# Eye in the Sky: Energy Efficient Model-Based Reinforcement Learning Aided Target Tracking Using UAVs

Yi Xia, *Member, IEEE*, Zekai Zhang, Jingzehua Xu, Pengfei Ren, Jingjing Wang, *Senior Member, IEEE*, and Zhu Han, *Fellow, IEEE*

***Abstract*—The rapid response and high energy efficiency of the unmanned aerial vehicle (UAV) are crucial prerequisites for enabling time-sensitive and long-endurance target tracking missions, such as search and rescue, area reconnaissance, and convoy monitoring. However, existing research in target tracking primarily focuses on enhancing tracking accuracy, which struggles to adapt to tasks considering strict time constraints and energy consumption. To address these issues, this paper introduces a model-based reinforcement learning tracking strategy (MRLTS) for the UAV to minimize control costs and achieve user-specified tracking performance, including a two-stage design. In the first stage, a steady-state robust tracking controller is developed based on available model knowledge that forces the UAV to asymptotically approximate a predefined observation path in spite of uncertainties. In the second stage, an intelligent component based on the soft actor-critic (SAC) algorithm is customized to empower the UAV to strike a trade-off between prescribed tracking performance and energy consumption, wherein a skilled barrier function is constructed to interpret specified time constraints. The proposed paradigm can provide a higher sampling efficiency than SAC-based strategy. Simulation results demonstrate that our strategy outperforms benchmarks and results in a 46.3% cost-effectiveness improvement at least.**



This research was in part supported by the National Natural Science Foundation of China under Grant No. 62222101, in part by the Aeronautical Science Foundation of China (ASFC) under Grant No. 2022Z071051013, in part by the Beijing Natural Science Foundation under Grant No. L232043 and No. L222039, and in part by the Fundamental Research Funds for the Central Universities, and in part by the NSF CNS-2107216, CNS-2128368, CMMI-2222810, ECCS-2302469, US Department of Transportation, Toyota, and in part by the Amazon and Japan Science and Technology Agency (JST) Adopting Sustainable Partnerships for Innovative Research Ecosystem (ASPIRE) JPMJAP2326. *(Corresponding author: Jingjing Wang.)*

Y. Xia is with the School of Instrument and Electronics, North University of China, Taiyuan 030051, and also with the State Key Laboratory of Integrated Services Networks (Xidian University), Xi'an, 710071, China. E-mail: xiayi26@hotmail.com.

Z. Zhang and J. Xu are with the Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, 518055, China. E-mail: {zhangzej21, xjzh23}@mails.tsinghua.edu.cn.

P. Ren is with the School of Cyber Science and Technology, Beihang University, Beijing 100191, China. Email: pfren@buaa.edu.cn.

J. Wang is with the School of Cyber Science and Technology, Beihang University, Beijing 100191, China, and also with the State Key Laboratory of Integrated Services Networks (Xidian University), Xi'an, 710071, China. Email: drwangjj@buaa.edu.cn.

Z. Han is with the Department of Electrical and Computer Engineering at the University of Houston, Houston, TX 77004 USA, and also with the Department of Computer Science and Engineering, Kyung Hee University, Seoul, South Korea, 446-701. Email: hanzhu22@gmail.com.

***Index Terms*—UAV, deep reinforcement learning, mobile target tracking, time constraints, energy consumption.**

## I. Introduction

RELYING on low manufacturing cost, convenient deployment and high maneuverability of unmanned aerial vehicles (UAVs), the UAV assistance paradigm, such as disaster rescue [1], data gathering [2], [3], and surveillance [4], [5], has spurred intensive interests. In these applications, robust tracking of ground mobile targets is a critical technology to fulfill the requirement for multi-directional information collection and efficient monitoring [6]. This technology attempts to steer UAVs to travel along a user-defined path centered at a ground target steadily, enabling a periodical and continuous observation of the interested object.

Recently, circular-guided target tracking strategies [7]–[15] have shown robust performance in tracking mobile targets, wherein UAVs are commanded to orbit the target at a constant radius, ensuring the reliability of tracking. A parameterized elliptical target tracking protocol [16], [17] is further developed to save energy while improving observation efficiency. It is worth declaring that reported schemes can merely guarantee relative distance disagreement to asymptotically decay toward the specific residual sets, resulting in a prolonged settling time inevitably. When confronted with tasks subject to stringent time constraints, existing results [7]–[17] are inadvisable.

Due to the need for rapid emergency response, the UAV needs to reach the desired position within a short time. Prescribed performance control (PPC) [18]–[21] offers an elegant solution to meet this requirement. It aims to transform inequality constraints into unconstrained ones by constructing an error transformation function, assuring strict confinement of the relative distance within prescribed regions. However, current PPC-based controllers typically achieve superior performance at the expense of excessive power consumption, which inevitably posts a challenge for long-endurance missions.

Characterized by powerful optimization ability, deep reinforcement learning (DRL) has attracted significant attention. It aims to approximate the optimal control policy iteratively by interacting with environment repeatedly, greatly overcoming the reliance on complete modeling knowledge in uncertain missions. Although DRL-driven target tracking methods have been successfully advocated in [22]–[24], they still face two serious drawbacks: *sluggish convergence and weak constraints*

TABLE I
COMPARISONS OF MRLTS AND OTHER ALGORITHMS

| Schemes | | Design concerns | | | Advantages | Disadvantages |
|---|---|---|---|---|---|---|
| | | Energy saving | Prescribed performance constraint | Anti-disturbance capability | | |
| Asymptotic-time controllers | [7]–[15] | ✗ | ✗ | ✗ | Conceptual intuitive; Easy for realization | Slow convergence rate; High energy usage; Inability to meet time constraints |
| PPC-based controllers | [18]–[21] | ✗ | ✓ | ✗ | Easy for realization; Designated convergence rate | High energy usage; Lack of optimization ability |
| Data-driven based schemes | [24], [25] | ✓ | ✗ | ✗ | Superior optimization ability; Eliminate the dependence on modeling information | Failure in handling time constraints |
| MRLTS | | ✓ | ✓ | ✓ | Designated convergence time; Optimality in tracking accuracy and energy usage; High sampling efficiency | Poor migration ability |

*handling ability*. On the one hand, due to the lack of high-quality model information to assist training, low sampling efficiency is always involved in current DRL. On the other hand, few efforts have been made along the line of DRL-related target tracking with designated tracking performance. These factors constitute the primary motivation to tailor the prevailing DRL to implement faster convergence and better compliance with constraints.

Enlightened by the previous analysis, this article proposes a nontrivial model-based reinforcement learning tracking strategy (MRLTS) for the UAV with optimal cost-effectiveness, which consists of a steady-state robust control item and an approximate optimal learning component. The salient merits of MRLTS are that under available modeling knowledge, a robust tracking controller is introduced into the DRL-based paradigm to generate confidential experienced datasets for DRL training, such that unnecessary explorations and redundant "trail-and-error" can be reduced effectively while achieving a better generalization ability. The learning component takes responsibility for addressing time constraints and minimizing control costs through a data-driven approach. The primary contributions are summarized as follows:

1) We propose a novel MRLTS to attain an optimal trade-off between tracking performance and energy consumption. Unlike non-DRL target tracking methods [7]–[21] suffering from exponential decaying rates, herein the proposed method can specify convergence time and confine error dynamics within predefined regions. Furthermore, in contrast to PPC-based designs [18]–[21] that prioritize accurate target tracking at the expense of excessive energy consumption, our strategy can promote an effective coordination between these two aspects.
2) Different from pure DRL results [22]–[24] confined to sluggish convergence property, a nontrivial model-based DRL paradigm is developed to greatly improve sampling efficiency and decrease the cost of sample collection, wherein an analytical control item is incorporated to stabilize error dynamics. Specifically, in order to recover uncertainties timely, a concise filtering named unknown system dynamics estimator (USDE) is devised in steady-state robust item. Moreover, a skilled barrier function that interprets specified-time restrictions is embedded in reward functions to be maximized, bringing significant state constraints handling ability into DRL framework.
3) We conduct comprehensive simulations and compare our algorithm with benchmark algorithms. Simulation results verify that our strategy achieves a 40.84% higher energy-efficient and reduces cost by 43.25% compared to PPC-based algorithm [21], respectively.

The rest of this paper is organized as follows. Related work is stated in Section II, while system model is provided in Section III. MRLTS is implemented in Section IV and simulation results are shown in Section V. Section VI concludes the paper.

*Notations:* The following notations are used throughout this paper. $|\cdot|$ denotes the absolute value. $\|\cdot\|$ represents the 2-norm. $[\cdot]^T$ denotes transpose of a matrix. $\mathbb{R}, \mathbb{R}^+, \mathbb{Z}$ denote the set of real numbers, positive real numbers and integer numbers respectively. $\mathbb{R}^n, \mathbb{R}^{n\times m}$ denote $n$-dimensional vector and $n\times m$ matrix, respectively.

## II. RELATED WORK

### A. Analytic-based Scheme

An asymptotic target-enclosing controller incorporating a consensus-based target observer was developed in [8] for UAVs with limited perceptual abilities to realize target enclosing by constructing a group of leader-follower interactive potentials. Relying on range and range rate measurements, Jia *et al.* [26] considered a distributed coordinate-free control scheme that can enforce UAVs to reach desired objective circle and then tightly surround the cooperative target at a preassigned distance. In particular, considering multi-targets distributed in a strip shape case, Chun *et al.* [16] devised an elliptical encirclement control protocol to enclose targets by inferring the non-orthogonal relationships between the axial and tangential unit vectors in terms of bearing angles.

Some researches focusing on prescribed performance can be found in numerous fields. Koksal *et al.* [18] employed

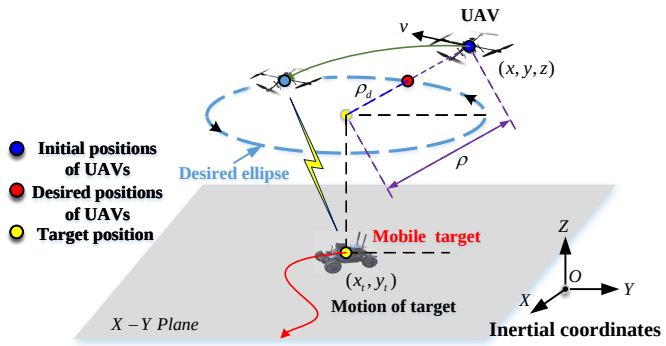


Fig. 1. Illustration of UAV-to-ground mobile target tracking scenario.

an adaptive control with prescribed performance limitations for UAVs to characterize preselected convergence rate and overshoot based on a strict-feedback form. Liu *et al.* [19] utilized PPC in nonlinear multi-agent systems to guarantee consensus errors evolving within predefined regions. Under the constraint of input saturation, a user-defined convergence rate of the line-of-sight angle can be ensured via resorting to a PPC-based adaptive sliding mode guidance law [20]. Zhang *et al.* [21] exploited an appointed-time performance function to construct the enclosing algorithm for UAVs, rendering that convergence speed of error profiles can be effortlessly governed by users.

### B. Data-driven Based Scheme

In recent years, more attentions have been gradually paid on seeking model-free solutions. Ma *et al.* [22] investigated a DRL technique to achieve a target capturing formation pattern with collision-free behaviors. To maximize average spectrum efficiency and improve convergence rate, Wu *et al.* [24] studied a federated multi-agent deep deterministic policy gradient algorithm to jointly optimize trajectories of multiple vehicles for an air-ground coordinated communications system. Messaoudi *et al.* [25] devised a multi-agent deep Q-network-based scheme for UAV data collection assisted by a mobile charging station, aiming to minimize the age of information and reduce energy consumption through trajectory optimization of the UAV.

A brief comparison between current control schemes and our proposed method is summarized in Table I.

## III. System Model And Problem Formulation

In this paper, we consider a quadrotor UAV steadily pursuing a ground cooperative target along a scheduled elliptical orbit at a fixed height, which can provide a reliable operational pattern, as shown in Fig. 1. To achieve efficient connectivity between UAVs and the target in cooperative tasks, we assume that the UAV can continuously receive the target's location, i.e., $\mathbf{p}_t = [x_t, y_t]^T \in \mathbb{R}^2$ using Global Positioning Systems and on-board antennas. In this section, we will provide the system model and formulate the problem of optimal tracking control for the UAV.

### A. System Modeling

Typically, the motion behaviors of the UAV are described at translational and rotational layers. Since rotational dynamics has a much smaller time constant than translational dynamics. Here, we assume that there exists a mature autopilot to follow translational references. Inspired by [27], a planar UAV motion can be formulated in the inertial frame as

$$\begin{cases} m\ddot{x} = U(\cos\varphi\cos\psi\sin\theta + \sin\varphi\sin\psi) + f_x + \Delta_x, \\ m\ddot{y} = U(\cos\varphi\sin\psi\sin\theta - \sin\varphi\cos\psi) + f_y + \Delta_y, \\ m\ddot{z} = U(\cos\varphi\cos\theta) - mg + f_z + \Delta_z, \end{cases} \tag{1}$$

where $\mathbf{p} = [x, y, z]^T \in \mathbb{R}^3$ determines the inertial position of the UAV, while $m, U, \varphi, \psi$ and $\theta$ stand for the mass, total driving force generated by motors, roll angle, yaw angle, and pitch angle of the UAV, respectively. Moreover, $f_x, f_y$ and $f_z$ represents unknown nonlinear damping function, while $\Delta_x, \Delta_y$, and $\Delta_z$ denote the ambient perturbations. $g$ is the gravitational acceleration. The main concentration of this paper is to devise a proper target tracking solution for the UAV. Thus, to facilitate controller design, the motion of a target is simplified as

$$\begin{cases} \dot{x}_t = v_{xt}, \\ \dot{y}_t = v_{yt}, \end{cases} \tag{2}$$

with $v_{xt}, v_{yt}$ being the target velocity.

Subsequently, define $u_x = U(\cos\varphi\cos\psi\sin\theta + \sin\varphi\sin\psi)/m$, $u_y = U(\cos\varphi\sin\psi\sin\theta - \sin\varphi\cos\psi)/m$, $u_z = U(\cos\varphi\cos\theta)/m - g$ as the control inputs. To promote the implement, model (1) is rewritten as

$$\begin{cases} \dot{\mathbf{p}} = \mathbf{v}, \\ \dot{\mathbf{v}} = \mathbf{u} + (\mathbf{f} + \mathbf{G})/m, \end{cases} \tag{3}$$

where $\mathbf{f} = \left[f_x, f_y, f_z\right]^T \in \mathbb{R}^3$, $\mathbf{G} = \left[\Delta_x, \Delta_y, \Delta_z\right]^T \in \mathbb{R}^3$, $\mathbf{u} = \left[u_x, u_y, u_z\right]^T \in \mathbb{R}^3$ and $\mathbf{v} = \left[v_x, v_y, v_z\right]^T \in \mathbb{R}^3$ is the linear velocity vector. Let $T_k$ be the unknown damping parameter with $k = x, y, z$, and thus $f_k$ can be expressed as $f_k = T_k v_k$.

### B. Problem Formulation

This paper aims to design a MRLTS for the UAV with constrained performances despite lumped disturbances $\mathbf{d}_m = (\mathbf{f} + \mathbf{G})/m = \left[D_x, D_y, D_z\right]^T$, which can be decomposed into the following objectives.

*1) Steady-State Tracking Objective:* Based on a simple filtering, lumped perturbations including damping uncertainties in terms of system states and exogenous disturbances caused by wind can be online counteracted, such that the robustness of systems can be improved, enabling that

$$\lim_{t\to\infty} \left\|\mathbf{d}_m - \hat{\mathbf{d}}_m\right\| \le \varepsilon_m, \tag{4}$$

where $\hat{\mathbf{d}}_m \in \mathbb{R}^3$ is the estimate value of $\mathbf{d}_m$, while $\varepsilon_m \in \mathbb{R}^+$ represents a sufficiently small positive real number.

*2) Optimized Tracking Objective:* Endow the UAV with a trade-off between performance and cost while conforming performance restrictions by MRLTS.

- *Reinforced performance assignment:* The relative range error $e_\rho = \rho - \rho_d$ and height error $e_z = z - z_d$ satisfy

$$-\underline{\lambda}_\rho S_\rho < e_\rho < \bar{\lambda}_\rho S_\rho, -\underline{\lambda}_z S_z < e_z < \bar{\lambda}_z S_z, \tag{5}$$

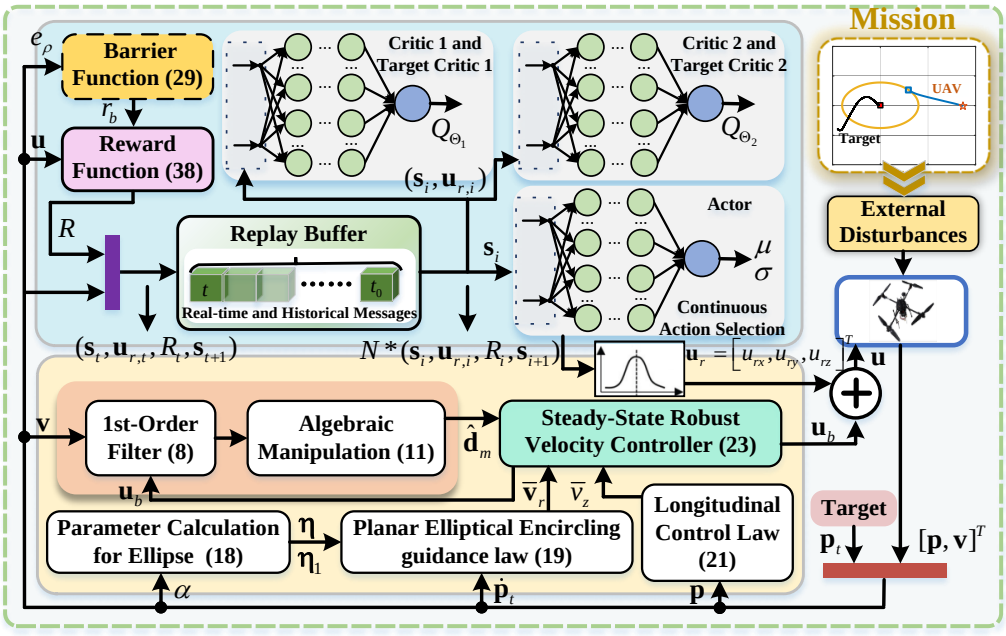


Fig. 2. Block diagram of model-based reinforcement learning ground target tracking control framework.

where $\rho = \sqrt{(x-x_t)^2+(y-y_t)^2}$, and $\rho_d$ is desired time-varying encircling radius. Moreover, $z_d$ defines the desired altitude, while $\underline{\lambda}_\rho, \bar{\lambda}_\rho, \underline{\lambda}_z, \bar{\lambda}_z \in \mathbb{R}^+$ are positive constants. $S_\rho, S_z$ denote two user-defined behavior functions, which are expounded in Section IV.

- *Optimizing assignment:* The control deviation $\mathbf{e} = [e_\rho, e_z]^T$ and energy consumption characterized by the control input $\mathbf{u}$ require to be minimized for the optimal control efficiency via iterative learning. The optimization problem can be formulated as

$$\min_{\mathbf{e},\mathbf{u}} \int_0^{T_f} \mathbf{e}^T \mathbf{Q}_{\rho,z} \mathbf{e} + \mathbf{u}^T \mathbf{P} \mathbf{u} dt, \forall t \in [0, T_f], \tag{6}$$

where $\mathbf{Q}_{\rho,z} > 0$ and $\mathbf{P} \in \mathbb{R}^{3\times3}$ are the weight matrices associated with errors and consumption, respectively. $T_f \in \mathbb{R}^+$ defines the terminal time of the task.

**Remark 1:** *By resorting to (3) and Newton's second law of motion, it is easy to obtain that the control input* $\mathbf{u}$ *means the thrust of the UAV, where thrust is powered by the on-board battery. The greater the thrust, the more energy of the UAV is consumed. Therefore, minimizing the* $\mathbf{u}^T\mathbf{P}\mathbf{u}$ *is equivalent to minimizing the energy consumption of the UAV.*

**Assumption 1:** *The derivative of lumped perturbations* $\dot{\mathbf{d}}_m$ *is bounded, and there is a positive number* $\bar{d}_m$*, fulfilling* $\|\dot{\mathbf{d}}_m\| \leq \bar{d}_m$.

**Remark 2:** *Assumption 1 has been recognized as a standard and sufficient condition widely applying in the prevailing disturbance estimators [28]–[30]. It is worth noting that the boundedness of continuous wind signals can be readily inferred, as wind field energy is typically constrained and can be represented by the superposition of sine and cosine waves with varying amplitudes, frequencies, and phases, according to the Fourier series theorem. Even sudden changes in wind, formulated as step signals, can be adequately characterized using high-order Taylor expansion polynomials to capture dramatic variations at an acceptable level. Consequently, Assumption 1 is reasonable and not too strong from engineering practices.*

## IV. Model-Based Reinforcement Learning Tracking Strategy Design

In this section, a model-based reinforcement learning elliptical tracking controller is investigated for the UAV enslaved to uncertainties and designated constraints. As observed from Fig. 2, it comprises a steady-state robust control item $\mathbf{u}_b$ and an approximate optimal component $\mathbf{u}_r$, i.e,

$$\mathbf{u} = \mathbf{u}_b + \mathbf{u}_r, \tag{7}$$

where $\mathbf{u}_b$ serves as a tracking error stabilization term that recovers nominal performance under the premise of lumped perturbations, while $\mathbf{u}_r$ is devised to strengthen encircling manners and deal with the contradiction among appointed time limits and control costs via learning.

### A. Steady-state Robust Control Policy

*1) Disturbance Observer Design:* In order to counteract uncertainties consisting of wind disturbances $\mathbf{G}$ and damping uncertainties $\mathbf{f}$, we construct a USDE [29] with a simple structure in the kinetic loop by imposing a series of filtering manipulations upon signals $\mathbf{v}$ and $\mathbf{u}_b$ yielding

$$\begin{cases} \kappa \dot{\mathbf{v}}_f + \mathbf{v}_f = \mathbf{v}, \mathbf{v}_f(0) = [0,0,0]^T, \\ \kappa \dot{\mathbf{u}}_{bf} + \mathbf{u}_{bf} = \mathbf{u}_b, \mathbf{u}_{bf}(0) = [0,0,0]^T, \end{cases} \tag{8}$$

where $\kappa \in \mathbb{R}^+$ is a filtering variable. According to an invariant manifold, a mathematical formulation is presented to link filtered signals $\mathbf{v}_f$, $\mathbf{u}_{bf}$ with disturbance $\mathbf{d}_m$.

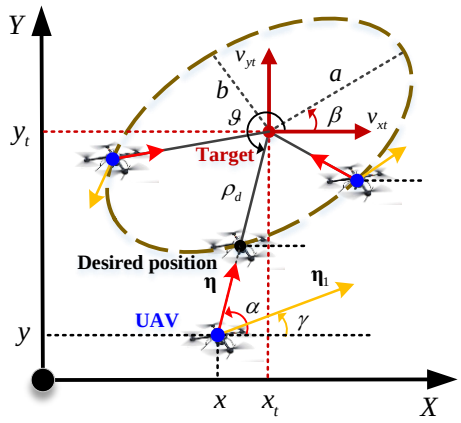


Fig. 3. Geometrical profile of elliptical tracking in the plane.

**Lemma 1:** *By invoking (3) and (8), an auxiliary variable is determined to explicitly describe a quantitative relationship between filtered signals* $\mathbf{v}_f$, $\mathbf{u}_{bf}$ *and disturbance* $\mathbf{d}_m$ *as*

$$\varepsilon = \left(\mathbf{v} - \mathbf{v}_f\right)/\kappa - \left(\mathbf{u}_{bf} + \mathbf{d}_m\right), \tag{9}$$

*where* $\varepsilon$ *is restricted and exponentially converge toward sufficiently small region around zero for* $\kappa \in \mathbb{R}^+$. *It further gets*

$$\lim_{\kappa\to 0}\left[\lim_{t\to\infty}\left(\left(\mathbf{v} - \mathbf{v}_f\right)/\kappa - \left(\mathbf{u}_{bf} + \mathbf{d}_m\right)\right)\right] = 0, \tag{10}$$

*inferring that* $\left(\mathbf{v} - \mathbf{v}_f\right)/\kappa - \left(\mathbf{u}_{bf} + \mathbf{d}_m\right) = 0$ *obeys an invariant manifold as* $\kappa \to 0$.

*Establishing a directed mapping property among filtered signals* $\mathbf{v}_f$, $\mathbf{u}_{bf}$ *and disturbance* $\mathbf{d}_m$, *in line with (10), a continuous-time USDE is devised to enhance the robustness of the system as follows*

$$\hat{\mathbf{d}}_m = -\mathbf{u}_{bf} + \left(\mathbf{v} - \mathbf{v}_f\right)/\kappa, \tag{11}$$

*with* $\hat{\mathbf{d}}_m$ *being the estimation of* $\mathbf{d}_m$.

*2) Robust Elliptical Tracking Policy:* To propel the UAV to circle around the a planned path centered at target, a parameterized elliptical trajectory, as illustrated in Fig. 3, relevant to a long semi-axis $a$, a short semi-axis $b$, and a counter clock wise rotation angle $\beta$ can be expounded as

$$\frac{(x\cos\beta + y\sin\beta)^2}{a^2} + \frac{(x\sin\beta - y\cos\beta)^2}{b^2} = 1. \tag{12}$$

Subsequently, define any point attached to the desired ellipse as $[\rho_d\cos\vartheta, \rho_d\sin\vartheta]^T$ in the polar framework corresponding to target $\mathbf{p}_t$, wherein $\vartheta$ is a polar angle from horizontal axis to the specific point along the ellipse. By using (12), one has

$$\rho_d(\vartheta) = \frac{ab}{\sqrt{a^2\sin^2(\vartheta-\beta) + b^2\cos^2(\vartheta-\beta)}}. \tag{13}$$

Let $\alpha$ be the bearing angle from the UAV to the target, calculated by

$$\alpha = \arctan\left(y_t - y, x_t - x\right), \tag{14}$$

with $\arctan(\cdot)$ being the arctangent function. Based on Fig. 3, geometrical relationship in terms of $\vartheta$ and $\alpha$ is written as

$$\theta - \alpha = \pi + 2i\pi, i \in \mathbb{Z}. \tag{15}$$

Substituting (15) into (14), one has

$$\rho_d(\alpha) = \frac{ab}{\sqrt{a^2\sin^2(\alpha-\beta) + b^2\cos^2(\alpha-\beta)}}. \tag{16}$$

Note that when the UAV is steered to the predetermined ellipse trajectory, i.e., $\rho = \rho_d$, the velocity direction of the UAV will be consistent with the unit tangent speed vector $\boldsymbol{\eta}_1 = [\cos\gamma, \sin\gamma]^T$ with $\gamma$ being the angle between the $x$-axis direction and the tangent direction of ellipse. Then, combining (15) and (16), the tangent of $\gamma$ yields

$$\begin{aligned}|\tan\gamma| &= \left|\frac{d\rho_d(\vartheta)\sin\vartheta}{d\vartheta}\frac{d\vartheta}{d\rho_d(\vartheta)\cos\vartheta}\right| \\ &= \left|\frac{a^2\sin(\alpha-\beta)\sin\beta - b^2\cos(\alpha-\beta)\cos\beta}{a^2\sin(\alpha-\beta)\cos\beta + b^2\cos(\alpha-\beta)\sin\beta}\right|.\end{aligned} \tag{17}$$

The following two equations implying two elements of the tangent unit vector $\boldsymbol{\eta}_1$ can be deduced from (17) as

$$\begin{cases}\cos\gamma = \dfrac{a^2\sin(\alpha-\beta)\cos\beta + b^2\cos(\alpha-\beta)\sin\beta}{\sqrt{a^4\sin^2(\alpha-\beta) + b^4\cos^2(\alpha-\beta)}}, \\ \sin\gamma = \dfrac{a^2\sin(\alpha-\beta)\sin\beta - b^2\cos(\alpha-\beta)\cos\beta}{\sqrt{a^4\sin^2(\alpha-\beta) + b^4\cos^2(\alpha-\beta)}}.\end{cases} \tag{18}$$

Since this paper considers a cooperative target for tracking, the states of the target are available to the UAV. To ensure that the UAV can pursue and enclose the mobile target along a prescribed ellipse, the velocity of the target is incorporated into a planar elliptical tracking guidance law. This allows the target to be followed regardless of its motion, as described below:

$$\bar{\mathbf{v}}_r = k_r\left(\rho - \rho_d\right)\boldsymbol{\eta} + v_d\boldsymbol{\eta}_1 + \dot{\mathbf{p}}_t, \tag{19}$$

where $\bar{\mathbf{v}}_r = \left[\bar{v}_{rx}, \bar{v}_{ry}\right]^T$ is a planar velocity command and $k_r$ is a nonnegative gain to be selected. Relative distance error $e_\rho = \rho - \rho_d$ in view of planar plane is introduced to drive the UAV ultimately approximating the desired elliptical orbit. $v_d \in \mathbb{R}^+$ is the value of the tangent velocity. Moreover, $\boldsymbol{\eta} = [\cos\alpha, \sin\alpha]^T$ is a unit radial vector pointing from the UAV to the target, aiming to sustain a stable surrounding pattern.

With regard to a pregiven flight height $z_d$, we attempt to establish a longitudinal control law to accurately stabilize the height tracking error $e_z = z - z_d$. By recalling (3), the differential of $e_z$ against time can be manufactured by

$$\dot{e}_z = v_z - \dot{z}_d. \tag{20}$$

The following virtual control law is raised to obtain the anticipated height changing rate

$$\bar{v}_z = -k_z e_z + \dot{z}_d, \tag{21}$$

where $k_z \in \mathbb{R}^+$ is a regulating coefficient. Next, we compound planar velocity command $\bar{\mathbf{v}}_r$ and vertical velocity instructions

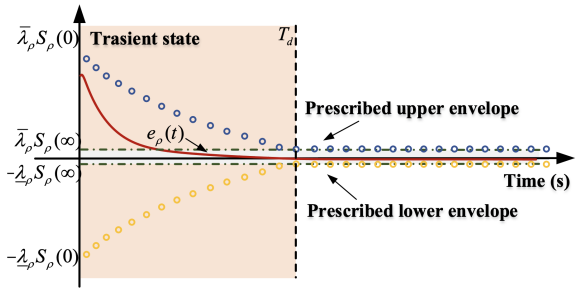


Fig. 4. Illustration of appointed-time performance envelope.

as an overall expected speed vector $\bar{\mathbf{v}}_e = \left[\bar{v}_{rx}, \bar{v}_{ry}, \bar{v}_z\right]^T$. Denote velocity tracking error as $\mathbf{e}_v = \mathbf{v} - \bar{\mathbf{v}}_e = \left[e_{vx}, e_{vy}, e_{vz}\right]^T$, and in line with (3), the time derivative of $\mathbf{e}_v$ is derived as

$$\dot{\mathbf{e}}_v = \mathbf{u}_b + \mathbf{d}_m - \dot{\bar{\mathbf{v}}}_e. \tag{22}$$

Based on the estimations generated by USDE, a steady-state elliptical tracking controller with anti-perturbation resistance is compiled as

$$\mathbf{u}_b = -k_u \mathbf{e}_v - \hat{\mathbf{d}}_m + \dot{\bar{\mathbf{v}}}_e, \tag{23}$$

where $\mathbf{u}_b = \left[u_{bx}, u_{by}, u_{bz}\right]^T$, while $\hat{\mathbf{d}}_m$ is a robust term that actively rejects the lumped disturbances. $k_u \in \mathbb{R}^+$ is a proportionality scalar to be assigned.

**Theorem 1:** *Given a UAV guided by a prescribed ellipse around a ground mobile target in (2), resorting to the steady-state robust target tracking control policy in (23), if Assumption 1 holds, the distance error $e_\rho, e_z$, velocity error $\mathbf{e}_v$, and observation error $\tilde{\mathbf{d}}_m = \mathbf{d}_m - \hat{\mathbf{d}}_m$ remain to be input-to-state stable (ISS).*

**Proof 1:** *See Appendix A.*

### *B. Approximate Optimal Compensator*

Soft actor-critic (SAC) is an off-policy DRL algorithm with a framework of actor-critic initially proposed by [31] containing one actor with parameters $\phi$, two critic and target critic networks with parameters $\Theta_1, \Theta_2$ and $\bar{\Theta}_1, \bar{\Theta}_2$. In order to overcome the hyperparameter sensitive problem existing in deep deterministic policy gradient (DDPG) [32], an entropy term is added in SAC to provide better learning robustness and sampling efficiency. In this part, the approximate optimal compensator based on SAC is constructed for the UAV.

*1) Appointed-time Constraints Design:* To ensure controlling errors evolve within user-designated envelopes, we firstly impose the following inequality:

$$-\underline{\lambda}_\rho S_\rho < e_\rho < \bar{\lambda}_\rho S_\rho, \tag{24}$$

where $\underline{\lambda}_\rho, \bar{\lambda}_\rho \in (0, 1]$ are the predefined parameters to be designed. $S_\rho$ is a behavior function to prescribe settling time and steady-state accuracy, as depicted in Fig. 4, which can be stated as

$$S_\rho(t) = \begin{cases} \left[\left(T_d - t\right)/T_d\right]^{1/\left(1-\xi_\rho\right)} \left(S_{\rho 0} - S_{\rho\infty}\right) + S_{\rho\infty}, & 0 \le t < T_d, \\ S_{\rho\infty}, & t \ge T_d, \end{cases} \tag{25}$$

where $T_d$ is the appointed convergence time. $\xi_\rho \in (0, 1)$ is applied to adjust the decaying rate of $S_\rho$. $S_{\rho\infty}$ represents the maximum permissible steady-state bound of $e_\rho$. $S_{\rho 0}$ is the initial allowable range, meeting $-\underline{\lambda}_\rho S_{\rho 0} < e_\rho(0) < \bar{\lambda}_\rho S_{\rho 0}$.

Inspired by the principle of PPC [19], [20], a transformed error function $F(z(t))$ is introduced to encode a constrained signal into an unrestrained one as

$$e_\rho(t) = S_\rho(t) F(z(t)), \tag{26}$$

with

$$F(z(t)) = \left(\bar{\lambda}_\rho e^{z(t)} - \underline{\lambda}_\rho e^{-z(t)}\right) / \left(e^{z(t)} + e^{-z(t)}\right). \tag{27}$$

By inversing the converted function (27), the transformed error $z(t)$ is calculated as

$$\begin{aligned} z(t) &= F^{-1}\left[e_\rho / S_\rho\right] \\ &= \frac{1}{2} \ln\left[\left(\underline{\lambda}_\rho + e_\rho/S_\rho\right) / \left(\bar{\lambda}_\rho - e_\rho/S_\rho\right)\right]. \end{aligned} \tag{28}$$

To prevent penalty tending to infinity and eliminate control singularity, a bounded barrier function $r_b$ incorporated with an saturation item is tailored according to (28), expressed by

$$r_b = -\frac{1}{2} \ln\left[\left(\underline{\lambda}_\rho + \mathrm{sat}\left(e_\rho/S_\rho\right)\right) / \left(\bar{\lambda}_\rho - \mathrm{sat}\left(e_\rho/S_\rho\right)\right)\right], \tag{29}$$

where the saturation function $\mathrm{sat}(\cdot)$ is denoted by

$$\mathrm{sat}(x) = \begin{cases} \bar{\lambda}_\rho - c_1, & x \ge \bar{\lambda}_\rho - c_1, \\ x, & -\underline{\lambda}_\rho + c_1 < x < \bar{\lambda}_\rho - c_1, \\ -\underline{\lambda}_\rho + c_1, & x \le -\underline{\lambda}_\rho + c_1, \end{cases} \tag{30}$$

with $c_1 \in \left(0, \min\left\{\bar{\lambda}_\rho, \underline{\lambda}_\rho, \bar{\lambda}_z, \underline{\lambda}_z\right\}\right)$, in which $\underline{\lambda}_z, \bar{\lambda}_z \in (0, 1]$. Similarly, we devise the barrier function for the height error $e_z$ as

$$r_z = -\frac{1}{2} \ln\left[\left(\underline{\lambda}_z + \mathrm{sat}\left(e_z/S_z\right)\right) / \left(\bar{\lambda}_z - \mathrm{sat}\left(e_z/S_z\right)\right)\right]. \tag{31}$$

where subscript $z$ corresponds to the height error $e_z$ and subscript $\rho$ represents the radial error $e_\rho$.

*2) Markov Decision Process Modeling:* To facilitate the implementation of DRL, interactions between agents and their environment can be described by a mathematical model called Markov decision process (MDP), commonly expressed by a five-tuple: $\langle \mathcal{S}, \mathcal{U}, \mathcal{P}, \mathcal{R}, \varsigma \rangle$, wherein $\mathcal{S}$, $\mathcal{U}$, $\mathcal{P}$ and $\mathcal{R}$ mean the state space, the action space, the state transition probability and the reward function, respectively. Additionally, $\varsigma \in (0, 1]$ is a discount factor.

The state vector $\mathbf{s}_t \in \mathcal{S}$ is defined at sampling time $t$:

$$\mathbf{s}_t = \left[v_{xp,t}, v_{yp,t}, e_{z,t}, \dot{e}_{z,t}, e_{\rho,t}, \dot{e}_{\rho,t}, -r_b, -r_z, u_{x,t}, u_{y,t}, u_{z,t}\right]^T, \tag{32}$$

where $v_{xp,t}$ and $v_{yp,t}$ are the practical radial and the tangent velocity, which can be combined as

$$\begin{bmatrix} v_{xp,t} \\ v_{yp,t} \end{bmatrix} = \begin{bmatrix} -\cos\vartheta_t & -\sin\vartheta_t \\ \cos\gamma_t & \sin\gamma_t \end{bmatrix} \begin{bmatrix} v_{xt,t} \\ v_{yt,t} \end{bmatrix}. \tag{33}$$

*3) Action Space:* In MDP, the action is updated at each time step $t$ based on the feedback of the environment, including the most recent action and current states. On the basis of the kinematics of the UAV, velocity commands of the $x, y$, and $z$

TABLE II
COMPARISONS OF DIFFERENT REWARD FUNCTIONS

| Literature | Design concerns of reward fucntion | | | |
|---|---|---|---|---|
| | Time constraints | Enhanced accuracy | Energy consumption | Prescribed performance |
| [22], [33] | ✗ | ✓ | ✗ | ✗ |
| [24], [25] | ✗ | ✓ | ✓ | ✗ |
| Our paper | ✓ | ✓ | ✓ | ✓ |

axis at the time step $t$ are chosen as the action vector $\mathbf{u}_{r,t} \in \mathcal{U}$, where $\mathbf{u}_{r,t}$ is a three-dimensional action space constituted by three continuous velocity commands, which is represented by

$$\mathbf{u}_{r,t} = \left[u_{rx,t}, u_{ry,t}, u_{rz,t}\right]^T. \tag{34}$$

Typically, the measured velocity of the UAV is always limited. Therefore, the selected actions are bounded, which satisfy $u_{rx,\min} \leqslant u_{rx} \leqslant u_{rx,\max}$, $u_{ry,\min} \leqslant u_{ry} \leqslant u_{ry,\max}$, $u_{rz,\min} \leqslant u_{rz} \leqslant u_{rz,\max}$, where $u_{rx,\min}$, $u_{rx,\max}$, $u_{ry,\min}$, $u_{ry,\max}$, $u_{rz,\min}$, $u_{rz,\max} \in \mathbb{R}$. Then, the Markov decision process for target tracking procedure is formulated as

$$\mathbf{s}_{t+1} \sim \mathcal{P}\left(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{u}_{r,t}\right), \tag{35}$$

with $\mathcal{P}\left(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{u}_{r,t}\right)$ being the transition probability from state $\mathbf{s}_t$ to $\mathbf{s}_{t+1}$ after selecting action $\mathbf{u}_{r,t}$.

*4) Reward Function Design:* Firstly, to confine relative distance error and height error without violating prescribed envelopes, we introduce the barrier functions into the reward function. Secondly, to reduce the energy consumption, we take control input and tangential speed offset $\tilde{v}_{yp} = v_{yp} - v_d$ into account. The reward function can be defined in a quadratic form as

$$R_1 = -\boldsymbol{\tau}^T \mathbf{Q}_\tau \boldsymbol{\tau} - \mathbf{u}^T \mathbf{Q}_u \mathbf{u}, \tag{36}$$

where $\boldsymbol{\tau} = \left[r_b, \tilde{v}_{yp}, r_z\right]^T$ and $\mathbf{Q}_\tau, \mathbf{Q}_u \in \mathbb{R}^{3\times3}$ are the error dynamics vector and weight matrices, respectively.

In particularly, when three controlling deviations remain at a high accuracy domain, a special sparse reward $R_2$ is involved to encourage agents to sustain such performance, as

$$R_2 = \begin{cases} r_\rho, & \text{if } |r_b| < \rho_{cir}, \\ r_v, & \text{if } \left|\tilde{v}_{yp}\right| < v_{cir}, \\ r_z, & \text{if } |r_z| < z_{cir}, \\ 0, & \text{otherwise}, \end{cases} \tag{37}$$

with $r_\rho, r_v, r_z$ being the positive reward value to be assigned.

To sum up, overall reward function defined to be

$$R = R_1 + R_2. \tag{38}$$

**Remark 3:** *Compared with existing reward functions [22]–[25], [33], [34] failing to address time constraints and prescribe tracking performance, a bounded barrier function that interprets time requirements is devised in this paper by converting the constrained error into an unconstrained one. This enables the DRL to effectively make a trade-off between performance and energy consumption. A brief comparison of different reward functions is presented in Table II.*

*5) Reinforcement Learning Based on SAC:* The objective of SAC is to learn a policy to maximize the expected return and the entropy of the policy utilizing

$$\pi^* = \arg\max_\pi \sum_t \mathbb{E}_{(\mathbf{s}_t, \mathbf{u}_{r,t})}\left[\varsigma^t \left(R_t + \delta \mathcal{H}\left(\pi\left(\cdot \mid \mathbf{s}_t\right)\right)\right)\right], \tag{39}$$

where $\pi$ is the agent's policy and $\pi^*$ is the agent's optimal policy. $\mathbb{E}$ specifies the expected value. $\delta$ denotes a temperature parameter. $\mathcal{H}\left(\pi\left(\cdot \mid \mathbf{s}_t\right)\right) = -\log \pi\left(\cdot \mid \mathbf{s}_t\right)$ means the entropy of the policy $\pi$ under the state $\mathbf{s}_t$.

Learning process of SAC can repeatedly execute policy evaluation and policy improvement. In the policy evaluation, the Q-function $Q_\Theta\left(\mathbf{s}_t, \mathbf{u}_{r,t}\right)$ is computed by applying a modified Bellman backup operator $\mathcal{T}^\pi$ [31], which can be given by

$$\begin{aligned} Q_\Theta\left(\mathbf{s}_{t+1}, \mathbf{u}_{r,t+1}\right) =& \mathcal{T}^\pi Q_\Theta\left(\mathbf{s}_t, \mathbf{u}_{r,t}\right) \\ =& R_t + \varsigma \mathbb{E}_{\mathbf{s}_{t+1}}\left[V\left(\mathbf{s}_{t+1}\right)\right], \end{aligned} \tag{40}$$

with

$$V\left(\mathbf{s}_t\right) = \mathbb{E}_\pi\left[Q_\Theta\left(\mathbf{s}_t, \mathbf{u}_{r,t}\right) - \delta \log \pi\left(\mathbf{u}_{r,t} \mid \mathbf{s}_t\right)\right], \tag{41}$$

being the state value function. In the policy improvement, policy is updated by

$$\pi_{new} = \arg\min_{\pi' \in \Pi} \mathcal{D}_{KL}\left(\pi'\left(\cdot \mid \mathbf{s}_t\right) \| \frac{\exp\left(Q^{\pi_{old}}\left(\mathbf{s}_t, \cdot\right)/\delta\right)}{Z^{\pi_{old}}\left(\mathbf{s}_t\right)}\right), \tag{42}$$

with $\mathcal{D}_{KL}, \Pi, \pi_{old}, Q^{\pi_{old}}, Z^{\pi_{old}}$ being the Kullback-Leibler (KL) divergence, a policy set, the policy from the last update, the Q-value of $\pi_{old}$, and the partition function, respectively.

**Lemma 2:** *Maximizing the objective function corresponds to minimizing the KL divergence between the policy distribution and Q-function distribution. The new policy in (42) satisfies* $Q^{\pi_{old}} \leq Q^{\pi_{new}}$.

**Proof 2:** *Please refer to [35].*

The network parameters $\Theta$ of $Q_\theta\left(\mathbf{s}_t, \mathbf{u}_{r,t}\right)$ are trained to minimize the following Bellman residual.

$$\begin{aligned} &J_Q(\Theta) = \\ &\mathbb{E}_{\left(\mathbf{s}_t, \mathbf{u}_{r,t}\right) \sim \mathcal{D}}\left[\frac{1}{2}\left(Q_\Theta\left(\mathbf{s}_t, \mathbf{u}_{r,t}\right) - \left(R_t + \varsigma \mathbb{E}_{\mathbf{s}_{t+1}}\left[V_{\bar{\Theta}}\left(\mathbf{s}_{t+1}\right)\right]\right)\right)^2\right], \end{aligned} \tag{43}$$

where $\mathcal{D}$ is the replay buffer, and $\bar{\Theta}$ is the network parameters of the target Q-function $Q_{\bar{\Theta}}\left(\mathbf{s}_t, \mathbf{u}_{r,t}\right)$. Similarly, the policy $\pi_\phi\left(\mathbf{u}_{r,t} \mid \mathbf{s}_t\right)$ that outputs the mean value $\mu$ and standard deviation $\sigma$ of a Gaussian distribution is learned by minimizing the KL divergence

$$J_\pi(\phi) = \mathbb{E}_{\left(\mathbf{s}_t, \mathbf{u}_{r,t}\right) \sim \mathcal{D}}\left[\delta \log \pi_\phi\left(\mathbf{u}_{r,t} \mid \mathbf{s}_t\right) - Q_\Theta\left(\mathbf{s}_t, \mathbf{u}_{r,t}\right)\right]. \tag{44}$$

**Remark 4:** *In order to facilitate the implementation of the MRLTS, a summary of tuning regulations and sensitivity analyses for fundamental parameters is provided as follows:*

1) *In view of steady-state robust controller,* $k_r$, $k_z$ *and* $k_u$ *are proportionality coefficients that the larger value of them will promote radial error and vertical error to converge to a smaller residual set.*
2) *For the USDE, We can find that the magnitude of observation error* $\tilde{\mathbf{d}}_m$ *corresponds to the filtering parameter*

**Algorithm 1** MRLTS Algorithm

Initialize the training environment, including the environmental information, target trajectory, as well as states of the UAV.
Assign values to the target network parameters $\bar{\Theta}_1 \leftarrow \Theta_1, \bar{\Theta}_2 \leftarrow \Theta_2, \mathcal{D} \leftarrow \varnothing, \mathcal{D}_0 \leftarrow \varnothing$.
Attain data set $\mathcal{D}_0$ by running $\mathbf{u}_b$ with $\mathbf{u}_r = 0$.
Train initial critic parameters $\Theta_1^0, \Theta_2^0$ using $\mathcal{D}_0$ according to (43).
Initialize the replay memory $\mathcal{D} \leftarrow \mathcal{D}_0$.
Assign values to critic network $\Theta_1 \leftarrow \Theta_1^0, \Theta_2 \leftarrow \Theta_2^0$ and their targets $\bar{\Theta}_1 \leftarrow \Theta_1^0, \bar{\Theta}_2 \leftarrow \Theta_2^0$.
**for** each episode **do**
  Select the UAV state from the given range.
  **for** each time step **do**
    Choose an action $\mathbf{u}_{r,t}$ in terms of $\mathbf{s}_t$.
    Obtain the overall control input $\mathbf{u}_t$ by combing $\mathbf{u}_{r,t}$ with $\mathbf{u}_b$.
    Calculate reward function $R_t$ by (38) and collect the next state $\mathbf{S}_{t+1}$ by (32) from environment.
    Store sampling tuple $\left(\mathbf{s}_t, \mathbf{u}_{r,t}, R_t, \mathbf{s}_{t+1}\right)$ into $\mathcal{D}$.
    Extract $N$ batches of historical data from $\mathcal{D}$.
    $\Theta_j \leftarrow \Theta_j - \lambda_Q \nabla_\theta J_Q\left(\Theta_j\right), j = 1, 2.$
    $\phi \leftarrow \phi - \lambda_\pi \nabla_\phi J_\pi(\phi).$
    $\delta \leftarrow \delta - \lambda_\delta \nabla_\delta J_\delta(\delta).$
    $\bar{\Theta}_j \leftarrow \kappa_\Theta \Theta_j + (1 - \kappa_\Theta)\, \bar{\Theta}_j, j = 1, 2.$
  **end for**
**end for**

*$\kappa$. Specifically, a minor $\kappa$ can contribute to an accurate estimation of lumped disturbances and an increased level of tracking accuracy.*

3) *Regarding the barrier function, the arguments $\underline{\lambda}_\rho$, $\bar{\lambda}_\rho$, $\underline{\lambda}_z$, $\bar{\lambda}_z$, $S_{\rho 0}$, and $S_{z0}$ should be chosen correctly to satisfy $-\underline{\lambda}_\rho S_{\rho 0} < e_\rho(0) < \bar{\lambda}_\rho S_{\rho 0}$ and $-\underline{\lambda}_z S_{z0} < e_z(0) < \bar{\lambda}_z S_{z0}$ An arbitrary arriving time can be theoretically realized by selecting the parameter $T_d$ $\xi_\rho$ and $\xi_z$ determines the decaying rate of radial and vertical errors. Specifically, larger $\xi_\rho$ and $\xi_z$ will lead to a faster convergence rate of involved errors.*
4) *For the reward function, $\mathbf{Q}_\tau$ and $\mathbf{Q}_u$ are weight matrices, adjusting the balance between performance and energy consumption metrics. A larger magnitude of $\mathbf{Q}_\tau$ indicates a higher priority given to tracking precision. $r_\rho$, $r_v$, and $r_z$ should be selected as a constant with minor magnitudes. Alternatively, an excessively large value typically leads to a sluggish reward convergence.*

## V. Implementation of MRLTS Algorithm

The flowchart of proposed MRLTS algorithm is displayed in Fig. 2, where a unique control input comprising of a steady-state robust tracking policy $\mathbf{u}_b$ and an approximate optimal compensator $\mathbf{u}_r$ is compounded to run the overall system. Specifically, for a given target tracking mission with time restrictions, the target motion, initial states of the UAV and environmental information should be determined for establishing a training environment. The entire offline training process is executed repeatedly, including a series of learning episodes and massive time steps. At each time step $t$, we collect experience samples via constantly interacting with the environment. Those historical data including the current state $\mathbf{s}_t$, the action $\mathbf{u}_{r,t}$, the reward $R_t$, and the state from the next time step $\mathbf{s}_{t+1}$ will be stored as a tuple $\left(\mathbf{s}_t, \mathbf{u}_{r,t}, R_t, \mathbf{s}_{t+1}\right)$ in a replay memory $\mathcal{D}$. At each policy evaluation and improvement step, we stochastically extract $N$ batches of historical data from the replay memory $\mathcal{D}$ to learn neural network parameters $\Theta$ and $\phi$. In initialization stage, a benchmark controller $\mathbf{u}_b$ is applied to generate starting data samples $\mathcal{D}_0$ for fitting an initial Q-value functions. As such, when learning is executed in the training stage, a good starting point with high-quality experiences can be provided for the agent, as shown in Algorithm 1, where $\lambda_Q, \lambda_\pi, \lambda_\alpha > 0$ are learning rates, while $\kappa_\Theta > 0$ is a soft updating constant. $\nabla(\cdot)$ is a gradient operation. When the initialization is over, both $\mathbf{u}_b$ and the latest updated policy $\pi_\phi\left(\mathbf{u}_{r,t} \mid \mathbf{s}_t\right)$ is exploited to run the UAV system again.

**Remark 5:** *It is worth emphasizing that the size of the sample sets has an impact on the convergence of the reward function. Specifically, a small sample size for DRL training will make it difficult to stabilize the reward function and will require more time for training. Conversely, a large sample size will lead to excessive memory occupation, which is intractable for scheduling limited onboard resources. Additionally, the simulation step is another crucial factor in evaluating the performance of MRLTS. A smaller time step can enhance the accuracy of target tracking and improve optimization capabilities. However, it unavoidably increases the computational burden, resulting in higher computational complexity. Therefore, it is important to select an appropriate sample size and time step to achieve a balance among various requirements.*

## VI. Simulation Results

In this section, to substantiate the viability and superiority of the involved MRLTS algorithm, four cases are executed via MATLAB/SIMULINK platform with a simulation time step configured by $0.02s$.

### A. Simulation Setup

Given a UAV pursuing a nonstationary ground target with a motion trajectory $\mathbf{p}_t = [0.4t, 0.2t + 6\sin(\pi t/50)]^T(m)$. The desired altitude $z_d$ of the UAV is set to be $10(m)$. Corresponding controller arguments are set in Table III.

In the initialization, the starting state of the UAV is assigned as $x(0) = 20(m), y(0) = 0(m)$ and $z(0) = 2(m)$. At the training stage, we repeat the training processes for 1,500 times, i.e., 1,500 episodes. For each episode, the tracking mission is performed for $100s$. Moreover, DRL configurations of MRLTS can be found in Table IV.

To straightforward prove the advantages of presented algorithms, various controllers are provided as below.

*1) Classical Elliptical Tracking Controller [16] (abbreviated as CETC).* It propels integrator agents to accomplish an

TABLE III
PARAMETERS FOR THE MRLTS

| Modules | | Values |
|---|---|---|
| Steady-state robust control policy | USDE | $\kappa = 0.125$ |
| | Elliptical circling action | $a = 8, b = 4, \beta = 0,$ $k_r = 0.7, v_d = 4,$ $k_z = 1.2, k_u = 20$ |
| Approximate optimal compensator | Barrier function | $\underline{\lambda}_\rho, \bar{\lambda}_\rho, \underline{\lambda}_z, \bar{\lambda}_z = 1,$ $S_{\rho 0} = 20, S_{z0} = 15, T_d = 3,$ $S_{\rho\infty}, S_{z\infty} = 0.5, \xi_\rho = 0.6,$ $\xi_z = 0.8, c_1 = 0.0001$ |
| | Reward function | $\mathbf{Q}_u = diag[0.05, 0.05, 0.05],$ $\mathbf{Q}_\tau = diag[1, 1, 1],$ $\rho_{cir}, v_{cir}, z_{cir} = 0.1,$ $r_v = 0.25, r_\rho, r_z = 0.75$ |

TABLE IV
DEEP REINFORCEMENT LEARNING CONFIGURATIONS

| Parameters | Values |
|---|---|
| Learning rate $\lambda_Q$ | 0.001 |
| Learning rate $\lambda_\pi$ | 0.0001 |
| Learning rate $\lambda_\delta$ | 0.0001 |
| Soft updating rate $\kappa_\Theta$ | 0.01 |
| Replay memory capacity | $1 \times 10^5$ |
| Sample batch size $N$ | 256 |
| Discount factor $\varsigma$ | 0.99 |
| Time steps per episode | 5000 |
| Training episodes | 1500 |
| Score averaging window length | 10 |
| Range of the action $[u_{rx,\min}, u_{rx,\max}]$ | $[-5, 5]$ $(m/s)$ |
| Range of the action $[u_{ry,\min}, u_{ry,\max}]$ | $[-5, 5]$ $(m/s)$ |
| Range of the action $[u_{rz,\min}, u_{rz,\max}]$ | $[-1, 1]$ $(m/s)$ |
| Dimensions of observations | 11 |

elliptical escorting mission without considering uncertainties and time restrictions, which is described as

$$\begin{cases} \bar{\mathbf{v}}_r = k_1 e_\rho \boldsymbol{\eta} + k_2 v_d \boldsymbol{\eta}_1, \\ \mathbf{u} = -k_u \mathbf{e}_v + \dot{\bar{\mathbf{v}}}_e, \end{cases} \tag{45}$$

where $k_1$ is a nonnegative coefficient and $k_2$ denotes a parameter to be designed. The related arguments are set as $k_1 = 0.7, k_2 = 1$.

*2) Soft Actor-Critic Based Elliptical Tracking Controller [31] (abbreviated as SAC-ETC).* It should be emphasized that DRL-based tracking results with time constraints are very rare. To reveal the merits of proposed learning rules, the prevailing SAC is tailored to address specified constraints, wherein we adopt approximate optimal compensator $\mathbf{u}_r$ as the target tracking strategy.

*3) PPC-based Elliptical Tracking Controller [21] (abbreviated as PPC-ETC).* By converting constrained error dynamics into an unconstrained one. The concrete expression of PPC-ETC follows

$$\begin{cases} \bar{\mathbf{v}}_r = k_3 z_\rho \boldsymbol{\eta} / \beta_\rho - e_\rho \dot{S}_\rho \boldsymbol{\eta} / S_\rho + v_d \boldsymbol{\eta}_1 + \dot{\mathbf{p}}_t, \\ \bar{v}_z = -k_4 z_z + \dot{z}_d + e_z \dot{S}_z / S_z, \\ \mathbf{u} = -k_u \mathbf{e}_v + \dot{\bar{\mathbf{v}}}_e - \hat{\mathbf{d}}_m, \end{cases} \tag{46}$$

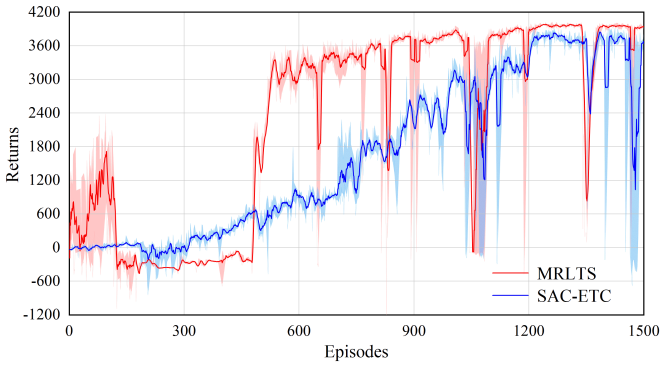


Fig. 5. Comparisons of learning curve using MRLTS and SAC-ETC.

where $\beta_\rho = \left[1/\left(\underline{\lambda}_\rho + e_\rho/S_\rho\right) + 1/\left(\bar{\lambda}_\rho - e_\rho/S_\rho\right)\right]/2S_\rho, k_3 = 0.7$, $k_4 = 1.2$. $z_i = 0.5 \ln\left(\left(\underline{\lambda}_i + e_i/S_i\right)/\left(\bar{\lambda}_i - e_i/S_i\right)\right), i = \rho, z.$

## B. Comparative Study

We compare the MRLTS with the SAC-ETC to demonstrate the improvement of sampling efficiency and reward convergence. Fig. 5 illustrates the learning curves of the MRLTS and SAC-ETC. It is clearly inspected that reward functions of both algorithms can converge to the optimal value as episodes augment. However, MRLTS results in a larger return and a faster convergence rate in comparison to the SAC-ETC, which is primarily profiting from the utilization of model-based controller to evade unnecessary and time-consuming explorations from mistakes.

To elucidate MRLTS's capability in addressing time constraints, comparisons are conducted. Figs. 6 and 7 depict the evaluation outcomes between CETC, SAC-ETC, and MRLTS, which comprise 3-D tracking curves, planar tracking orbit, transient profiles under relative coordinates, and response of errors. It can be observed that all involved strategies can drive the UAV to arrive at the desired ellipse while maintaining an anticipated tracking pattern around a mobile target. Unfortunately, the CETC fails to handle performance constraints and lumped disturbances. Moreover, although SAC-ETC can drive the UAV to realize the appointed time requirements, due to the lack of the assistance of modeling information, SAC-ETC results in an inferior transient tracking performance and a worse steady-state accuracy than MRLTS.

In contrast, benefiting from the barrier function that interprets inequality constraints, the MRLTS can confine the relative range error and the height error strictly within user-defined performance regions. The UAV can be driven to arrive the desired ellipse before designated time and retain a stable encirclement around the target. Subsequently, to validate the effectiveness of the USDE, various lumped perturbations $\mathbf{d}_m$ including time-varying wind disturbances $\mathbf{G}$ and damping uncertainties $\mathbf{f}$ are imposed on the UAV, which are presented in (47), (48), and (49). It can be observed from Fig. 8 that uncertainties can be estimated promptly by USDE and the

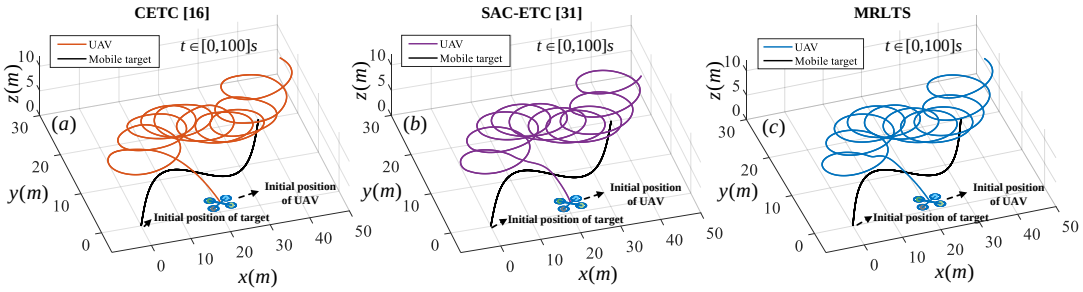


Fig. 6. 3D curves of tracking a ground mobile target. (a) CETC. (b) SAC-ETC. (c) MRLTS.

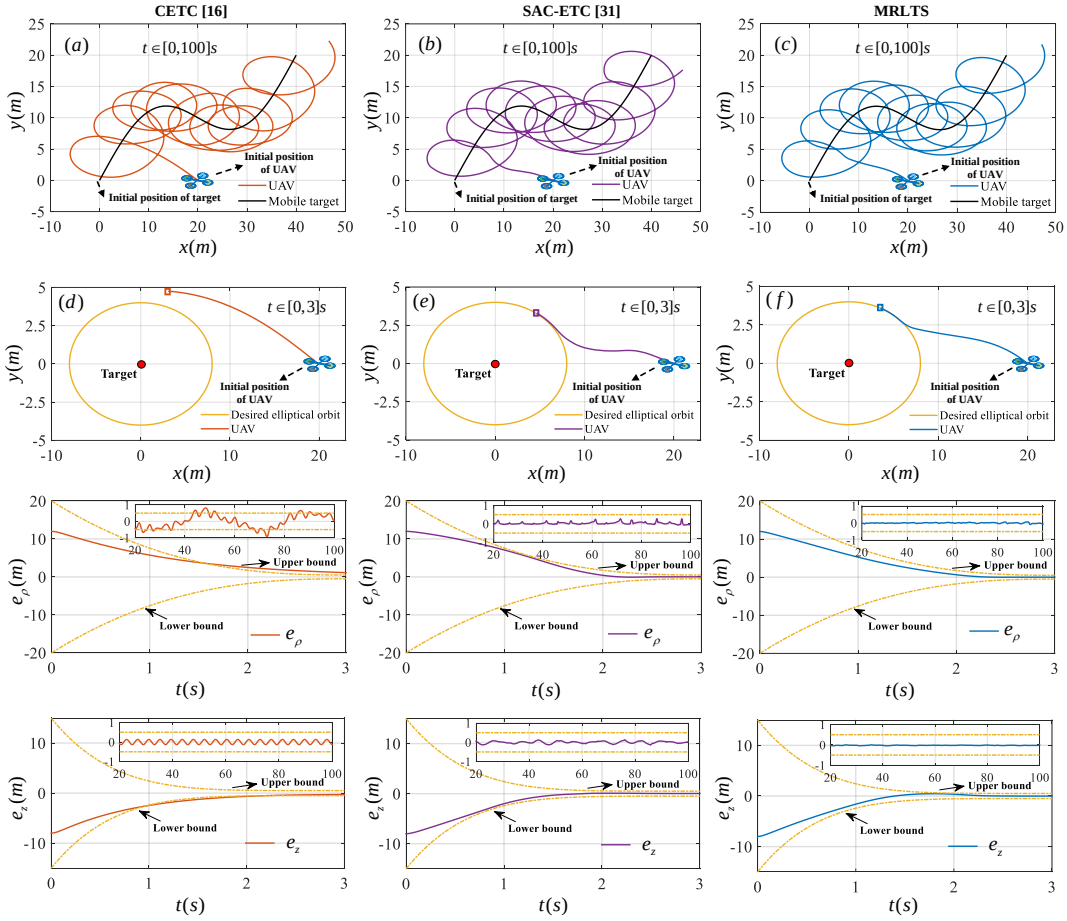


Fig. 7. Planar tracking path, transient profiles of relative positions as well as relative range errors and height errors under CETC, SAC-ETC, and MRLTS.

robustness of the system can be ensured.

$$\text{Disturbance 1:}\begin{cases}\mathbf{f}=\left[0.008v_x,0.008v_y,0.008v_z\right]^T,\\ \mathbf{G}=[5(\sin t+\sin 1.2t-\cos 0.8t),\\ \quad 5(\cos 1.2t+\sin 0.5t-\cos 0.8t),\\ \quad 5\sin 0.6t]^T.\end{cases}\tag{47}$$

$$\text{Disturbance 2:}\begin{cases}\mathbf{f}=\left[0.012v_x,0.012v_y,0.012v_z\right]^T,\\ \mathbf{G}=[10(\sin t+\sin 0.5t-\cos 0.8t),\\ \quad 10(\cos t+\sin 0.5t-\cos 0.8t),\\ \quad 10\sin 1.5t]^T.\end{cases}\tag{48}$$

$$\text{Disturbance 3:}\begin{cases}\mathbf{f}=\left[0.02v_x,0.02v_y,0.02v_z\right]^T,\\ \mathbf{G}=[15(\sin 1.2t+\sin 0.8t-\cos t),\\ \quad 15(\cos 1.2t+\sin 0.5t-\cos t),\\ \quad 15\sin 1.8t]^T.\end{cases}\tag{49}$$

### *C. Effectiveness Confirmation of MRLTS by Setting Various Profiles*

In this section, to certify the effectiveness of the MRLTS in addressing different performance constraints. We conduct numerous simulations considering various parameters of profiles, and related parameters are listed in Table V. As elaborated

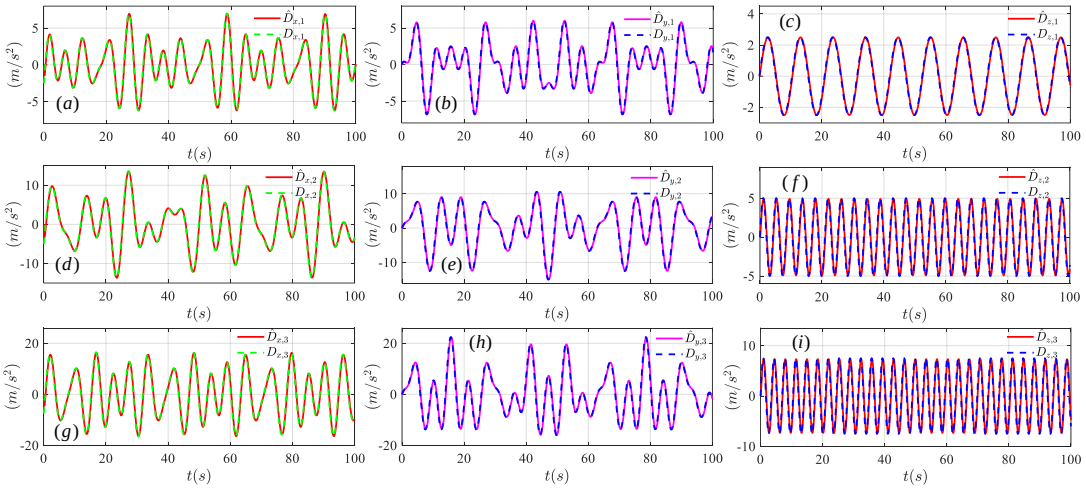

Fig. 8. Estimations of lumped disturbances by USDE. (a) $x$ direction of $\mathbf{d}_{m1}$. (b) $y$ direction of $\mathbf{d}_{m1}$. (c) $z$ direction of $\mathbf{d}_{m1}$. (d) $x$ direction of $\mathbf{d}_{m2}$. (e) $y$ direction of $\mathbf{d}_{m2}$. (f) $z$ direction of $\mathbf{d}_{m2}$. (g) $x$ direction of $\mathbf{d}_{m3}$. (h) $y$ direction of $\mathbf{d}_{m3}$. (i) $z$ direction of $\mathbf{d}_{m3}$.

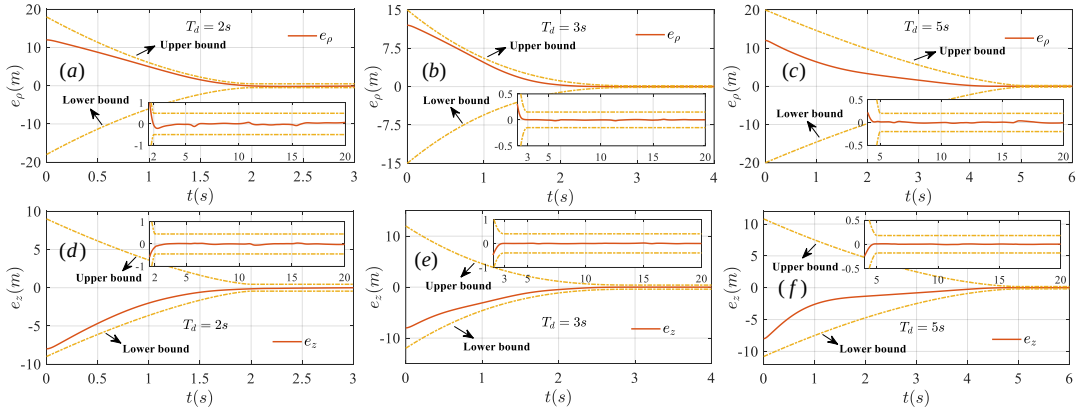

Fig. 9. Evolutions of controlling errors under different profiles. (a) Profile 1. (b) Profile 2. (c) Profile 3. (d) Profile 4. (e) Profile 5. (f) Profile 6.

in Fig. 9, the convergence performances of radial error $e_\rho$ and vertical error $e_z$, including arriving time, steady-state precision, and decaying rate can be readily regulated by setting corresponding arguments. It further demonstrates the superior ability of MRLTS in handling appointed-time constraints and implementing designated performance requirements. Moreover, from Table V and Fig. 9, we can see that the convergence time of errors are no more than prescribed arriving time, while error profiles can be confined within designed boundaries, implying that the MRLTS can ensure the stability and maintain the robustness of the system.

TABLE V
DIFFERENT PROFILES SETTING

| | Design arguments | Convergence time of the error |
|---|---|---|
| Profile 1 | $\underline{\lambda}_\rho = 1, \bar{\lambda}_\rho = 1, S_{\rho 0} = 18,$ $S_{\rho\infty} = 0.5, T_d = 2, \xi_\rho = 0.4.$ | $1.98s$ |
| Profile 2 | $\underline{\lambda}_\rho = 0.5, \bar{\lambda}_\rho = 0.5, S_{\rho 0} = 30,$ $S_{\rho\infty} = 0.3, T_d = 3, \xi_\rho = 0.6.$ | $2.52s$ |
| Profile 3 | $\underline{\lambda}_\rho = 1, \bar{\lambda}_\rho = 1, S_{\rho 0} = 20,$ $S_{\rho\infty} = 0.2, T_d = 5, \xi_\rho = 0.3.$ | $4.6s$ |
| Profile 4 | $\underline{\lambda}_z = 0.9, \bar{\lambda}_z = 0.9, S_{z0} = 10,$ $S_{z\infty} = 0.5, T_d = 2, \xi_z = 0.3.$ | $2s$ |
| Profile 5 | $\underline{\lambda}_z = 0.8, \bar{\lambda}_z = 0.8, S_{z0} = 15,$ $S_{z\infty} = 0.5, T_d = 3, \xi_z = 0.6.$ | $2.72s$ |
| Profile 6 | $\underline{\lambda}_z = 0.6, \bar{\lambda}_z = 0.6, S_{z0} = 18,$ $S_{z\infty} = 0.3, T_d = 5, \xi_z = 0.4.$ | $4.84s$ |

### *D. Performance Analysis*

To explicitly declare the superiority of MRLTS in reconciling contradictions between performance and energy consumption, different strategies are executed.

Above all, we construct a cost function, to assess the comprehensive performance of various controllers, which can be defined as $C_{\cos t}(\mathbf{e},\mathbf{u}) = \int_0^{T_f} \mathbf{e}^T\mathbf{Q}_{\rho,z}\mathbf{e} + \mathbf{u}^T\mathbf{P}\mathbf{u}dt, \forall t \in [0, T_f]$. Here, $T_f$ is the terminal time, $\mathbf{Q}_{\rho,z} = diag[1,1]$, and $\mathbf{P} = diag[0.05, 0.05, 0.05]$. Fig. 10 illustrates outcomes of the PPC-ETC method. Compared with MRLTS, although PPC-ETC can effectively guide the UAV to complete time-sensitive tracking tasks with high accuracy, due to the lack of learning capability, it results in unnecessary maneuvering

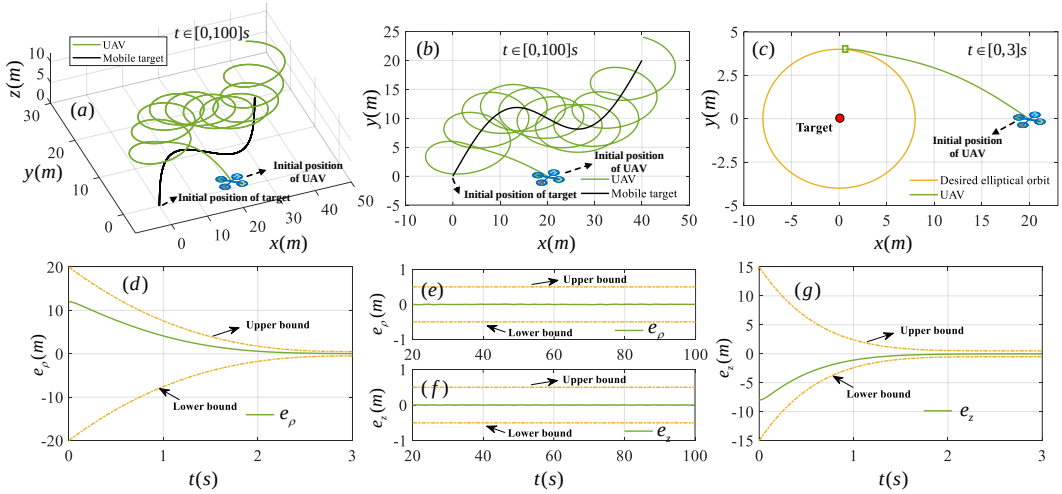


Fig. 10. Evolution of UAV's elliptical tracking a ground mobile target. (a) 3-D curve of tracking during $t \in [0, 100]s$. (b) Planar motion trajectory of the UAV during $t \in [0, 100]s$. (c) Transient profile under relative coordinates during $t \in [0, 3]s$. (d) Evolution of transient relative range error. (e) Evolution of steady-state relative range error. (f) Evolution of steady-state relative height error. (g) Evolution of transient relative height error.

TABLE VI
QUANTITATIVE INDICATORS COMPARISONS

| Index \ Controller | CETC [16] | PPC-ETC [21] | SAC-ETC [31] | **MRLTS** |
|---|---|---|---|---|
| RMSE | 1.0859 | 0.8633 | 1.0874 | 0.9509 |
| STD | 1.0666 | 0.8573 | 1.0748 | 0.9434 |
| Energy cost | 27083 | 8501 | 14495 | 5029 |
| Cost values | 84274 | 9714 | 15485 | 5216 |
| Optimization ability | No | No | Yes | Yes |
| Constraint handling ability | No | Yes | Yes | Yes |
| Computational complexity [a] | 1.02% | 1.56% | 9.43% | 9.38% |

$a$. Computational complexity is defined by the ratio of accumulated operational time to the entire simulation time.

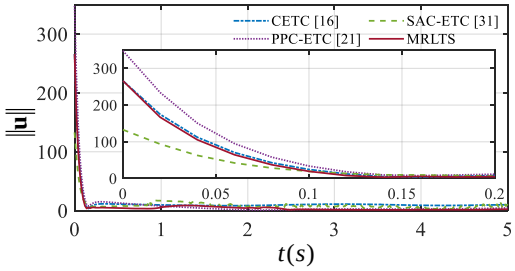


Fig. 11. Comparisons of transient control efforts about different controllers.

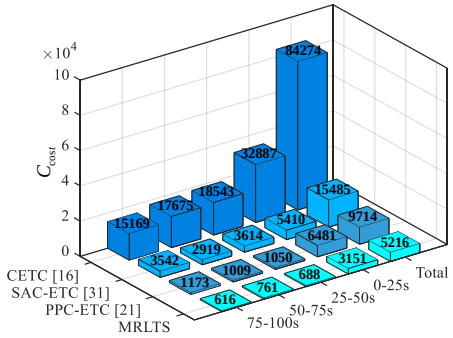


Fig. 12. Comparisons of cost functions concerning different controllers.

and excessive energy consumption, which is evident from its motion trajectory originating from starting point to the elliptical path as well as the steady-state error profiles concerning relative distance error $e_\rho$ and altitude error $e_z$.

Furthermore, Figs. 11 and 12 plot the comparisons of the transient control efforts and the cost value of various schemes in different intervals. Figs. 11 and 12 reveal that MRLTS is advantageous in making a trade-off between tracking performance and energy consumption. Different from non-DRL tracking methods that prioritize tracking performance, incurring excessive energy waste, our method can attain the lowest cost value during each time phase.

Next, to quantitatively demonstrate strengths and weaknesses among involved strategies. Table VI summarizes cost values, root mean square error RMSE $= \sqrt{\frac{1}{p}\sum_{i=1}^{p}\left(e_\rho(i)\right)^2}$, energy cost, optimization ability, standard deviation STD $= \sqrt{\frac{1}{p}\sum_{i=1}^{p}\left(e_\rho(i)-e_\rho^*\right)^2}$, constraints handling ability as well as computational complexity, where $p$ is the number of the data elements and $e_\rho^*$ is the mean value of the error. Compared with other strategies, the MRLTS can achieve an optimal target tracking while sternly submitting to prescribed time limitations without incurring great computational burden.

### *E. Evaluation of Generalization Ability*

In this section, we conduct simulations of two different scenarios to demonstrate the strong generalization ability of MRLTS. In this first scenario, we compare MRLTS with the DDPG [32] by involving the initial position sets of the UAV that have not been considered during the training process, wherein DDPG has been trained under the same environment as MRLTS prior to evaluation. In the second scenario, we introduce a more maneuverable target that has not been considered during pre-training for UAV tracking.

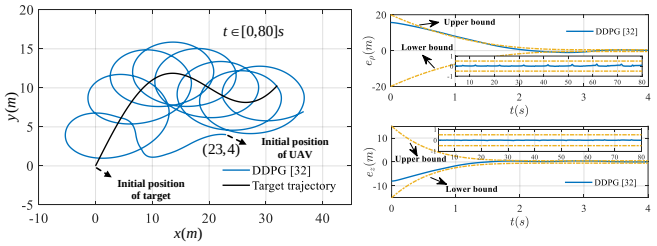


Fig. 13. Mobile target tracking with different initial positions of the UAV by DDPG [32] (scenario 1).

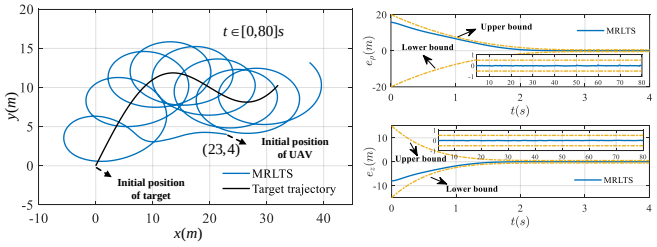


Fig. 14. Mobile target tracking with different initial positions of the UAV by MRLTS (scenario 1).

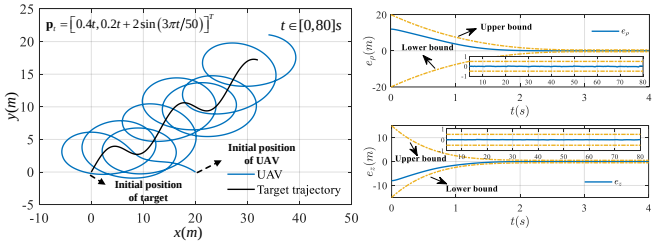


Fig. 15. Evolution of a more complicated target tracking $\mathbf{p}_t = [0.4t, 0.2t + 2\sin(3\pi t/50)]^T(m)$ (scenario 2).

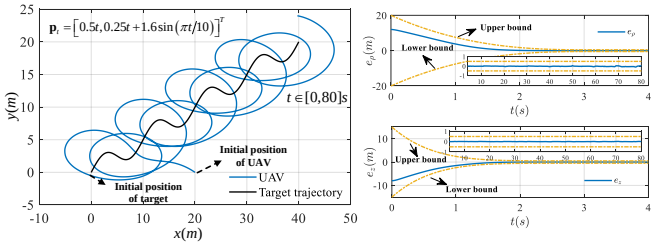


Fig. 16. Evolution of a more complicated target tracking $\mathbf{p}_t = [0.5t, 0.25t + 1.6\sin(\pi t/10)]^T(m)$ (scenario 2).

In view of scenario 1, we select starting locations of the UAV as $x(0) = 23(m), y(0) = 4(m)$, and $z(0) = 2(m)$ that are not involved in the training stage. Fig. 13 displays the tracking trajectory of the UAV with altered initial positions using DDPG, revealing that the controlling error exceeds the prescribed performance boundaries and fails to meet specific time requirements. In contrast, as shown in Fig. 14, MRLTS demonstrates superior adaptability to untrained scenarios, which retains the capability to address designated state constraints while maintaining a high level of tracking accuracy.

In scenario 2, the motion trajectories of a more maneuverable target are selected as $\mathbf{p}_t = [0.4t, 0.2t + 2\sin(3\pi t/50)]^T(m)$ and $\mathbf{p}_t = [0.5t, 0.25t+1.6\sin(\pi t/10)]^T(m)$, and corresponding results are exhibited in Fig. 15 and Fig. 16. It can be inferred that MRLTS is capable of stabilizing tracking errors within designated profiles and has the potential to handle more complex target motion.

## VII. Conclusion

In this paper, a special learning-based paradigm is investigated for the UAV to achieve fast reward convergence and a balance between performance and consumption. First, a steady-state robust item is designed to driven the UAV to approximate the specified tracking trajectory despite unknown uncertainties. Then, a complementary learning part is explored to empower the UAV with optimization ability and constraint handling capability by incorporating a skilled barrier function into DRL framework. Simulation results verify the effectiveness of MRLTS. However, the migration capability of MRLTS is limited when the motion trajectory of the target significantly deviates from the trajectory trained during the learning stage.

In the near future, it is valuable and meaningful to implement and validate the target tracking approach in real-world UAV systems.

## Appendix A
## Proof of Theorem 1

**Proof:** *In terms of (3), (8) and (11), the derivative of error dynamics* $\dot{\tilde{\mathbf{d}}}_m$ *is deduced as*

$$\begin{aligned}\dot{\tilde{\mathbf{d}}}_m &= \dot{\mathbf{d}}_m - \dot{\hat{\mathbf{d}}}_m = \dot{\mathbf{d}}_m - \left[\left(\dot{\mathbf{v}} - \dot{\mathbf{v}}_f\right)/\kappa - \dot{\mathbf{u}}_{bf}\right] \\ &= -\left(\mathbf{d}_m - \hat{\mathbf{d}}_m\right)/\kappa + \dot{\mathbf{d}}_m = -\tilde{\mathbf{d}}_m/\kappa + \dot{\mathbf{d}}_m.\end{aligned} \tag{50}$$

*Then, define a Lyapunov function as*

$$V_1 = \frac{1}{2}\tilde{\mathbf{d}}_m^T\tilde{\mathbf{d}}_m. \tag{51}$$

*Consider (50) and recall Young's inequality [17], there has*

$$\begin{aligned}\dot{V}_1 &= -\tilde{\mathbf{d}}_m^T\tilde{\mathbf{d}}_m/\kappa + \tilde{\mathbf{d}}_m\dot{\mathbf{d}}_m \\ &\le -\left[(2-\kappa)\tilde{\mathbf{d}}_m^T\tilde{\mathbf{d}}_m\right]/2\kappa + \bar{d}_m^2/2 = -K_1V_1 + \varpi_1,\end{aligned} \tag{52}$$

*where* $K_1 = (2-\kappa)/\kappa, \kappa \in (0,2)$ *and* $\varpi_1 = \bar{d}_m^2/2$. *Taking the integration of (52) over* $(0,t)$ *yields*

$$0 \le V_1(t) \le \varpi_1\left(1 - e^{-K_1t}\right)/K_1 + V_1(0)e^{-K_1t}. \tag{53}$$

*Consequently, by resorting to* $\sqrt{m+n} \le \sqrt{m} + \sqrt{n}$, $m > 0, n > 0$, *one can get that*

$$\left\|\tilde{\mathbf{d}}_m(t)\right\| \le \sqrt{2\varpi_1\left(1 - e^{-K_1t}\right)/K_1} + \left\|\tilde{\mathbf{d}}_m(0)\right\|\sqrt{e^{-K_1t}}. \tag{54}$$

*Therefore, under the premise of Assumption 1, the estimation error is ISS with upper bound of* $\left\|\tilde{\mathbf{d}}_m(t)\right\| \le \max\left\{\left\|\tilde{\mathbf{d}}_m(0)\right\|\sqrt{e^{-K_1t}}, \sqrt{2\varpi_1/K_1}\right\}$.

*Let* $\mathbf{p}_1 = [x,y]^T \in \mathbb{R}^+, \mathbf{v}_1 = \left[v_x, v_y\right]^T \in \mathbb{R}^+$ *be the planar location and velocity of the UAV. Subsequently, notice that*

$\boldsymbol{\eta} = [\cos\alpha, \sin\alpha]^T \triangleq (\mathbf{p}_t - \mathbf{p})/\rho$ *and differentiating* $e_\rho$ *over time based on (19) renders*

$$\begin{aligned}\dot{e}_\rho &= \left[\left(\dot{\mathbf{p}}_t^T - \dot{\mathbf{p}}_1^T\right)(\mathbf{p}_t - \mathbf{p}_1)/\rho\right] - \dot{\rho}_d = -\left(\mathbf{v}_1^T - \dot{\mathbf{p}}_t^T\right)\boldsymbol{\eta} - \dot{\rho}_d \\ &= -k_r e_\rho - \left(e_{vx}\cos\alpha + e_{vy}\sin\alpha\right) - v_d\boldsymbol{\eta}_1^T\boldsymbol{\eta} - \dot{\rho}_d.\end{aligned} \tag{55}$$

*In view of (55), (22) substituting (21) into (20), the error dynamics of translational system can be restated as*

$$\begin{cases}\dot{e}_\rho = -k_r e_\rho - \left(e_{vx}\cos\alpha + e_{vy}\sin\alpha\right) - v_d\boldsymbol{\eta}_1^T\boldsymbol{\eta} - \dot{\rho}_d, \\ \dot{e}_z = -k_z e_z + e_{vz}, \\ \dot{\mathbf{e}}_v = \mathbf{u}_b + \mathbf{d}_m - \dot{\bar{\mathbf{v}}}_e.\end{cases} \tag{56}$$

*For the tracking system, taking velocity loop and trajectory loop, construct the Lyapunov candidate as below:*

$$V_2 = \frac{1}{2}\left(\mathbf{e}_v^T\mathbf{e}_v + e_\rho^2 + e_z^2\right). \tag{57}$$

*Invoking (29), the derivative of* $V_2$ *with regard to time follows that*

$$\begin{aligned}\dot{V}_2 \le &- k_r e_\rho^2 - k_z e_z^2 - k_u\|\mathbf{e}_v\|^2 \\ &+ \left|e_\rho\right|\left[\left(|e_{vx}| + \left|e_{vy}\right|\right) + v_d + |\dot{\rho}_d|\right] \\ &+ |e_z e_{vz}| + \left\|\mathbf{e}_v^T\tilde{\mathbf{d}}_m\right\|^2.\end{aligned} \tag{58}$$

*According to (16), one can readily infer that* $\dot{\rho}_d$ *is constrained, satisfying* $|\dot{\rho}_d| \le \bar{\rho}$*. By virtue of Young's inequality, one has*

$$\begin{cases}\left|e_\rho\right|\left(|e_{vx}| + \left|e_{vy}\right|\right) \le e_\rho^2 + e_{vx}^2/2 + e_{vy}^2/2, \\ \left|e_\rho\right|\left(v_d + |\dot{\rho}_d|\right) \le e_\rho^2/2 + \left(v_d + \bar{\rho}\right)^2/2, \\ |e_z e_{vz}| \le e_{vz}^2/2 + e_z^2/2, \\ \left\|\mathbf{e}_v^T\tilde{\mathbf{d}}_m\right\|^2 \le \|\mathbf{e}_v\|^2/2 + \left\|\tilde{\mathbf{d}}_m\right\|^2/2.\end{cases} \tag{59}$$

*Afterwards, based on (59), (58) can be rewritten as the following simplified inequation.*

$$\dot{V}_2 \le -K_2 V_2 + \varpi_2, \tag{60}$$

*where* $K_2 = \min\left(2k_\rho - 3, 2k_z - 1, 2k_u - 2\right) > 0$ *and* $\varpi_2 = \left(v_d + \bar{\rho}\right)^2/2 + \left\|\tilde{\mathbf{d}}_m\right\|^2/2$*. Then taking the integration of (60) over* $(0, t)$*, yields*

$$0 \le V_2(t) \le \varpi_2\left(1 - e^{-K_2 t}\right)/K_2 + V_2(0)e^{-K_2 t}. \tag{61}$$

*Therefore, prompted by Theorem 1, when* $t \to \infty$*,* $e_\rho, e_z, \mathbf{e}_v$ *are terminally upper bounded by* $e_\rho \le \sqrt{2\varpi_2/K_2}, e_z \le \sqrt{2\varpi_2/K_2}$*,* $\|\mathbf{e}_v\| \le \sqrt{2\varpi_2/K_2}$*, and the system is ISS.* ■

## REFERENCES


[1] T. M. Ho, K. K. Nguyen, and M. Cheriet, "UAV control for wireless service provisioning in critical demand areas: A deep reinforcement learning approach," *IEEE Trans. Veh. Technol.*, vol. 7, no. 70, pp. 7138–7152, Jul. 2021.

[2] B. Alzahrani, O. S. Oubbati, A. Barnawi, A. Atiquzzaman, and D. Alghazzawi, "UAV assistance paradigm: State-of-the-art in applications and challenges," *J. Netw. Comput. Appl.*, vol. 166, Sep. 2020, Art. no. 102706.

[3] Y. Wu, W. Yang, X. Guan, and Q. Wu, "UAV-enabled relay communication under malicious jamming: Joint trajectory and transmit power optimization," *IEEE Trans. Veh. Technol.*, vol. 8, no. 70, pp. 8275–8279, Aug. 2021.

[4] S. Li, F. Wu, S. Luo, Z. Fan, J. Chen, and S. Fu, "Dynamic online trajectory planning for a UAV-enabled data collection system," *IEEE Trans. Veh. Technol.*, vol. 12, no. 71, pp. 13 332–13 343, Dec. 2022.

[5] G. Nie, T. Ma, Z. Zhang, H. Tian, S. Mumtaz, and Z. Ding, "Coarse closed-loop trajectory design of multiple UAVs for parallel data collection," *IEEE Trans. Veh. Technol.*, vol. 3, no. 72, pp. 4026–4039, Mar. 2023.

[6] K. Messaoudi, O. S. Oubbati, A. Rachedi, A. Lakas, T. Bendouma, and N. Chaib, "A survey of UAV-based data collection: Challenges, solutions and future perspectives," *J. Netw. Comput. Appl.*, vol. 216, Jul. 2023, Art. no. 103670.

[7] A. Sen, S. R. Sahoo, and M. Kothari, "Circumnavigation on multiple circles around a nonstationary target with desired angular spacing," *IEEE Trans. Cybern.*, vol. 51, no. 1, pp. 222–232, Jan. 2021.

[8] D. Zhang, H. Duan, and Z. Zeng, "Leader-follower interactive potential for target enclosing of perception-limited UAV groups," *IEEE Syst. J.*, vol. 16, no. 1, pp. 856–867, Aug. 2021.

[9] S. Ju, J. Wang, and L. Dou, "MPC-based cooperative enclosing for nonholonomic mobile agents under input constraint and unknown disturbance," *IEEE Trans. Cybern.*, vol. 53, no. 2, pp. 845–858, Feb. 2023.

[10] F. Dong, K. You, and J. Zhang, "Flight control for UAV loitering over a ground target with unknown maneuver," *IEEE Trans. Control Syst. Technol.*, vol. 28, no. 6, pp. 2461–2473, Nov. 2020.

[11] Y. Wang, H. Wang, J. Wu, Y. Liu, and Y. Lun, "UAV standoff tracking for narrow-area target in complex environment," *IEEE Syst. J.*, vol. 16, no. 3, pp. 4583–4594, Sep. 2022.

[12] Y. Jiang, Z. Peng, D. Wang, Y. Yin, and Q.-L. Han, "Cooperative target enclosing of ring-networked underactuated autonomous surface vehicles based on data-driven fuzzy predictors and extended state observers," *IEEE Trans. Fuzzy Syst.*, vol. 30, no. 7, pp. 2515–2528, Jul. 2022.

[13] Y. Jiang, Z. Peng, D. Wang, and C. P. Chen, "Line-of-sight target enclosing of an underactuated autonomous surface vehicle with experiment results," *IEEE Trans. Industr. Inform.*, vol. 16, no. 2, pp. 832–841, Feb. 2020.

[14] X. Peng, K. Guo, X. Li, and Z. Geng, "Cooperative moving-target enclosing control for multiple nonholonomic vehicles using feedback linearization approach," *IEEE Trans. Syst. Man Cybern. Syst.*, vol. 51, no. 8, pp. 4929–4935, Aug. 2021.

[15] H. Yang and Y. Wang, "Cyclic pursuit-fuzzy PD control method for multi-agent formation control in 3D space," *Int. J. Fuzzy Syst.*, vol. 23, no. 6, pp. 1904–1913, Sep. 2021.

[16] S. Chun and Y. Tian, "Multi-targets localization and elliptical circumnavigation by multi-agents using bearing-only measurements in two-dimensional space," *Int. J. Robust Nonlinear Control*, vol. 30, no. 8, pp. 3250–3268, May 2020.

[17] X. Yue, X. Shao, and W. Zhang, "Elliptical encircling of quadrotors for a dynamic target subject to aperiodic signals updating," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 9, pp. 14 375–14 388, Sep. 2022.

[18] N. Koksal, H. An, and B. Fidan, "Backstepping-based adaptive control of a quadrotor UAV with guaranteed tracking performance," *ISA Trans.*, vol. 105, pp. 98–110, Oct. 2020.

[19] Y. Liu and G.-H. Yang, "Prescribed performance-based consensus of nonlinear multiagent systems with unknown control directions and switching networks," *IEEE Trans. Syst. Man Cybern. Syst.*, vol. 50, no. 2, pp. 609–616, Feb. 2020.

[20] M. Ma, L. Tan, and S. Song, "Three-dimensional sliding mode guidance law for maneuvering target with prescribed performance and input saturation," *Trans. Inst. Meas. Control*, vol. 43, no. 5, pp. 1176–1190, Mar. 2021.

[21] F. Zhang, X. Shao, Y. Xia, and W. Zhang, "Elliptical encirclement control capable of reinforcing performances for UAVs around a dynamic target," *Def. Technol.*, doi: 10.1016/j.dt.2023.03.014.

[22] J. Ma, H. Lu, J. Xiao, Z. Zeng, and Z. Zheng, "Multi-robot target encirclement control with collision avoidance via deep reinforcement learning," *J. Intell. Robot Syst.*, vol. 99, no. 2, pp. 371–386, Aug. 2020.

[23] Z. Xia, J. Du, J. Wang, C. Jiang, Y. Ren, G. Li, and Z. Han, "Multi-agent reinforcement learning aided intelligent UAV swarm for target tracking," *IEEE Trans. Veh. Technol.*, vol. 71, no. 1, pp. 931–945, Jan. 2022.

[24] S. Wu, W. Xu, F. Wang, G. Li, and M. Pan, "Distributed federated deep reinforcement learning based trajectory optimization for air-ground cooperative emergency networks," *IEEE Trans. Veh. Technol.*, vol. 71, no. 8, pp. 9107–9112, Aug. 2022.

[25] K. Messaoudi, O. S. Oubbati, A. Rachedi, and T. Bendouma, "UAV-UGV-based system for AoI minimization in IoT networks," *ICC 2023-IEEE Int. Conf. Commun.*, pp. 4743–4748, 2023.

[26] J. Jia, X. Chen, W. Wang, and M. Zhang, "Distributed control of target cooperative encirclement and tracking using range-based measurements," *Asian J. Control*, vol. 25, no. 6, pp. 4595–4608, May 2023.

[27] X. Shao, L. Xu, and W. Zhang, "Quantized control capable of appointed-time performances for quadrotor attitude tracking: Experimental validation," *IEEE Trans. Ind. Electron.*, vol. 69, no. 5, pp. 5100–5110, May 2022.
[28] J. Na, J. Yang, S. Wang, G. Gao, and C. Yang, "Unknown dynamics estimator-based output-feedback control for nonlinear pure-feedback systems," *IEEE Trans. Syst. Man Cybern. Syst.*, vol. 51, no. 6, pp. 3832–3843, Jun. 2021.
[29] J. Na, B. Jing, Y. Huang, G. Gao, and C. Zhang, "Unknown system dynamics estimator for motion control of nonlinear robotic systems," *IEEE Trans. Ind. Electron.*, vol. 67, no. 5, pp. 3850–3859, May 2020.
[30] Y. Huang, J. Wu, J. Na, S. Han, and G. Gao, "Unknown system dynamics estimator for active vehicle suspension control systems with time-varying delay," *IEEE Trans. Cybern.*, vol. 52, no. 8, pp. 8504–8514, Aug. 2022.
[31] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," *Int. Conf. Mach. Learn.*, pp. 1861–1870, Jul. 2018.
[32] L. Luo, J. Zhang, S. Chen, X. Zhang, B. Ai, and D. W. K. Ng, "Downlink power control for cell-free massive MIMO with deep reinforcement learning," *IEEE Trans. Veh. Technol.*, vol. 71, no. 6, pp. 6772–6777, Jun. 2022.
[33] J. Wang, Z. Jiao, J. Chen, X. Hou, T. Yang, and D. Lan, "Blockchain-aided secure access control for UAV computing networks," *IEEE Trans. Netw. Sci. Eng.*, pp. 1–14, 2023, doi: 10.1109/TNSE.2023.3324639.
[34] H. Feng, J. Wang, Z. Fang, J. Chen, and D.-T. Do, "Evaluating AoI-centric HARQ protocols for UAV networks," *IEEE Trans. Commun.*, vol. 72, no. 1, pp. 288–301, Jan. 2024.
[35] C. Sun, X. Li, J. Wen, X. Wang, Z. Han, and V. C. M. Leung, "Federated deep reinforcement learning for recommendation-enabled edge caching in mobile edge-cloud computing networks," *IEEE J. Sel. Area. Comm.*, vol. 41, no. 3, pp. 690–705, Mar. 2023.

**Yi Xia** (Member, IEEE) received the B.S. and M.S. degrees in instrument science and technology from the North University of China, Taiyuan, China in 2021 and 2024, respectively. He is currently pursuing the Ph.D. degree in control science and engineering at the College of Intelligence Science and Technology, National University of Defense Technology, Changsha, China, and also in State Key Laboratory of Integrated Services Networks (Xidian University), Xi'an, China. His research interests include coordinated encircling design, robust sliding mode control, deep reinforcement learning, and anti-disturbance control for multi-agent systems.

**Zekai Zhang** was born in Nanjing, Jiangsu, China, in 2000. He received the B.S. degree in electronic engineering from North University of China, Shanxi, China, in 2021. He is currently pursuing the M.S. degree in electronic engineering at Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, China. His research interests include robot simulation technology, multi-agent cooperation and industrial applications.

**Jingzehua Xu** was born in Xuzhou, China in 2001. He is currently pursuing the M.S. Degree in Tsinghua Shenzhen International Graduate School in Tsinghua University of China. His main research interest is deep reinforcement learning and robotic manipulation.

**Pengfei Ren** received the B.E. degree in software engineering from Xidian University, Xi'an, China in 2023. He is currently pursuing the Ph.D. degree in the School of Cyber Science and Technology in Beihang University, Beijing, China. His research interests include deep learning and semantic communication.

**Jingjing Wang** (Senior Member, IEEE) received his B.S. degree in electronic information engineering from the Dalian University of Technology, Liaoning, China in 2014 and the Ph.D. degree in information and communication engineering from the Tsinghua University, Beijing, China in 2019, both with the highest honors. From 2017 to 2018, he visited the next generation wireless group chaired by Prof. Lajos Hanzo in the University of Southampton, UK. Dr. Wang is currently a Professor at the School of Cyber Science and Technology, Beihang University, Beijing, China, and also at State Key Laboratory of Integrated Services Networks (Xidian University), Xi'an, China. His research interests include AI enhanced next-generation wireless networks, UAV networking and swarm intelligence. He has published over 100 IEEE Journal/Conference papers. He is currently serving as an Editor for the IEEE Wireless Communications Letter and the IEEE Open Journal of the Communications Society. He has served as a Guest Editor for IEEE Internet of Things Journal. Dr. Wang was a recipient of the Best Journal Paper Award of IEEE ComSoc Technical Committee on Green Communications & Computing in 2018, the Best Paper Award of the IEEE ICC and the IEEE IWCMC in 2019.

**Zhu Han** (Fellow, IEEE) received the B.S. degree in electronic engineering from Tsinghua University, in 1997, and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of Maryland, College Park, in 1999 and 2003, respectively. From 2000 to 2002, he was an R&D Engineer of JDSU, Germantown, Maryland. From 2003 to 2006, he was a Research Associate at the University of Maryland. From 2006 to 2008, he was an assistant professor at Boise State University, Idaho. Currently, he is a John and Rebecca Moores Professor in the Electrical and Computer Engineering Department as well as in the Computer Science Department at the University of Houston, Texas. Dr. Han's main research targets on the novel game-theory related concepts critical to enabling efficient and distributive use of wireless networks with limited resources. His other research interests include wireless resource allocation and management, wireless communications and networking, quantum computing, data science, smart grid, carbon neutralization, security and privacy. Dr. Han received an NSF Career Award in 2010, the Fred W. Ellersick Prize of the IEEE Communication Society in 2011, the EURASIP Best Paper Award for the Journal on Advances in Signal Processing in 2015, IEEE Leonard G. Abraham Prize in the field of Communications Systems (best paper award in IEEE JSAC) in 2016, IEEE Vehicular Technology Society 2022 Best Land Transportation Paper Award, and several best paper awards in IEEE conferences. Dr. Han was an IEEE Communications Society Distinguished Lecturer from 2015 to 2018 and ACM Distinguished Speaker from 2022 to 2025, AAAS fellow since 2019, and ACM Fellow since 2024. Dr. Han is a 1% highly cited researcher since 2017 according to Web of Science. Dr. Han is also the winner of the 2021 IEEE Kiyo Tomiyasu Award (an IEEE Field Award), for outstanding early to mid-career contributions to technologies holding the promise of innovative applications, with the following citation: "for contributions to game theory and distributed management of autonomous communication networks."