2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

Optimal Containment Control of Multiple Quadrotors via Reinforcement Learning*

Ming Cheng[1], Hao Liu[2], Deyuan Liu[3], Haibo Gu[3], and Xiangke Wang[4]

Abstract— **This paper explores the optimal containment control problem for nonlinear and underactuated quadrotors with multiple team leaders governed by nonlinear dynamics, employing the reinforcement learning. A cascade controller is formulated, comprising a position control component to ensure containment achievement and an attitude control component to govern rotational channel. The proposed optimal control protocols derived from historical data collected from quadrotor systems without requirement for exact knowledge of vehicle dynamics. The simulation illustrates the effectiveness of the proposed controller in managing a quadrotor team with multiple leaders.**

I. INTRODUCTION

In recent years, the cooperative control of unmanned aerial vehicles (UAVs) has garnered considerable attention, driven by their practical utility across various sectors such as agriculture, logistics, and remote sensing (see [1]–[3]). Within cooperative control, containment control emerges as a focal point, managing to guide each vehicle within a networked system towards the convex region formed by team leaders. This particular challenge becomes a focal point for extensive research within both the control and robotics domains because of its potential applications in surveillance, load transportation, and cooperative source seeking.

Various methods have been developed to deal with the containment control problem for networked linear systems in the last decade. In [4], a distributed adaptive observer technique was used to handle a bipartite containment control problem of linear multi-agent systems. A decentralized control strategy for formation-containment tracking of multi-agent systems using neighboring relative information was presented in [5]. In [6], two intermittent control schemes were developed to achieve containment control for multi-agent systems subject to communication delays. In [7], a distributed containment control strategy was designed for multiple UAVs under fixed and switching interaction topologies. However, these works [4]–[7] assumed that the agents were linear systems, which did not involve the complex nonlinear rotational dynamics and external disturbance in the controller design. Moreover, the team leaders in [4]–[7] did not have control inputs for performing various complicated geometric configurations and autonomous behaviors to avoid unexpected threats. In [8], an adaptive fuzzy event-triggered approach was studied for the containment control of a strict-feedback system. In [9] and [10], state feedback approaches were implemented to deal with the output containment control problems of heterogeneous multi-agent systems. However, none of these works [8]–[10] discussed the scenario that the team leaders are governed by nonlinear dynamics.

In practical applications, the dynamical parameters of the quadrotors are uncertain due to various loads and equipped devices. In [11], a standard Riccati design method was utilized to solve the containment control problem of discrete-time single-input linear multi-agent systems. In [12], a distributed LQR-based consensus control algorithm was proposed for heterogeneous multiagent systems and experimental results were provided to demonstrate the effectiveness.However, the classical optimal control laws in [11] and [12], based on accurate dynamical modeling, are challenging to apply to multi-vehicle systems that suffer from uncertain and nonlinear dynamics. Recently, advancements in the reinforcement learning (RL) theory have emerged to address optimal control challenges encountered by teams of multiple vehicles operating under partially or completely unknown dynamical models, as described in [13]. In [14], a partially model-free controller was designed for the containment of a multi-agent system, but the dynamics of each agent was simplified as linear systems. In [15], the containment control problem for multiple unmanned surface vessels with completely unknown kinetic models was discussed, but external disturbance was not further considered. Therefore, the challenge persists in achieving optimal containment control for quadrotors with nonlinear leaders, parametric uncertainties, underactuation, nonlinear couplings, and external disturbances.

This paper proposes an approach based on a reinforcement learning to address the optimal containment control challenge inherent in the quadrotor vehicles with multiple nonlinear leaders. An optimal hierarchical controller is devised to guarantee the position containment and regulate the attitude dynamics, by developing RL-based algorithms utilizing the historical data obtained from the quadrotor trajectories. The resulting control law can ensure both the achievement of containment flight and optimal control performance for the quadrotors.

*This work was supported by the Beijing Natural Science Foundation under Grant 4232045, the National Natural Science Foundation of China under Grants 62273015 and U23B2032, the China Post-Doctoral Science Foundation under Grant 2021M700336.

[1]Ming Cheng is with the School of Astronautics, Beihang University, Beijing 100191, China mingcheng@buaa.edu.cn

[2]Hao Liu is with the Institute of Artificial Intelligence, Beihang University, Beijing, China liuhao13@buaa.edu.cn

[3]Deyuan Liu and Haibo Gu are with the School of Automation Science and Electrical Engineering, Beihang University, Beijing 100191, China liudeyuan@buaa.edu.cn, guhaibo@buaa.edu.cn

[4]Xiangke Wang is with the College of Intelligence Science and Technology, National University of Defense Technology, Changsha 410073, China xkwang@nudt.edu.cn

II. PROBLEM FORMULATION

A. Preliminaries

Let $I_n \in R^{n\times n}$ denote the identity matrix, $1_n \in R^n$ a column vector with all elements set to 1, $c_{a,b} \in R^a$ a column vector having 0s in all elements except the b-th one, which is 1, and $0_{m\times n} \in R^{m\times n}$ the zero matrix. The Kronecker product is denoted by the symbol $\otimes$. The Euclidean norm distance from a vector $x \in R^n$ to a set $\mathcal{C}$ is written as $dist\,(x,\mathcal{C})$, which means $dist\,(x,\mathcal{C}) = \inf_{y\in\mathcal{C}}\|x-y\|_2$. The minimal convex set that contains all the points in $\mathcal{U} = \{u_1, u_2, \cdots, u_n\}$ with finite elements is expressed as $Co\,(\mathcal{U})$.

A set of UAVs composed of $\mathfrak{n}_f$ followers, defined as $\mathscr{F} \triangleq \{1, 2, \cdots, \mathfrak{n}_f\}$ and $\mathfrak{n}_l$ leaders labeled as $\mathscr{L} \triangleq \{\mathfrak{n}_f+1, \cdots, \mathfrak{n}_f+\mathfrak{n}_l\}$ is considered. The followers are quadrotors which are control plants in this paper, while the leaders can be other types of UAVs. The interaction topology is described by two weighted directed graphs: $\mathscr{G}_e = \{\mathscr{V}_e, \mathscr{E}_e, \mathscr{A}_e\}$ and $\mathscr{G}_f = \{\mathscr{V}_f, \mathscr{E}_f, \mathscr{A}_f\}$, encompassing the entire team of $\mathfrak{n}_f+\mathfrak{n}_l$ UAVs and the quadrotors, respectively. $\mathscr{V}_f = \{v_{fi}\}, (i = 1, 2, \cdots, \mathfrak{n}_f)$ constitutes the set of nodes, where v_{fi} denotes the *i*-th quadrotor. $\mathscr{E}_f \subset \mathscr{V}_f \times \mathscr{V}_f$ represents the set of edges and $A_f = [a_{ij}] \in R^{\mathfrak{n}_f\times\mathfrak{n}_f}$, where $a_{ij} > 0$, if $(v_{fi}, v_{fj}) \in \mathscr{E}_f$, and $a_{ij} = 0$, otherwise. The UAV leaders have no incoming edges, while the quadrotors have incoming edges. The neighbors of the i-th quadrotor are denoted by $N_{fi} = \{v_{fj} | (v_{fi}, v_{fj}) \in \mathscr{E}_f\}$. A directed path from the *i*-th node to the *j*-th node is a series of consecutive edges, *i.e.*, $\{(v_{fi}, v_{fk}), (v_{fk}, v_{fl}), \cdots, (v_{fm}, v_{fj})\}$. The connection indicator of the *v*-th leader is defined by $W_v = \mathrm{diag}\left\{\rho_1^v, \rho_2^v, \cdots, \rho_{\mathfrak{n}_f}^v\right\}, (v \in \mathscr{L})$, where ρ_i^v is 1, if a path from the *v*-th leader to the *i*-th follower exists, and 0 otherwise. Besides, let $\bar{W}_f = \sum_{v=1}^{\mathfrak{n}_l} W_v$.

B. Quadrotor Model

This paper considers completely nonlinear dynamics of the quadrotors. The inertial frame attached to the earth is represented by $\hat{\mathbb{E}}_I$, and the body-fixed frame fixed to the *i*-th quadrotor is represented by $\hat{\mathbb{E}}_B$. The spatial coordinates of the i-th quadrotor in $\hat{\mathbb{E}}_I$ is denoted by $p_{fi} = \begin{bmatrix} p_{fi}^x & p_{fi}^y & p_{fi}^z \end{bmatrix}^T \in R^3$, and the attitude in Euler angle is denoted by $\Theta_i = \begin{bmatrix} \phi_i & \theta_i & \psi_i \end{bmatrix}^T \in R^3$. The completely nonlinear dynamical model of each quadrotor is given by the following expression, according to [16],

$$\begin{aligned} m_i\ddot{p}_{fi} &= \mathscr{R}_{IB}\mathcal{F}_i + d_{pi}, \\ \mathcal{J}_i\ddot{\Theta}_i &= -\mathscr{C}\left(\Theta_i, \dot{\Theta}_i\right)\dot{\Theta}_i + \tau_i + d_{\Theta i}, \end{aligned} \tag{1}$$

where m_i and $\mathcal{J}_i$ represent the mass and inertial matrix of the i-th quadrotor, respectively, with $\mathcal{J}_i = \mathrm{diag}\left\{\mathcal{J}_i^\phi, \mathcal{J}_i^\theta, \mathcal{J}_i^\psi\right\} \in R^{3\times3}$. The matrix $\mathscr{R}_{IB} \in R^{3\times3}$ transforms the coordinates from $\hat{\mathbb{E}}_B$ to $\hat{\mathbb{E}}_I$ and $\mathscr{C}\left(\Theta_i, \dot{\Theta}_i\right) \in R^{3\times3}$ represents the nonlinear Coriolis term [17]. The external forces and torques generated by electric rotors in $\hat{\mathbb{E}}_I$ are denoted by $\mathcal{F}_i \in R^3$ and $\tau_i \in R^3$. The external disturbances on translational and rotational motion in $\hat{\mathbb{E}}_I$ and $\hat{\mathbb{E}}_B$ are represented by d_{pi} and $d_{\Theta i}$, respectively. $\mathcal{F}_i$ and τ_i are given by $\mathcal{F}_i = c_{3,3}k_{\omega i}\sum_{k=1}^4 \omega_{k,i}^2 - \mathscr{R}_{IB}^T c_{3,3} m_i g$, and $\tau_i = \begin{bmatrix} \tau_i^x & \tau_i^y & \tau_i^z \end{bmatrix}^T$, respectively. Specifically, $\tau_{i,x} = l_{ti}k_{\omega i}\left(\omega_{1,i}^2 - \omega_{3,i}^2\right)$, $\tau_{i,y} = l_{ti}k_{\omega i}\left(\omega_{2,i}^2 - \omega_{4,i}^2\right)$, $\tau_{i,z} = k_{\tau i}k_{\omega i}\sum_{k=1}^4 (-1)^{k+1}\omega_{k,i}^2$, g signifies the gravity constant, $\omega_{j,i}$ denotes the spinning rate of the j-th rotor of the i-th quadrotor, and l_{ti}, $k_{\omega i}$, and $k_{\tau i}$ are scaling factors of the i-th quadrotor. Define the control input commands as $u_{zi} = \sum_{k=1}^4 \omega_{k,i}^2$, $u_{\phi i} = \omega_{2,i}^2 - \omega_{4,i}^2$, $u_{\theta i} = \omega_{1,i}^2 - \omega_{3,i}^2$, and $u_{\psi i} = \sum_{k=1}^4 (-1)^{k+1}\omega_{k,i}^2$. Due to the inherent underactuation in quadrotor dynamics, formulate a virtual input for position control denoted as $u_{pi} \in R^3$ in the following manner

$$u_{pi} = u_{zi}\begin{bmatrix} \sin\phi_{ri}\sin\psi_{ri} + \cos\phi_{ri}\cos\psi_{ri}\sin\theta_{ri} \\ \cos\phi_{ri}\sin\psi_{ri}\sin\theta_{ri} - \cos\psi_{ri}\sin\psi_{ri} \\ \cos\phi_{ri}\cos\theta_{ri} \end{bmatrix}, \tag{2}$$

where ϕri, θri, and ψri represent the attitude reference for the *i*-th quadrotor. Let $b_{pi} = k_{\omega i}I_3/m_i\,(i \in \mathscr{F})$ and $b_{\Theta i} = \mathrm{diag}\left\{b_{\Theta i}^1, b_{\Theta i}^2, b_{\Theta i}^3\right\} = \mathrm{diag}\left\{l_{ti}k_{\omega i}, l_{ti}k_{\omega i}, k_{\tau i}\right\}$. Under these conditions, the quadrotor dynamics in (1) can be expressed as

$$\begin{aligned} \ddot{p}_{fi} &= b_{pi}u_{pi} - gc_{3,3} + \Delta_{pi}, \\ \ddot{\Theta}_i &= -\mathcal{J}_i^{-1}\left(\mathscr{C}\left(\Theta_i, \dot{\Theta}_i\right)\dot{\Theta}_i + b_{\Theta i}u_{\Theta i} + d_{\Theta i}\right), \end{aligned} \tag{3}$$

where $u_{\Theta i} = \begin{bmatrix} u_{\phi i} & u_{\theta i} & u_{\psi i} \end{bmatrix}^T \in R^3$, Δ_{pi} represents external disturbance given by $\Delta_{pi} = b_{pi}\tilde{u}_{pi} + d_{pi}$, where $\tilde{u}_{pi} = u_{zi}\mathscr{R}_{IB}c_{3,3} - u_{pi}$.

C. Problem Statement

Let p_{lv} be the position of the v-th leader and $\zeta_{lv} = \begin{bmatrix} p_{lv}^T & \dot{p}_{lv}^T \end{bmatrix}^T \in R^{6\times1}$ becomes the corresponding translational state. Without loss of generosity, the dynamical model of each UAV leader is described as:

$$\dot{\zeta}_{lv} = f_l\left(\zeta_{lv}\right), p_{lv} = N_l\zeta_{lv}, v \in \mathcal{L}, \tag{4}$$

where $f_l\left(\zeta_{lv}\right) \in R^{6\times1}$ is a Lipschitz continuous function with $f_l(0) = 0$, $N_l = \begin{bmatrix} c_{3,1} & c_{3,2} & c_{3,3} & 0_{3\times3} \end{bmatrix}$. In this paper, it is assumed that the directed communication graph $\mathscr{G}_f$ of the quadrotors is strongly connected and each UAV leader has access to all quadrotor followers. In this case, there exists a matrix $\Psi_f = diag\left(v_{f1}, v_{f2}, \cdots, v_{f\mathfrak{n}_f}\right) > 0$ such that $\Pi_f = \left(\Psi_f\left(\mathcal{L}_f + \bar{W}_f\right) + \left(\mathcal{L}_f + \bar{W}_f\right)^T\Psi_f\right)/2 > 0$. It should be noted that, unlike the previous works in [18], where the dynamical models of the leaders are linear systems, this paper considers a more generalized issue with $f_l\left(\zeta_{lv}\right)$ serving as the right-hand function. In this paper, the generalized dynamic function $f_l\left(\zeta_{lv}\right)$ can be selected as complex nonlinear ones, which increases the potential maneuverability of the UAV team.

Specify the subsequent $e_{pi} \in R^3 (i \in \mathscr{F})$ as the local containment error for the i-th quadrotor agent

$$e_{pi} = \sum_{j=1}^{\mathfrak{n}_f} a_{ij}\left(p_{fj} - p_{fi}\right) + \sum_{v=\mathfrak{n}_f+1}^{\mathfrak{n}_f+\mathfrak{n}_l} \rho_i^v\left(p_{lv} - p_{fi}\right). \tag{5}$$

One can obtain the compact form of (5) as follows

$$e_p = -\sum_{v=\mathfrak{n}_f+1}^{\mathfrak{n}_f+\mathfrak{n}_l} \left(\Phi_v \otimes I_3\right)\left(\bar{p}_f - \tilde{p}_{lv}\right), \tag{6}$$

where $e_p = \left[e_{p1}^T, e_{p2}^T, \cdots, e_{p\mathfrak{n}_f}^T\right]^T \in R^{3\mathfrak{n}_f}$, $\tilde{p}_{lv} = 1_{\mathfrak{n}_l} \otimes p_{lv}$ and $\Phi_v = (1/n_f) L_f + W_v$, where L_f is the Laplacian matrix of $\mathscr{G}_f$. The objective of this paper is to design the control laws u_{pi} and $u_{\Theta i}, (i \in \mathscr{F})$ for the quadrotor followers, such that the global closed-loop quadrotor systems are stable and the containment flight can be achieved.

In fact, from [19], if the global error e_p converges to 0, the containment can be achieved regardless of the type of agent dynamics.

III. RL-BASED OPTIMAL CONTAINMENT CONTROLLER

In the following section, the containment controller is introduced for the quadrotor followers, including the RL-based optimal position control component to guarantee the trajectories of the quadrotors inside the convex hull formed by the team leaders and produce the attitude reference, and the RL-based optimal rotational control component to track these attitude references.

A. RL-based Optimal Position Control Law

Containment observers are firstly constructed to generate the trajectory reference for the position controller quadrotor followers. Let $\xi_{fi} = \left[\hat{p}_{ri}^T \quad \dot{\hat{p}}_{ri}^T\right]^T \in R^6$ be the state of the *i*-th observer, where $\hat{p}_{ri} \in R^3$ is the position reference for the *i*-th quadrotor to track. Let $\epsilon_{oi} \in R^6, (i \in \mathscr{F})$ be the local estimation error of the *i*-th observer. By replacing p_{fi} with ξ_{fi} in (6), it follows that $\epsilon_{oi} = \sum_{j=1}^{\mathfrak{n}_f} a_{ij}\left(\xi_{fj} - \xi_{fi}\right) + \sum_{v=\mathfrak{n}_f+1}^{\mathfrak{n}_f+\mathfrak{n}_l} \rho_i^v \left(\zeta_{lv} - \xi_{fi}\right)$. To ensure the convergence of the global estimation error, design the following observer for generating positional reference.

$$\dot{\xi}_{fi} = f_l\left(\xi_{fi}\right) + \varrho_l \epsilon_{oi}, \tag{7}$$

where ϱ_l is a positive constant that is large enough and satisfies that $\varrho_l \geq (\chi_l + 0.5)\,\bar{\upsilon}_f / \underline{\lambda}\left(\Pi_f\right)$, where χ_l denote the Lipschitz constant of $f_l(\cdot)$, $\bar{\upsilon}_f$ denotes the maximum of $\left\{\upsilon_{f1}, \upsilon_{f2}, \cdots, \upsilon_{f\mathfrak{n}_f}\right\}$, $\underline{\lambda}\left(\Pi_f\right)$ denotes the minimum eigenvalue of Π_f. Using the observer in (7), the containment estimation error of each quadrotor can converge to zero and the proof is similar to [20], which is omitted there.

Moreover, the optimal position control law is formulated to follow the trajectory reference derived from the observer and generate attitude reference. The translational dynamics can be reformulated as (3) as

$$\begin{aligned} \dot{z}_{pi} &= M_{pi} z_{pi} + \mathcal{G}_{pi} u_{pi} - g c_{6,6} + D_{pi} \Delta_{pi}, \\ y_{pi} &= N_{pi} z_{pi}, \end{aligned} \tag{8}$$

where $z_{pi} = \left[p_{fi}^T \quad \dot{p}_{fi}^T\right]^T \in R^6$, $\mathcal{G}_{pi} = \left[0 \quad b_{pi}^T\right]^T$, $D_{pi} = \left[0 \quad I_3\right]^T$, $M_{pi} = \left[c_{6,4} \quad c_{6,5} \quad c_{6,6} \quad 0_{6\times 3}\right]^T$, and $N_{pi} = N_l$. Combining (7) and (8) leads to the following position augmented system

$$\begin{aligned} \dot{Z}_{pi} =& \bar{M}_{pi}\left(Z_{pi}\right) + \bar{G}_{pi} u_{pi} - c_{12,6} g + \bar{D}_{pi} \Delta_{pi} \\ &+ T_{pi} \epsilon_{oi}, \\ \delta_{pi} =& \bar{N}_{pi} Z_{pi}, \end{aligned} \tag{9}$$

where $Z_{pi} = \left[z_{pi}^T \quad \xi_{fi}^T\right]^T \in R^{12}$, $\bar{M}_{pi}(\cdot) = \left[\left(M_{pi} z_{pi}\right)^T \quad f_l^T\left(\xi_{fi}\right)\right]^T$, $\bar{G}_{pi} = \left[\mathcal{G}_{pi}^T \quad 0\right]$, $T_{pi} = diag\left\{0_{6\times 6}, \rho_l I_6\right\}$, $\bar{N}_{pi} = \left[N_{pi} - N_l\right]$, $\bar{D}_{pi} = \left[D_{pi}^T \quad 0\right]^T$. The corresponding disturbance Δ_{pi} induces uncertain effects on the system (9) and must be attenuated in the augmented system. Additionally, by employing the devised containment observer in (7) for each quadrotor, the estimation error ϵ_{oi} can converge to 0. To address the impact of external disturbance on the augmented system (9), according to [21], one can apply the disturbance attenuation condition and obtain that

$$\frac{\int_t^\infty e^{-\alpha_i(\tau - t)}\left(\delta_{pi}^T Q_{pi} \delta_{pi} + u_{pi}^T R_{pi} u_{pi}\right) d\tau}{\int_t^\infty e^{-\alpha_i(\tau - t)} \Delta_{pi}^T \Delta_{pi} d\tau} \leq \gamma_p^2, \tag{10}$$

where $\delta_{pi} = p_{fi} - \hat{p}_{ri}$, $\alpha_i > 0$ denotes a discount scalar, $Q_{pi} > 0$, $R_{pi} > 0$, $\gamma_p \geq 0$. In (10), it can be obtained that γ_p signifies the magnitude of attenuation from the impact of Δ_{pi} to the performance of the translational system. Consider the subsequent performance index for the position augmented system as

$$J_{pi} = \int_t^\infty e^{-\alpha_i(\tau - t)} r_p\left(\delta_{pi}, u_{pi}, \Delta_{pi}\right) d\tau, \tag{11}$$

where $r_p\left(\delta_{pi}, u_{pi}, \Delta_{pi}\right) = \delta_{pi}^T Q_{pi} \delta_{pi} + u_{pi}^T R_{pi} u_{pi} - \gamma_p^2 \Delta_{pi}^T \Delta_{pi}$. The Nash solution of the differential game can be obtained as

$$J_{pi}^* = \min_{u_{pi}} \max_{\Delta_{pi}} \int_t^\infty e^{-\alpha_i(\tau - t)} r_p\left(\delta_{pi}, u_{pi}, \Delta_{pi}\right) d\tau, \tag{12}$$

where $J_{pi}^*\left(\delta_{pi}, u_{pi}, \Delta_{pi}\right)$ represents the optimal value. The optimal position control part focuses on formulating the position control law u_{pi} that complies with inequality in (10) ensuring that p_{fi} tracks the position reference produced by the containment observer, and simultaneously minimizing the value given by (12). One can derive the Hamiltonian function as follows

$$\begin{aligned} H\left(J_{pi}, u_{pi}, \Delta_{pi}\right) \triangleq& r_p\left(\delta_{pi}, u_{pi}, \Delta_{pi}\right) - \alpha_i J_{pi} \\ &+ \Delta J_{pi}^T\left(\bar{M}_{pi}\left(Z_{pi}\right) + \bar{G}_{pi} u_{pi}\right) \\ &- \Delta J_{pi}^T\left(c_{12,6} g - \bar{D}_{pi} \Delta_{pi}\right), \end{aligned} \tag{13}$$

where $\Delta J_{pi} = \partial J_{pi} / \partial Z_{pi}$. By applying the stationary condition theorem as shown in [21], differentiating equation (13) with respect to u_{pi} and Δ_{pi}, *i.e.*, $\partial H\left(J_{pi}^*, u_{pi}, \Delta_{pi}\right) / \partial u_{pi} = 0$ and $\partial H\left(J_{pi}^*, u_{pi}, \Delta_{pi}\right) / \partial \Delta_{pi} = 0$, results the subsequent optimal position control protocol and the disturbance input

$$u_{pi}^* = -R_{pi}^{-1} \bar{G}_{pi}^T \Delta J_{pi}^* / 2, \Delta_{pi}^* = \bar{D}_{pi}^T \Delta J_{pi}^* / \left(2\gamma_p^2\right). \tag{14}$$

Substituting (14) into (13) leads to

$$\delta_{pi}^T Q_{pi}\delta_{pi} - \alpha_i J_{pi} + \left(\Delta J_{pi}^*\right)^T \left(\bar{M}_{pi}\left(Z_{pi}\right) - c_{12,6}g\right) - \frac{1}{4}\left(\Delta J_{pi}^*\right)^T \left[\bar{G}_{pi}R_{pi}^{-1}\bar{G}_{pi}^T - \frac{1}{\gamma_p^2}\bar{D}_{pi}\bar{D}_{pi}^T\right]\Delta J_i^* = 0. \tag{15}$$

However, equation (14) requires exact knowledge of the system dynamics, which limits its application. In this case, the reinforcement learning algorithm is developed without accurate quadrotor model in Algorithm 1. By introducing the updated terms, one can rewrite the extended system in (9) as

$$\begin{aligned}\dot{Z}_{pi} =& \bar{M}_{pi}\left(Z_{pi}\right) + \bar{G}_{pi}u_{pi}^n - c_{12,6}g + \bar{D}_{pi}\Delta_{pi}^n + T_{pi}\epsilon_{pi} \\ &+ \bar{G}_{pi}\left(u_{pi} - u_{pi}^n\right) + \bar{D}_{pi}\left(\Delta_{pi} - \Delta_{pi}^n\right).\end{aligned} \tag{16}$$

One can differentiate J_{pi} with respect to system dynamics (16) and utilize (15), leading to

$$\begin{aligned}\dot{J}_{pi}^n =& \alpha_i J_{pi}^n - 2\left(u_{pi}^{n+1}\right)^T R_{pi}\left(u_{pi} - u_{pi}^n\right) \\ &+ \left(\Delta J_{pi}^n\right)^T T_{pi}\mu_{pi} - r_p\left(\delta_{pi}, u_{pi}^n, \Delta_{pi}^n\right) \\ &+ \left(2\gamma_p^2\Delta_{pi}^{n+1}\right)^T\left(\Delta_{pi} - \Delta_{pi}^n\right).\end{aligned} \tag{17}$$

Algorithm 1: Optimal Position Control Law via Reinforcement Learning

Step 1. Given a disturbance input Δ_{pi}, implement a permissible position control protocol u_{pi}^a with an exploratory noise u_{pi}^e to the vehicle position system. Record the historical data of Z_{pi}, u_{pi} and Δ_{pi}.

Step 2. Start with any initial control law u_{pi}^0 and disturbance input $\Delta_{\Theta i}^0$, and apply them into the Bellman equation (18). Update J_{pi}^{n+1}, u_{pi}^{n+1}, and Δ_{pi}^{n+1} by solving the Bellman equation (18).

Step 3. Update n with $n+1$ and go to **Step 2** until $\left\|u_{pi}^{n+1} - u_{pi}^n\right\| \leq \epsilon_p^u$ and $\left\|\Delta_{pi}^{n+1} - \Delta_{pi}^n\right\| \leq \epsilon_p^\Delta$, where $\epsilon_p^u \geq 0$ and $\epsilon_p^\Delta \geq 0$ are predetermined positive constants.

Step 4.Let $u_{pi}^* = u_{pi}^{n+1}$, $\Delta_{pi}^* = \Delta_{pi}^{n+1}$ and output u_{pi}^*, Δ_{pi}^*.

For acquiring the temporal difference of the value function J_{pi}, one can apply the multiplication of $e^{-\alpha_i t}$ on both sides of (17) and integrate both sides, resulting the Bellman equation as

$$\begin{aligned}&\int_t^{t+\delta_T} \frac{d}{d\tau}\left(e^{-\alpha_i(\tau-t)}J_{pi}^n\left(Z_{pi}\left(\tau\right)\right)\right)d\tau \\ =& -\int_t^{t+\delta_T} e^{\alpha_i(t-\tau)}2\left(u_{pi}^{n+1}\right)^T R_{pi}\left(u_{pi} - u_{pi}^n\right)d\tau \\ &-\int_t^{t+\delta_T} e^{\alpha_i(t-\tau)}r_p\left(\delta_{pi}, u_{pi}^n, \Delta_{pi}^n\right)d\tau \\ &+\int_t^{t+\delta_T} e^{\alpha_i(t-\tau)}\left(\Delta_{pi}^n\right)^T T_{pi}\mu_{pi}d\tau \\ &+2\gamma_p^2\int_t^{t+\delta_T} e^{\alpha_i(t-\tau)}\left(\Delta_{pi}^{n+1}\right)^T\left(\Delta_{pi} - \Delta_{pi}^n\right)d\tau.\end{aligned} \tag{18}$$

From Algorithm 1, one can see that the Bellman equation (18) combines the policy evaluation with policy improvement. From Weierstrass theorem, approximate the performance index J_{pi}^n, the control command u_{pi}^{n+1} and the disturbance input Δ_{pi}^{n+1} through the following neural networks:

$$\begin{aligned}\hat{J}_{pi}^n\left(Z_{pi}\right) &= \hat{K}_{pi1}\sigma_{pi1}\left(Z_{pi}\right), \\ \hat{u}_{pi}^{n+1}\left(Z_{pi}\right) &= \hat{K}_{pi2}\sigma_{pi2}\left(Z_{pi}\right), \\ \hat{\Delta}_{pi}^{n+1}\left(Z_{pi}\right) &= \hat{K}_{pi3}\sigma_{pi3}\left(Z_{pi}\right), i \in \mathscr{F},\end{aligned} \tag{19}$$

where $\hat{J}_{pi}^n\left(Z_{pi}\right)$, $\hat{u}_{pi}^{n+1}\left(Z_{pi}\right)$, and $\hat{\Delta}_{pi}^{n+1}$ are the approximated values, $\sigma_{pi1}\left(Z_{pi}\right) \in R^{n_{p1}}$, $\sigma_{pi2}\left(Z_{pi}\right) \in R^{n_{p2}}$, and $\sigma_{pi3} \in R^{n_{p3}}$ are the basis functions with n_{p1}, n_{p2}, and n_{p3} neurons, $\hat{K}_{pi1} \in R^{1\times n_{p1}}$, $\hat{K}_{pi2} \in R^{3\times n_{p2}}$, and $\hat{K}_{pi3} \in R^{3\times n_{p3}}$ are the weighted matrices.

B. RL-based Optimal Attitude Control Law

Consider $\Theta_{ri} = \begin{bmatrix}\phi_{ri} & \theta_{ri} & \psi_{ri}\end{bmatrix}^T$ as the reference for the vehicle rotational system, resulting in

$$\begin{aligned}u_{zi} &= u_{pi}^z / \left(\cos\theta_i \cos\phi_i\right), \\ \phi_{ri} &= \sin^{-1}\left[\frac{\left(\sin\theta_i \sin\psi_i \cos\phi_i - \frac{u_{pi}}{u_{zi}^y}\right)}{\cos\psi_i}\right], \\ \theta_{ri} &= \sin^{-1}\left[\frac{\left(\frac{u_{pi}^x}{u_{zi}} - \sin\phi_i \sin\psi_i\right)}{\left(\cos\psi_i \cos\phi_i\right)}\right].\end{aligned} \tag{20}$$

The optimal attitude control law is devised to trace the attitude reference Θ_{ri} for each quadrotor. By using the same design method of the translational system, the attitude control protocol is given by $\partial H\left(J_{\Theta i}^*, u_{\Theta i}, \Delta_{\Theta i}\right)/\partial u_{\Theta i} = 0$, $\partial H\left(J_{\Theta i}^*, u_{\Theta i}, \Delta_{\Theta i}\right)/\partial \Delta_{\Theta i} = 0$. Consequently, one can derive the optimal attitude control law $u_{\Theta i}^*$ and the attitude disturbance input $\Delta_{\Theta i}^*$ as

$$u_{\Theta i}^* = -\frac{1}{2}R_{\Theta i}^T\bar{G}_{\Theta i}^T\Delta J_{\Theta i}^*, \Delta_{\Theta i}^* = \frac{1}{2\gamma_\Theta^2}\bar{D}_{\Theta i}^T\Delta J_{\Theta i}^*, \tag{21}$$

where $R_{\Theta i} = R_{\Theta i}^T > 0$, $\gamma_\Theta \geq 0$, $\bar{D}_{\Theta i} = \begin{bmatrix}D_{\Theta i}^T & 0\end{bmatrix}^T$, and $\bar{G}_{\Theta i} = [G_{\Theta i} - G_{\Theta ri}].\left(\Delta J_{\Theta i}^*\right)^T$ is the optimal performance index satisfying that

$$\begin{aligned}&\left(\Delta J_{\Theta i}^*\right)^T\left(\bar{M}_{\Theta i}\left(Z_{\Theta i}\right) + \bar{G}_{\Theta i}u_{\Theta i} + \bar{D}_{\Theta i}\Delta_{\Theta i}\right) \\ =& \beta_i J_{\Theta i} - \frac{1}{4\gamma_\Theta^2}\left(\Delta J_{\Theta i}^*\right)^T\bar{D}_{\Theta i}\bar{D}_{\Theta i}^T\Delta J_{\Theta i}^* \\ &- \delta_{\Theta i}^T Q_{\Theta i}\delta_{\Theta i} + \frac{1}{4}\left(\Delta J_{\Theta i}^*\right)^T\bar{G}_{\Theta i}R_{\Theta i}^{-1}\bar{G}_{\Theta i}^T\Delta J_{\Theta i}^*.\end{aligned} \tag{22}$$

Note that equation (22) exhibits nonlinearity with respect to $J_{\Theta i}^*$ and needs precise information of the rotational dynamics. In order to derive the optimal attitude control law, one can employ three neural networks to approximate the performance index $J\Theta i$, the control law $u_{\Theta i}^{n+1}$, and the disturbance input $\Delta_{\Theta i}^{n+1}$. These neural networks utilize weight matrices, namely $\hat{K}_{\Theta i1} \in R^{1\times n_{\Theta 1}}$, $\hat{K}_{\Theta i2} \in R^{1\times n_{\Theta 2}}$, and $\hat{K}_{\Theta i3} \in R^{1\times n_{\Theta 3}}$, where $n_{\Theta 1}$, $n_{\Theta 2}$, and $n_{\Theta 3}$ represent the neuron counts, respectively. The utilized weight matrices $\hat{K}_{\Theta i1}$, $\hat{K}_{\Theta i2}$, and $\hat{K}_{\Theta i3}$ can be tuned using the least-squares method under persistent excitation.

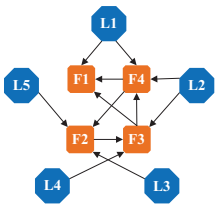

Fig. 1. Communication relationship among the UAV team.

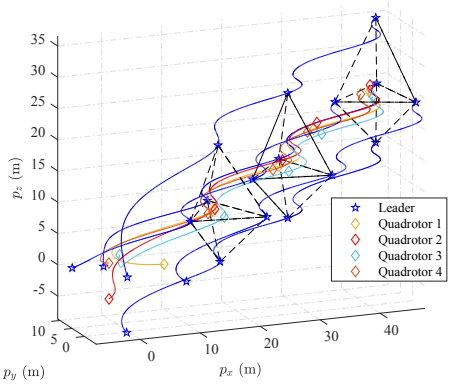

Fig. 2. Three dimensional trajectories of the containment system.

IV. SIMULATION RESULTS

In this section, a UAV system consisting of 5 team leaders modeled as (4) and 4 quadrotor followers modeled as (3) is constructed with the following system configurations: $b_{pi} = \text{diag}\{1,1,1\}$, $g = 9.81$ m/s^2, $J_i = \text{diag}\{0.013, 0.015, 0.0076\} \times 10^{-3}$ kg $\cdot$ m^2 and $b_{\Theta i} = \text{diag}\{43.65, 44.18, 118.54\}$ $(i = 1,2,3,4)$. The communication relationship is shown in Fig. 1. The introduced external disturbances on the position and attitude dynamics are selected as follows: $d_{pi} = \left[0.3\mathscr{E}_{c,i}(t) \quad 0.2\mathscr{E}_{s,i}(t) \quad 0.3\mathscr{E}_{c,i}(t)\right]^T$ and $d_{\Theta i} = \left[0.9\mathscr{E}_{s,i}(t) \quad 0.8\mathscr{E}_{c,i}(t) \quad 0.9\mathscr{E}_{s,i}(t)\right]^T$, $(i = 1,2,3)$, where $\mathscr{E}_{c,i}(t) = (-1)^i \cos(t)$ and $\mathscr{E}_{s,i}(t) = (-1)^i \sin(t)$. The dynamics of each UAV leader is set as $f_l(\zeta_{lv}) = [\zeta_{lv}(4)$ $\zeta_{lv}(5)$ $\zeta_{lv}(6)$ $5\sin(0.8\zeta_{lv}(2))$ $0.5\cos(0.8\zeta_{lv}(1))$ $0]$.

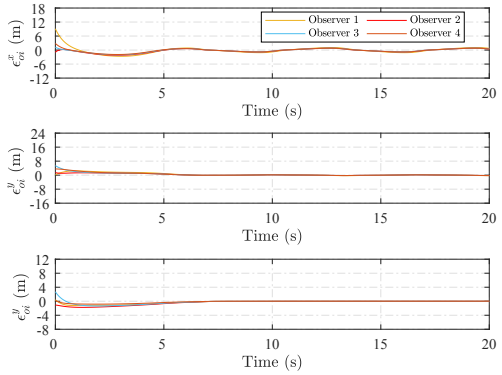

Fig. 3. Estimation errors of the containment observers.

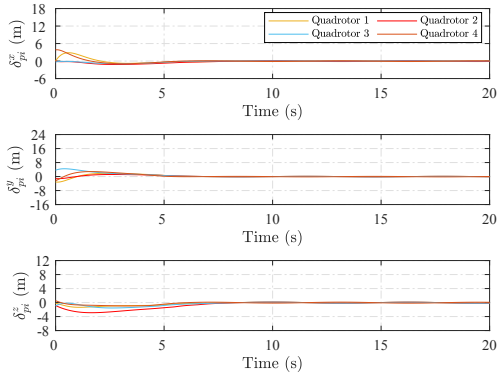

Fig. 4. Position tracking errors of the 4 quadrotors.

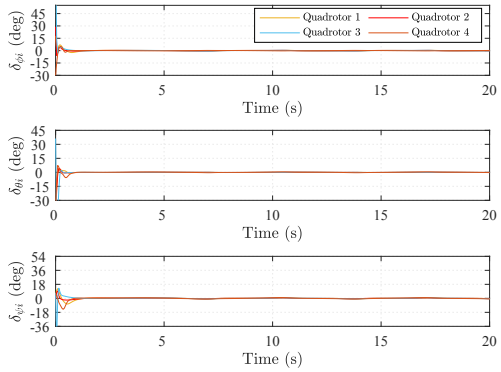

Fig. 5. Attitude tracking errors of the 4 quadrotors.

The RL-based algorithm is implemented to derive the optimal control laws without requiring accurate quadrotor dynamics with the following selected parameters $\alpha_i = 0.04$, $\gamma_p^2 = 4$, $R_{pi} = 2I_3$, $Q_{pi} = 17I_6$, $Q_{\Theta i} = 90I_6$, $R_{\Theta i} = I_3$, and $\gamma_\Theta^2 = 6$. The time interval δ_T is set to 0.06 s. The persisting excitation noise is produced by the summation of a series trigonometric signals. The position and attitude states of the UAV team are initialized as $p_{l1}(0) = \left[-0.5 \quad -0.4 \quad -8.7\right]^T$ m, $p_{l2}(0) = \left[-7.7 \quad -4.8 \quad -0.2\right]^T$ m, $p_{l3}(0) = \left[-0.1 \quad 0.2 \quad 0\right]^T$ m, $p_{l4}(0) = \left[-0.5 \quad 9.0 \quad -0.5\right]^T$ m, $p_{l5}(0) = \left[-7.3 \quad 5.0 \quad 1.0\right]^T$ m, $(v = 1,2,\cdots,5)$, $p_{f1}(0) = \left[8.0 \quad 5.0 \quad 0\right]^T$ m, $p_{f2}(0) = \left[-2.0 \quad 3.0 \quad -4.0\right]^T$ m, $p_{f3}(0) = \left[1.0 \quad 6.0 \quad 2.0\right]^T$ m, $p_{f4}(0) = \left[0.2 \quad 8.4 \quad 0.1\right]^T$ m, $\dot{p}_{fi}(0) = 0_{3\times1}$ m/s, $\Theta_i = 0_{3\times1}$, $\dot{\Theta}_i = 0_{3\times1}$ $(i = 1,2,3,4)$.

The trajectories of the containment system in three dimensions are presented in Fig. 2 and the estimation errors of the distributed observers are displayed in Fig. 3. The position and attitude tracking errors, δ_{pi} and $\delta_{\Theta i}$, are illustrated in Fig. 4 and 5, respectively. The figures clearly demonstrate that the quadrotors successfully fly into the hexahedron spanned

by the team leaders. Besides, the position tracking errors are less than 0.5 m within 7 s and the attitude tracking errors are less than 0.2 deg within 3 s. These results validate the effectiveness of the proposed containment control law, which guarantees convergence of the containment error and ultimately achieves containment.

V. CONCLUSIONS

In this paper, a reinforcement learning-based optimal containment control law is devised for the quadrotors, without requiring information of quadrotor dynamics. The algorithm to learn the optimal control law is presented. For implementation, the optimal performance indices are approximated by several neural works, and the weights are updated from the trajectory data of the vehicle systems. The simulation results verify the effectiveness of the proposed optimal containment control law.

REFERENCES

[1] A. Furchi, M. Lippi, R. F. Carpio, and A. Gasparri, "Route optimization in precision agriculture settings: a multi-steiner TSP formulation," *IEEE Transactions on Automation Science and Engineering*, in press, vol. 20, no. 4, Oct. 2023.

[2] C. Liu and T. Sziranyi, "Road condition detection and emergency rescue recognition using on-board UAV in the wildness," *Remote Sensing*, vol. 14, no. 17, Sep. 2022.

[3] H. Huang, C. Hu, J. Zhu, M. Wu, and R. Malekian, "Stochastic task scheduling in uav-based intelligent on-demand meal delivery system," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 8, pp. 13040-13054, Aug. 2022.

[4] H. Zhang, Y. Zhou, Y. Liu, and J. Sun, "Cooperative bipartite containment control for multiagent systems based on adaptive distributed observer," *IEEE Transactions on Cybernetics*, vol. 52, no. 6, pp. 5432-5440, Jun. 2022.

[5] Y. Hua, X. Dong, L. Han, Q. Li, and Z. Ren, "Formation-containment tracking for general linear multi-agent systems with a tracking-leader of unknown control input," *Systems & Control Letters*, vol. 122, pp. 67-76, Dec. 2018.

[6] Q. Xiao, F. L. Lewis and Z. Zeng, "Containment control for multiagent systems under two intermittent control schemes," *IEEE Transactions on Automatic Control*, vol. 64, no. 3, pp. 1236-1243, Mar. 2019.

[7] T. Han, M. Chi, Z. Guan, B. Hu, J. Xiao, and Y. Huang, "Distributed three-dimensional formation containment control of multiple unmanned aerial vehicle systems," *Asian Journal of Control*, vol. 19, no. 3, pp. 1103-1113, May 2017.

[14] Y. Yang, H. Modares, D. C. Wunsch, and Y. Yin, "Optimal containment control of unknown heterogeneous systems with active leaders," *IEEE Transactions on Control Systems Technology*, vol. 27, no. 3, pp. 1228-1236, May 2019.

[8] W. Wang and S. Tong, "Distributed adaptive fuzzy event-triggered containment control of nonlinear strict-feedback systems," *IEEE Transactions on Cybernetics*, vol. 50, no. 9, pp. 3973-3983, Sep. 2020.

[9] H. Haghshenas, M. A. Badamchizadeh, and M. Baradarannia. "Containment control of heterogeneous linear multi-agent systems," *Automatica*, vol. 54, pp. 210-216, Apr. 2015.

[10] S. Zuo, Y. Song, F. L. Lewis, and A. Davoudi, "Output containment control of linear heterogeneous multi-agent systems using internal model principle," *IEEE Transactions on Cybernetics*, vol. 47, no. 8, pp. 2099-2109, Aug. 2017.

[11] J. Zhang, F. Yan, T. Feng, T. Deng, and Y. Zhao, "Fastest containment control of discrete-time multi-agent systems using static linear feedback protocol," *Information Sciences*, vol. 614, pp. 362-373, Oct. 2022.

[12] B. Mu and Y. Shi, "Distributed LQR consensus control for heterogeneous multiagent systems: theory and experiments," *IEEE/ASME Transactions on Mechatronics*, vol. 23, no. 1, pp. 434-443, Feb. 2018.

[13] W. Zhao, H. Liu, and F. L. Lewis, "Data-driven fault-tolerant control for attitude synchronization of nonlinear quadrotors," *IEEE Transactions on Automatic Control*, vol. 66, no. 11, pp. 5584-5591, Nov. 2021.

[15] N. Gu, D. Wang, Z. Peng, T. Li, and S. Song, "Model-free containment control of underactuated surface vessels under switching topologies based on guiding vector fields and data-driven neural predictors," *IEEE Transactions on Cybernetics*, vol. 52, no. 10, pp. 10843-10854, Oct 2022.

[16] H. Liu, T. Ma, F. L. Lewis, and Y. Wan, "Robust formation control for multiple quadrotors with nonlinearities and disturbances," *IEEE Transactions on Cybernetics*, vol. 50, no. 4, pp. 1362-1371, Apr. 2020.

[17] H. Liu, T. Ma, F. L. Lewis, and Y. Wan, "Robust formation trajectory tracking control for multiple quadrotors with communication delays," *IEEE Transactions on Control Systems Technology*, vol. 28, no. 6, pp. 2633-2640, Nov. 2020.

[18] M. Cheng, H. Liu, Q. Gao, J. Lü, and X. Xia, "Optimal containment control of a quadrotor team with active leaders via reinforcement learning," *IEEE Transactions on Cybernetics*, in press, DOI: 10.1109/TCYB.2023.3284648.

[19] Z. Li, Z. Duan, W. Ren, and G. Feng, "Containment control of linear multi-agent systems with multiple leaders of bounded inputs using distributed continuous controllers," *International Journal of Robust and Nonlinear Control*, vol. 25, no. 13, pp. 2101-2121, Sep. 2015.

[20] L. Xia, Q. Li, R. Song, and H. Modares, "Optimal synchronization control of heterogeneous asymmetric input-constrained unknown nonlinear mass via reinforcement learning," *IEEE/CAA Journal of Automatica Sinica*, vol. 9, no. 3, pp. 520-532, 2022.

[21] H. Modares, F. L. Lewis, and Z. P. Jiang, "H_∞ tracking control of completely unknown continuous-time systems via off-policy reinforcement learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 10, pp. 2550-2562, Oct. 2015.