3D UAV Localization Optimization under Jamming Attacks: A Mixture Gaussian Distribution based Collaborative Reinforcement Learning

Yujiao Zhu, *Student Member, IEEE*, Mingzhe Chen, *Member, IEEE*, Sihua Wang, *Member, IEEE*, Yuchen Liu, *Member, IEEE*, Changchuan Yin, *Senior Member, IEEE*, and Tony Q. S. Quek, *Fellow, IEEE*

***Abstract*—In this paper, the optimization of unmanned aerial vehicle (UAV) localization under jamming attacks is studied. In the considered network, a base station (BS) collaborates with an active UAV to localize a target UAV. During this positioning process, a jamming UAV transmits discontinuous signals to passive UAVs to interfere the distance information measurement. To localize the target UAV under jamming attacks, the BS jointly use two localization methods: 1) generative adversarial network (GAN) based positioning method and 2) time difference of arrival (TDOA) based positioning method. Since GAN-based method cannot defense against a strong jamming signal while TDOA-based method may consume more energy and sacrifice localization accuracy, the BS must select an appropriate positioning method (GAN-based or TDOA-based methods) and four distance measurement information of passive UAVs to localize the target UAV. This problem is formulated as an optimization problem whose goal is to minimize the positioning error between the estimated and the ground truth positions of the target UAV while considering jamming attacks and the trajectory of passive UAVs. To solve this problem, we propose a mixture Gaussian distribution model based collaborative reinforcement learning (RL) method which enables the active UAV to optimize its transmit power and trajectory, and enables the BS to select the most appropriate subsets of distance measurement information and the optimal positioning method according to the UAVs movement and the unknown jamming attack pattern. Simulation results show the proposed method can reduce the positioning error of the target UAV by up to 36.5% compared to the method that does not consider the GAN-based positioning method.**

***Index Terms*—UAV localization, jamming attacks, GAN-based positioning method, TDOA-based positioning method.**

I. Introduction

Localization of unmanned aerial vehicles (UAVs) has been widely used in military and civilian applications [1]–[3]. Radio frequency (RF) based passive localization methods can accurately localize unknown target UAVs in scenarios where the global navigation satellite systems (GNSSs) are not available [4]–[6]. However, using passive radio frequency localization methods to localize target UAVs faces many challenges [7]. First, the high-speed mobility of UAVs makes it difficult to estimate the real-time three-dimensional (3D) coordinates of UAVs [8]. Second, the interference of the dynamic wireless environments and attacks of jamming objects affect the transmission signals used for UAV localization.

A. Related Works

Existing works [9]–[16] have studied several problems of using radio frequency for localizing the target UAV. Specificallly, the authors in [9] investigated to use WiGig devices for estimating UAV positions according to the beam fingerprinting information. In [10] and [11], the authors used the time of arrival (TOA) information obtained by ground sensors to estimate the position of the target UAV. [12] focused on angle measurements of signals to determine UAV positions. In [13], the authors obtained the distance information based on the signal strength for calculating UAV positions. Additionally, periodic communication signals transmitted by UAVs were investigated in [14] for UAV positions estimations and UAV tracking. However, the works in [10]–[14] ignored the impacts of the positions of sensors on localization accuracy. The authors in [15] studied the relationship between the deployment of sensors and the UAV localization accuracy. The authors in [16] optimized the selection of sensors for real-time UAV localization. However, most of these works [9]–[16] localized UAVs by using static stations, which may not be applied for UAVs with high-speed movement. In addition, the above works [9]–[16] did not consider how dynamic jamming attacks affect the UAV localization performance.

A number of existing works such as [17]–[21] have studied the problem of UAV localization while avoiding jamming attacks, including machine learning (ML) [22] based positioning methods and radio frequency based positioning methods. The authors in [17] proposed a novel deep neural network (DNN) model to generate an image of received signals amplitude and phase to improve the positioning accuracy of the UAV by using noise and interference in the environment. The authors in [18] used a convolutional neural network (CNN) to analyze the

Y. Zhu, S. Wang, and C. Yin are with Beijing Laboratory of Advanced Information Network, and the Beijing Key Laboratory of Network System Architecture and Convergence, Beijing University of Posts and Telecommunications, Beijing, 100876, China (E-mail: 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. Liu is with the Department of Computer Science, North Carolina State University, Raleigh, NC, 27695, USA (Email: yuchen.liu@ncsu.edu).

Y. Zhu and T. Q. S. Quek are with the Department of Information Systems Technology and Design, Singapore University of Technology and Design, Singapore, 487372 (Email: tonyquek@sutd.edu.sg).

received RF signals to prevent interference and estimate the angle of arrival, thus decreasing the positioning error of the UAV. In [19], the authors used a DNN to recognize the visual information of UAVs under a scenario with high interference and improve the positioning accuracy of the UAV. However, these ML based jamming attack defense methods in [17]–[19] cannot defense strong jamming attacks. In [20], the authors proposed a UAV grouping scheme to reduce the influence of interference on UAV localization. The authors in [21] analyzed the relationship between the UAV localization performance and the number of participating BSs under different signal-to-interference-plus-noise radio (SINR) conditions. However, these RF based positioning methods in [20], [21] require to frequently adjust the position of UAVs or the number of BSs, thus increasing energy consumption and affecting localization accuracy. Moreover, these works [17]–[21] did not consider to jointly use the ML based positioning method and traditional RF based positioning method to defense jamming attacks and improve the localization accuracy.

B. Contributions

The primary contribution of this research lies in introducing a passive 3D UAV localization framework that jointly uses the base station (BS) and an active UAV to localize a target UAV under jamming attacks. Our key contributions are summarized as follows:

- We consider a 3D passive UAV localization network. In the considered network, an active UAV transmits signals towards the target UAV. Passive UAVs will receive the signals reflected by the target UAV and calculate the sum of distance between the active UAV and the target UAV, and the distance between the target UAV and the passive UAV. These distance information will be transmitted to the BS for localizing the target UAV.
- During the localization process, the jamming UAV transmits interference signals to passive UAVs to interfere with the accuracy of measured distance information. To improve the localization accuracy under jamming attacks, the BS jointly use the generative adversarial network (GAN)-based positioning method and time difference of arrival (TDOA)-based positioning method. Since GAN-based positioning method cannot accurately localize the target UAV when the jamming signal is strong and TDOA-based positioning method may require more energy to adjust UAV positions and sacrifice localization accuracy, the BS must select the optimal positioning method and four distance information to localize the target UAV. We formulate the minimization of positioning error for the target UAV as an optimization problem, which involves jointly optimizing the transmit power and trajectory of the active UAV, as well as optimizing the selection positioning methods and the subset of distance measurement information at the BS.
- To solve this problem, we propose a mixture Gaussian distribution model based collaborative reinforcement learning (RL) method which enables the BS to select the optimal positioning method and the optimal subset of distance information, and the active UAV to optimize its trajectory and transmit power under unknown jamming attack patterns. Compared to tradition RL methods that estimate the expected values of value functions directly, the proposed method approximates the probability distribution of value functions by mixed Gaussian distributions thus estimating the expected values of value functions with low complexity and high accuracy.
- To further reduce the effect of jamming attacks on the localization performance, we analyze how the jamming attacks from the jamming UAV affect the positioning error of the target UAV. Meanwhile, we derive the expression of the positioning error of the target UAV in a scenario where the jamming UAV transmits signals in a fixed pattern.

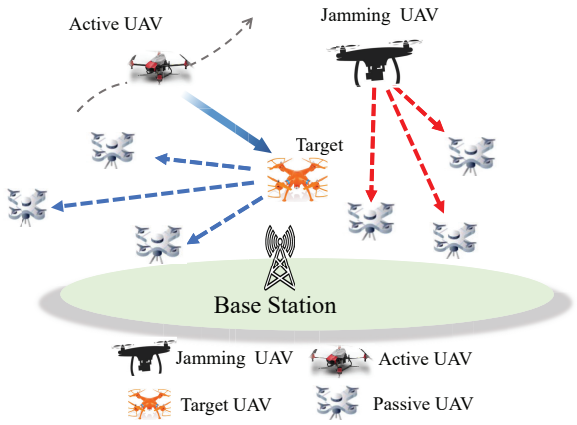


Fig. 1. The studied UAV localization network.

The subsequent sections of this paper are structured as follows. Section II presents the system model and outlines problem formulation. The mixture Gaussian distribution based collaborative RL method proposed in this paper is introduced in Section III. In Sections IV and V, the target UAV localization performance and simulation results are analyzed, respectively. Conclusions are drawn in Section VI.

II. System Model and Problem Formulation

Consider a wireless network in which one BS, one active UAV and a set $\mathcal{U}$ of U passive UAVs to cooperatively localize a target UAV under attacks from a jamming UAV. In the considered network, signals are transmitted from the active UAV to the target UAV. These signals, upon reflection by the target UAV, are received by passive UAVs. Utilizing the time of signal transmission, distances are calculated by passive UAVs and transmitted from passive UAVs to the BS. Subsequently, the BS calculates the position of the target UAV based on these information. To reduce the influence of the jamming UAV attacks and the mobility of UAVs, the active UAV requires to optimize its trajectory and transmit power. In addition, the

TABLE I. List of Notations

Notation	Description	Notation	Description
U	Number of passive UAVs	$\boldsymbol{l}_{0,t}$	Position of the active UAV
$\boldsymbol{l}_{u,t}$	Position of passive UAV u	v	Flight speed of the active UAV
Δ_t	Duration of a time slot	α_t	Yaw angle of the active UAV
β_t	Pitch angle of the active UAV	$s_{0,t}^{\mathrm{G}}$	SNR from the active UAV to the BS
p_t^{F}	Aerodynamic power consumption of the active UAV	E_t^{F}	Aerodynamic energy consumption of the active UAV
$\boldsymbol{l}_t^{\mathrm{J}}$	Position of the jamming UAV	P^{J}	Transmit power of the jamming UAV
$\boldsymbol{l}_t$	Position of the target UAV	$\hat{\boldsymbol{l}}_t$	Estimated position of the target UAV
p_t^{T}	Transmit power of the active UAV	f_{J}	Probability of the jamming attacks
$D_{u,t}^{\mathrm{G}}$	Data size of passive UAV u	$p_{u,t}$	Transmit power of passive UAV u
$x_{u,t}$	Reflection coefficient from the active UAV to passive u	β_0	Path loss of LoS links at a reference distance
$s_{u,t}^{\mathrm{A}}$	SINR of passive UAV u	$s_{u,t}^{\mathrm{G}}$	SNR of the BS from passive UAV u
ϵ^2	Power of Gaussian noise	$I_{u,t}^{\mathrm{J}}$	Jamming signals received by passive UAV u
$\boldsymbol{l}_{\mathrm{B}}$	Position of the BS	$\chi_{u,t}$	Elevation angle of passive UAV u
L_{FS}	Path loss in a free space	$\Pr\left(l_{u,t}^{\mathrm{LoS}}\right)$	Probability of LoS
W_1	Bandwidth of UAV links	ϵ^2	Variance of Gaussian noise
W_2	Bandwidth of UAV-BS links	g_t	Jamming attack defense selection
$\hat{\boldsymbol{d}}_t$	Measured distance information	g	Gravitational acceleration
e_t	Positioning error of the target UAV	$q_{m,t}$	Indicator for selecting distance m for localizing the target UAV

BS requires to select the most appropriate subset of distance information and the optimal position estimation method to accurately localize the target UAV in real time. Next, we first introduce the mobility patterns of UAVs and the jamming pattern of the jamming UAV. Then, the signal transmission model and the positioning model which is used to estimate the position of the target UAV are introduced. Finally, we formulate the optimization problem.

A. UAV Aerodynamic Model

The mobility patterns and propulsion energy consumption of each UAV depend on its position, pitch angle, yaw angle, and flying speed. Next, we present the UAV movement model and UAV flight energy consumption model, respectively.

1) UAV movement model: We define the 3D coordinate of the active and passive UAVs at time slot t as $\boldsymbol{l}_t = \{\boldsymbol{l}_{0,t}, \cdots, \boldsymbol{l}_{U,t}\}$, where $\boldsymbol{l}_{0,t} = [x_{0,t}, y_{0,t}, z_{0,t}]^T$ is the coordinate of the active UAV and $\boldsymbol{l}_{u,t} = [x_{u,t}, y_{u,t}, z_{u,t}]^T, u \in \{1, \cdots, U\}$ is the position of passive UAV u. Hereinafter, an index 0 indicates the active UAV and an index $u \in \{1, \cdots, U\}$ represents passive UAV u. Given the active UAV yaw angle α_t, speed v, and pitch angle β_t, its coordinates $\boldsymbol{l}_{0,t+1}$ at time slot $t+1$ is [23]

$$\boldsymbol{l}_{0,t+1}(\alpha_t, \beta_t) = \boldsymbol{l}_{0,t} + v\Delta_t \begin{bmatrix} \cos\alpha_t \cos\beta_t \\ \sin\alpha_t \cos\beta_t \\ \sin\beta_t \end{bmatrix}, \quad (1)$$

where Δ_t is the time duration of each time slot.

2) UAV propulsion energy consumption model: The aerodynamic power consumption $p_t^{\mathrm{F}}(\alpha_t, \beta_t)$ of the active UAV at time slot t is [24]

$$p_t^{\mathrm{F}}(\alpha_t, \beta_t) = \frac{C_1}{\sqrt{\left(v_t^{\mathrm{L}}\right)^2 + \sqrt{\left(v_t^{\mathrm{L}}\right)^4 + 4\left(v_t^{\mathrm{H}}\right)^4}}} + Mgv_t^{\mathrm{Z}} + C_2\left(v_t^{\mathrm{L}}\right)^3, \quad (2)$$

where C_1 and C_2 are coefficients [24], $v_t^{\mathrm{L}} = v\cos\beta_t$ is the horizontal flight speed, $v_t^{\mathrm{Z}} = v\sin\beta_t$ is the vertical flight speed, g is the acceleration of gravity, M is the weight of each UAV, and v_t^{H} is the power required to hover [24]. Then, the required instantaneous propulsion energy at time slot t is

$$E_t^{\mathrm{F}}(\alpha_t, \beta_t) = p_t^{\mathrm{F}}(\alpha_t, \beta_t)\Delta_t. \quad (3)$$

B. Jamming Model

To interfere with the localization of the target UAV, the jamming UAV transmits discontinuous interference signals to active and passive UAVs. We use an indicator j_t to represent whether the jamming UAV transmits signals at time slot t. $j_t = 1$ implies that the jamming UAV transmits signals and $j_t = 0$, otherwise. Let f_{J} be the probability that the jamming UAV transmits jamming signals. The jamming power received by passive UAV u is given by

$$I_{u,t}^{\mathrm{J}} = j_t P^{\mathrm{J}} |h_{\mathrm{J},u,t}|^2, \quad (4)$$

where P^{J} is the transmit power of the jamming UAV, $|h_{\mathrm{J},u,t}|^2 = \beta_0 \left\|\boldsymbol{l}_t^{\mathrm{J}} - \boldsymbol{l}_{u,t}\right\|^{-2}$ is the path loss from the jamming UAV to passive UAV u with $\boldsymbol{l}_t^{\mathrm{J}}$ being the position of the jamming UAV at time slot t and β_0 being the path loss at a reference distance.

C. Transmission Model

In the studied model, the transmission links consist of a) links between UAVs that are used to transmit signals to calculate the distance information and b) links from passive UAVs to the BS that are used to transmit distance information measured by passive UAVs.

1) UAV Links: We assume that a link between any two UAVs is line-of-sight (LoS). Due to the interference introduced by channel noise and the jamming UAV, the signal-to-interference-plus-noise ratio (SINR) of the signals received by passive UAV u at time slot t is [25]

$$s_{u,t}^{\mathrm{A}}\left(\boldsymbol{l}_{0,t}, p_t^{\mathrm{T}}\right) = \frac{p_t^{\mathrm{T}} |h_{u,t} x_{u,t} h_t|^2}{\epsilon^2 + I_{u,t}^{\mathrm{J}}}, \quad (5)$$

where p_t^{T} is the transmit power of the active UAV, ϵ^2 is the power of the Gaussian noise, $I_{u,t}^{\mathrm{J}}$ is the interference caused by the jamming UAV, and $x_{u,t}$ is the reflection coefficient of the target UAV at time slot t [26]. Furthermore, $h_t^2 = \beta_0 \left\| \boldsymbol{l}_{0,t} - \boldsymbol{l}_t \right\|^{-2}$ and $h_{u,t}^2 = \beta_0 \left\| \boldsymbol{l}_t - \boldsymbol{l}_{u,t} \right\|^{-2}$ represent the path loss from the active UAV to the target UAV and the path loss from the target UAV to passive UAV u with $\boldsymbol{l}_t$ being the position of the target UAV [27], respectively.

2) UAV-BS links: Given the position $\boldsymbol{l}_{u,t}$ of passive UAV u, $\boldsymbol{l}_{0,t}$ of the active UAV and the position $\boldsymbol{l}_{\mathrm{B}}$ of the BS, the probabilistic LoS and non-line-of-sight (NLoS) channel model is used to model the UAV-BS link [28]. To be specific, the LoS path loss $g_{u,t}^{\mathrm{LoS}}$ and NLoS path loss $g_{u,t}^{\mathrm{NLoS}}$ from the passive UAV u to the BS at time slot t can be given by [29]

$$g_{u,t}^{\mathrm{LoS}} = L_{\mathrm{FS}}\left(l_0\right) + 10\mu_{\mathrm{LoS}} \log\left(\left\|\boldsymbol{l}_{u,t} - \boldsymbol{l}_{\mathrm{B}}\right\|\right) + \varphi_{\sigma_{\mathrm{LoS}}}, \quad (6)$$

$$g_{u,t}^{\mathrm{NLoS}} = L_{\mathrm{FS}}\left(l_0\right) + 10\mu_{\mathrm{NLoS}} \log\left(\left\|\boldsymbol{l}_{u,t} - \boldsymbol{l}_{\mathrm{B}}\right\|\right) + \varphi_{\sigma_{\mathrm{NLoS}}}, \quad (7)$$

where l_0 is the free-space reference distance, $L_{\mathrm{FS}}\left(l_0\right) = 20\log\left(l_0 f_0 4\pi/c\right)$ is the path loss in a free space with f_0 being the carrier frequency. $\varphi_{\sigma_{\mathrm{LoS}}}$ and $\varphi_{\sigma_{\mathrm{NLoS}}}$ are the shadowing random variables with zero mean and σ_{LoS}^2, σ_{NLoS}^2 dB variances. Given (6) and (7), the path loss from passive UAV u to the BS at time slot t is given by [30]

$$\bar{g}_{u,t} = \Pr\left(g_{u,t}^{\mathrm{LoS}}\right) \times g_{u,t}^{\mathrm{LoS}} + \left(1 - \Pr\left(g_{u,t}^{\mathrm{LoS}}\right)\right) \times g_{u,t}^{\mathrm{NLoS}}, \quad (8)$$

where $\Pr\left(g_{u,t}^{\mathrm{LoS}}\right) = \left(1 + \zeta \exp\left(-\eta\left[\chi_{u,t} - \zeta\right]\right)\right)^{-1}$ is the probability of LoS with ζ and η being constants which depend on the environment factors, and $\chi_{u,t} = \arcsin\frac{z_{u,t}}{\left\|\boldsymbol{l}_{u,t} - \boldsymbol{l}_{\mathrm{B}}\right\|}$ being the elevation angle of passive UAV u.

Given the transmit power $p_{u,t}$ of passive UAV u, the signal-to-noise ratio (SNR) of the BS from passive UAV u to the BS at time slot t is given by

$$s_{u,t}^{\mathrm{G}} = \frac{p_{u,t}}{\epsilon^2} 10^{-\bar{g}_{u,t}/10}. \quad (9)$$

The delay required for transmitting distance information from passive UAV u to the BS at time slot t is [31]

$$T_{u,t}^{\mathrm{G}} = \frac{D_{u,t}^{\mathrm{G}}}{W_2 \log_2\left(1 + s_{u,t}^{\mathrm{G}}\right)}, \quad (10)$$

where W_2 is the bandwidth of each passive UAV to transmit distance to the BS and $D_{u,t}^{\mathrm{G}}$ is the data size of the distance information.

D. Localization Model

After receiving signals transmitted from the active UAV, passive UAVs estimate the distance measurement information $\hat{\boldsymbol{d}}_t = \left\{\hat{d}_{1,t}, \cdots, \hat{d}_{U,t}\right\}$. Then, the measured distances are transmitted to the BS. Due to the mobility of UAVs and the jamming attacks, the BS requires to select a subset of distance information for localizing the target UAV. Let $\boldsymbol{q}_t = \{q_{1,t}, \cdots, q_{U,t}\}$ be the select indicator vector, where $q_{u,t} \in \{0,1\}$ with $q_{u,t} = 1$ indicating that distance information measured by passive UAV u is selected to estimate the position of the target UAV at time slot t, otherwise, we have $q_{u,t} = 0$. Based on distance information $\hat{\boldsymbol{d}}_t$, positions of passive UAVs $\boldsymbol{l}_t^{\mathrm{P}} = \{\boldsymbol{l}_{1,t}, \cdots, \boldsymbol{l}_{U,t}\}$, and the UAV selection scheme $\boldsymbol{q}_t$, the BS can determine the position $\hat{\boldsymbol{l}}_t$ of the target UAV by using a) GAN-based position estimation method and b) a standard time difference of arrival (TDOA) method. Here, we consider GAN-based positioning method since GANs can eliminate the distance measurement error caused by interference signals via training with noisy samples and estimate the position of the target UAV accurately even in scenarios with input errors. We also consider the TDOA-based position estimation method since TDOA-based method can use the mobility of the active and passive UAVs to avoid jamming attacks. The BS will determine the method (i.e., GAN or TDOA) used for UAV positioning according to distance measurement information and the positions of the active and passive UAVs.

E. Jamming Attack Defense Methods

Next, we introduce the GAN-based position estimation method and traditional TDOA-based position estimation method as follows:

- GAN-based position estimation method: The input of the GAN is the received distance measurement information and positions of passive UAVs i.e., $\left\{\hat{\boldsymbol{d}}_{1,t}, \cdots, \hat{\boldsymbol{d}}_{4,t}, \boldsymbol{l}_{1,t}, \cdots, \boldsymbol{l}_{4,t}\right\}$. The output of the GAN is the estimated position $\hat{\boldsymbol{l}}_t^{\mathrm{G}}$ of the target UAV. However, GAN-based positioning method cannot defense a strong jamming signal. To this end, we introduce a TDOA-based positioning method to avoid strong jamming signals.
- TDOA-based position estimation method: Here, we defense the jamming attacks by adjusting the trajectory of the active UAV. The estimated position of the target UAV $\hat{\boldsymbol{l}}_t^{\mathrm{W}}\left(\boldsymbol{l}_{0,t}, p_t^{\mathrm{T}}, \boldsymbol{q}_t\right)$ is obtained by the TDOA-based method based on the distance information $\hat{\boldsymbol{d}}_t$ and positions of passive UAVs $\boldsymbol{l}_t^{\mathrm{P}} = \{\boldsymbol{l}_{1,t}, \cdots, \boldsymbol{l}_{U,t}\}$ [32]. However, this method requires to adjust the trajectory of the active UAV frequently, which will increase energy consumption of the active UAV and affect localization accuracy.

Having introduced the position estimation methods, We use $g_t \in \{0,1\}$ to represent the positioning method selection indicator with $g_t = 1$ implying that the BS uses a GAN-based positioning method to localize the target UAV and $g_t = 0$ implying that the BS uses the TDOA-based method. Then, the estimated position of the target UAV is

$$\hat{\boldsymbol{l}}_t\left(\boldsymbol{l}_{0,t}, p_t^{\mathrm{T}}, \boldsymbol{q}_t, g_t\right) = g_t \hat{\boldsymbol{l}}_t^{\mathrm{G}} + \left(1 - g_t\right) \hat{\boldsymbol{l}}_t^{\mathrm{W}}\left(\boldsymbol{l}_{0,t}, p_t^{\mathrm{T}}, \boldsymbol{q}_t\right) \quad (11)$$

The positioning error of the target UAV is the error between the estimated position $\hat{\boldsymbol{l}}_t\left(\boldsymbol{l}_{0,t}, p_t^{\mathrm{T}}, \boldsymbol{q}_t\right)$ and the ground truth position $\boldsymbol{l}_t$ of the target UAV at time slot t, which is given by $e_t\left(\boldsymbol{l}_{0,t}, p_t^{\mathrm{T}}, \boldsymbol{q}_t, g_t\right) = \sqrt{\left\|\hat{\boldsymbol{l}}_t\left(\boldsymbol{l}_{0,t}, p_t^{\mathrm{T}}, \boldsymbol{q}_t, g_t\right) - \boldsymbol{l}_t\right\|^2}$.

F. Problem Formulation

After defining the system model, our goal is to accurately estimate the real-time position of the target UAV under the

interference introduced by wireless channel noise and the jamming UAV. We formulate an optimization problem whose goal is to minimize the positioning error $e_t\left(\boldsymbol{l}_{0,t}, p_t^{\mathrm{T}}, \boldsymbol{q}_t, g_t\right)$ over T time slots by determining the trajectory α_t, β_t and transmit power p_t^{T} of the active UAV, the UAV selection scheme $\boldsymbol{q}_t$, and the position estimation method selection g_t under the attacks of the jamming UAV. Then, the minimization problem is given by

$$\min_{\alpha_t,\beta_t,p_t^{\mathrm{T}},\boldsymbol{q}_t,g_t} \quad \sum_{t=1}^{T} e_t\left(\boldsymbol{l}_{0,t}\left(\alpha_t,\beta_t\right), p_t^{\mathrm{T}}, \boldsymbol{q}_t, g_t\right), \tag{12}$$

$$\text{s.t.} \quad E_t^{\mathrm{F}} \leqslant E_{\max}^{\mathrm{F}}, \tag{12a}$$

$$q_{u,t} T_{u,t}^{\mathrm{G}} \leqslant T_{\max},\ \forall u \in \mathcal{U}, \tag{12b}$$

$$\|\boldsymbol{l}_{u,t} - \boldsymbol{l}_{0,t}\| \geqslant L_{\min}, \forall u \in \mathcal{U}, \tag{12c}$$

$$\beta_{\min} \leqslant \beta_t \leqslant \beta_{\max}, \tag{12d}$$

$$\alpha_{\min} \leqslant \alpha_t \leqslant \alpha_{\max}, \tag{12e}$$

$$\sum_{u=0}^{U} q_{u,t} = 4, \tag{12f}$$

where $E_{\max}^{\mathrm{F}}$ is the maximum active UAV propulsion energy consumption, $T_{\max}$ is the maximum transmission delay, and $L_{\min}$ is the safe distance between any two UAVs. Constraints (12a) is the flight energy constraint of the active UAV. Constraints (12b) is the delay requirements for distance measurement information transmission from passive UAVs to the BS, and constraint (12c) is the safe distance requirement between any two UAVs. Constraints (12d) and (12e) are the movement constraint of the active UAV. Constraint (12f) is the number of distance measurement information required to localize the target UAV[1].

Problem (12) is difficult to solve by traditional convex algorithms due to the following reasons. First, the relationship between the estimated position of the target UAV obtained by the BS and the optimization variables in (12) cannot be accurately characterized due to the unknown jamming pattern and positions of the jamming UAV. Second, traditional optimization algorithms require the BS to calculate the positioning error $e_t\left(\boldsymbol{l}_{0,t}, p_t^{\mathrm{T}}, \boldsymbol{q}_t, g_t\right)$ based on the ground truth position $\boldsymbol{l}_t$ of the target UAV. However, $\boldsymbol{l}_t$ is unknown in practice. To this end, we investigate a collaborative RL method to jointly optimize the UAV selection scheme, the trajectory and transmit power of the active UAV, and the position estimation method selection method according to the observation of the active UAV and the BS.

III. The Proposed Collaborative RL Method

To solve (12), we introduce a collaborative RL method. The proposed method empowers the active UAV to adjust its trajectory and transmit power and the BS to select the most appropriate subset of distance measurement information and the optimal positioning method to thereby jointly maximize the target positioning accuracy. Compared to traditional RL methods that use the DNN to output the estimated expected value of future rewards directly (such as [33]), the proposed mixture Gaussian distribution based RL method can use mixed Gaussian distributions to approximate the distribution of the sum of future rewards and use neural networks to predict the variance, means, and weights parameters of the Gaussian distributions thus reducing training complexity and improving convergence speed. Next, the components and training process of the proposed RL method are introduced first. Then, we analyze the convergence, implementation, and complexity of the proposed collaborative RL method.

A. Components of the collaborative RL method

The proposed method consists of the following five components:

- *Agents*: The agents are the BS and the active UAV. In particular, the BS requires to select an appropriate subset of the distance information and determine a positioning method (the GAN or the TDOA) for estimating the position of the target UAV. Meanwhile, the active UAV needs to optimize its transmit power and trajectory.
- *States*: A state of the BS is $\boldsymbol{o}_t^{\mathrm{B}} = \left[\hat{\boldsymbol{d}}_t, \boldsymbol{l}_t^{\mathrm{P}}, \boldsymbol{s}_t\right]$ that captures the distance measurement information, the deployment of active and passive UAVs, and the SINR at each passive UAV. A state of the active UAV is $\boldsymbol{o}_t = [\boldsymbol{l}_{0,t}]$ that captures its position at time slot t. Hereinafter, we use $\boldsymbol{o}_t = \left[\boldsymbol{o}_t, \boldsymbol{o}_t^{\mathrm{A}}\right]$ to represent the global state at time slot t.
- *Actions*: An action of the active UAV is represented by $\boldsymbol{a}_t^{\mathrm{A}} = \left[\alpha_t, \beta_t, p_t^{\mathrm{T}}\right]$ that optimize its trajectory and the transmit power. An action of the BS is $\boldsymbol{a}_t^{\mathrm{B}} = [\boldsymbol{q}_t, g_t]$, where $\boldsymbol{q}_t$ is the distance subset selection strategy and g_t is the position estimation method selection indicator. We define the actions of all agents at time slot t as $\boldsymbol{a}_t = \left[\boldsymbol{a}_t^{\mathrm{A}}, \boldsymbol{a}_t^{\mathrm{B}}\right]$.
- *Reward*: The reward of agents can be represented by $r_t\left(\boldsymbol{o}_t, \boldsymbol{a}_t\right) = -e_t\left(\boldsymbol{l}_{0,t}, p_t^{\mathrm{T}}, \boldsymbol{q}_t, g_t\right)$, where $e_t\left(\boldsymbol{l}_{0,t}, p_t^{\mathrm{T}}, \boldsymbol{q}_t, g_t\right)$ is the positioning error of the target UAV at time slot t. Since the BS calculates the estimated position $\hat{\boldsymbol{l}}_t\left(\boldsymbol{l}_{0,t}, p_t^{\mathrm{T}}, \boldsymbol{q}_t\right)$ of the target UAV after obtaining all distance measurement information from passive UAVs, the BS and the active UAV share a reward $r_t\left(\boldsymbol{o}_t, \boldsymbol{a}_t\right)$ at time slot t.
- *Value function*: Under a given state $\boldsymbol{o}_t$, a selected action $\boldsymbol{a}_t$, and a policy π, the value function of the active UAV is $v\left(\boldsymbol{o}_t^{\mathrm{A}}, \boldsymbol{a}_t^{\mathrm{A}}\right) = \sum_{t=0}^{\infty} \gamma^t r_t\left(\boldsymbol{o}_t, \boldsymbol{a}_t\right)$ with γ being the discount factor and the value function of the BS is $v\left(\boldsymbol{o}_t^{\mathrm{B}}, \boldsymbol{a}_t^{\mathrm{B}}\right) = \sum_{t=0}^{\infty} \gamma^t r_t\left(\boldsymbol{o}_t, \boldsymbol{a}_t\right)$. Compared with traditional RL methods that estimate the expected values of value functions directly [34], the proposed method first approximates the probability distributions of $v\left(\boldsymbol{o}_t^{\mathrm{A}}, \boldsymbol{a}_t^{\mathrm{A}}\right)$ and $v\left(\boldsymbol{o}_t^{\mathrm{B}}, \boldsymbol{a}_t^{\mathrm{B}}\right)$ by using mixture Gaussian distributions and then estimate the expected values of value functions by sampling the mixture Gaussian distributions. Next, the process of approximating the probability distribution of the sum of future rewards by mixture Gaussian distributions is introduced. Each mixture Gaussian distribution

[1]Estimating 3D coordinates of the target UAV requires at least four UAVs that receives signals reflected by the target UAV [32].

consists of several Gaussian distributions parameterized by variances and means. The active UAV and the BS approximate the probability distribution of their value functions (i.e., $v\left(\boldsymbol{o}_t^{\mathrm{A}}, \boldsymbol{a}_t^{\mathrm{A}}\right)$ and $v\left(\boldsymbol{o}_t^{\mathrm{B}}, \boldsymbol{a}_t^{\mathrm{B}}\right)$) by using DNNs parameterized by $\boldsymbol{w}_{\mathrm{A}}$ and $\boldsymbol{w}_{\mathrm{B}}$. The input of DNN at each agent is its state and action and the output is the variances $\boldsymbol{\lambda}^{\mathrm{A}} = \{\lambda_1^{\mathrm{A}}, \cdots, \lambda_k^{\mathrm{A}}, \cdots, \lambda_K^{\mathrm{A}}\}$, the means $\boldsymbol{\xi}^{\mathrm{A}} = \{\xi_1^{\mathrm{A}}, \cdots, \xi_k^{\mathrm{A}}, \cdots, \xi_K^{\mathrm{A}}\}$ of K Gaussian distributions, and their weight parameters $\boldsymbol{\phi}^{\mathrm{A}} = \{\phi_1^{\mathrm{A}}, \cdots, \phi_k^{\mathrm{A}}, \cdots, \phi_K^{\mathrm{A}}\}$. Here, the weight parameters ϕ_k^{A} represent the importance of Gaussian distribution k in the mixture Gaussian distribution. Given the mixture Gaussian distribution parameters, we sample S samples from the mixture Gaussian distribution so as to calculate the expected values of the value functions. Hence, the expected values of the approximated value functions of the active UAV and the BS can be written as

$$\bar{v}\left(\boldsymbol{o}_t^{\mathrm{A}}, \boldsymbol{a}_t^{\mathrm{A}}\right) = \frac{1}{S}\sum_{i=1}^{S} s_i^{\mathrm{A}}, \tag{13}$$

$$\bar{v}\left(\boldsymbol{o}_t^{\mathrm{B}}, \boldsymbol{a}_t^{\mathrm{B}}\right) = \frac{1}{S}\sum_{i=1}^{S} s_i^{\mathrm{B}}, \tag{14}$$

where $\bar{v}\left(\boldsymbol{o}_t^{\mathrm{A}}, \boldsymbol{a}_t^{\mathrm{A}}\right)$ and $\bar{v}\left(\boldsymbol{o}_t^{\mathrm{B}}, \boldsymbol{a}_t^{\mathrm{B}}\right)$ are the approximated expected values of value functions (i.e., $v\left(\boldsymbol{o}_t^{\mathrm{A}}, \boldsymbol{a}_t^{\mathrm{A}}\right)$ and $v\left(\boldsymbol{o}_t^{\mathrm{B}}, \boldsymbol{a}_t^{\mathrm{B}}\right)$), s_i^{A} and s_i^{B} are samples sampled from the mixture Gaussian distribution of the active UAV and the BS, respectively.

B. Training process of the collaborative RL approach

In this section, we introduce the collaborative implementation of the proposed method by the BS and the active UAV, aiming at minimizing the positioning error of the target UAV and mitigating jamming attacks. We begin by presenting the loss function associated with the proposed method, followed by a comprehensive description of the entire training process.

The loss function of the proposed method is [35]

$$\begin{aligned}\rho\left(\boldsymbol{w}_{\mathrm{A}}, \boldsymbol{w}_{\mathrm{B}}\right) =& \mathbb{E}\left[\left(r_t\left(\boldsymbol{o}_t, \boldsymbol{a}_t\right) + \gamma \max_{\boldsymbol{a}_{t+1}} \bar{v}\left(\boldsymbol{o}_{t+1}, \boldsymbol{a}_{t+1}\right)\right.\right. \\ & \left.\left. -\bar{v}\left(\boldsymbol{o}_t, \boldsymbol{a}_t\right)\right)^2\right],\end{aligned} \tag{15}$$

where $\bar{v}\left(\boldsymbol{o}_t, \boldsymbol{a}_t\right) = \bar{v}\left(\boldsymbol{o}_t^{\mathrm{A}}, \boldsymbol{a}_t^{\mathrm{A}}\right) + \bar{v}\left(\boldsymbol{o}_t^{\mathrm{B}}, \boldsymbol{a}_t^{\mathrm{B}}\right)$ and γ is the discounted factor.

Since the proposed method is collaboratively trained by the BS and the active UAV. The training process can be divided into the training process at the BS and the training process at the active UAV. Next, we will introduce these two training process in detail.

- Training process at the BS: From (15), the total loss requires passive UAVs to transmit a) the distance information measured by passive UAVs and real-time positions of passive UAVs for calculating the reward $r_t\left(\boldsymbol{o}_t, \boldsymbol{a}_t\right)$, and b) value function of the active UAV to the BS for calculating value functions $v_t\left(\boldsymbol{o}_{t+1}, \boldsymbol{a}_{t+1}, \boldsymbol{\zeta}\right)$ and $v_t\left(\boldsymbol{o}_t, \boldsymbol{a}_t, \boldsymbol{\zeta}\right)$. Based on $r_t\left(\boldsymbol{o}_t, \boldsymbol{a}_t\right)$, $v_t\left(\boldsymbol{o}_{t+1}, \boldsymbol{a}_{t+1}, \boldsymbol{\zeta}\right)$, and $v_t\left(\boldsymbol{o}_t, \boldsymbol{a}_t, \boldsymbol{\zeta}\right)$, the BS calculates the loss function $\rho\left(\boldsymbol{w}_0, \boldsymbol{w}_{\mathrm{B}}\right)$ based on (15) and updates its DNN parameters as follows

$$\boldsymbol{w}_{\mathrm{B}} = \boldsymbol{w}_{\mathrm{B}} + \alpha \bigtriangledown_{\boldsymbol{w}_{\mathrm{B}}} \rho\left(\boldsymbol{w}_{\mathrm{A}}, \boldsymbol{w}_{\mathrm{B}}\right), \tag{16}$$

  where α is the step size.
- Training process at the active UAV: The active UAV requires to update its DNN based on the loss function $\rho\left(\boldsymbol{w}_k, \boldsymbol{w}_{\mathrm{B}}\right)$. The update of the active UAV is given by

$$\boldsymbol{w}_{\mathrm{A}} = \boldsymbol{w}_{\mathrm{A}} + \alpha \bigtriangledown_{\boldsymbol{w}_{\mathrm{A}}} \rho\left(\boldsymbol{w}_{\mathrm{A}}, \boldsymbol{w}_{\mathrm{B}}\right). \tag{17}$$

Algorithm 1 summarizes the training procedure of the proposed method.

C. Convergence, Implementation, and Complexity Analysis

This section focuses on the proof of convergence, the implementation process, and the analysis of training complexity of the proposed mixture Gaussian distribution based RL method.

1) Convergence Analysis: We first define the optimal expected value $M^*\left(\boldsymbol{o}_t, \boldsymbol{a}_t\right)$ of the sum of future rewards. Then, the gap between $M^*\left(\boldsymbol{o}_t, \boldsymbol{a}_t\right)$ and $\bar{v}\left(\boldsymbol{o}_t, \boldsymbol{a}_t\right)$ can be given by $e\left(\boldsymbol{o}_t, \boldsymbol{a}_t\right) = M^*\left(\boldsymbol{o}_t, \boldsymbol{a}_t\right) - \bar{v}\left(\boldsymbol{o}_t, \boldsymbol{a}_t\right)$. To this end, to prove the convergence of the proposed collaborative RL method, we only need to prove that the gap $e\left(\boldsymbol{o}_t, \boldsymbol{a}_t\right)$ converges to zero.

Lemma 1. The proposed RL method is ensured to converge to zero when the gap $e_t\left(\boldsymbol{o}_t, \boldsymbol{a}_t\right)$ satisfies the following conditions [36]:

1) The gap $e\left(\boldsymbol{o}_t, \boldsymbol{a}_t\right)$ satisfies

$$e_{k+1}\left(\boldsymbol{o}_t, \boldsymbol{a}_t\right) = \left(1-\alpha\right) e_k\left(\boldsymbol{o}_t, \boldsymbol{a}_t\right) + \alpha F_k\left(\boldsymbol{o}_t, \boldsymbol{a}_t\right), \tag{18}$$

   where $F_k\left(\boldsymbol{o}_t, \boldsymbol{a}_t\right) = r\left(\boldsymbol{o}_t, \boldsymbol{a}_t\right) + \gamma M_k\left(\boldsymbol{o}_{t+1}, \boldsymbol{a}_{t+1}\right) - M^*\left(\boldsymbol{o}_t, \boldsymbol{a}_t\right)$.
2) $||\mathbb{E}\left[F_k\left(\boldsymbol{o}_t, \boldsymbol{a}_t\right)\right]||_\infty \leqslant \gamma||e_k\left(\boldsymbol{o}_t, \boldsymbol{a}_t\right)||_\infty, \forall \gamma \in (0,1)$, where $||\cdot||_\infty$ represents the maximal absolute value of elements, $\mathbb{E}\left[F_k\left(\boldsymbol{o}_t, \boldsymbol{a}_t\right)\right]$ is the expected value of $F_k\left(\boldsymbol{o}_t, \boldsymbol{a}_t\right)$ with respect to the state transition probability.
3) $\mathrm{Var}\left(F_k\left(\boldsymbol{o}_t, \boldsymbol{a}_t\right)\right) \leqslant C_{\mathrm{F}}\left(1 + ||e_k\left(\boldsymbol{o}_t, \boldsymbol{a}_t\right)||_\infty^2\right)$, where $\mathrm{Var}\left(F_k\left(\boldsymbol{o}_t, \boldsymbol{a}_t\right)\right)$ is the variance of $F_k\left(\boldsymbol{o}_t, \boldsymbol{a}_t\right)$, and C_{F} is some constant.

   Proof: See Appendix A. □

2) Implementation Analysis: The implementation of the proposed mixture Gaussian distribution based RL method includes 1) the off-policy training stage and 2) the on-policy decision making stage. In the off-policy stage, the BS first requires to collect the transmit power, distance measurement information, and the position of the active UAV as well as distance information measurement and positions of passive UAVs for calculating the positioning error of the target UAV. In addition, the BS also needs to collect the expected values $\bar{v}\left(\boldsymbol{o}_t^{\mathrm{A}}, \boldsymbol{a}_t^{\mathrm{A}}\right)$ of the sum of future rewards at the active UAV to calculate the expected values of the global value function $\bar{v}\left(\boldsymbol{o}_t, \boldsymbol{a}_t\right)$. Then, in terms of the active UAV, it requires to collect the approximated expected values of the global value function and the value of reward to update parameters of

Algorithm 1 Mixture Gaussian distribution based RL Method

1: Initialize the DNN parameters $\boldsymbol{w}_{\mathrm{A}}$ and $\boldsymbol{w}_{\mathrm{B}}$.
2: **for** each iteration **do**
3. **for** $t = 1, \cdots, T$ **do**
5. Observe the environment $\boldsymbol{o}_t^{\mathrm{A}}$ and $\boldsymbol{o}_t^{\mathrm{B}}$.
6: According to a ϵ-greedy scheme, agents select actions.
7: Output mixture Gaussian distributions and sample to calculate the expected value function values $\bar{v}\left(\boldsymbol{o}_t^{\mathrm{A}}, \boldsymbol{a}_t^{\mathrm{A}}\right)$and $\bar{v}\left(\boldsymbol{o}_t^{\mathrm{B}}, \boldsymbol{a}_t^{\mathrm{B}}\right)$ at time slot t and $t+1$.
8: **end for**
9: The active UAV transmits $\boldsymbol{o}_t^{\mathrm{A}}$, $\bar{v}\left(\boldsymbol{o}_t^{\mathrm{A}}, \boldsymbol{a}_t^{\mathrm{A}}\right)$, and $\bar{v}\left(\boldsymbol{o}_{t+1}^{\mathrm{A}}, \boldsymbol{a}_{t+1}^{\mathrm{A}}\right)$ to the BS.
10: **end for**
11: The BS calculates the value of loss function and transmits it to the active UAV.
12: **for** each agent u **do**
13: Update $\boldsymbol{w}_{\mathrm{A}}$ and $\boldsymbol{w}_{\mathrm{B}}$ based on (16) and (17).
14: **end for**
15: **end for**

DNN parameters according to (15) and (17). In the on-policy stage, the proficiently trained DNNs can be employed directly to ascertain the distance measurement information selection scheme, the positioning method selection scheme, the active UAV transmit power, and the active UAV trajectory.

3) Complexity Analysis: The complexity of training the proposed mixture Gaussian distribution based RL method consists of the complexity of training DNN parameters at the active UAV and the complexity of training DNN parameters of the BS at each iteration.

In terms of complexity of training the DNN at the active UAV, the proposed RL method used the mixture Gaussian distribution to approximate the probability distribution of value functions. The mixture Gaussian distribution consists of K Gaussian distributions parameterized by variances, means, and weights. Hence, the time-complexity of training the DNN at the active UAV is $\mathcal{O}\left(\left|\boldsymbol{o}_t^{\mathrm{A}}\right| l_1^{\mathrm{A}} + \sum_{i=1}^{L_{\mathrm{A}}} l_i^{\mathrm{A}} l_{i+1}^{\mathrm{A}} + 3Kl_{L_{\mathrm{A}}}^{\mathrm{A}} + 3\left|\boldsymbol{a}_t^{\mathrm{A}}\right| K\right)$, where L_{A} is the number of hidden layers at the active UAV, l_i^{A} is the number of neurons in the i-th hidden layer, $\left|\boldsymbol{a}_t^{\mathrm{A}}\right|$ is the dimension of $\boldsymbol{a}_t^{\mathrm{A}}$, and $\left|\boldsymbol{o}_t^{\mathrm{A}}\right|$ is the dimension of $\boldsymbol{o}_t^{\mathrm{A}}$.

Similarly, the time-complexity of training the DNN at the BS is $\mathcal{O}\left(\left|\boldsymbol{o}_t^{\mathrm{B}}\right| l_1^{\mathrm{B}} + \sum_{i=1}^{L_{\mathrm{B}}} l_i^{\mathrm{B}} l_{i+1}^{\mathrm{B}} + 3Kl_{L_{\mathrm{B}}} + 3\left|\boldsymbol{a}_t^{\mathrm{B}}\right| K\right)$, where L_{B} is the number of hidden layers of the DNN at the BS. Hence, the entire complexity of training the proposed RL by both BS and the active UAV is

$$\mathcal{O}\left(\max\left(\left|\boldsymbol{o}_t^{\mathrm{A}}\right| l_1^{\mathrm{A}} + \sum_{i=1}^{L_{\mathrm{A}}} l_i^{\mathrm{A}} l_{i+1}^{\mathrm{A}} + 3Kl_{L_{\mathrm{A}}} + 3\left|\boldsymbol{a}_t^{\mathrm{A}}\right| K, \left|\boldsymbol{o}_t^{\mathrm{B}}\right| l_1^{\mathrm{B}} + \sum_{i=1}^{L_{\mathrm{B}}} l_i^{\mathrm{B}} l_{i+1}^{\mathrm{B}} + 3Kl_{L_{\mathrm{B}}} + 3\left|\boldsymbol{a}_t^{\mathrm{B}}\right| K\right)\right). \tag{19}$$

IV. Analysis of Target UAV Localization Performance

This section aims to analyze the impact of jamming attacks on the positioning errors of the target UAV. The distance measured by passive UAV u can be expressed as

$$\hat{d}_{u,t} = r_{u,t} + r_{0,t} + \Delta d_{u,t}, \tag{20}$$

where $r_{u,t} = \sqrt{\|\boldsymbol{l}_{u,t} - \boldsymbol{l}_t\|^2}$ represents the true distance between the target UAV and passive UAV u and $r_{0,t} = \sqrt{\|\boldsymbol{l}_{0,t} - \boldsymbol{l}_t\|^2}$ represents the true distance between the target UAV and the active UAV. $\Delta d_{u,t}$ is the distance measurement error between the true distance information $r_{u,t} + r_{0,t}$ and the measured distance information $\hat{d}_{u,t}$ and $\Delta d_{u,t}$ follows the Gaussian distribution, where the mean of the Gaussian distribution is zero and the variance dependent on $s_{u,t}^{\mathrm{A}}\left(\boldsymbol{l}_{0,t}, p_t^{\mathrm{T}}\right)$. Take the derivative of both sides of (20) with respect to $\boldsymbol{l}_t$, we have

$$\begin{aligned}\partial\hat{d}_{u,t} = &\left(\frac{x_t - x_{u,t}}{r_{u,t}} + \frac{x_t - x_{0,t}}{r_{0,t}}\right)\partial x_t \\ &+ \left(\frac{y_t - y_{u,t}}{r_{u,t}} + \frac{y_t - y_{0,t}}{r_{0,t}}\right)\partial y_t \\ &+ \left(\frac{z_t - z_{u,t}}{r_{u,t}} + \frac{z_t - z_{0,t}}{r_{0,t}}\right)\partial z_t.\end{aligned} \tag{21}$$

Based on four distance measurement information selected by the BS, the position of the target UAV can be estimated. We denote the selected distance subset as $\hat{\boldsymbol{d}}_t^{\mathrm{S}} = \left[\hat{d}_{m_1,t}, \hat{d}_{m_2,t}, \hat{d}_{m_3,t}, \hat{d}_{m_4,t}\right]^T \subset \hat{\boldsymbol{d}}_t$, where $m_1, m_2, m_3, m_4 \in \{1, \cdots, U\}$ are the passive UAVs. Based on (21), the relationship between the selected distance subset and the estimated target UAV position can be given by

$$\partial\hat{\boldsymbol{d}}_t^{\mathrm{S}} = \boldsymbol{M}\partial\boldsymbol{l}_t, \tag{23}$$

where $\partial\hat{\boldsymbol{d}}_t^{\mathrm{S}} = \left[\partial\hat{d}_{m_1,t}, \partial\hat{d}_{m_2,t}, \partial\hat{d}_{m_3,t}, \partial\hat{d}_{m_4,t}\right]^T$, $\partial\hat{\boldsymbol{l}}_t = [\partial x_t, \partial y_t, \partial z_t]^T$, and $\boldsymbol{M}$ is given by (22).

From (23), we have

$$\partial\boldsymbol{l}_t = \left(\boldsymbol{M}^T\boldsymbol{M}\right)^{-1}\boldsymbol{M}^T\partial\hat{\boldsymbol{d}}_t^{\mathrm{S}}, \tag{24}$$

where $\left(\boldsymbol{M}^T\boldsymbol{M}\right)^{-1}$ is the inverse matrix of $\boldsymbol{M}^T\boldsymbol{M}$. The positioning error of the target UAV can be written as

$$e_t = \sqrt{(\partial x_t)^2 + (\partial y_t)^2 + (\partial z_t)^2} = \mathrm{tr}\left(\mathbb{E}\left[\partial\boldsymbol{l}_t\left(\partial\boldsymbol{l}_t\right)^T\right]\right), \tag{25}$$

where $\mathrm{tr}(\cdot)$ is the trace of the matrix $\mathbb{E}\left[\partial\boldsymbol{l}_t\left(\partial\boldsymbol{l}_t\right)^T\right]$. Then, we analyze how the jamming attacks affect the positioning Proposition 1.

Proposition 1. The positioning error of the target UAV is

$$e_t = \mathrm{tr}\left(\left(\boldsymbol{M}^T\boldsymbol{M}\right)^{-1}\boldsymbol{M}^T\boldsymbol{J}\boldsymbol{M}\left(\boldsymbol{M}^T\boldsymbol{M}\right)^{-1}\right), \tag{26}$$

where $\boldsymbol{J} = \mathrm{diag}\left(\frac{k_1}{\epsilon^2 + j_t P^{\mathrm{J}}|h_{\mathrm{J},m_1,t}|^2}, \cdots, \frac{k_4}{\epsilon^2 + j_t P^{\mathrm{J}}|h_{\mathrm{J},m_4,t}|^2}\right)$ is the distance measurement variance matrix with k_i being a

$$\boldsymbol{M}=\begin{bmatrix} \frac{x_t-x_{m_1,t}}{r_{m_1,t}}+\frac{x_t-x_{0,t}}{r_{0,t}} & \frac{y_t-y_{m_1,t}}{r_{m_1,t}}+\frac{y_t-y_{0,t}}{r_{0,t}} & \frac{z_t-z_{m_1,t}}{r_{m_1,t}}+\frac{z_t-z_{0,t}}{r_{0,t}} \\ \frac{x_t-x_{m_2,t}}{r_{m_2,t}}+\frac{x_t-x_{0,t}}{r_{0,t}} & \frac{y_t-y_{m_2,t}}{r_{m_2,t}}+\frac{y_t-y_{0,t}}{r_{0,t}} & \frac{z_t-z_{m_2,t}}{r_{m_2,t}}+\frac{z_t-z_{0,t}}{r_{0,t}} \\ \frac{x_t-x_{m_3,t}}{r_{m_3,t}}+\frac{x_t-x_{0,t}}{r_{0,t}} & \frac{y_t-y_{m_3,t}}{r_{m_3,t}}+\frac{y_t-y_{0,t}}{r_{0,t}} & \frac{z_t-z_{m_3,t}}{r_{m_3,t}}+\frac{z_t-z_{0,t}}{r_{0,t}} \\ \frac{x_t-x_{m_4,t}}{r_{m_4,t}}+\frac{x_t-x_{0,t}}{r_{0,t}} & \frac{y_t-y_{m_4,t}}{r_{m_4,t}}+\frac{y_t-y_{0,t}}{r_{0,t}} & \frac{z_t-z_{m_4,t}}{r_{m_4,t}}+\frac{z_t-z_{0,t}}{r_{0,t}} \end{bmatrix}. \tag{22}$$

TABLE II. PARAMETERS

Parameters	Values	Parameters	Values
ϵ^2	-95 dBm	Δ_t	1 s
$v_{u,t}^{\mathrm{H}}$	9.43 m/s	$p_{u,t}$	5 W
C_1	4929	W	1 MHz
C_2	0.002	M	4 kg
σ_{LoS}^2	8.41	σ_{NLoS}^2	33.78
$E_{\max}^{\mathrm{F}}$	500 J	D_{B}	5 bit
$L_{\min}$	80 m	$L_{\max}$	10 km
$\beta_{\min}$	-15^o	$\beta_{\max}$	15^o
$\alpha_{\min}$	-15^o	$\alpha_{\max}$	15^o
X	11.9	Y	0.13
T	10	f_{J}	0.5
μ_{LoS}^B	2	μ_{NLoS}^B	2.4

coefficient, j_t being the jamming indicator, and P^{J} being the jamming power.

Proof: See Appendix B. □

From Proposition 1, we see that the positioning error depends on the distance measurement variance matrix and the position of the active and passive UAVs. In particular, when the jamming UAV transmits signals in real time at a fixed jamming power P^{J} and the distance between passive UAVs and the target UAV satisfy $r_{m_1,t}=r_{m_2,t}=r_{m_3,t}=r_{m_4,t}$, we have $\boldsymbol{J}=\mathrm{diag}\left(\frac{k}{\epsilon^2+P^{\mathrm{J}}|h_{\mathrm{J},m_1,t}|^2},\cdots,\frac{k}{\epsilon^2+P^{\mathrm{J}}|h_{\mathrm{J},m_1,t}|^2}\right)=\frac{k}{\epsilon^2+P^{\mathrm{J}}|h_{\mathrm{J},m_1,t}|^2}\boldsymbol{I}$ with $k=k_i$. Then, e_t is

$$\begin{aligned} e_t &= \mathrm{tr}\left(\left(\boldsymbol{M}^T\boldsymbol{M}\right)^{-1}\boldsymbol{M}^T\frac{k}{\epsilon^2+P^{\mathrm{J}}|h_{\mathrm{J},m_1,t}|^2}\boldsymbol{I}\boldsymbol{M}\left(\boldsymbol{M}^T\boldsymbol{M}\right)^{-1}\right) \\ &= \frac{k}{\epsilon^2+P^{\mathrm{J}}|h_{\mathrm{J},m_1,t}|^2}\mathrm{tr}\left(\left(\boldsymbol{M}^T\boldsymbol{M}\right)^{-1}\boldsymbol{M}^T\boldsymbol{M}\left(\boldsymbol{M}^T\boldsymbol{M}\right)^{-1}\right) \\ &= \frac{k}{\epsilon^2+P^{\mathrm{J}}|h_{\mathrm{J},m_1,t}|^2}\mathrm{tr}\left(\left(\boldsymbol{M}^T\boldsymbol{M}\right)^{-1}\right). \end{aligned} \tag{27}$$

V. Simulation Results and Analysis

For simulations, we consider that the jamming UAV, the active UAV, and passive UAVs are randomly distributed in a 3D space. The system parameters of the simulations are listed in Table I. Next, we first introduce the models of GAN and the proposed RL. Then, we analyze the simulation results.

A. Models of GAN and the Proposed RL Method

The GAN consists of a generator network and a discriminator network. The generator network includes an input layer with 16 neurons representing the selected four distance information and 3D positions of 4 UAVs, an output layer with 3 neurons representing the estimated 3D position of the target UAV, and five fully connected hidden layers with 4096, 2048, 1024, 512, 256, and 64 neurons, respectively. The discriminator network consists of an input layer with 22 neurons representing the input and output of the generator network, an output layer with one neuron outputting 0 or 1, and four fully hidden layers with 1024, 512, 256, 128, and 64 neurons. In the proposed RL method, the DNN of each agent consists of one input layer, two hidden layers (a fully connected layer and a gated recurrent unit (GRU) recurrent layer), and an output layer. Each hidden layer has 64 neurons.

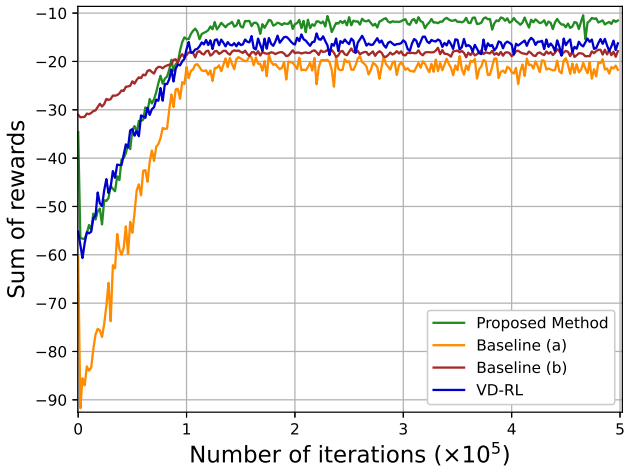


Fig. 2. The convergence of the proposed method.

B. Simulation Results

For comparison purpose, we consider three baseline methods: a) an algorithm that uses the proposed RL method and uses only the GAN-based position estimation method to avoid attacks without considering trajectory design of the active UAV to avoid jamming attacks, b) an algorithm that uses the proposed RL and uses only trajectory design of the active UAV to avoid jamming attacks but does not consider the GAN-based position estimation method, and c) VD-RL method that selects the optimal position estimation method by using a value decomposition based deep Q network [34].

In Fig. 2, we show the convergence of the proposed method and three baseline methods. From this figure, we see that the proposed method can achieve up to 36.5%, 27.4%, and 12.7% gains in terms of the sum of rewards compared to the baselines a), b), and c). The 36.5% and 27.4% gains stem from the fact that baselines a) and b) use only one position estimation method (GAN-based or TDOA-based method) but

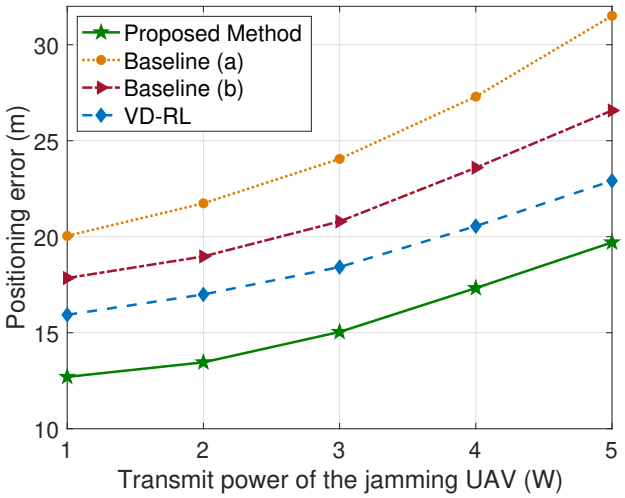


Fig. 3. Positioning error varies with fluctuations in the jamming power of the jamming UAV.

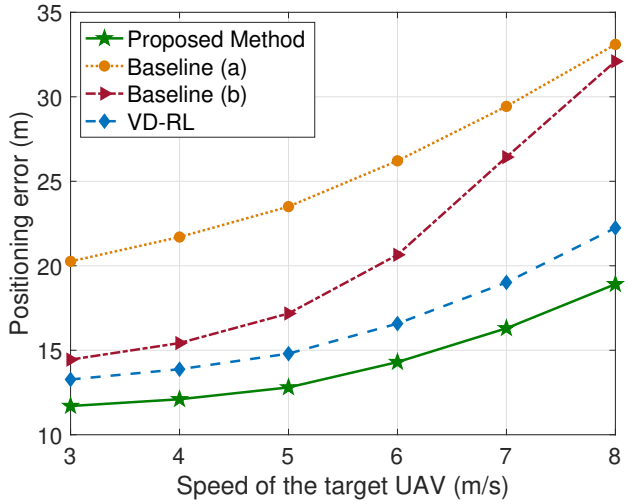


Fig. 4. Positioning error varies with fluctuations in the speed of the target UAV.

the proposed method can dynamically select the optimal position estimation method according to the positions of UAVs and jamming attack patterns. The 12.7% gain stems from the fact that the proposed mixture Gaussian distribution based collaborative RL method uses value function probability distribution to estimate the expected value thus approximating the value function accurately.

Fig. 3 shows the positioning errors of the target UAV under varying transmit power of the jamming UAV. In Fig. 3, the positioning error of the target UAV obtained by these methods increase as the transmit power of the jamming UAV increases. The reason is that as the jamming power increases, the SINR of received signals at passive UAVs decreases. In addition, compared to baselines a), b), and c), the proposed method can achieve up to 37.4%, 24.8%, and 14.0% gains in terms of the positioning error of the target UAV with the jamming power to be 5 W. These gains stem from the fact that the proposed method can use mixture Gaussian distributions to approximate the distribution of value functions and estimate the expected values accurately. Based on the accurate expected values, the proposed method can optimally adjust the trajectory and transmit power of the active UAV and enable the BS to select the optimal position estimation method while baselines a) and b) use a fixed position estimation method.

Fig. 4 shows the positioning errors of the target UAV at different flying speeds. From Fig. 4, as the speed of the target UAV increases, the positioning errors of the target UAV obtained by all methods increase. This is because the active UAV cannot follow the target UAV with increasing speed in time, thus increasing the distance information error. In Fig. 4, we also shows when the speed of the target UAV is 6 m/s, compared to baselines (a), (b), and the VD-RL method, the proposed method can reduce the positioning error of the target UAV by up to 45.4%, 30.8% and 13.7%. From Fig. 4, we can also see that the increasing rate of the positioning

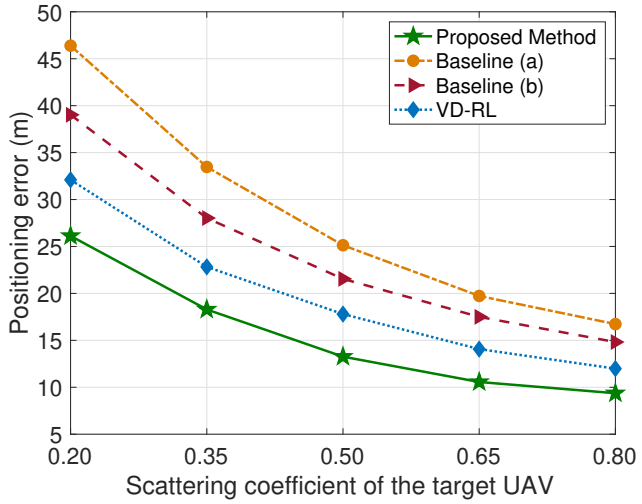


Fig. 5. Positioning error as the scattering coefficient varies.

error obtained by baseline (b) is the fastest. The reason is that baseline (b) only uses the GAN-based position estimation method to avoid attack. Since GAN-based positioning method depends on the training data samples. When the speed of target UAV increases, the limited number of data samples cannot cover the UAV moving range, thus the positioning error increases rapidly.

Fig. 5 shows the positioning errors under different scattering coefficients of the target UAV. In Fig. 5, it is observed that with the increase in the scattering coefficient of the target UAV, the positioning errors decrease. This trend occurs due to as the scattering coefficient increases, the signals strength received by passive UAVs increase and the SINR at passive UAVs increase, thus improving the accuracy of the distance measurement information. Additionally, Fig. 5 illustrates that the positioning error decreases first quickly and then becomes slowly. The

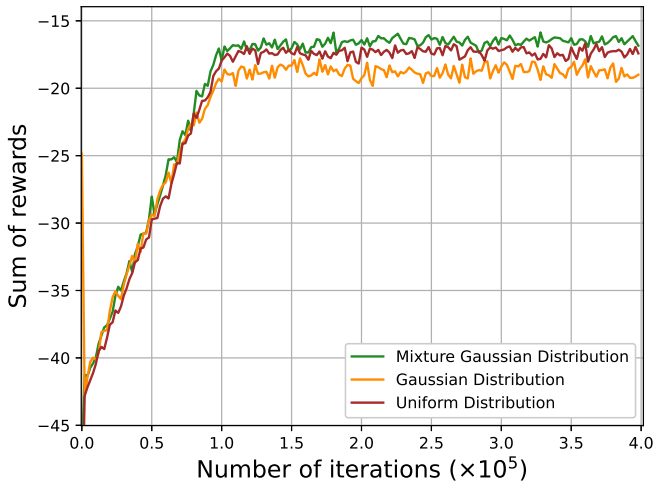


Fig. 6. Positioning error under different distributions.

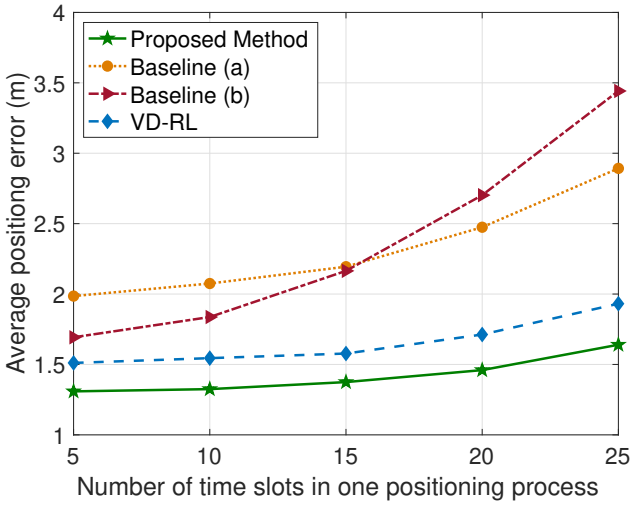


Fig. 7. Positioning error varies with fluctuations in the number of time slots.

reason is when the scattering coefficient of the target UAV is small, the signal strength received by passive UAVs are the main factor of limiting the localization performance. When the scattering coefficient is large enough, the scattering coefficient is no longer the main factor, and the localization performance is mainly affected by the other factors such as the jamming attacks and the deployment of UAVs. In addition, Fig. 5 demonstrates the capability of the proposed method that can diminish the positioning error of the target UAV by as much as 41.9%, 24.5%, and 18.8% compared to baselines (a), (b), and (c) when the scattering coefficient of the target UAV is 0.35. The 41.9% gain is because that baseline (a) only uses the TDOA-based position estimation method to avoid attacks. Since TDOA-based positioning method depends on the SINRs of received signals at passive UAVs. When the scattering coefficient is small, the SINRs of passive UAVs are small and the localization accuracy obtained by TDOA-based positioning method is worse than other methods.

In Fig. 6, we show the convergence of the proposed RL method that uses uniform distribution, Gaussian distribution, and mixture Gaussian distribution to approximate the probability distribution of the sum of future rewards when the speed of the target UAV is 7 m/s, respectively. In particular, the uniform distribution can be represented by two parameters: the lower and upper bounds. The DNN at each agent outputs the values of the two bounds to approximate the probability distribution of value functions by a uniform distribution. In addition, each DNN can output the variance and the mean to approximate the probability distribution of value functions by a Gaussian distribution. Fig. 6 shows the localization performance of the proposed RL method with mixture Gaussian distribution is better than that of other distributions. This is because the mixture Gaussian distribution can adjust the parameters of each individual Gaussian components flexibly, thus providing a feasible and comprehensive representation of the distribution of value functions.

Fig. 7 illustrates the variation in the average positioning error of the target UAV, denoted as $\bar{e}_t = \frac{1}{T}\sum_{t=1}^{T} e_t$, with respect to the number of time slots T. As depicted in Figure 7, it is observed that as T increases, the average positioning errors of the target UAV obtained by all considered methods increase.This is because the increasing number of time slots leads to a larger movement range of the target UAV. The GAN-based positioning method cannot always localize the target UAV accurately due to the limitation of the training datasets. Fig. 7 also shows that the as the number of time slots in one positioning process increase, the increase of the average positioning error obtained by baseline (b) that only uses GAN-based positioning error is the fastest and the average positioning error obtained by the proposed mixture Gaussian distribution based RL method increases slower than other baseline methods, this is because as the number of time slots increase, the moving range of the target UAV becomes larger and GAN-based positioning error cannot estimate the real-time position of the target UAV when the target UAV is not in the coverage of the training data samples while the proposed RL method can adaptively select the optimal positioning method from GAN-based and TDOA-based positioning methods.

Fig. 8 shows the convergence performance of the proposed mixture Gaussian distribution based RL method as the number of Gaussian distributions used to approximate the probability distribution of value functions. From Fig. 8, we can see that the positioning error of the target UAV decreases as the number of Gaussian distributions used to approximate the distribution of the sum of future rewards increases. This phenomenon occurs because when the number of Gaussian distributions increases, the mixture Gaussian distribution can represent the distribution of the sum of future rewards more accurately. Furthermore, Fig. 8 indicates that as the number of Gaussian distributions

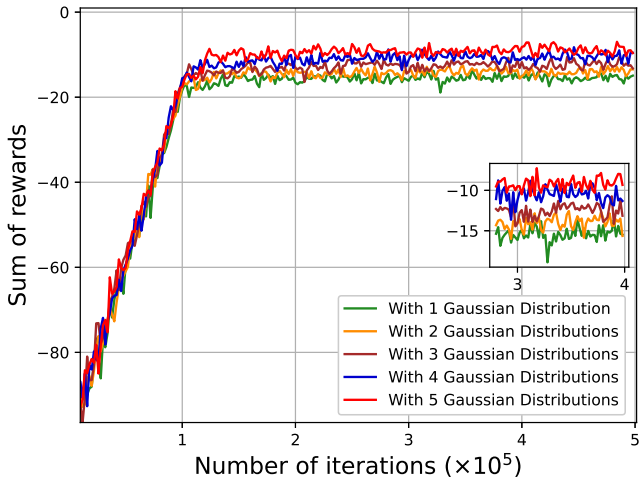
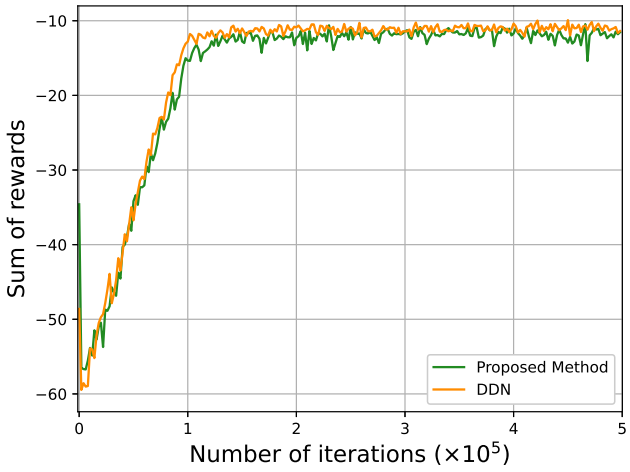


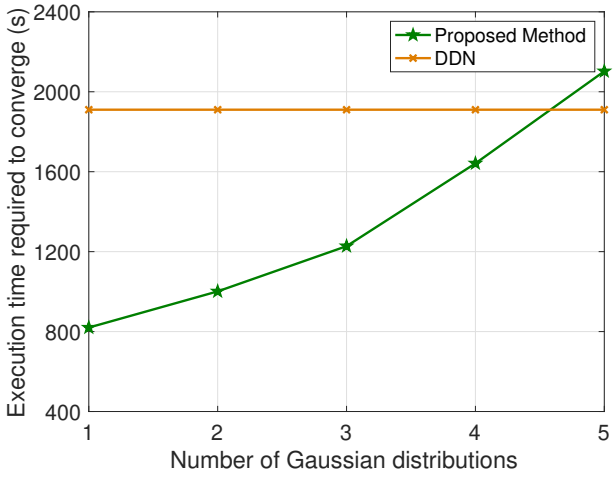
Fig. 8. Positioning error varies with fluctuations in the number of Gaussian distributions.

increases, the proposed mixture Gaussian distribution based RL method requires more iterations to converge. This stems from the fact that as the number of Gaussian distributions increases, the number of neurons at the output layer of each DNN increases.

Fig. 9 shows the convergence of the proposed mixture Gaussian distribution based RL method and the DDN method [37]. The DDN method approximates the distribution of the sum of future rewards by using the sum of future rewards under different probabilities. The input of each DNN in DDN method is a set of probability values and the output is the approximated distribution of the sum of future rewards. In Fig. 9(a), we see the proposed method has the similar localization accuracy as the DDN method when the number of Gaussian distributions in the proposed method is 3. Moreover, we have tested the execution time per iteration of the DDN method, which is 0.0186 s, and the proposed method with $\{1, 2, 3, 4, 5\}$ Gaussian distributions, which are 0.0083s, 0.0094 s, 0.0103 s, 0.0167 s. The number of iterations required by the DDN method to reach convergence is 10270 and the number of iterations required by the proposed method with $\{1, 2, 3, 4, 5\}$ Gaussian distributions are 98800, 106500, 112600, 118100, and 125900. Hence, the execution time required for these methods to reach convergence is shown in Fig. 9(b). From Fig. 9(b), we see as the number of number of iterations increases, the execution time required by the proposed method to reach converge increase and it is smaller than that of the DDN method when the number of Gaussian distributions is smaller than 5. This is because the proposed RL method uses mixture Gaussian distribution to approximate the probability distribution of value functions. By adjusting the parameters of the mixture Gaussian distribution, the proposed method can capture the features of the distribution of value functions more accurately.



(a) Convergence performance



(b) Execution time required to converge

Fig. 9. Performance of the proposed method and the DDN method

VI. Conclusion

In this paper, we have proposed a novel framework that enables an active UAV and a BS cooperatively localize the target UAV under jamming attacks. In the proposed framework, the BS can jointly use the GAN-based positioning method and TDOA-based positioning method to improve localization accuracy and avoide jamming attacks. We have formulated an optimization problem whose goal is to minimize the positioning error of the target UAV while considering the jamming attacks and UAV trajectory. To address this problem, a mixture Gaussian distribution model based collaborative RL method is proposed. This method empowers the active UAV to optimize its transmit power and trajectory and the BS to select an appropriate subset of distance information and the optimal positioning method. Simulation results demonstrate the significant reduction in the positioning error of the target UAV achieved by our proposed method compared to baseline methods.

APPENDIX

A. Proof of Lemma 1

Here, we first prove that the gap $e_t(\boldsymbol{o}_t, \boldsymbol{a}_t)$ satisfies condition 1). From (17), $e_{k+1}(\boldsymbol{o}_t, \boldsymbol{a}_t)$ can be written as

$$\begin{aligned}
&e_{k+1}(\boldsymbol{o}_t, \boldsymbol{a}_t) \\
&= (1-\alpha) M_k(\boldsymbol{o}_t, \boldsymbol{a}_t) - M^*(\boldsymbol{o}_t, \boldsymbol{a}_t) \\
&\quad + \alpha \left(r_t(\boldsymbol{o}_t, \boldsymbol{a}_t) + \gamma \max_{\boldsymbol{a}_{t+1}} M_k(\boldsymbol{o}_{t+1}, \boldsymbol{a}_{t+1}) \right) \\
&= (1-\alpha)\left(M_k(\boldsymbol{o}_t, \boldsymbol{a}_t) - M^*(\boldsymbol{o}_t, \boldsymbol{a}_t)\right) \\
&\quad + \alpha \left(r_t(\boldsymbol{o}_t, \boldsymbol{a}_t) + \gamma \max_{\boldsymbol{a}_{t+1}} M_k(\boldsymbol{o}_{t+1}, \boldsymbol{a}_{t+1}) - M^*(\boldsymbol{o}_t, \boldsymbol{a}_t) \right) \\
&= (1-\alpha) e_k(\boldsymbol{o}_t, \boldsymbol{a}_t) + \alpha F_k(\boldsymbol{o}_t, \boldsymbol{a}_t).
\end{aligned} \tag{28}$$

Hence, condition 1) is proved to be satisfied. Then, we prove condition 2) is satisfied. Since $F_k(\boldsymbol{o}_t, \boldsymbol{a}_t) = r(\boldsymbol{o}_t, \boldsymbol{a}_t) + \gamma M_k(\boldsymbol{o}_t, \boldsymbol{a}_t) - M^*(\boldsymbol{o}_t, \boldsymbol{a}_t)$, the expected value of $F_k(\boldsymbol{o}_t, \boldsymbol{a}_t)$ is given by

$$\begin{aligned}
&\mathbb{E}\left[F_k(\boldsymbol{o}_t, \boldsymbol{a}_t)\right] \\
&= \mathbb{E}\left[r(\boldsymbol{o}_t, \boldsymbol{a}_t) + \gamma M_k(\boldsymbol{o}_t, \boldsymbol{a}_t) - M^*(\boldsymbol{o}_t, \boldsymbol{a}_t)\right] \\
&= \sum_{\boldsymbol{o}_{t+1}} P_{\boldsymbol{a}_t}(\boldsymbol{o}_t, \boldsymbol{o}_{t+1}) \left[r(\boldsymbol{o}_t, \boldsymbol{a}_t) + \gamma M_k(\boldsymbol{o}_{t+1}, \boldsymbol{a}_{t+1}) \right. \\
&\quad \left. - M^*(\boldsymbol{o}_t, \boldsymbol{a}_t) \right] \\
&\overset{(a)}{=} \sum_{\boldsymbol{o}_{t+1}} P_{\boldsymbol{a}_t}(\boldsymbol{o}_t, \boldsymbol{o}_{t+1}) \left[r(\boldsymbol{o}_t, \boldsymbol{a}_t) + \gamma M_k(\boldsymbol{o}_{t+1}, \boldsymbol{a}_{t+1}) \right. \\
&\quad \left. - r(\boldsymbol{o}_t, \boldsymbol{a}_t) - \gamma M^*(\boldsymbol{o}_{t+1}, \boldsymbol{a}_{t+1}) \right] \\
&= \gamma \sum_{\boldsymbol{o}_{t+1}} P_{\boldsymbol{a}_t}(\boldsymbol{o}_t, \boldsymbol{o}_{t+1}) \left[M_k(\boldsymbol{o}_{t+1}, \boldsymbol{a}_{t+1}) - M^*(\boldsymbol{o}_{t+1}, \boldsymbol{a}_{t+1}) \right],
\end{aligned} \tag{29}$$

where (a) stems from the fact that $M^*(\boldsymbol{o}_t, \boldsymbol{a}_t) = r(\boldsymbol{o}_t, \boldsymbol{a}_t) + \gamma M^*(\boldsymbol{o}_{t+1}, \boldsymbol{a}_{t+1})$. Since $M_k(\boldsymbol{o}_{t+1}, \boldsymbol{a}_{t+1}) = \max_{\boldsymbol{a}'_t} M_k(\boldsymbol{o}_{t+1}, \boldsymbol{a}'_t)$, we have

$$\begin{aligned}
&\left\|\mathbb{E}\left[F_k(\boldsymbol{o}_t, \boldsymbol{a}_t)\right]\right\|_\infty \\
&= \max_{\boldsymbol{o}_t, \boldsymbol{a}_t} \left| \gamma \sum_{\boldsymbol{o}_{t+1}} P_{\boldsymbol{a}_t}(\boldsymbol{o}_t, \boldsymbol{o}_{t+1}) \left[\max_{\boldsymbol{a}'_t} M_k(\boldsymbol{o}_{t+1}, \boldsymbol{a}'_t) \right. \right. \\
&\quad \left. \left. - M^*(\boldsymbol{o}_{t+1}, \boldsymbol{a}_{t+1}) \right] \right| \\
&= \max_{\boldsymbol{o}_t, \boldsymbol{a}_t} \gamma \sum_{\boldsymbol{o}_{t+1}} P_{\boldsymbol{a}_t}(\boldsymbol{o}_t, \boldsymbol{o}_{t+1}) \left| \max_{\boldsymbol{a}'_t} M_k(\boldsymbol{o}_{t+1}, \boldsymbol{a}'_t) \right. \\
&\quad \left. - M^*(\boldsymbol{o}_{t+1}, \boldsymbol{a}_{t+1}) \right| \\
&\leqslant \max_{\boldsymbol{o}_t, \boldsymbol{a}_t} \gamma \sum_{\boldsymbol{o}_{t+1}} P_{\boldsymbol{a}_t}(\boldsymbol{o}_t, \boldsymbol{o}_{t+1}) \max_{\boldsymbol{a}'_t} \left| M_k(\boldsymbol{o}_{t+1}, \boldsymbol{a}'_t) \right. \\
&\quad \left. - M^*(\boldsymbol{o}_{t+1}, \boldsymbol{a}_{t+1}) \right| \\
&= \max_{\boldsymbol{o}_t, \boldsymbol{a}_t} \gamma \sum_{\boldsymbol{o}_{t+1}} P_{\boldsymbol{a}_t}(\boldsymbol{o}_t, \boldsymbol{o}_{t+1}) \left\| M_k(\boldsymbol{o}_{t+1}, \boldsymbol{a}'_t) \right. \\
&\quad \left. - M^*(\boldsymbol{o}_{t+1}, \boldsymbol{a}_{t+1}) \right\|_\infty \\
&\overset{(a)}{=} \gamma \left\| M_k(\boldsymbol{o}_t, \boldsymbol{a}_t) - M^*(\boldsymbol{o}_t, \boldsymbol{a}_t) \right\|_\infty \\
&= \gamma \left\| e_k(\boldsymbol{o}_t, \boldsymbol{a}_t) \right\|_\infty,
\end{aligned} \tag{30}$$

where (a) is derived in [35, Lemma 1]. Hence, condition 2) is satisfied. The variance of $F_k(\boldsymbol{o}_t, \boldsymbol{a}_t)$ is given by

$$\begin{aligned}
&\mathrm{Var}\left(F_k(\boldsymbol{o}_t, \boldsymbol{a}_t)\right) \\
&= \mathbb{E}\left[\left(F_k(\boldsymbol{o}_t, \boldsymbol{a}_t) - \mathbb{E}\left(F_k(\boldsymbol{o}_t, \boldsymbol{a}_t)\right) \right)^2 \right] \\
&= \mathbb{E}\left[\left(r(\boldsymbol{o}_t, \boldsymbol{a}_t) + M_k(\boldsymbol{o}_t, \boldsymbol{a}_t) - M^*(\boldsymbol{o}_t, \boldsymbol{a}_t) \right. \right. \\
&\quad \left. \left. - \left(r(\boldsymbol{o}_t, \boldsymbol{a}_t) + \mathbb{E}\left(M_k(\boldsymbol{o}_t, \boldsymbol{a}_t)\right) + M^*(\boldsymbol{o}_t, \boldsymbol{a}_t) \right) \right)^2 \right] \\
&= \mathbb{E}\left[\left(M_k(\boldsymbol{o}_t, \boldsymbol{a}_t) - \mathbb{E}\left(M_k(\boldsymbol{o}_t, \boldsymbol{a}_t)\right) \right)^2 \right] \\
&= \mathrm{Var}\left(r_t(\boldsymbol{o}_t, \boldsymbol{a}_t) + \gamma \max_{\boldsymbol{a}'_t} M_k(\boldsymbol{o}_{t+1}, \boldsymbol{a}'_t) \right) \\
&\leqslant C_{\mathrm{F}} \left(1 + \left\| e_k(\boldsymbol{o}_t, \boldsymbol{a}_t) \right\|_\infty^2 \right),
\end{aligned} \tag{31}$$

where (a) is satisfied since $r_t(\boldsymbol{o}_t, \boldsymbol{a}_t)$ and $\max_{\boldsymbol{a}'_t} M_k(\boldsymbol{o}_{t+1}, \boldsymbol{a}'_t)$ are both bounded. Hence, condition 3) is proved to be satisfied. This completes the proof.

B. Proof of Proposition 1

Based on (24), e_t can be rewritten as

$$\begin{aligned}
e_t &= \mathrm{tr}\left(\mathbb{E}\left[\boldsymbol{\partial l}_t \left(\boldsymbol{\partial l}_t \right)^T \right] \right) \\
&= \mathrm{tr}\left(\mathbb{E}\left[\left(\boldsymbol{M}^T \boldsymbol{M} \right)^{-1} \boldsymbol{M}^T \boldsymbol{\partial} \hat{\boldsymbol{d}}_t^{\mathrm{S}} \left(\left(\boldsymbol{M}^T \boldsymbol{M} \right)^{-1} \boldsymbol{M}^T \boldsymbol{\partial} \hat{\boldsymbol{d}}_t^{\mathrm{S}} \right)^T \right] \right) \\
&= \mathrm{tr}\left(\left(\boldsymbol{M}^T \boldsymbol{M} \right)^{-1} \boldsymbol{M}^T \mathbb{E}\left[\boldsymbol{\partial} \hat{\boldsymbol{d}}_t^{\mathrm{S}} \boldsymbol{\partial} \hat{\boldsymbol{d}}_t^{\mathrm{S}T} \right] \left(\left(\boldsymbol{M}^T \boldsymbol{M} \right)^{-1} \boldsymbol{M}^T \right)^T \right) \\
&= \mathrm{tr}\left(\left(\boldsymbol{M}^T \boldsymbol{M} \right)^{-1} \boldsymbol{M}^T \boldsymbol{J} \left(\left(\boldsymbol{M}^T \boldsymbol{M} \right)^{-1} \boldsymbol{M}^T \right)^T \right) \\
&= \mathrm{tr}\left(\left(\boldsymbol{M}^T \boldsymbol{M} \right)^{-1} \boldsymbol{M}^T \boldsymbol{J} \left(\boldsymbol{M}^T \right)^T \left(\left(\boldsymbol{M}^T \boldsymbol{M} \right)^{-1} \right)^T \right) \\
&\overset{(a)}{=} \mathrm{tr}\left(\left(\boldsymbol{M}^T \boldsymbol{M} \right)^{-1} \boldsymbol{M}^T \boldsymbol{J} \boldsymbol{M} \left(\boldsymbol{M}^T \boldsymbol{M} \right)^{-1} \right),
\end{aligned} \tag{32}$$

where equation (a) is obtained due to the fact that $\left(\boldsymbol{M}^T \right)^T = \boldsymbol{M}$ and $\left(\left(\boldsymbol{M}^T \boldsymbol{M} \right)^{-1} \right)^T = \left(\boldsymbol{M}^T \boldsymbol{M} \right)^{-1}$. $\boldsymbol{J}$ is the variance matrix of the selected distance subset, which is given by

$$\begin{aligned}
\boldsymbol{J} &= \mathbb{E}\left[\boldsymbol{\partial} \hat{\boldsymbol{d}}_t^{\mathrm{S}} \left(\boldsymbol{\partial} \hat{\boldsymbol{d}}_t^{\mathrm{S}} \right)^T \right] \\
&= \mathrm{tr}\left(\begin{bmatrix} \sigma_{m_1,t}^2 & 0 & 0 & 0 \\ 0 & \sigma_{m_2,t}^2 & 0 & 0 \\ 0 & 0 & \sigma_{m_3,t}^2 & 0 \\ 0 & 0 & 0 & \sigma_{m_4,t}^2 \end{bmatrix} \right)
\end{aligned}$$

with $\sigma_{m_i,t}^2, i = 1, \cdots, 4$ being the variance of the distance measurement $\hat{d}_{m_i,t}$. Based on [38], the variance $\sigma_{m_i,t}^2$ depends on the SINR of the signals transmitted from the active UAV and received by passive UAV m_i, then we have $\sigma_{m_i,t}^2 = \frac{k_i}{\epsilon^2 + j_t P^{\mathrm{J}} \left| h_{\mathrm{J},m_i,t} \right|^2}$ where k_i is the coefficient in [38]. This completes the proof.

REFERENCES

[1] 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, Mar. 2019.

[2] F. Wen, D. Ren, X. Zhang, G. Gui, B. Adebisi, H. Sari, and F. Adachi, "Fast localizing for anonymous uavs oriented toward polarized massive MIMO systems," *IEEE Internet of Things Journal*, vol. 10, no. 22, pp. 20 094–20 106, Nov. 2023.

[3] 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, 2018.

[4] J. Shen, A. F. Molisch, and J. Salmi, "Accurate passive location estimation using TOA measurements," *IEEE Transactions on Wireless Communications*, vol. 11, no. 6, pp. 2182–2192, 2012.

[5] Y.-E. Chen, H.-H. Liew, J.-C. Chao, and R.-B. Wu, "Decimeter-accuracy positioning for drones using two-stage trilateration in a GPS-denied environment," *IEEE Internet of Things Journal*, vol. 10, no. 9, pp. 8319–8326, May 2023.

[6] I. Guvenc and C.-C. Chong, "A survey on TOA based wireless localization and NLOS mitigation techniques," *IEEE Communications Surveys Tutorials*, vol. 11, no. 3, pp. 107–124, 2009.

[7] P. Yang, X. Cao, T. Q. Quek, and D. O. Wu, "Networking of internet of UAVs: Challenges and intelligent approaches," *IEEE Wireless Communications*, vol. 31, no. 1, pp. 156–163, Feb. 2024.

[8] W. Yi, Y. Liu, Y. Deng, and A. Nallanathan, "Clustered UAV networks with millimeter wave communications: A stochastic geometry view," *IEEE Transactions on Communications*, vol. 68, no. 7, pp. 4342–4357, July 2020.

[9] P.-Y. Hong, C.-Y. Li, H.-R. Chang, Y. Hsueh, and K. Wang, "WBF-PS: WiGig beam fingerprinting for UAV positioning system in GPS-denied environments," in *Proc. IEEE INFOCOM 2020 - IEEE Conference on Computer Communications*, pp. 1778–1787, Toronto, ON, Canada, Aug. 2020.

[10] 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.

[11] 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*, Aug. 2021.

[12] M. T. Dabiri, M. Rezaee, L. Mohammadi, F. Javaherian, V. Yazdanian, M. O. Hasna, and M. Uysal, "Modulating retroreflector based free space optical link for UAV-to-ground communications," *IEEE Transactions on Wireless Communications*, vol. 21, no. 10, pp. 8631–8645, Apr. 2022.

[13] B. R. Stojkoska, J. Palikrushev, K. Trivodaliev, and S. Kalajdziski, "Indoor localization of unmanned aerial vehicles based on RSSI," *IEEE EUROCON 2017 -17th International Conference on Smart Technologies*, pp. 120–125, Aug. 2017.

[14] F. Mason, M. Capuzzo, D. Magrin, F. Chiariotti, A. Zanella, and M. Zorzi, "Remote tracking of UAV swarms via 3D mobility models and LoRaWAN communications," *IEEE Transactions on Wireless Communications*, vol. 21, no. 5, pp. 2953–2968, Oct. 2022.

[15] A. N. Bishop, B. Fidan, B. D. Anderson, K. Dogancay, and P. N. Pathirana, "Optimality analysis of sensor-target geometries in passive localization: Part 1 - bearing-only localization," in *Proc. 2007 3rd International Conference on Intelligent Sensors, Sensor Networks and Information*, pp. 7–12, Melbourne, VIC, Australia, Apr. 2007.

[16] Y. Zhao, Z. Li, B. Hao, P. Wan, and L. Wang, "How to select the best sensors for TDOA and TDOA/AOA localization?" *China Communications*, vol. 16, no. 2, pp. 134–145, Feb. 2019.

[17] K. Gao, H. Wang, H. Lv, and P. Gao, "A DL-based high-precision positioning method in challenging urban scenarios for B5G CCUAVs," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 6, pp. 1670–1687, June 2023.

[18] R. Akter, M. Golam, V.-S. Doan, J.-M. Lee, and D.-S. Kim, "IoMT-Net: Blockchain-integrated unauthorized UAV localization using lightweight convolution neural network for internet of military things," *IEEE Internet of Things Journal*, vol. 10, no. 8, pp. 6634–6651, Apr. 2023.

[19] H. Luo, T. Chen, X. Li, S. Li, C. Zhang, G. Zhao, and X. Liu, "Keepedge: A knowledge distillation empowered edge intelligence framework for visual assisted positioning in UAV delivery," *IEEE Transactions on Mobile Computing*, vol. 22, no. 8, pp. 4729–4741, Aug. 2023.

[20] R. Chen, B. Yang, and W. Zhang, "Distributed and collaborative localization for swarming UAVs," *IEEE Internet of Things Journal*, vol. 8, no. 6, pp. 5062–5074, Mar. 2021.

[21] I. A. Meer, M. Ozger, and C. Cavdar, "On the localization of unmanned aerial vehicles with cellular networks," in in *Proc. of 2020 IEEE Wireless Communications and Networking Conference (WCNC)*, Seoul, Korea (South), May 2020, pp. 1–6.

[22] X. Liu, Y. Liu, and Y. Chen, "Machine learning empowered trajectory and passive beamforming design in UAV-RIS wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 7, pp. 2042–2055, July 2021.

[23] Y. Zhu, M. Chen, S. Wang, Y. Hu, Y. Liu, and C. Yin, "Collaborative reinforcement learning based unmanned aerial vehicle (UAV) trajectory design for 3D UAV tracking," *IEEE Transactions on Mobile Computing*, pp. 1–16, Mar. 2024.

[24] 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 Transactions on Communications*, vol. 67, no. 6, pp. 4281–4298, Feb. 2019.

[25] 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.

[26] 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.

[27] 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, Jan. 2018.

[28] Y. Zeng, J. Xu, and R. Zhang, "Energy minimization for wireless communication with rotary-wing UAV," *IEEE Transactions on Wireless Communications*, vol. 18, no. 4, pp. 2329–2345, Apr. 2019.

[29] M. Chen, W. Saad, and C. Yin, "Liquid state machine learning for resource and cache management in LTE-U unmanned aerial vehicle (UAV) networks," *IEEE Transactions on Wireless Communications*, vol. 18, no. 3, pp. 1504–1517, Jan. 2019.

[30] C.-W. Fu, M.-L. Ku, Y.-J. Chen, and T. Q. S. Quek, "UAV trajectory, user association, and power control for multi-UAV-enabled energy-harvesting communications: Offline design and online reinforcement learning," *IEEE Internet of Things Journal*, vol. 11, no. 6, pp. 9781–9800, Mar. 2024.

[31] M. Chen, Z. Yang, W. Saad, C. Yin, H. V. Poor, and S. Cui, "A joint learning and communications framework for federated learning over wireless networks," *IEEE Transactions on Wireless Communications*, vol. 20, no. 1, pp. 269–283, Jan. 2021.

[32] 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.

[33] P. Sunehag, G. Lever, A. Gruslys, W. M. Czarnecki, V. Zambaldi, M. Jaderberg, M. Lanctot, N. Sonnerat, J. Z. Leibo, and K. Tuyls, "Value-decomposition networks for cooperative multi-agent learning," *https://arxiv.org/abs/1706.05296*, June 2017.

[34] 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.

[35] 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.

[36] 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.

[37] 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*, vol. 139, pp. 9945–9954, Dec. 2021.

[38] 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.