(\mathbf{o}_{m,t}, \mathbf{a}_{m,t}) + \gamma \mathbb{E}[M(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})])^2] \\ &= \gamma^2 \mathbb{E}[(M(\mathbf{o}_{m,t+1}, \mathbf{a}_{m,t+1}) - \mathbb{E}[M(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})])^2] \\ &= \gamma^2 \text{Var}(\mathbb{E}[M(\mathbf{o}_{m,t+1}, \mathbf{a}_{m,t+1})]) \\ &\leq \gamma^2 \mathbb{E}[M(\mathbf{o}_{m,t+1}, \mathbf{a}_{m,t+1})^2] \\ &\leq \gamma^2 \max_{\mathbf{o}_{m,t}} \max_{\mathbf{a}_{m,t}} (M(\mathbf{o}_{m,t}, \mathbf{a}_{m,t}))^2 \\ &\leq \gamma^2 ||e(\mathbf{o}_{m,t}, \mathbf{a}_{m,t}) + M^*(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})||_\infty^2 \\ &= \gamma^2 ||e(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})||_\infty^2 + \gamma^2 ||M^*(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})||_\infty^2 \\ &\quad + 2\gamma^2 ||e(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})||_\infty ||M^*(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})||_\infty. \end{aligned} \quad (38)$$

Since the value of $\text{Var}(\mathbb{E}[F(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})])$ depends on $||e(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})||_\infty$, next, we calculate the maximum value of $\text{Var}(\mathbb{E}[F(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})])$ according to the value of $||e(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})|| \leq 1$. In particular, when $||e(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})||_\infty \leq 1$, (38) can be written as

$$\begin{aligned} &\gamma^2 ||e(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})||_\infty^2 + \gamma^2 ||M^*(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})||_\infty^2 \\ &\quad + 2\gamma^2 ||e(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})||_\infty ||M^*(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})||_\infty \\ &\leq \gamma^2 ||e(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})||_\infty^2 + 2\gamma^2 ||M^*(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})||_\infty \\ &\quad + \gamma^2 ||M^*(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})||_\infty^2 \\ &\leq \gamma^2 (||M^*(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})||_\infty^2 + 2||M^*(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})||_\infty) \\ &\quad \times (1 + ||e(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})||_\infty^2). \end{aligned} \quad (39)$$

If $\|e(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})\|_\infty \geq 1$, we have $\|e(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})\|_\infty \leq \|e(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})\|_\infty^2$ and (38) can be rewritten as

$$\begin{aligned} & \gamma^2 \|e(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})\|_\infty^2 + \gamma^2 \|M^*(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})\|_\infty^2 \\ & + 2\gamma^2 \|e(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})\|_\infty \|M^*(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})\|_\infty \\ & \leq \gamma^2 (1 + 2\|M^*(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})\|_\infty) \|e(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})\|_\infty^2 \\ & + \gamma^2 \|M^*(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})\|_\infty^2 \\ & \leq \gamma^2 (1 + 2\|M^*(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})\|_\infty) (1 + \|e(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})\|_\infty^2). \end{aligned} \quad (40)$$

Based on (39) and (40), we have

$$\text{Var}(\mathbb{E}[F(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})]) \leq C_F (1 + \|e(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})\|_\infty^2), \quad (41)$$

where C_F is the maximal value of $2\gamma^2 \|M^*(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})\|_\infty + \gamma^2 \|M^*(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})\|_\infty^2$ and $\gamma^2 (1 + 2\|M^*(\mathbf{o}_{m,t}, \mathbf{a}_{m,t})\|_\infty)$. Hence, condition 3) is satisfied. This completes the proof.

B. Proof of Theorem 2

Since $e_t = \sqrt{\text{tr}(\mathbb{E}[\mathbf{d}\mathbf{s}_t \mathbf{d}\mathbf{s}_t^T])}$, we first calculate the value of $\mathbb{E}[\mathbf{d}\mathbf{s}_t \mathbf{d}\mathbf{s}_t^T]$. From (28), we have

$$\mathbf{d}\mathbf{s}_t = (\mathbf{M}^T \mathbf{M})^{-1} \mathbf{M}^T d\mathbf{r}_t, \quad (42)$$

and the positioning error e_t of the target UAV at time slot t can be rewritten as

$$\begin{aligned} & \mathbb{E}[\mathbf{d}\mathbf{s}_t \mathbf{d}\mathbf{s}_t^T] \\ &= \mathbb{E}\left[(\mathbf{M}^T \mathbf{M})^{-1} \mathbf{M}^T d\mathbf{r}_t \left((\mathbf{M}^T \mathbf{M})^{-1} \mathbf{M}^T d\mathbf{r}_t\right)^T\right] \\ &= \mathbb{E}\left[(\mathbf{M}^T \mathbf{M})^{-1} \mathbf{M}^T d\mathbf{r}_t d\mathbf{r}_t^T \left((\mathbf{M}^T \mathbf{M})^{-1} \mathbf{M}^T\right)^T\right] \\ &= (\mathbf{M}^T \mathbf{M})^{-1} \mathbf{M}^T \mathbb{E}[d\mathbf{r}_t d\mathbf{r}_t^T] \left((\mathbf{M}^T \mathbf{M})^{-1} \mathbf{M}^T\right)^T, \end{aligned} \quad (43)$$

where \mathbf{M}^T is a transpose matrix of \mathbf{M} , $(\mathbf{M}^T \mathbf{M})^{-1}$ is an inverse matrix of $\mathbf{M}^T \mathbf{M}$, $\mathbb{E}[d\mathbf{r}_t d\mathbf{r}_t^T] = \text{diag}(\sigma_{1,t}^2, \sigma_{2,t}^2, \sigma_{3,t}^2, \sigma_{4,t}^2)$ with $\sigma_{m,t}^2 = k(d_{m,t} + d_{0,t})^2$ being the variance of the independent Gaussian measurement error of passive UAV m at time slot t and k being a coefficient [24]. Since $d_{1,t} = d_{2,t} = d_{3,t} = d_{4,t}$, $\mathbb{E}[d\mathbf{r}_t d\mathbf{r}_t^T]$ can be rewritten as

$$\mathbb{E}[d\mathbf{r}_t d\mathbf{r}_t^T] = k(d_{m,t} + d_{0,t})^2 \mathbf{I}, \quad (44)$$

where $\mathbf{I} = \text{diag}(1, 1, 1, 1)$. Substituting (44) into (43), we have

$$\begin{aligned} & \mathbb{E}[\mathbf{d}\mathbf{s}_t \mathbf{d}\mathbf{s}_t^T] \\ &= k(d_{m,t} + d_{0,t})^2 (\mathbf{M}^T \mathbf{M})^{-1} \mathbf{M}^T \left((\mathbf{M}^T \mathbf{M})^{-1} \mathbf{M}^T\right)^T \\ &= k(d_{m,t} + d_{0,t})^2 (\mathbf{M}^T \mathbf{M})^{-1} \mathbf{M}^T \mathbf{M} \left(\mathbf{M}^T \mathbf{M}\right)^{-1} \\ &= k(\mathbf{M}^T \mathbf{M})^{-1}. \end{aligned} \quad (45)$$

Based on (45), the positioning error e_t of the target UAV can be given by

$$\begin{aligned} e_t &= \sqrt{\text{tr}\left(k(d_{m,t} + d_{0,t})^2 (\mathbf{M}^T \mathbf{M})^{-1}\right)} \\ &= \sqrt{k(d_{m,t} + d_{0,t})^2 \text{tr}\left((\mathbf{M}^T \mathbf{M})^{-1}\right)} \\ &\stackrel{(a)}{\geq} \sqrt{4kL_{\min}^2 \text{tr}\left((\mathbf{M}^T \mathbf{M})^{-1}\right)}, \end{aligned} \quad (46)$$

where equation (a) stems from the fact that the distance $d_{m,t}$ between each controlled UAV and the target UAV satisfy $d_{m,t} \geq L_{\min}, m = 0, \dots, 4$, according to constraint (13e). Therefore, equation (a) is hold when $d_{m,t} = d_{0,t} = L_{\min}$. This completes the proof.

C. Proof of Lemma 2

Given the positions \mathbf{s}_t and $\mathbf{u}_{0,t}$, the distance $d_{0,t}$ between the target UAV and the active UAV is a constant and (27) can be rewritten as

$$\begin{aligned} d\mathbf{r}_{m,t} &= \frac{x_t - x_{m,t}}{d_{m,t}} dx_t + \frac{y_t - y_{m,t}}{d_{m,t}} dy_t + \frac{z_t - z_{m,t}}{d_{m,t}} dz_t, \\ m &= 1, 2, 3, 4. \end{aligned} \quad (47)$$

Then, the value of \mathbf{M} in Theorem 2 can be rewritten as

$$\mathbf{M} = \frac{1}{d_{m,t}} \begin{bmatrix} x_t - x_{1,t} & y_t - y_{1,t} & z_t - z_{1,t} \\ x_t - x_{2,t} & y_t - y_{2,t} & z_t - z_{2,t} \\ x_t - x_{3,t} & y_t - y_{3,t} & z_t - z_{3,t} \\ x_t - x_{4,t} & y_t - y_{4,t} & z_t - z_{4,t} \end{bmatrix}. \quad (48)$$

From (46), the positioning error e_t can be written as $e_t = \sqrt{k(d_{m,t} + d_{0,t})^2 \text{tr}\left((\mathbf{M}^T \mathbf{M})^{-1}\right)}$. Since $\text{tr}\left((\mathbf{M}^T \mathbf{M})^{-1}\right) = \sum_{i=1}^3 \frac{1}{\varrho_i}$ with ϱ_i being the eigenvalue of $\mathbf{M}^T \mathbf{M}$ [34], e_t can be rewritten as

$$\begin{aligned} e_t &= \sqrt{k(d_{m,t} + d_{0,t})^2 \sum_{i=1}^3 \frac{1}{\varrho_i}} \\ &\stackrel{(a)}{\geq} \sqrt{k(d_{m,t} + d_{0,t})^2 3 \left(\prod_{i=1}^3 \frac{1}{\varrho_i}\right)^{\frac{1}{3}}} \\ &\stackrel{(b)}{=} \sqrt{k(d_{m,t} + d_{0,t})^2 3 \left(\frac{3}{\text{tr}(\mathbf{M}^T \mathbf{M})}\right)}, \end{aligned} \quad (49)$$

where equation (a) is achieved by the triangle-inequality and equation (a) is hold when $\varrho_1 = \varrho_2 = \varrho_3$, equation (b) stems from the fact that $\varrho_1 + \varrho_2 + \varrho_3 = \text{tr}(\mathbf{M}^T \mathbf{M})$ and

$\varrho_i = \frac{1}{3}\text{tr}(\mathbf{M}^T \mathbf{M})$ when $\varrho_1 = \varrho_2 = \varrho_3$. Based on (48), $\text{tr}(\mathbf{M}^T \mathbf{M})$ is given by

$$\begin{aligned} & \text{tr}(\mathbf{M}^T \mathbf{M}) \\ &= \frac{1}{d_{m,t}^2} \left(\sum_{m=1}^4 (x_t - x_{m,t})^2 + \sum_{m=1}^4 (y_t - y_{m,t})^2 \right. \\ &\quad \left. + \sum_{m=1}^4 (z_t - z_{m,t})^2 \right) \\ &= \frac{1}{d_{m,t}^2} \sum_{m=1}^4 ((x_t - x_{m,t})^2 + (y_t - y_{m,t})^2 + (z_t - z_{m,t})^2) \\ &= \frac{1}{d_{m,t}^2} \left(\sum_{m=1}^4 d_{m,t}^2 \right) \\ &\stackrel{(a)}{=} 4, \end{aligned} \quad (50)$$

where equation (a) stems from the fact that $d_{1,t} = d_{2,t} = d_{3,t} = d_{4,t}$. Substituting (50) into (49), we have

$$\begin{aligned} e_t &= \sqrt{k(d_{m,t} + d_{0,t})^2 3 \left(\frac{3}{4} \right)} \\ &= \sqrt{\frac{9}{4} k (d_{m,t} + d_{0,t})^2} \\ &\stackrel{(a)}{\geq} \frac{3}{2} (L_{\min} + d_{0,t}) \sqrt{k}, \end{aligned} \quad (51)$$

where equation (a) stems from the fact that $d_{m,t} \geq L_{\min}$ as shown in (13e). This completes the proof.

REFERENCES

- [1] Y. Zhu, M. Chen, S. Wang, Y. Liu, and C. Yin, "Trajectory design for 3D UAV localization in UAV based networks," submitted to *IEEE International Global Communications Conference (GLOBECOM)*, Kuala Lumpur, Malaysia, Dec. 2023.
- [2] I. Guvenc, F. Koohifar, S. Singh, M. L. Sichitiu, and D. Matolak, "Detection, tracking, and interdiction for amateur drones," *IEEE Communications Magazine*, vol. 56, no. 4, pp. 75–81, Apr. 2018.
- [3] 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 Communications Surveys & Tutorials*, vol. 21, no. 3, pp. 2334–2360, Thirdquarter. 2019.
- [4] Z. Yang, C. Pan, M. Shikh-Bahaei, W. Xu, M. Chen, M. Elkashlan, and A. Nallanathan, "Joint altitude, beamwidth, location, and bandwidth optimization for UAV-enabled communications," *IEEE Communications Letters*, vol. 22, no. 8, pp. 1716–1719, June 2018.
- [5] Q. Wu, Y. Zeng, and R. Zhang, "Joint trajectory and communication design for multi-UAV enabled wireless networks," *IEEE Transactions on Wireless Communications*, vol. 17, no. 3, pp. 2109–2121, May 2018.
- [6] P. Sinha and I. Guvenc, "Impact of antenna pattern on TOA based 3D UAV localization using a terrestrial sensor network," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 7, pp. 7703–7718, Apr. 2022.
- [7] Z. Yang, C. Pan, K. Wang, and M. Shikh-Bahaei, "Energy efficient resource allocation in UAV-enabled mobile edge computing networks," *IEEE Transactions on Wireless Communications*, vol. 18, no. 9, pp. 4576–4589, Sept. 2019.
- [8] H. Sier, X. Yu, I. Catalano, J. P. Queralta, Z. Zou, and T. Westerlund, "UAV tracking with Lidar as a camera sensors in GNSS-denied environments," <https://arxiv.org/abs/2303.00277>, Mar. 2023.
- [9] Z. Xu, X. Zhan, Y. Xiu, C. Suzuki, and K. Shimada, "Onboard dynamic-object detection and tracking for autonomous robot navigation with RGB-D camera," <https://arxiv.org/abs/2303.00132>, Feb. 2023.
- [10] U. Bhattacherjee, E. Ozturk, O. Ozdemir, I. Guvenc, M. L. Sichitiu, and H. Dai, "Experimental study of outdoor UAV localization and tracking using passive RF sensing," <https://arxiv.org/abs/2108.07857>, Sept. 2022.
- [11] F. Wen, J. Shi, G. Gui, H. Gacanin, and O. A. Dobre, "3-D positioning method for anonymous UAV based on bistatic polarized MIMO radar," *IEEE Internet of Things Journal*, vol. 10, no. 1, pp. 815–827, Sept. 2023.
- [12] M. Sadeghi, F. Behnia, and R. Amiri, "Optimal geometry analysis for TDOA-based localization under communication constraints," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 57, no. 5, pp. 3096–3106, Oct. 2021.
- [13] A. Gendia, O. Muta, S. Hashima, and K. Hatano, "UAV positioning with joint NOMA power allocation and receiver node activation," in *Proc. IEEE Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, pp. 240–245, Kyoto, Japan, Dec. 2022.
- [14] V. Saj, B. Lee, D. Kalathil, and M. Benedict, "Robust reinforcement learning algorithm for vision-based ship landing of UAVs," <https://arxiv.org/abs/2209.08381>, Sept. 2022.
- [15] V. Tilwari and S. Pack, "Autonomous 3D UAV localization using taylor series linearized TDOA-based approach with machine learning algorithms," in *Proc. International Conference on Information and Communication Technology Convergence (ICTC)*, pp. 783–785, Jeju Island, Korea, Nov. 2022.
- [16] B. Joshi, D. Kapur, and H. Kandath, "Sim-to-real deep reinforcement learning based obstacle avoidance for UAVs under measurement uncertainty," <https://arxiv.org/abs/2303.07243>, Mar. 2023.
- [17] C. Wang, J. Wang, Y. Shen, and X. Zhang, "Autonomous navigation of UAVs in large-scale complex environments: A deep reinforcement learning approach," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 3, pp. 2124–2136, Mar. 2019.
- [18] Y. Hu, M. Chen, W. Saad, H. V. Poor, and S. Cui, "Distributed multi-agent meta learning for trajectory design in wireless drone networks," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 10, pp. 3177–3192, Oct. 2021.
- [19] M. Chen, Y. Wang, and H. V. Poor, "Performance optimization for wireless semantic communications over energy harvesting networks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8647–8651, Singapore, May 2022.
- [20] A. Albanese, P. Mursia, V. Sciancalepore, and X. Costa-Perez, "PAPIR: Practical RIS-aided localization via statistical user information," in *Proc. International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, pp. 531–535, Lucca, Italy, Nov. 2021.
- [21] Y. Zeng and R. Zhang, "Energy-efficient UAV communication with trajectory optimization," *IEEE Transactions on Wireless Communications*, vol. 16, no. 6, pp. 3747–3760, Mar. 2017.
- [22] X. Tong, Z. Zhang, Y. Zhang, Z. Yang, C. Huang, K.-K. Wong, and M. Debbah, "Environment sensing considering the occlusion effect: A multi-view approach," *IEEE Transactions on Signal Processing*, vol. 70, pp. 3598–3615, June 2022.
- [23] Y. Chan and K. Ho, "A simple and efficient estimator for hyperbolic location," *IEEE Transactions on Signal Processing*, vol. 42, no. 8, pp. 1905–1915, Aug. 1994.
- [24] A. Quazi, "An overview on the time delay estimate in active and passive systems for target localization," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 3, pp. 527–533, June 1981.
- [25] Y. Su, H. Zhou, Y. Deng, and M. Dohler, "Energy-efficient cellular-connected UAV swarm control optimization," <https://arxiv.org/abs/2303.10398>, Mar. 2023.
- [26] W. Dabney, G. Ostrovski, D. Silver, and R. Munos, "Implicit quantile networks for distributional reinforcement learning," in *Proc. International Conference on Machine Learning (ICML)*, pp. 2640–3498, Stockholm, Sweden, Jun. 2018.
- [27] W.-F. Sun, C.-K. Lee, and C.-Y. Lee, "DFAC framework: Factorizing the value function via quantile mixture for multi-agent distributional Q-learning," in *Proc. International Conference on Machine Learning (ICML)*, pp. 9945–9954, Vienna, Austria, Dec. 2021.
- [28] W. Dabney, M. Rowland, M. Bellemare, and R. Munos, "Distributional reinforcement learning with quantile regression," in *Proc. the AAAI Conference on Artificial Intelligence*, 32(1), pp. 2892–2901, New Orleans, USA, Oct. 2018.
- [29] J. Zhao, Y. Zhu, X. Mu, K. Cai, Y. Liu, and L. Hanzo, "Simultaneously transmitting and reflecting reconfigurable intelligent surface (STAR-RIS) assisted UAV communications," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 10, pp. 3041–3056, Oct. 2022.

- [30] M. G. Bellemare, W. Dabney, and R. Munos, "A distributional perspective on reinforcement learning," <https://arxiv.org/abs/1707.06887>, Jul. 2017.
- [31] T. Jaakkola, M. I. Jordan, and S. P. Singh, "On the convergence of stochastic iterative dynamic programming algorithms," *Neural Computation*, vol. 6, no. 6, pp. 1185–1201, Nov. 1994.
- [32] S. Wang, M. Chen, Z. Yang, C. Yin, W. Saad, S. Cui, and H. V. Poor, "Distributed reinforcement learning for age of information minimization in real-time IoT systems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 3, pp. 501–515, Jan. 2022.
- [33] H. Godrich, A. M. Haimovich, and R. S. Blum, "Target localization accuracy gain in MIMO radar-based systems," *IEEE Transactions on Information Theory*, vol. 56, no. 6, pp. 2783–2803, May 2010.
- [34] M. Zhang and J. Zhang, "A fast satellite selection algorithm: Beyond four satellites," *IEEE Journal of Selected Topics in Signal Processing*, vol. 3, no. 5, pp. 740–747, Oct. 2009.



Yujiao Zhu received the B.S. degree in Communication Engineering from Harbin Institute of Technology, Shandong, China. She is currently pursuing the Ph.D. degree with the Information and Communication Engineering Department, Beijing University of Posts and Telecommunications, Beijing, China. Her research interests include unmanned aerial vehicles, reinforcement learning, joint sensing and communications, machine learning in wireless networks.



Mingzhe Chen is currently an Assistant Professor with the Department of Electrical and Computer Engineering and Institute of Data Science and Computing at University of Miami. His research interests include federated learning, reinforcement learning, virtual reality, unmanned aerial vehicles, and Internet of Things. He has received four IEEE Communication Society journal paper awards including the IEEE Marconi Prize Paper Award in Wireless Communications in 2023, the Young Author Best Paper Award in 2021 and 2023, and the Fred W. Ellersick Prize

Award in 2022, and four conference best paper awards at ICCCN in 2023, IEEE WCNC in 2021, IEEE ICC in 2020, and IEEE GLOBECOM in 2020. He currently serves as an Associate Editor of IEEE Transactions on Mobile Computing, IEEE Wireless Communications Letters, IEEE Transactions on Green Communications and Networking, and IEEE Transactions on Machine Learning in Communications and Networking.



Sihua Wang received the Ph.D. degree from the Beijing University of Posts and Telecommunications, Beijing, China, in 2021. He is currently a Hong Kong Scholar Fellow with the Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong. He is also a Post-Doctoral Researcher at School of Computer Science (National Pilot Software Engineering School), Beijing University of Posts and Telecommunications, Beijing, P. R. China. His research interests include mobile edge computing, resource allocation, and machine learning in wireless networks.



Ye Hu is an Assistant Professor in the Industrial and System Engineering Department at University of Miami. She received her Ph.D. degree from Virginia Tech, VA, USA, in 2021. After graduation, she has served as a Postdoctoral Research Scientist at the Columbia University, and the North Carolina State University. Her research interests span from unmanned aerial vehicle networks, low earth orbit satellite, cyber physical human system, network security to distributed machine learning. She is also the recipient of the best paper award at IEEE GLOBECOM 2020 for her work on meta-learning for drone-based communications.



Yuchen Liu is currently an Assistant Professor with the Department of Computer Science at North Carolina State University, USA. He got his Ph.D. degree at the Georgia Institute of Technology, USA. His research interests include wireless networking, generative AI, reinforcement learning, mobile computing, and software simulation. He has received several best paper awards at IEEE and ACM conferences. He currently serves as an Associate Editor of IEEE Transactions on Green Communications and Networking and International Journal of Sensors, Wireless Communications and Control.



Changchuan Yin received the Ph. D. degree in signal and information processing from Beijing University of Posts and Telecommunications, Beijing, P. R. China, in 1998. In 2004, he was a visiting scholar in the Faculty of Science, the University of Sydney, Sydney, NSW, Australia. From 2007 to 2008, he held a visiting position with the Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX, USA. He is currently a professor with the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications. His research interests include wireless networks and statistical signal processing. He was the co-recipient of the IEEE Guglielmo Marconi Prize Paper Award in 2023 and the IEEE International Conference on Wireless Communications and Signal Processing Best Paper Award in 2009. He has served as the symposium co-chair and TPC member for many IEEE conferences.