Differential Game-Based Deep Reinforcement Learning in Underwater Target Hunting Task

Wei Wei, *Student Member, IEEE*, Jingjing Wang, *Senior Member, IEEE*, Jun Du, *Senior Member, IEEE*, Zhengru Fang, *Student Member, IEEE*, Yong Ren, *Senior Member, IEEE*, and C. L. Philip Chen, *Fellow, IEEE*

***Abstract*— To meet requirements for real-time trajectory scheduling and distributed coordination, underwater target hunting task is challenging in terms of turbulent ocean environments and dynamic adversarial environment. Despite the existing research in game-based target hunting area, few approaches have considered dynamic environmental factors, such as sea currents, winds, and communication delay. In this article, we focus on a target hunting system consisted of multiple unmanned underwater vehicles (UUVs) and a target with high maneuverability. Besides, differential game theory is leveraged to analyze adversarial behaviors between hunters and the escapee. However, it is intractable that UUVs have to deploy an adaptive scheme to guarantee the consistency and avoid the escape of the target without collision. Therefore, we conceive the Hamiltonian function with Leibniz's formula to obtain feedback control policies. In addition, it proves that the target hunting system is asymptotically stable in the mean, and the system can satisfy Nash equilibrium relying on the proposed control policies. Furthermore, we design a modified multiagent reinforcement learning (MARL) to facilitate the underwater target hunting task under the constraints of energetic flows and acoustic propagation delay. Simulation results show that the proposed scheme is superior to the typical MARL algorithm in terms of reward and success rate.**

***Index Terms*— Differential game, hamiltonian function, multiagent reinforcement learning (MARL), Nash equilibrium, underwater target hunting.**

Manuscript received 29 October 2022; revised 30 May 2023 and 13 August 2023; accepted 15 October 2023. This work was supported in part by the National Key Research and Development Program of China under Grant 2020YFD0901000 and in part by "The Verification Platform of Multi-Tier Coverage Communication Network for Oceans" of Peng Cheng Laboratory under Grant LZC0020. This work of Jingjing Wang was supported in part by the National Natural Science Foundation of China under Grant 62071268, in part by the Young Elite Scientist Sponsorship Program by the China Association for Science and Technology under Grant 2020QNRC001, in part by the Beijing Natural Science Foundation under Grant L222039, and in part by the Fundamental Research Funds for the Central Universities. An earlier version of this paper was presented at the IEEE International Conference on Communications (ICC), Seoul, South Korea. *(Corresponding author: Jingjing Wang.)*

Wei Wei is with the Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong, China (e-mail: weiwei@eee.hku.hk).

Jingjing Wang is with the School of Cyber Science and Technology, Beihang University, Beijing 100191, China (e-mail: drwangjj@buaa.edu.cn).

Jun Du is with the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China (e-mail: jundu@tsinghua.edu.cn).

Zhengru Fang is with the Department of Computer Science, City University of Hong Kong, Hong Kong, China (e-mail: zhefang4-c@my.cityu.edu.hk).

Yong Ren is with the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China, and also with the Peng Cheng Laboratory, Shenzhen 518055, China (e-mail: reny@tsinghua.edu.cn).

C. L. Philip Chen is with the School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China (e-mail: Philip.Chen@ieee.org).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TNNLS.2023.3325580.

Digital Object Identifier 10.1109/TNNLS.2023.3325580

I. Introduction

A. Background

AS THE complexity of underwater tasks increases, a single unmanned underwater vehicle (UUV) is unable to fulfill the high-efficiency and large-scale missions' requirements due to limited operational capability and sophisticated environment. For instance, UUV must support comprehensive underwater services in terms of dynamic adversarial scenarios, ranging from real-time path planning, distributed coordination, optimal control, and varied task scheduling. Hence, to achieve efficient underwater task execution, we explore the utilization of UUV swarm, which imitates the organized behavior of insects: using centralized/decentralized control to coordinate complex behavior among varied UUVs. UUVs with swarm intelligence are capable of being deployed quickly to complete a wide variety of underwater tasks [1]. However, the existing studies of UUV swarm intelligence show promise but require more study.

Recently, swarm intelligence mainly focuses on three fields, Lyapunov analysis, computer simulations, and model-based theory. Specifically, Lyapunov analysis remains confined to boundary problems. Moreover, simulating methods suffer from difficulties, such as accuracy, convergence, and complexity analysis. Model-based approaches, especially game theories, provide proper frameworks for analyzing conflicting interests of intelligent agents in one or more teams and reveal inner structural properties with comprehensive theoretical analysis. Thus, differential game is more suitable framework to model target hunting tasks, where swarm agents can be separated into adversarial teams: hunters and targets. Specifically, hunters apply tracking and hunting strategies and aim to encircle targets in their attacking scope. By contrast, targets make efforts to escape from hunters' searching area [2].

However, there are many challenges in differential game-based underwater target hunting task [3]. On the one hand, when adversarial scenarios change over time causing by target's escaping behaviors and natural ecological processes,

UUVs may spend more efforts replanning hunting paths. Such procedure can be seen as a multistage hunting process and is going to be intractable, because UUVs must find feasible solutions under the constraints of appropriate control laws and limited time. In addition, the design of guidance laws must consider the maneuverability of targets and their response to disturbances in the context of turbulent ocean, especially for a relatively small-sized UUV. At the same time, those challenges also lead to the instability of the optimal solutions to the differential game in terms of UUVs' behaviors. On the other hand, it is inevitable that the large communication delay of acoustic signal propagation deteriorates the noncausality of control policy. In the following section, the aforementioned problems are solved by introducing states containing past information and are corrected by predictive effects with current control laws. Specifically, along with batch learning, experience replay, and batch normalization, multiagent reinforcement learning (MARL) performs superior to obtain predictive effects and handles intricate tasks without much prior knowledge [4].

B. State of the Art

In this section, we review the related literature of differential game-based target hunting tasks, MARL-assisted target hunting tasks, and underwater target hunting tasks, respectively.

1) Differential Game-Based Target Hunting Tasks: Numerous differential game-based schemes have been designed for target hunting tasks [6], [7], [8], [9]. In [5] and [6], researchers investigated differential game among multiple hunters and targets and obtained stable team formation strategies satisfying Nash equilibrium. However, their work did not consider the stability of the game system. Fuchs and Khargonekar [7] discouraged hunters from attacking through developing equilibrium open loop policies and encouraged evacuating by tackling the differential game of engagement. Obviously, the above strategies had a detrimental impact to hunt a target with multiple hunters. Moreover, in [6], [7], and [8], the target's intelligence was relatively low, which did not meet with the real pursuit–evasion condition. As a further development, Chen et al. [8] addressed plenty of multihunter pursuit–evasion games with one superior target moving faster than hunters.

2) MARL-Assisted Target Hunting Tasks: Kamalapurkar et al. [9] sought to combine the actor–critic–identifier architecture with differential game theory to approximate the optimal controllers for formation tracking in multiagent system. Chen et al. [10] constructed an autonomous tracking model to localize a mobile target with multiple hunters and proposed an enhanced MARL method to coordinate a swarm of hunters performing real-time target tracking. Moreover, Moon et al. [11] focused their attention on the optimal control policies quantified by the reward function composed of the entire system reward and each hunter's contribution. Furthermore, Xia et al. [12] proposed an end-to-end cooperative MARL scheme to track and hunt the target. However, since all target's behaviors above were set simple without considering their adaptability with antihunting strategies, the proposed target hunting systems were unable to extend to more complex underwater adversarial scenarios.

3) Underwater Target Hunting Tasks: Sun et al. [13] dealt with the target hunting differential game with a reachability-based approach between two players influenced by dynamic disturbances, e.g., winds and sea currents. Besides, Ni et al. [14] focused on target searching and hunting and proposed a novel spinal neural system-based approach. Moreover, Yan et al. [15] designed an adaptive-nonsingular fast terminal sliding mode tracking controller to drive UUV to the target. However, the above schemes may not meet the problem with multiple hunters in a swarm team. Yu et al. [16] formulated target escaping problem with multiple UUVs and performed environment simulation, map building, voyage scheduling, and trajectory planning with a hybrid heuristic approach. However, their work did not consider the communication mechanism among multiple UUVs and underwater dynamic environment. As a further development, Li et al. [17] formulated UUVs' path planning into a convex optimization problem with physical constraints including communicating conditions, collision avoidance, and kinematics simulations. Recently, Wei et al. [18] designed a joint unmanned aerial vehicle (UAV)–unmanned surface vehicle (USV)–UUV network and provided improved the efficiency of underwater cooperative target hunting task. However, the aforementioned work have not considered adversarial relationships between UUVs and targets.

C. Motivation and New Contributions

Most of the existing treatises ignored target's intelligence in hunting task. By contrast, target can sense the approaching of hunters and adapt adaptive escaping strategies in many realistic scenarios. Moreover, the challenge lies in UUVs must conduct control policies in consideration of not only target's escaping behaviors, but also collision avoidance and consistency among swarm hunters. Specifically, UUVs have to balance competing objectives of maintaining consistency, while avoiding collisions during the entire mission. However, for all aforementioned literature in game-based target hunting area, little research had been conducted into the impact of dynamic environmental factors on the outperformance of differential game. In comparison with the aforementioned work based on underwater target hunting task in a relatively ideal underwater scenario, we establish a differential game-based model to analyze the underwater target hunting task with multiple UUVs and an intelligent target with high maneuverability. Against the above background, the main contributions lie in the following.

1) To the best of our knowledge, we first exploit a differential game framework to tackle underwater target hunting task with consistency-aware payoff functions, the collision avoidance of target hunting team, and adversarial behaviors between UUVs and the target. Under the assumption that UUVs and the target have knowledge (velocity, state, and so on) of each agent, we conceive the Hamiltonian function with Leibniz's formula to obtain stable feedback control strategies. Furthermore, it proves that the proposed target hunting system is asymptotically stable in the mean, and the system is in Nash equilibrium accordance with the feedback control policies.

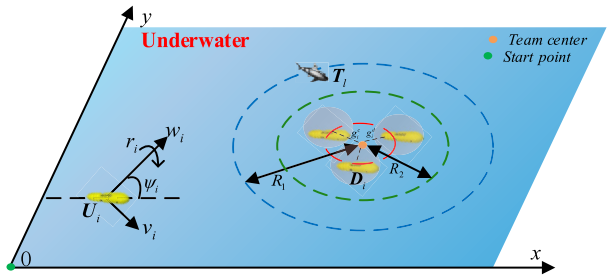


Fig. 1. Underwater differential game with multiple agents on the horizontal plane.

2) We design an MARL-assisted scheme for striking a trade-off between the cooperation of hunters and noncooperation of hunters against target under the constraints of underwater disturbances and acoustic propagation delay. Specifically, we construct a differential game-based multiagent twin-delayed deep deterministic policy gradient (MATD3) framework to tackle the underwater target hunting task.
3) Simulation results show that our proposed scheme improves the reward and success rate in comparison to typical multiagent deep deterministic policy gradient (MADDPG) algorithms. Moreover, the MATD3 scheme makes the global optimization of noncooperative differential game possible and achieves Nash equilibrium by applying exploitation and exploration polices.

D. Organization

The remainder is outlined as follows. First, we present the system model and problem formulation relying on the differential game for underwater target hunting task in Section II. Then, we obtain the feedback control policies with Hamiltonian function and conceive the modified MARL-assisted framework in Section III. In Section IV, simulation results are provided for characterizing the proposed differential game model with acoustic propagation delay and disturbances. Finally, in Section V, we conclude our work.

II. System Model

This section describes the differential game-based system model, associated definitions, and assumptions relating to the adversarial environment. Moreover, we consider an underwater target hunting differential game with multiple UUVs on a 2-D horizontal plane at d meters below the water surface. Specifically, the differential game is established to reveal cooperative properties among M UUVs and noncooperative relationship between UUVs' hunting team and the target. Furthermore, Fig. 1 shows that the target $\boldsymbol{T}_l$ appears on the horizontal plane randomly, and the swarm team $\boldsymbol{U}$ composed of multiple UUVs disperses around the start point $\boldsymbol{O} = (O_x, O_y, d)$ initially. The coordinates are further expressed as $\boldsymbol{T}_l = (t_{x_l}, t_{y_l}, d)$ and $\boldsymbol{U} = \cup_{i=1}^{M}\{\boldsymbol{U}_i = (u_{x_i}, u_{y_i}, d),\ i \in (0, M]\}$, respectively. Besides, we assume that UUVs and the target have perfect knowledge (velocity, state, policies, and so on) of each agent in the differential game-based model [19].

A. Dynamics of UUVs and Target

When a hunter in UUV team detects the target, UUVs first track the target on the 2-D plane. Herein, we formulate a three-degrees-of-freedom underactuated dynamic model with the aid of an earth-fixed reference coordinate system $\boldsymbol{\eta}_i = [u_{x_i}, u_{y_i}, \psi_i]^{\mathrm{T}}$ and a body-fixed coordinate system $\boldsymbol{v}_i = [w_i, v_i, r_i]^{\mathrm{T}}$ [20]. Specifically, ψ_i represents the yaw angle. Besides, w_i, v_i, and r_i are defined as velocities when surging, swaying, and yawing, respectively. Moreover, the maximum speed of $\boldsymbol{v}_i$ is limited to V_1 due to UUV's kinetic characteristic, i.e., $\|\boldsymbol{v}_i\| \le V_1$. Thus, the dynamics of the ith UUV is formulated as follows:

$$\dot{\boldsymbol{\eta}}_i = \boldsymbol{J}(\boldsymbol{\eta}_i)\boldsymbol{v}_i \tag{1}$$

$$\boldsymbol{M}\dot{\boldsymbol{v}}_i + \boldsymbol{C}(\boldsymbol{v}_i)\boldsymbol{v}_i + \boldsymbol{B}(\boldsymbol{v}_i)\boldsymbol{v}_i + \boldsymbol{G}(\boldsymbol{v}_i) = \boldsymbol{p}_i + \boldsymbol{\tau}_d \tag{2}$$

where $\boldsymbol{M}$ and $\boldsymbol{C}(\boldsymbol{v}_i)$ represent system inertia parameters, such as added mass and the Coriolis-centripetal matrices. Moreover, $\boldsymbol{B}(\boldsymbol{v}_i)$ denotes the damping matrix, and $\boldsymbol{G}(\boldsymbol{v}_i)$ represents the resultant matrix of gravity and buoyancy. Besides, the ith UUV takes $\boldsymbol{p}_i$ as the control input, while affected by the underwater disturbance $\boldsymbol{\tau}_d$. Furthermore, the transformation matrix $\boldsymbol{J}(\boldsymbol{\eta}_i)$ can be expressed as follows:

$$\boldsymbol{J}(\boldsymbol{\eta}_i) = \begin{bmatrix} \cos\psi_i & -\sin\psi_i & 0 \\ \sin\psi_i & \cos\psi_i & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{3}$$

Furthermore, we assume that target's velocity $\boldsymbol{v}_T$ is subject to the maximum speed V_2, i.e., $\|\boldsymbol{v}_T\| \le V_2$. For simplicity, we use the underactuated model above to formulate the dynamics of the target as follows:

$$\dot{\boldsymbol{\eta}}_T = \boldsymbol{J}(\boldsymbol{\eta}_T)\boldsymbol{v}_T \tag{4}$$

$$\boldsymbol{M}\dot{\boldsymbol{v}}_T + \boldsymbol{C}(\boldsymbol{v}_T)\boldsymbol{v}_T + \boldsymbol{B}(\boldsymbol{v}_T)\boldsymbol{v}_T + \boldsymbol{G}(\boldsymbol{v}_T) = \boldsymbol{p}_T + \boldsymbol{\tau}_d. \tag{5}$$

Considering UUV swarm team performs better chasing ability by implementing cooperating and coordinating strategies, we make an underlying assumption that the maximum velocities satisfy $V_1 > V_2$. Moreover, due to the escaping target performing more maneuverability compared with a self-organized swarm team, there is an assumption that the target has a wider range of steering angle, i.e., $\|\Delta\psi_{T\max} - \Delta\psi_{T\min}\| > \|\Delta\psi_{i\max} - \Delta\psi_{i\min}\|$.

B. Underwater Acoustic Propagation Delay

In general, underwater acoustic propagation delay has a decisive influence on UUVs and target in terms of obtaining the current knowledge of each agent [21]. Specifically, the information exchange process is dependent on the underwater acoustic transmission, while the speed $\|\boldsymbol{v}_w\|$ of sounds in water is provided according to the empirical formula

$$\|\boldsymbol{v}_w\| = 1450 + 4.21T - 0.037T^2 + 1.14(S-35) + 0.175P \tag{6}$$

where T, S, and P denote the temperature, the salinity, and the pressure, respectively [19]. For simplicity, it is assumed that underwater acoustic signals transmit in straight lines.

Let $\boldsymbol{e}_i = \boldsymbol{T}_l - \boldsymbol{U}_i$ stand for the positional vector from the current coordinate of the ith UUV to the target's position. Therefore, the acoustic propagation delay from UUV to the target is defined as $\delta_{i\to T} = \boldsymbol{e}_i/(\boldsymbol{v}_T + \boldsymbol{v}_w)$, while $\delta_{T\to i} = -\boldsymbol{e}_i/(\boldsymbol{v}_i + \boldsymbol{v}_w)$ represents the acoustic propagation delay caused by the transmission from the target to UUV. Accordingly, the average one-way acoustic propagation delay δ is given by

$$\delta = \frac{1}{2M}\sum_{i=1}^{M}\left(\left\|\frac{\boldsymbol{e}_i}{\boldsymbol{v}_T + \boldsymbol{v}_w}\right\| + \left\|\frac{-\boldsymbol{e}_i}{\boldsymbol{v}_i + \boldsymbol{v}_w}\right\|\right). \tag{7}$$

C. Problem Formulation

We assume that each UUV's searching range and attacking range are circular area with the radius R_1 and R_2. The target can be detected as long as there exists some UUV satisfying $\|\boldsymbol{e}_i\| < R_1$ and can be caught when $\|\boldsymbol{e}_i\| < R_2$. Accordingly, the multi-UUV cooperative target hunting problem focuses on obtaining feedback control policies $\boldsymbol{\eta}_i$ and $\boldsymbol{v}_i$, which navigate UUVs from starting point to the target, avoid collisions, and maintain the team's consistency. Herein, we consider a dynamic game model denoted by the ordinary differential equation, and thus, the state function of multiple UUVs and an escaping target can be further given by

$$\begin{aligned}\dot{\boldsymbol{s}}(t) &= \boldsymbol{F}_s\boldsymbol{s}(t-\delta) + \sum_{i=1}^{M}\boldsymbol{G}_{12i}\boldsymbol{p}_i(t) + \boldsymbol{G}_{21}\boldsymbol{q}(t)\\ \boldsymbol{s}(0) &= \boldsymbol{s}_0, \quad t \in [0,\ T_h]\end{aligned} \tag{8}$$

where $\boldsymbol{s}(t) = [\boldsymbol{U}_1^{\mathrm{T}}, \boldsymbol{U}_2^{\mathrm{T}}, \ldots, \boldsymbol{U}_M^{\mathrm{T}}, \boldsymbol{T}_l^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{R}^{n\times 1}$ $(n = M+1)$ denotes the positional matrix of M UUVs and a target. Moreover, $\boldsymbol{p}_i(t) = [\dot{w}_i, \dot{v}_i, \dot{r}_i]^{\mathrm{T}} \in \mathbb{R}^{3\times 1}$ represents the control input of the ith UUV with choices of velocity and heading direction. Furthermore, control inputs of UUVs can be given by $\boldsymbol{p}(t) = [\boldsymbol{p}_1^{\mathrm{T}}, \boldsymbol{p}_2^{\mathrm{T}}, \ldots, \boldsymbol{p}_M^{\mathrm{T}}]^{\mathrm{T}}$, while $\boldsymbol{q}(t) = [\dot{w}_T, \dot{v}_T, \dot{r}_T]^{\mathrm{T}} \in \mathbb{R}^{3\times 1}$ stands for target's control input. Besides, T_h denotes the maximum time when UUVs tracking and hunting the target. Furthermore, $\boldsymbol{s}_0$ denotes the initial state, and $\boldsymbol{s}_f$ represents the final state when $\|\boldsymbol{e}_i\| > R_1$ or $\|\boldsymbol{e}_i\| < R_2$. In addition, $\boldsymbol{F}_s \in \mathbb{R}^{n\times n}$, $\boldsymbol{G}_{12i} \in \mathbb{R}^{n\times 3}$, $\boldsymbol{G}_{12} = [\boldsymbol{G}_{12_1}^{\mathrm{T}}, \boldsymbol{G}_{12_2}^{\mathrm{T}}, \ldots, \boldsymbol{G}_{12_M}^{\mathrm{T}}]^{\mathrm{T}}$, and $\boldsymbol{G}_{21} \in \mathbb{R}^{n\times 3}$ are coefficient matrices.

1) UUV's Payoff Function: Considering a swarm hunting team organized by M UUVs $(i \in (0, M])$, the ith UUV's payoff function consisting of an integral cost function and a terminal state can be expressed as follows:

$$P_i(\boldsymbol{p}_i, \boldsymbol{q}, \boldsymbol{s}_0) = \frac{1}{2}\int_0^{T_h}\boldsymbol{p}_i^{\mathrm{T}}\left(\alpha_i^d g_i^d(t) + \beta_i^c g_i^c(t)\right)\boldsymbol{p}_i \mathrm{d}t - \boldsymbol{s}_f^{\mathrm{T}}\phi_i(\boldsymbol{s}_f)\boldsymbol{s}_f \tag{9}$$

where constants α_i^d and β_i^c are nonnegative and can be used to adjust the formation of UUVs' hunting team.

1) *Collision Avoidance Constraint:* Let r stand for the safety radius of each UUV to avoid collisions with team members. Specifically, the ith UUV is recognized as colliding with the jth UUV when existing an instant satisfying $\|\boldsymbol{U}_i - \boldsymbol{U}_j\|^2 \le r$, for $i, j \in (0, M]$, and $i \ne j$. Thus, we denote the avoidance region of the ith UUV as $D_i = \cup_{j=1, j\ne i}^{M} D_{ij}$, where $D_{ij} = \{\|\boldsymbol{U}_i - \boldsymbol{U}_j\|^2 \le r, i \in (0, M]\}$. Furthermore, a penalty function $g_i^d(t)$ is defined to prevent collision when the ith UUV drawing near to its teammates [22]

$$g_i^d(t) = \sum_{j=1, j\ne i}^{M}\left(\frac{1}{\|\boldsymbol{U}_i(t) - \boldsymbol{U}_j(t)\|^2 - r^2}\right)^l \tag{10}$$

where the constant l satisfying $0 < l < 1$, and it is obvious that $\lim_{U_i\to\partial D_i} g_i^d = +\infty$.

2) *Communication Constraint:* Communication is one of the key technologies for UUVs to coordinate and cooperate with teammates when conducting the complicated underwater hunting task. Specifically, UUVs are considered to maintain the reliability of communication and realize the consistency of task execution, such that $\lim_{t\to\infty}\|U_i(t) - U_j(t)\| = 0$ $(\forall i, j \in (0, M]$, and $i \ne j)$ holds under any initial conditions [23]. Therefore, we define $g_i^c(t)$ as a penalty function avoiding the remote distance between the ith UUV and other UUVs, which is formulated as follows:

$$g_i^c(t) = \sum_{j=1, j\ne i}^{M}\|\boldsymbol{U}_i(t) - \boldsymbol{U}_j(t)\|^2. \tag{11}$$

3) *Terminal Constraint:* When some target enters some UUV's attacking scope (a circular area of radius R_2) or run off all UUVs' sensing range (area composed of M circles whose radius are R_1), the differential hunting game comes to an end and leads to a final condition $\boldsymbol{s}(f)$. Thus, the terminal value function $\phi_i(\boldsymbol{s}_f)$ is defined as follows:

$$\phi_i(\boldsymbol{s}_f) = \begin{cases} a, & \forall\ \|\boldsymbol{e}_i\| > R_1 \\ b, & \exists\ \|\boldsymbol{e}_i\| < R_2 \end{cases} \tag{12}$$

where a and b are constants to distinguish the success or failure of the UUVs' hunting task.

2) Target's Payoff Function: To avoid being caught by UUV team, the swarm target has the following individual interests: 1) maximizing the distance away from UUVs when escaping; 2) scheduling an optimal route to minimize energy consumption; and 3) changing the direction and speed flexibly to cope with UUVs' hunting strategy. By doing so, the UUVs would spend more efforts to complete hunting task, while the target gains more opportunities to escape. Therefore, the target's payoff with respect to the ith UUV is defined as follows:

$$P_T^i(\boldsymbol{p}_i, \boldsymbol{q}, \boldsymbol{s}_0) = \frac{1}{2}\int_0^{T_h}\boldsymbol{q}^{\mathrm{T}}\left(\frac{1}{\|\boldsymbol{U}_i(t) - \boldsymbol{T}_l(t)\|^2}\right)^l\boldsymbol{q}\mathrm{d}t. \tag{13}$$

3) System's Payoff Function: Accordingly, the payoff function of the hunting system is obtained with the combination of UUV's payoff function (9) and target's payoff function (13),

which can be expressed as follows:

$$\begin{aligned} P_E(\boldsymbol{p},\boldsymbol{q},\boldsymbol{s}_0) &= \sum_{i=1}^{M}\left\{P_i\left(\boldsymbol{p}_i,\boldsymbol{q},\boldsymbol{s}_0\right) - P_T^i(\boldsymbol{p}_i,\boldsymbol{q},\boldsymbol{s}_0)\right\} \\ &= \frac{1}{2}\sum_{i=1}^{M}\left\{\int_0^{T_h}\boldsymbol{p}_i^{\mathrm{T}}\left(\alpha_i^d g_i^d(t)+\beta_i^c g_i^c(t)\right)\boldsymbol{p}_i \mathrm{d}t \right. \\ &\quad \left. -\int_0^{T_h}\boldsymbol{q}^{\mathrm{T}}\left(\frac{1}{\|\boldsymbol{U}_i(t)-\boldsymbol{T}_l(t)\|^2}\right)\boldsymbol{q}\mathrm{d}t\right\} \\ &\quad -\boldsymbol{s}_f^{\mathrm{T}}\phi_i(\boldsymbol{s}_f)\boldsymbol{s}_f. \end{aligned} \tag{14}$$

4) Nash Equilibrium: Considering system's payoff function formulated in (14), the differential game is defined based on the motivation that each player aims to minimize their respective payoff function in some initial state [7], which is given by

$$P_E^*(\boldsymbol{s}_0) = \min_{\boldsymbol{p}}\max_{\boldsymbol{q}}\{P_E(\boldsymbol{p},\boldsymbol{q},\boldsymbol{s}_0)\}. \tag{15}$$

In addition, the equilibrium value $P_E^*(\boldsymbol{s}_0)$ should satisfy the following Nash equilibrium condition:

$$P_E\left(\boldsymbol{p},\boldsymbol{q}^*,\boldsymbol{s}_0\right) \geq P_E\left(\boldsymbol{p}^*,\boldsymbol{q}^*,\boldsymbol{s}_0\right) = P_E^*(\boldsymbol{s}_0) \geq P_E\left(\boldsymbol{p}^*,\boldsymbol{q},\boldsymbol{s}_0\right). \tag{16}$$

5) Problem Definition: The problem takes the state function (8) and system's payoff (14) into account [19]. Moreover, it is obvious that solving the noncooperative underwater differential game between UUVs and the target lies in finding a feasible pair of feedback strategies ($\boldsymbol{p}^*$ and $\boldsymbol{q}^*$), such that $P_E(\boldsymbol{p},\boldsymbol{q}^*,\boldsymbol{s}_0) \geq P_E(\boldsymbol{p}^*,\boldsymbol{q}^*,\boldsymbol{s}_0) = P_E^*(\boldsymbol{s}_0) \geq P_E(\boldsymbol{p}^*,\boldsymbol{q},\boldsymbol{s}_0)$. Furthermore, UUVs terminate the target hunting task when catching the target, while minimizing the payoff during the hunting process. By contrast, the target aims to maximize UUVs' payoff throughout the course of differential game. Thus, in accordance with agents' respective payoff (9) and (13) and underwater dynamics (1)–(5), the differential game can be formulated as follows:

$$\begin{aligned} &V_E^*(\boldsymbol{s}_0) \\ &:= \min_{\boldsymbol{p}}\max_{\boldsymbol{q}}\{P_E(\boldsymbol{p},\boldsymbol{q},\boldsymbol{s}_0)\} \\ &\text{s.t. (a)} \quad \dot{\boldsymbol{\eta}}_i = \boldsymbol{J}(\boldsymbol{\eta}_i)\boldsymbol{v}_i \\ &\quad\ \text{(b)} \quad \boldsymbol{M}\dot{\boldsymbol{v}}_i + \boldsymbol{C}(\boldsymbol{v}_i)\boldsymbol{v}_i + \boldsymbol{B}(\boldsymbol{v}_i)\boldsymbol{v}_i + \boldsymbol{G}(\boldsymbol{\eta}_i) = \boldsymbol{p}_i + \boldsymbol{\tau}_d \\ &\quad\ \text{(c)} \quad \dot{\boldsymbol{\eta}}_T = \boldsymbol{J}(\boldsymbol{\eta}_T)\boldsymbol{v}_T \\ &\quad\ \text{(d)} \quad \boldsymbol{M}\dot{\boldsymbol{v}}_T + \boldsymbol{C}(\boldsymbol{v}_T)\boldsymbol{v}_T + \boldsymbol{B}(\boldsymbol{v}_T)\boldsymbol{v}_T + \boldsymbol{G}(\boldsymbol{\eta}_T) \\ &\qquad\quad = \boldsymbol{p}_T + \boldsymbol{\tau}_d \\ &\quad\ \text{(e)} \quad \dot{\boldsymbol{s}}(t) = \boldsymbol{F}_s\boldsymbol{s}(t-\delta) + \sum_{i=1}^{M}\boldsymbol{G}_{12i}\boldsymbol{p}_i(t) + \boldsymbol{G}_{21}\boldsymbol{q}(t) \end{aligned} \tag{17}$$

with the final condition $\boldsymbol{s}_f$ satisfying the terminal constraint (12), i.e., $\forall\ \|\boldsymbol{e}_i\| > R_1$ or $\exists\ \|\boldsymbol{e}_i\| < R_2$.

III. Solution Technique

In this section, we exploit the Hamiltonian function to obtain feedback control strategies of target hunting task. Moreover, the MATD3 approach is further applied to analyze impacts of propagation delay and underwater disturbances.

A. Optimal Control Policies for Underwater Target Hunting Task Without Acoustic Propagation Delay

Let $V_E(\boldsymbol{s}_0)$ stand for the equilibrium value of the differential game beginning with the initial state $\boldsymbol{s}_0$. Furthermore, the equilibrium control strategies $\boldsymbol{p}^*$ and $\boldsymbol{q}^*$ can be expressed as follows:

$$\boldsymbol{p}^*,\ \boldsymbol{q}^* = \arg\min_{\boldsymbol{p}}\max_{\boldsymbol{q}}\{P_E(\boldsymbol{p},\boldsymbol{q},\boldsymbol{s}_0)\}. \tag{18}$$

According to the feedback strategies, the cost $V_E(\boldsymbol{s}(t))$ of the policy pair $(\boldsymbol{p},\boldsymbol{q})$ can be formulated as follows [24]:

$$\begin{aligned} &V_E(\boldsymbol{s}(t)) \\ &= \sum_{i=1}^{M}\left\{\frac{1}{2}\int_t^{T_h}\boldsymbol{p}_i^{\mathrm{T}}\left(\alpha_i^d g_i^d(t)+\beta_i^c g_i^c(t)\right)\boldsymbol{p}_i \mathrm{d}t \right. \\ &\quad \left. -\int_t^{T_h}\boldsymbol{q}^{\mathrm{T}}\left(\frac{1/2}{\|\boldsymbol{U}_i(t)-\boldsymbol{T}_l(t)\|^2}\right)\boldsymbol{q}\mathrm{d}t - \boldsymbol{s}_f^{\mathrm{T}}\phi_i(\boldsymbol{s}_f)\boldsymbol{s}_f\right\}. \end{aligned} \tag{19}$$

Since the value of $V_E(\boldsymbol{s}(t))$ is finite, the Hamiltonian function can be constructed with Leibniz's formula, which is formulated as follows [7]:

$$\begin{aligned} 0 &= \nabla V_E^{\mathrm{T}}\cdot\dot{\boldsymbol{s}}(t) + \frac{1}{2}\sum_{i=1}^{M}\left\{\boldsymbol{p}_i^{\mathrm{T}}\left(\alpha_i^d g_i^d(t)+\beta_i^c g_i^c(t)\right)\boldsymbol{p}_i \right. \\ &\quad \left. -\boldsymbol{q}^{\mathrm{T}}\frac{1}{\|\boldsymbol{U}_i(t)-\boldsymbol{T}_l(t)\|^2}\boldsymbol{q}\right\} \\ &:= H_E(\boldsymbol{s},\boldsymbol{p},\boldsymbol{q},\triangledown V_E) \end{aligned} \tag{20}$$

where the gradient vector ∇V_E is defined as $\nabla V_E = \partial V_E/\partial\boldsymbol{s}$.

In addition, UUVs and the target seek optimal control policies to maximize or minimize the Hamiltonian function. Then, for each feedback policy pair $(\boldsymbol{p},\boldsymbol{q})$, the necessary condition in (17) is further defined as follows:

$$\boldsymbol{p}^*,\boldsymbol{q}^* = \arg\min_{\boldsymbol{p}}\max_{\boldsymbol{q}}\{H_E(\boldsymbol{s},\boldsymbol{p},\boldsymbol{q},\nabla V_E)\}. \tag{21}$$

Moreover, when reaching Nash equilibrium, the stationary conditions of the Hamiltonian function are formulated as follows:

$$\partial H_E/\partial\boldsymbol{s} = \boldsymbol{F}_s^{\mathrm{T}}\nabla V_E = -\nabla^2 V_E \tag{22a}$$

$$\partial H_E/\partial\nabla V_E = \boldsymbol{F}_s\boldsymbol{s} + \sum_{i=1}^{M}\boldsymbol{G}_{12i}\boldsymbol{p}_i + \boldsymbol{G}_{21}\boldsymbol{q} = \dot{\boldsymbol{s}} \tag{22b}$$

$$\frac{\partial H_E}{\partial\boldsymbol{p}_i} = \left[\alpha_i^d g_i^d + \beta_i^c g_i^c\right]\boldsymbol{p}_i + \boldsymbol{G}_{12_i}^{\mathrm{T}}\nabla V_E^* = 0 \tag{22c}$$

$$\frac{\partial H_E}{\partial\boldsymbol{q}} = -\sum_{i=1}^{M}\|\boldsymbol{U}_i - \boldsymbol{T}_l\|^{-2}\boldsymbol{q} + \boldsymbol{G}_{21}^{\mathrm{T}}\nabla V_E^* = 0. \tag{22d}$$

Thus, we can obtain the optimal control laws by substituting (22c) and (22d) to the Hamiltonian function defined in (20), which are expressed as follows:

$$\begin{aligned} \boldsymbol{p}_i^*(\boldsymbol{s}) &= -\left[\alpha_i^d g_i^d(t)+\beta_i^c g_i^c(t)\right]^{-1}\boldsymbol{G}_{12_i}^{\mathrm{T}}(\boldsymbol{s})\nabla V_E^*(\boldsymbol{s}) \\ \boldsymbol{q}^*(\boldsymbol{s}) &= \frac{\boldsymbol{G}_{21}^{\mathrm{T}}(\boldsymbol{s})\nabla V_E^*(\boldsymbol{s})}{\sum_{i=1}^{M}\|\boldsymbol{U}_i(t)-\boldsymbol{T}_l(t)\|^{-2}}. \end{aligned} \tag{23}$$

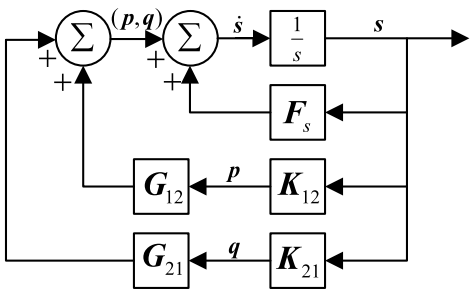


Fig. 2. Feedback control diagram.

Accordance with the assumption $\nabla V_E = \boldsymbol{P}\cdot\boldsymbol{s}$ ($\boldsymbol{P}\in R^{n\times n}$), the optimal policies in (23) is reformulated as follows:

$$\begin{aligned}\boldsymbol{p}_i^*(\boldsymbol{s}) &= -\left[\alpha_i^d g_i^d(t)+\beta_i^c g_i^c(t)\right]^{-1}\boldsymbol{G}_{12_i}^{\mathrm{T}}(s)\boldsymbol{P}\boldsymbol{s}\\ \boldsymbol{q}^*(\boldsymbol{s}) &= \frac{\boldsymbol{G}_{21}^{\mathrm{T}}(s)\boldsymbol{P}\boldsymbol{s}}{\sum_{i=1}^{M}\|\boldsymbol{U}_i(t)-\boldsymbol{T}_l(t)\|^{-2}}.\end{aligned}\tag{24}$$

Substituting $\nabla V_E = \boldsymbol{P}\cdot\boldsymbol{s}$ to the stationary condition (22a), we can obtain the equality $\boldsymbol{F}_s^{\mathrm{T}}\boldsymbol{P}\boldsymbol{s}+\dot{\boldsymbol{P}}\boldsymbol{s}+\boldsymbol{P}\dot{\boldsymbol{s}}=0$, where $\boldsymbol{P}$ is a symmetric matrix. Thus, Riccati equality can be gained by substituting (24)–(8), which is given by

$$\begin{aligned}\boldsymbol{F}_s^{\mathrm{T}}\boldsymbol{P}+\dot{\boldsymbol{P}}+\boldsymbol{P}\boldsymbol{F}_s-\boldsymbol{P}\sum_{i=1}^{M}\boldsymbol{G}_{12i}\frac{1}{\alpha_i^d g_i^d(t)+\beta_i^c g_i^c(t)}\boldsymbol{G}_{12_i}^{\mathrm{T}}\boldsymbol{P}\\ +\boldsymbol{P}\boldsymbol{G}_{21}\frac{\boldsymbol{G}_{21}^{\mathrm{T}}(s)\boldsymbol{P}}{\sum_{i=1}^{M}\|\boldsymbol{U}_i(t)-\boldsymbol{T}_l(t)\|^{-2}}=0.\end{aligned}\tag{25}$$

Specifically, $\boldsymbol{P}$ should satisfy the terminal state equation $\boldsymbol{P}_f=\boldsymbol{s}_f$ for $\forall\boldsymbol{s}$.

Accordingly, the feedback control policies [19] can be defined as $\boldsymbol{p}=\boldsymbol{K}_{12}\boldsymbol{s}$ and $\boldsymbol{q}=\boldsymbol{K}_{21}\boldsymbol{s}$, while feedback constraints $\boldsymbol{K}_{12}=[\boldsymbol{K}_{12}^{1\,\mathrm{T}},\boldsymbol{K}_{12}^{2\,\mathrm{T}},\ldots,\boldsymbol{K}_{12}^{M\,\mathrm{T}}]^{\mathrm{T}}$ and $\boldsymbol{K}_{21}$ presented in Fig. 2 satisfying

$$\begin{aligned}\boldsymbol{K}_{12}^i &= -\left[\alpha_i^d g_i^d(t)+\beta_i^c g_i^c(t)\right]^{-1}\boldsymbol{G}_{12_i}^{\mathrm{T}}(s)\boldsymbol{P}\\ \boldsymbol{K}_{21} &= \frac{\boldsymbol{G}_{21}^{\mathrm{T}}(s)\boldsymbol{P}}{\sum_{i=1}^{M}\|\boldsymbol{U}_i(t)-\boldsymbol{T}_l(t)\|^{-2}}.\end{aligned}\tag{26}$$

B. Analyses of Stability and Nash Equilibrium

Substituting (23) into (20), the following Hamiltonian equation is obtained:

$$\begin{aligned}H_E(\boldsymbol{s},\boldsymbol{p}^*,\boldsymbol{q}^*,\nabla V_E{}^*) &= \nabla V_E^{*\mathrm{T}}\boldsymbol{F}_s\boldsymbol{s}-\sum_{i=1}^{M}\frac{\nabla V_E^{*\mathrm{T}}\boldsymbol{G}_{12i}\boldsymbol{G}_{12_i}^{\mathrm{T}}\nabla V_E^*}{2\left[\alpha_i^d g_i^d(t)+\beta_i^c g_i^c(t)\right]}\\ &\quad+\frac{\nabla V_E^{*\mathrm{T}}\boldsymbol{G}_{21}\boldsymbol{G}_{21}^{\mathrm{T}}\nabla V_E^*}{2\sum_{i=1}^{M}\|\boldsymbol{U}_i(t)-\boldsymbol{T}_l(t)\|^{-2}}.\end{aligned}\tag{27}$$

Lemma 1: For any admissible control policies $\boldsymbol{p}$ and $\boldsymbol{q}$, the following equation holds:

$$\begin{aligned}&H_E(\boldsymbol{s},\boldsymbol{p},\boldsymbol{q},\nabla V_E)\\ &=H_E(\boldsymbol{s},\boldsymbol{p}^*,\boldsymbol{q}^*,\nabla V_E)\\ &\quad+\frac{\sqrt{2}}{2}\sum_{i=1}^{M}\Big\{\left[\alpha_i^d g_i^d+\beta_i^c g_i^c\right]\left(-\boldsymbol{p}_i^*+\boldsymbol{p}_i\right)^{\mathrm{T}}\left(-\boldsymbol{p}_i^*+\boldsymbol{p}_i\right)\\ &\qquad-\|\boldsymbol{U}_i(t)-\boldsymbol{T}_l(t)\|^{-2}\left(\boldsymbol{q}_i^*+\boldsymbol{q}\right)^{\mathrm{T}}\left(-\boldsymbol{q}_i^*+\boldsymbol{q}\right)\Big\}\\ &\quad-\frac{1}{2}\sum_{i=1}^{M}\Big\{\left[\alpha_i^d g_i^d+\beta_i^c g_i^c\right]\boldsymbol{p}_i^{*\mathrm{T}}\boldsymbol{p}_i-\|\boldsymbol{U}_i(t)-\boldsymbol{T}_l(t)\|^{-2}\boldsymbol{q}^{*\mathrm{T}}\boldsymbol{q}\Big\}\\ &\quad+\frac{1}{2}\sum_{i=1}^{M}\Big\{\left[\alpha_i^d g_i^d+\beta_i^c g_i^c\right]\boldsymbol{p}_i^{\mathrm{T}}\boldsymbol{p}_i^*-\|\boldsymbol{U}_i(t)-\boldsymbol{T}_l(t)\|^{-2}\boldsymbol{q}^{\mathrm{T}}\boldsymbol{q}^*\Big\}.\end{aligned}\tag{28}$$

Proof 1: See Appendix A.

Theorem 1: Let $V_E(\boldsymbol{s}(t))>0$ be a smooth function satisfying the Hamiltonian equation described in (28), and then, there exist the following.

1) System (8) is asymptotically stable in the mean with the control policies $\boldsymbol{p}$ and $\boldsymbol{q}$ derived in (23).
2) The policies $\boldsymbol{p}$ and $\boldsymbol{q}$ described in (23) provide a saddle point to the underwater differential game, and the system is in Nash equilibrium accordance with the optimal control policies.

Proof 2: See Appendix B.

Obviously, if only the coupled Hamilton equations can be solved, we will obtain the Nash equilibrium for target hunting system. However, due to the nonlinear nature of the Hamilton equation, it is complicated to obtaining its analytical solution. Therefore, we introduce MATD3-assisted algorithm to solve the differential game-based underwater target hunting problem.

C. MATD3-Assisted Algorithm for Underwater Target Hunting Task With Acoustic Propagation Delay

The connection between MARL and differential game has prompted majority research toward developing the optimal control policies online in time. Herein, we use MATD3 to save past information and pass back predictive impacts through target network [25]. Meanwhile, we jointly simulate the acoustic propagation delay and environmental distributions. As shown in Fig. 3, the MATD3 framework consists of the adversarial environment and n agents (M UUVs and a target), and each player has two stages: 1) the centralized training and 2) decentralized execution. Specifically, UUVs and the target execute actions through observation in a decentralized was. Agents collect experiences through interacting with the adversarial environment and then transmit the collected information to the central controller. After centralized training, the updated parameters are sent back to UUVs and the target, respectively. We first model our proposed underwater multitarget hunting task as a Markov decision process by a tuple $(\mathcal{S},\mathcal{A},\mathcal{R},\mathcal{P},\chi)$, where $\mathcal{S}$ denotes the state space, $\mathcal{A}$ represents the action space, $\mathcal{R}$ is defined as the reward space, $\mathcal{P}(\boldsymbol{s}(k+1)|\boldsymbol{s}(k),\boldsymbol{p}(k),\boldsymbol{q}(k))$ stands for the transition probability, and $\chi\in(0,1)$ is an MARL discount parameter, respectively. At time slot k, agents observe a state $\boldsymbol{s}(k)\in\mathcal{S}$ and take actions $[\boldsymbol{p}(k),\boldsymbol{q}(k)]\in\mathcal{A}$ based on certain policies, further producing a new state $\boldsymbol{s}(k+1)$ subject to the transition probability. The framework and key elements $\mathcal{S}$, $\mathcal{A}$, and $\mathcal{R}$ of MATD3 are provided as follows.

1) *MATD3 Framework:* As illustrated in Fig. 3, MATD3 framework is composed of the adversarial environment, M UUVs, and an escaping target, where each agent is implemented by twin-delayed deep deterministic policy gradient (TD3) method. Specifically, for each UUV or

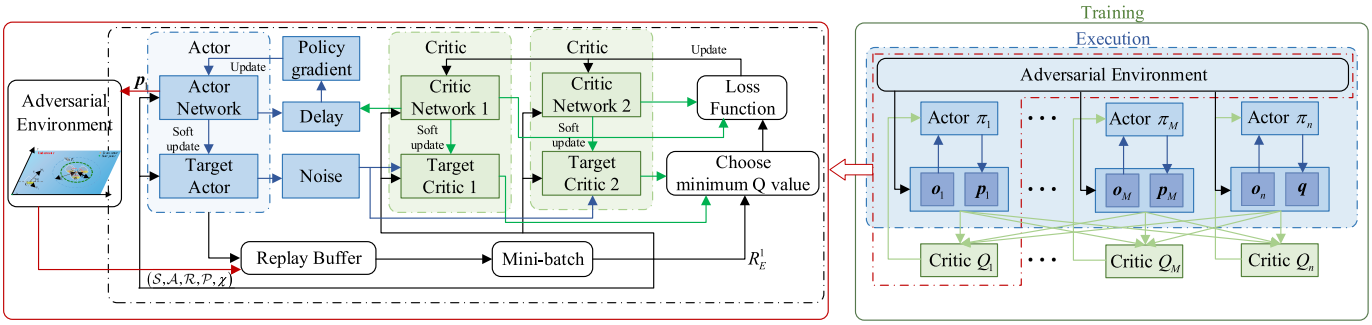


Fig. 3. MATD3 framework in underwater target hunting task.

target, the actor network and critic network are designed to satisfy policy function and Q function. On the one hand, the actor network treats environmental observation $\boldsymbol{o}_i$ as the input and action $\boldsymbol{p}_i$ (or $\boldsymbol{q}$) as the output. On the other hand, the critic network evaluates actor networks' outputs based on the observation $\boldsymbol{o}$ and control polices $\boldsymbol{p}$ and $\boldsymbol{q}$. Accordingly, these two networks enable the centralized training and decentralized execution.

2) *TD3 Algorithm:* In Fig. 3, the TD3 algorithm adopted by a UUV is illustrated in the left, which tackles the problem of overestimation caused by traditional DDPG. Specifically, TD3 considers the interplay between policy and value updates in the approximation errors of function and applies proximal policy optimization and soft actor–critic to address the overestimation bias problem.
3) *Action Space:* The action for each UUV or target is the output of its actor network that contains three accelerated velocities $\boldsymbol{p}_i(k) = [\dot{w}_i, \dot{v}_i, \dot{r}_i]$ or $\boldsymbol{q}(k) = [\dot{w}_T, \dot{v}_T, \dot{r}_T]$. The UUV chooses accelerated velocity to hunt the target, while the target correspondingly applies escaping policy. Meanwhile, the applied actions are limited by $\|\boldsymbol{v}_i\|_{\max} > \|\boldsymbol{v}_T\|_{\max}$ and $\Delta\psi_i < \Delta\psi_T$.
4) *State Space:* At each time slot k, UUVs or target in the core backbone network will make decisions according to the accepted state information $\mathcal{S} = (\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_n)$, which denotes the concatenation of environmental observation.
5) *Reward Function:* Each UUV or target selects an action from $\mathcal{A}$ at time slot k and then receives its reward according to $\mathcal{S}$, thus consisting of the reward space $\mathcal{R} = (\mathcal{R}_1, \mathcal{R}_2, \ldots, \mathcal{R}_M, \mathcal{R}_T)$. Furthermore, the reward function motivating the ith UUV to accomplish the target hunting task is negative to the payoff function defined in (14) and is formulated as follows:

$$\mathcal{R}_i = \begin{cases} -P_i(\boldsymbol{p}_i, \boldsymbol{q}, \boldsymbol{s}_0), & \|\boldsymbol{e}_i\| \in [R_2, R_1],\ i \in (0, M] \\ -a_f, & \|\boldsymbol{e}_i\| > R_1,\ i \in (0, M] \\ -b_f, & \|\boldsymbol{e}_i\| < R_2,\ i \in (0, M]. \end{cases} \tag{29}$$

Moreover, the reward function of the target can be expressed as $\sum_{i=1}^{M} \mathcal{R}_T^i$. Specifically, the reward function of the target related to the ith UUV is $\mathcal{R}_T^i$, which can be expressed as follows:

$$\mathcal{R}_T^i = \begin{cases} -P_T^i(\boldsymbol{p}_i, \boldsymbol{q}, \boldsymbol{s}_0), & \|\boldsymbol{e}_i\| \in [R_2, R_1],\ i \in (0, M] \\ -a, & \|\boldsymbol{e}_i\| > R_1,\ i \in (0, M] \\ -b, & \|\boldsymbol{e}_i\| < R_2,\ i \in (0, M]. \end{cases} \tag{30}$$

Herein, Q value (R_E^*) can be updated iteratively by conducting control strategies with the Bellman equation, which is provided by

$$R_E^* = \mathrm{E}_{s' \sim s}\left[\mathcal{R}(t) + \chi \max_{\boldsymbol{p}'} \min_{\boldsymbol{q}'} \mathcal{R}(\boldsymbol{s}', \boldsymbol{p}', \boldsymbol{q}') | \boldsymbol{s}(t-\delta), \boldsymbol{p}, \boldsymbol{q}\right] \tag{31}$$

where $0 \le \chi \le 1$ denotes the MARL discounting parameter to adjust the weights of future rewards. Besides, $\boldsymbol{s}'$, $\boldsymbol{p}'$, and $\boldsymbol{q}'$ represent the state and policies in the following time slot, respectively. Different from conventional MADDPG overestimating the max operator of the Bellman equation, our modified MATD3 approach formulates double Q learning with two critic networks [26]. The corresponding estimation R_{Ei}^* of the ith UUV is calculated by

$$\begin{aligned} R_{Ei}^* &= \chi \min\left\{ Q_i^{1'}(\boldsymbol{s}_1', \boldsymbol{s}_2', \ldots, \boldsymbol{s}_M', \boldsymbol{s}_n'; \boldsymbol{p}_1', \boldsymbol{p}_2', \ldots, \boldsymbol{p}_M', \boldsymbol{p}_n' | \boldsymbol{w}_i^{a'}), \right. \\ &\qquad \left. Q_i^{2'}(\boldsymbol{s}_1', \boldsymbol{s}_2', \ldots, \boldsymbol{s}_M', \boldsymbol{s}_n'; \boldsymbol{p}_1', \boldsymbol{p}_2', \ldots, \boldsymbol{p}_M', \boldsymbol{p}_n' | \boldsymbol{w}_i^{b'}) \right\} \\ &\quad + \mathcal{R}_i(t) \end{aligned} \tag{32}$$

where $\boldsymbol{w}_i^{a'}$ and $\boldsymbol{w}_i^{b'}$ denote the parameters of the pair of critic networks of the ith UUV, respectively. Moreover, $Q_i^{1'}$ and $Q_i^{2'}$ represent the corresponding outputs of the critic networks. Similarly, the corresponding estimation R_{ET}^* of the escaping target is calculated by

$$\begin{aligned} R_{ET}^* &= \chi \min\left\{ Q_n^{1'}(\boldsymbol{o}_1', \boldsymbol{o}_2', \ldots, \boldsymbol{o}_M', \boldsymbol{o}_n'; \boldsymbol{p}_1', \boldsymbol{p}_2', \ldots, \boldsymbol{p}_M', \boldsymbol{p}_n' | \boldsymbol{w}_n^{a'}), \right. \\ &\qquad \left. Q_n^{2'}(\boldsymbol{o}_1', \boldsymbol{o}_2', \ldots, \boldsymbol{o}_M', \boldsymbol{o}_n'; \boldsymbol{p}_1', \boldsymbol{p}_2', \ldots, \boldsymbol{p}_M', \boldsymbol{p}_n' | \boldsymbol{w}_n^{b'}) \right\} \\ &\quad + \mathcal{R}_T(t) \end{aligned} \tag{33}$$

where $\boldsymbol{w}_T^{a'}$ and $\boldsymbol{w}_T^{b'}$ denote parameters of the pair of critic networks of the escaping target, respectively. Moreover, $Q_T^{1'}$

Algorithm 1 MATD3-Based Algorithm for Underwater Target Hunting Task

Input: Admissible control policy $(\boldsymbol{p}_0, \boldsymbol{q}_0)$, initial state $\boldsymbol{s}_0$, disturbances $\boldsymbol{\tau}_d$, $V_E = 0$, time slot $k = 0$
Output: $\dot{\boldsymbol{s}}(k)$

```
for episode = 1, 2, ..., 4000 do
    for k = 1, 2, ..., k_max do
        while not ∀‖e_i(k)‖ > R_1 or ∃‖e_i(k)‖ < R_2 do
            Calculate the underwater acoustic
             propagation delay ⌊δ⌋ accroading to (7);
            if k − ⌊δ⌋ ≤ t_0 then
                Applying randomly stratigies with
                 initial settings;
            else
                Update control policies p(k), q(k) with
                 the past state s(k − ⌊δ⌋);
                Calculate the reward R_i based on (29)
                 or the reward R_T based on (30);
                Update Q-value R*_Ei based on (32) or
                 update Q-value R*_ET based on (33);
                Update the parameters of the network;
            end
        end
        k ← k + 1
    end
end
```

and $Q_T^{2'}$ represent the corresponding outputs of the critic networks. Algorithm 1 shows the updating strategies of varied time slots to complete underwater target hunting task with acoustic propagation delay and environmental disturbance.

IV. Simulation Results

In our simulations, results demonstrate the performance of underwater differential game model proposed in Section II for target hunting. More specifically, $M = 4$ UUVs are assigned around $(0, 0, -200)$ m initially to hunt the adversarial target by the cooperation of the clustering team and the Markov process. Nash equilibrium is reached by the MATD3 scheme. Moreover, the target distributes 100 m away from UUV team, i.e., $\|\boldsymbol{T}_l - \boldsymbol{O}\| = 100$ m. At the beginning, UUVs and target start at the speed of 8.5[1] and 3 knot, respectively. The speed of UUVs varies between 4 and 15 knot, while the speed of the target ranges from 1 to 5 knot. In addition, the heading angle of UUVs during each time slot has a range of $(0, \pi/2)$, while the heading angle of the target has a range of $(0, \pi)$.

The dynamics of UUVs and the target in (1)–(5) can be rewritten as the following equations:

$$\begin{aligned}
\dot{\boldsymbol{\eta}}_i &= \boldsymbol{J}(\boldsymbol{\eta}_i)\boldsymbol{v}_i \\
\dot{\boldsymbol{\eta}}_T &= \boldsymbol{J}(\boldsymbol{\eta}_T)\boldsymbol{v}_T \\
\dot{\boldsymbol{v}}_i &= -\boldsymbol{M}^{-1}\left[\boldsymbol{C}\boldsymbol{v}_i + \boldsymbol{B}\boldsymbol{v}_i + \boldsymbol{G}\boldsymbol{v}_i - \boldsymbol{p}_i\right] + \boldsymbol{M}^{-1}\boldsymbol{\tau}_d \\
\dot{\boldsymbol{v}}_T &= -\boldsymbol{M}^{-1}\left[\boldsymbol{C}\boldsymbol{v}_T + \boldsymbol{B}\boldsymbol{v}_T + \boldsymbol{G}\boldsymbol{v}_T - \boldsymbol{p}_T\right] + \boldsymbol{M}^{-1}\boldsymbol{\tau}_d.
\end{aligned}$$

[1] 1 knot = 1.852 km/h.

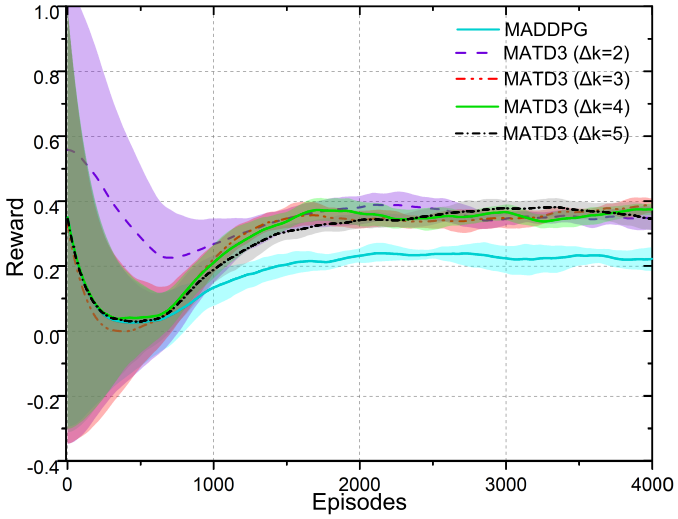


Fig. 4. Average reward comparison among the MADDPG and MATD3s with different updating frequencies without considering acoustic propagation delay.

Furthermore, we can obtain an approximative discrete-time model based on the first-order Taylor expansion, considering MARL is only applicable for discrete-time system [27], which can be expressed as follows:

$$\begin{aligned}
\boldsymbol{\eta}_i^{k+1} &= \boldsymbol{\eta}_i^k + \Delta k \boldsymbol{J}\left(\boldsymbol{\eta}_i^k\right)\boldsymbol{v}_i^k \\
\boldsymbol{\eta}_T^{k+1} &= \boldsymbol{\eta}_T^k + \Delta k \boldsymbol{J}\left(\boldsymbol{\eta}_T^k\right)\boldsymbol{v}_T^k \\
\boldsymbol{v}_i^{k+1} &= \boldsymbol{v}_i^k - \Delta k \boldsymbol{M}^{-1}\left[(\boldsymbol{C} + \boldsymbol{B} + \boldsymbol{G})\boldsymbol{v}_i - \boldsymbol{p}_i\right] + \boldsymbol{M}^{-1}\boldsymbol{\tau}_d^k \\
\boldsymbol{v}_T^{k+1} &= \boldsymbol{v}_T^k - \Delta k \boldsymbol{M}^{-1}\left[(\boldsymbol{C} + \boldsymbol{B} + \boldsymbol{G})\boldsymbol{v}_T - \boldsymbol{p}_T\right] + \boldsymbol{M}^{-1}\boldsymbol{\tau}_d^k
\end{aligned}$$

where $\boldsymbol{\eta}_i^k$, $\boldsymbol{v}_i^k$, $\boldsymbol{\eta}_T^k$, $\boldsymbol{v}_T^k$, and $\boldsymbol{\tau}_d(k)$ are the sampled values of $\boldsymbol{\eta}_i$, $\boldsymbol{v}_i$, $\boldsymbol{\eta}_T$, $\boldsymbol{v}_T$, and $\boldsymbol{\tau}_d$ at the kth sampling time, respectively.

Moreover, the safe radius r, the sensing radius of UUV (R_1), and the radius of attacking scope (R_2) are initialed with 3, 200, and 15 m, respectively. Moreover, considering that the target hunting task should be completed within finite time, we define the maximum time slots of one episode as 1500. In addition, constraints a, b, and l are set to -0.01, 1, and 0.5, respectively. In the simulations, MADDPG is used as the benchmark method to verify the effectiveness of our proposed MATD3 scheme. Furthermore, we implement the MATD3/MADDPG algorithm with PyTorch and execute 4000 trining episodes to test the performance, where the structure of actor network or critic network is constructed with a fully connected neural network including two hidden layers to stand for the policy function or the Q function for each UUV. As to both Q function and policy networks, hidden layers are activated by the rectified linear unit (ReLU) function. Meanwhile, the tanh activating function is used to limit outputs by activating the output layer of the policy network. Then, they can be subject to dynamic properties of the UUV or target. Moreover, a linear function is used to activate the output layer of the Q network. Herein, the inputs of the above networks are normalized to $[-1, 1]$ to equally capture changes with varied input features. Specifically, Table I depicts system's parameters and initial settings of MATD3/MADDPG.

TABLE I
PARAMETERS OF ALGORITHMS AND SYSTEM

	Parameters	Values
MATD3/ MADDPG parameters	Time interval between each step	0.3 s
	Learning rate of actor network	1.5×10^{-6}
	Learning rate of critic network	3×10^{-6}
	Number of training episodes (ℓ)	4×10^{3}
	Discounting factor (χ)	0.99
	Batch size	32
	Memory capacity	1×10^{6}
	Target network update speed	4×10^{-3}
	Maximum number of time slots ($k_{\max}$)	1.5×10^{3}
System parameters	Start point of UUVs ($\boldsymbol{O}$)	(0, 0, -200) m
	Number of UUVs (M)	4
	Initial distance $\|\boldsymbol{T}_l - \boldsymbol{O}\|$	100 m
	Range of UUV's speed	5-15 knot
	Range of target's speed	1-5 knot
	The maximum speed of UUVs (V_1)	15 knot
	The maximum speed of target (V_2)	5 knot
	Movement range of UUV ($\Delta\psi_i$)	$[0, \pi/2]$ rad/s
	Movement range of target ($\Delta\psi_T$)	$[0, \pi]$ rad/s
	Radius of UUV's safe area (r)	3 m
	Radius of UUV's sensing area (R_1)	200 m
	Radius of UUV's attacking area (R_2)	15 m
	UUV's initial speed (V_G)	8.5 knot
	Target's initial speed (V_t)	3 knot
	Constraint ratio ($\alpha_i^d : \beta_i^c$)	2.2
	Constraints (a, b, l)	0.01, -1, 0.5

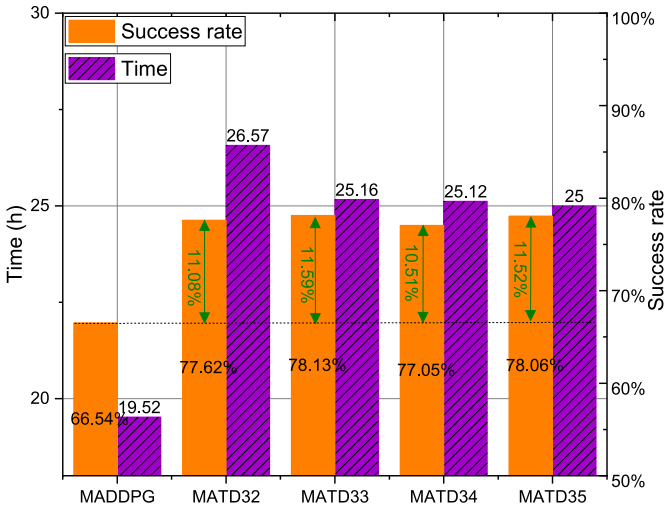


Fig. 5. Success rate and time consumption among the MADDPG and MATD3s with different updating frequencies without considering acoustic propagation delay.

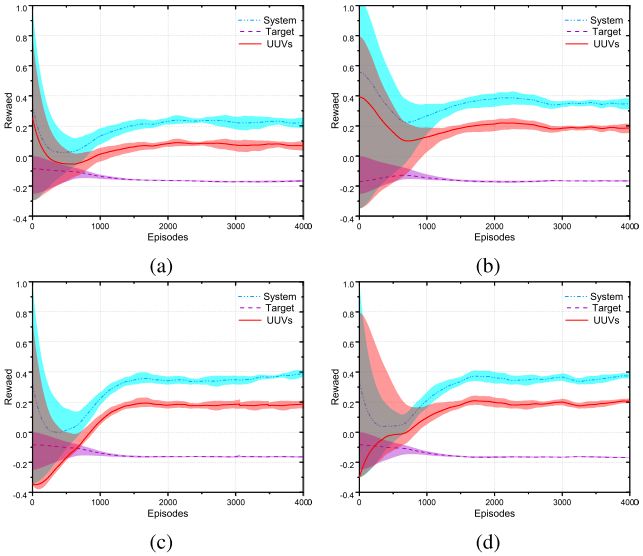


Fig. 6. Comparison of average system reward, UUVs' reward, and target reward without considering acoustic propagation delay. (a) MADDPG. (b) MATD3 ($\Delta k = 2$). (c) MATD3 ($\Delta k = 3$). (d) MATD3 ($\Delta k = 4$).

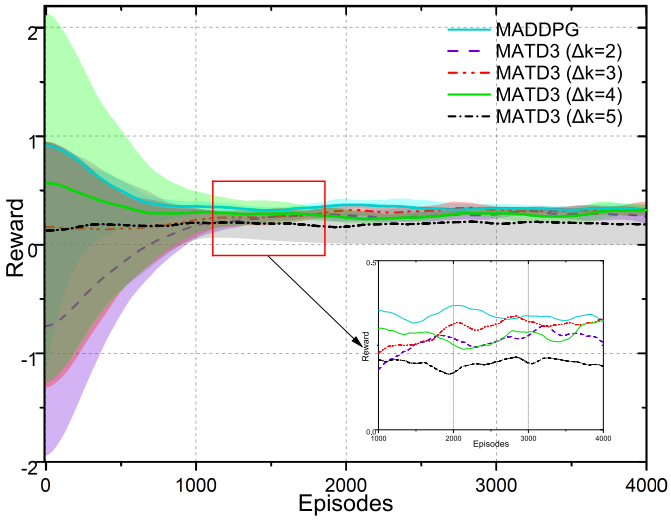


Fig. 7. Average reward comparison among the MADDPG and MATD3s with different updating frequencies considering acoustic propagation delay.

In this article, we define a successful target hunting task as occurring when the target enters the attacking area of at least one of the UUVs. We measure the success rate of the target hunting task as the ratio of successful experiments out of 4000. Figs. 4–6 simulate underwater target hunting game without considering the acoustic propagation delay using the proposed MATD3 algorithm (the network updating frequency ranging from every two time slots to every five time slots), compared with the common MADDPG algorithm. Fig. 4 shows all the reward curves reach Nash equilibrium after converged. Specifically, the area enclosed by some color represents the fluctuation range of rewards of multiple experiments of the same type. The corresponding curve of the same color stands for the average reward curve of multiple experiments. It can be seen that, under the same experimental setting, the fluctuation range of reward functions gradually narrows, and the average reward curve converges to an intermediate reward value after about 1500 episodes between the maximized value $R_E(\boldsymbol{p}^*, \boldsymbol{q}, \boldsymbol{s}_0)$ and the minimized value $R_E(\boldsymbol{p}, \boldsymbol{q}^*, \boldsymbol{s}_0)$, and it further satisfies the condition $P_E(\boldsymbol{p}, \boldsymbol{q}^*, \boldsymbol{s}_0) \geq P_E(\boldsymbol{p}^*, \boldsymbol{q}^*, \boldsymbol{s}_0) = P_E^*(\boldsymbol{s}_0) \geq P_E(\boldsymbol{p}^*, \boldsymbol{q}, \boldsymbol{s}_0)$. Moreover, coping with the overestimating problem of MADDPG, the proposed MATD3 gains higher system reward than the conventional MADDPG. We also draw the conclusion from Fig. 5 that the proposed MATD3 has a prompting in success rate (more than 10%) conducting the adversarial target hunting game at the cost of more time consumption compared with the traditional MADDPG method. Specifically, the training time of MADDPG is shorter than the training time of MATD3, which is in line with the characteristics of better fitting of reinforcement learning with longer training time.

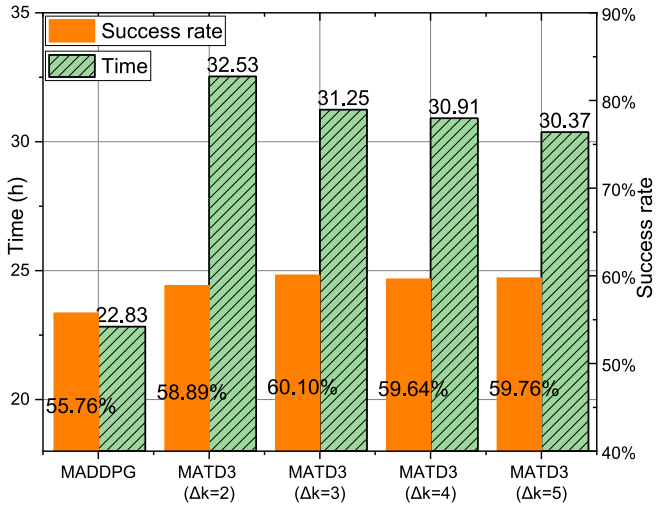


Fig. 8. Success rate and time consumption among the MADDPG and MATD3s with different updating frequencies considering communication delay.

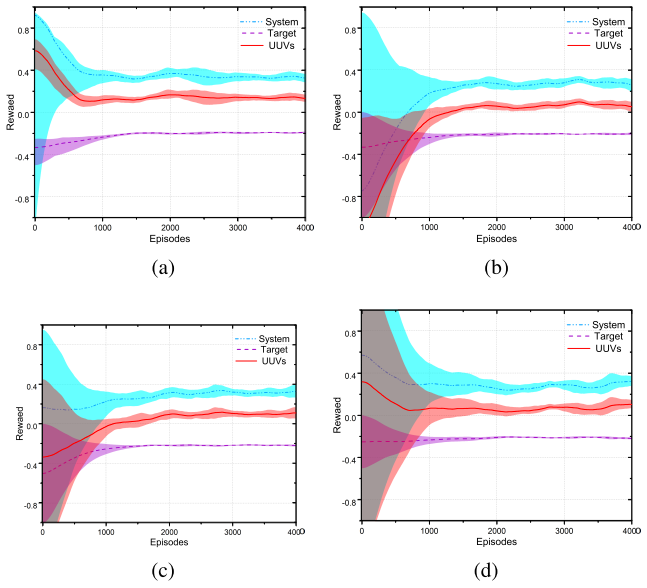


Fig. 9. Comparison of average system reward, UUVs' reward, and target reward considering communication delay. (a) MADDPG. (b) MATD3 ($\Delta k = 2$). (c) MATD3 ($\Delta k = 3$). (d) MATD3 ($\Delta k = 4$).

Moreover, MATD3 has a high success rate when hunting the high intelligent target, which represents the rationality of our game-based target hunting model and the design of the reward functions. In summary, both the high success rate and the long training time show the superiority of our proposed MATD3. Furthermore, Fig. 6 depicts the average system reward, UUVs' reward, and target reward, respectively. Results validate all the system components converge in the adversarial environment, the UUVs' reward is maximized globally, and the target is minimized correspondingly.

Figs. 7–9 simulate underwater target hunting game considering the acoustic propagation delay using the proposed MATD3 algorithm, compared with the common MADDPG algorithm. Fig. 7 shows all the reward curves reach Nash equilibrium after converged even affected by the noncausality of control strategies caused by acoustic propagation delay between UUVs and the target. Compared with Fig. 4, results also indicate that responses of convergence perform hysteresis property due to the acoustic propagation delay. We also draw the conclusion from Fig. 8 that the proposed MATD3 has a promotion in success rate (more than 5% than MADDPG) conducting the adversarial target hunting game. Compared with Fig. 5, the time consumption is larger, since MARL methods need more time to learn the lagging information exchange caused by underwater acoustic propagation delay. Furthermore, Fig. 9 depicts the average system reward UUVs' reward and target reward considering underwater acoustic propagation delay.

V. Conclusion

This article proposes an underwater target hunting differential game to analyze adversarial behaviors of agents and swarm team's cooperation strategies. In addition, we conceive the Hamiltonian function to investigate the optimal policies of UUVs accordance with the aim of minimizing the system's payoff function. Meanwhile, the feedback control policies have been successfully solved, and the analysis shows that the target hunting system is asymptotically stable on average. In addition, the system can achieve Nash equilibrium by relying on the control policies that have been obtained. Furthermore, the proposed MATD3-assisted method analyzes the impact of acoustic propagation delay and environmental disturbances on target hunting system. Simulation results reveal that the proposed scheme is superior to the typical MADDPG algorithms in terms of the reward and success rate. The method proposed in this article is only suitable for tracking and hunting a single target in underwater environment. The future work can focus on the multitarget hunting task, utilizing an auction mechanism and mean-field game theory as the basis for the approach.

Appendix A
Proof of Lemma 1

Proof: Applying the state function (8) between UUVs and the target to the defined Hamiltonian function, (20) can be reformulated as follows:

$$\begin{aligned} &H_E(\boldsymbol{s}, \boldsymbol{p}, \boldsymbol{q}, \nabla V_E) \\ &= \nabla V_E^{\mathrm{T}} \cdot \boldsymbol{F}_s \boldsymbol{s} + \sum_{i=1}^{M} \left\{ \frac{\boldsymbol{p}_i^{\mathrm{T}} \left[\alpha_i^d g_i^d + \beta_i^c g_i^c\right] \boldsymbol{p}_i}{2} + \nabla V_E^{\mathrm{T}} \cdot \boldsymbol{G}_{12}^i \boldsymbol{p}_i \right\} \\ &\quad - \frac{1}{2} \sum_{i=1}^{M} \boldsymbol{q}^{\mathrm{T}} \frac{1}{\|\boldsymbol{U}_i - \boldsymbol{T}_l\|^2} \boldsymbol{q} + \nabla V_E^{\mathrm{T}} \cdot \boldsymbol{G}_{21} \boldsymbol{q}. \end{aligned} \tag{34}$$

Substituting (27) into (34), the Hamiltonian equation is further expressed as follows:

$$\begin{aligned} &H_E(\boldsymbol{s}, \boldsymbol{p}, \boldsymbol{q}, \nabla V_E) \\ &\quad = H_E(\boldsymbol{s}, \boldsymbol{p}^*, \boldsymbol{q}^*, \nabla V_E) \\ &\qquad + \sum_{i=1}^{M} \frac{\nabla V_E{}^{\mathrm{T}} \boldsymbol{G}_{12i} \boldsymbol{G}_{12_i}{}^{\mathrm{T}} \nabla V_E}{2\left[\alpha_i^d g_i^d + \beta_i^c g_i^c\right]} - \frac{\nabla V_E{}^{\mathrm{T}} \boldsymbol{G}_{21} \boldsymbol{G}_{21}^{\mathrm{T}} \nabla V_E}{2 \sum_{i=1}^{M} \|\boldsymbol{U}_i - \boldsymbol{T}_l\|^{-2}} \\ &\qquad + \sum_{i=1}^{M} \left\{ \frac{\boldsymbol{p}_i^{\mathrm{T}} \left(\alpha_i^d g_i^d + \beta_i^c g_i^c\right) \boldsymbol{p}_i}{2} + \nabla V_E^{\mathrm{T}} \cdot \boldsymbol{G}_{12i} \boldsymbol{p}_i \right\} \end{aligned}$$

$$
\begin{aligned}
&-\frac{1}{2}\sum_{i=1}^{M}\boldsymbol{q}^{\mathrm{T}}\frac{1}{\|\boldsymbol{U}_i-\boldsymbol{T}_l\|^2}\boldsymbol{q}+\nabla V_E^{\mathrm{T}}\cdot\boldsymbol{G}_{21}\boldsymbol{q}\\
&=H_E(\boldsymbol{s},\boldsymbol{p}^*,\boldsymbol{q}^*,\triangledown V_E)\\
&\quad+\frac{\sqrt{2}}{2}\sum_{i=1}^{M}\Big\{\big[\alpha_i^d g_i^d+\beta_i^c g_i^c\big]\big(\boldsymbol{p}_i-\boldsymbol{p}_i^*\big)^{\mathrm{T}}\big(\boldsymbol{p}_i-\boldsymbol{p}_i^*\big)\\
&\qquad\qquad-\|\boldsymbol{U}_i-\boldsymbol{T}_l\|^{-2}\big(\boldsymbol{q}^*+\boldsymbol{q}\big)^{\mathrm{T}}\big(\boldsymbol{q}-\boldsymbol{q}^*\big)\Big\}\\
&\quad-\frac{1}{2}\sum_{i=1}^{M}\Big\{\big[\alpha_i^d g_i^d+\beta_i^c g_i^c\big]\boldsymbol{p}_i^{*\mathrm{T}}\boldsymbol{p}_i-\|\boldsymbol{U}_i-\boldsymbol{T}_l\|^{-2}\boldsymbol{q}^{*\mathrm{T}}\boldsymbol{q}\Big\}\\
&\quad+\frac{1}{2}\sum_{i=1}^{M}\Big\{\big[\alpha_i^d g_i^d+\beta_i^c g_i^c\big]\boldsymbol{p}_i^{\mathrm{T}}\boldsymbol{p}_i^*-\|\boldsymbol{U}_i-\boldsymbol{T}_l\|^{-2}\boldsymbol{q}^{\mathrm{T}}\boldsymbol{q}^*\Big\}.
\end{aligned}
\tag{35}
$$

APPENDIX B
PROOF OF THEOREM 1

A. Stability

Here, we choose the Lyapunov function candidate $V_E(\boldsymbol{s}(t))$, and there exists $V_E(\boldsymbol{s}(t))=0$ if and only if $\boldsymbol{s}_e=\boldsymbol{0}$. Thus, the Lyapunov function candidate $V_E(\boldsymbol{s}(t))$ is expressed as follows [28]:

$$
\begin{aligned}
V_E(\boldsymbol{s}(t))=\frac{1}{2}\sum_{i=1}^{M}\Big\{&\int_t^{T_h}\boldsymbol{p}_i^{\mathrm{T}}\big[\alpha_i^d g_i^d+\beta_i^c g_i^c\big]\boldsymbol{p}_i\mathrm{d}t\\
&-\int_t^{T_h}\frac{\boldsymbol{q}^{\mathrm{T}}\boldsymbol{q}}{\|\boldsymbol{U}_i-\boldsymbol{T}_l\|^2}\mathrm{d}t-\boldsymbol{s}_f^{\mathrm{T}}\phi_i(\boldsymbol{s}_f)\boldsymbol{s}_f\Big\}.
\end{aligned}
\tag{36}
$$

Therefore, we denote the derivation of $V_E(\boldsymbol{s}(t))$ with respect to time, which is denoted by

$$
\begin{aligned}
\dot{V}_E&=\frac{\partial V_E}{\partial\boldsymbol{s}}\dot{\boldsymbol{s}}=\nabla V_E\bigg(\boldsymbol{F}_s\boldsymbol{s}+\sum_{i=1}^{M}\boldsymbol{G}_{12i}\boldsymbol{p}_i+\boldsymbol{G}_{21}\boldsymbol{q}\bigg)\\
&=H_E(\boldsymbol{s},\boldsymbol{p},\boldsymbol{q},\triangledown V_E)\\
&\quad-\frac{1}{2}\sum_{i=1}^{M}\Big\{\big[\alpha_i^d g_i^d+\beta_i^c g_i^c\big]\boldsymbol{p}_i^{\mathrm{T}}\boldsymbol{p}_i-\frac{\boldsymbol{q}^{\mathrm{T}}\boldsymbol{q}}{\|\boldsymbol{U}_i-\boldsymbol{T}_l\|^2}\Big\}.
\end{aligned}
\tag{37}
$$

Applying (35) to (37), $\dot{V}_E$ can be reformulated as follows:

$$
\begin{aligned}
\dot{V}_E&=H_E(\boldsymbol{s},\boldsymbol{p}^*,\boldsymbol{q}^*,\triangledown V_E)\\
&\quad+\frac{\sqrt{2}}{2}\sum_{i=1}^{M}\Big\{\big[\alpha_i^d g_i^d+\beta_i^c g_i^c\big]\big(\boldsymbol{p}_i-\boldsymbol{p}_i^*\big)^{\mathrm{T}}\big(\boldsymbol{p}_i-\boldsymbol{p}_i^*\big)\\
&\qquad\qquad-\|\boldsymbol{U}_i(t)-\boldsymbol{T}_l(t)\|^{-2}\big(\boldsymbol{q}^*+\boldsymbol{q}\big)^{\mathrm{T}}\big(\boldsymbol{q}-\boldsymbol{q}^*\big)\Big\}\\
&\quad-\frac{1}{2}\sum_{i=1}^{M}\Big\{\big[\alpha_i^d g_i^d+\beta_i^c g_i^c\big]\boldsymbol{p}_i^{*\mathrm{T}}\boldsymbol{p}_i-\|\boldsymbol{U}_i-\boldsymbol{T}_l\|^{-2}\boldsymbol{q}^{*\mathrm{T}}\boldsymbol{q}\Big\}\\
&\quad+\frac{1}{2}\sum_{i=1}^{M}\Big\{\big[\alpha_i^d g_i^d+\beta_i^c g_i^c\big]\boldsymbol{p}_i^{\mathrm{T}}\boldsymbol{p}_i^*-\|\boldsymbol{U}_i-\boldsymbol{T}_l\|^{-2}\boldsymbol{q}^{\mathrm{T}}\boldsymbol{q}^*\Big\}\\
&\quad-\frac{1}{2}\sum_{i=1}^{M}\Big\{\big[\alpha_i^d g_i^d+\beta_i^c g_i^c\big]\boldsymbol{p}_i^{\mathrm{T}}\boldsymbol{p}_i-\frac{\boldsymbol{q}^{\mathrm{T}}\boldsymbol{q}}{\|\boldsymbol{U}_i-\boldsymbol{T}_l\|^2}\Big\}.
\end{aligned}
\tag{38}
$$

Since $(\|\boldsymbol{U}_i(t)-\boldsymbol{U}_j(t)\|^2-r^2)^l\ll\|\boldsymbol{U}_i-\boldsymbol{T}_l\|^2$, the equation with $\boldsymbol{p}=\boldsymbol{p}^*$ and $\boldsymbol{q}=\boldsymbol{q}^*$ in (38) is reformulated as follows:

$$
\begin{aligned}
\dot{V}_E&=-\frac{1}{2}\sum_{i=1}^{M}\Big\{\big[\alpha_i^d g_i^d+\beta_i^c g_i^c\big]\boldsymbol{p}_i^{\mathrm{T}}\boldsymbol{p}_i-\frac{\boldsymbol{q}^{\mathrm{T}}\boldsymbol{q}}{\|\boldsymbol{U}_i-\boldsymbol{T}_l\|^2}\Big\}\\
&=-\frac{1}{2}\sum_{i=1}^{M}\Bigg\{\sum_{j=1,j\neq i}^{M}\Bigg\{\frac{\alpha_i^d}{\big(\|\boldsymbol{U}_i(t)-\boldsymbol{U}_j(t)\|^2-r^2\big)^l}\\
&\qquad\qquad+\beta_i^c\|\boldsymbol{U}_i(t)-\boldsymbol{U}_j(t)\|^2\Bigg\}\boldsymbol{p}_i^{\mathrm{T}}\boldsymbol{p}_i\\
&\qquad-\frac{\boldsymbol{q}^{\mathrm{T}}\boldsymbol{q}}{\|\boldsymbol{U}_i-\boldsymbol{T}_l\|^2}\Bigg\}\le 0.
\end{aligned}
\tag{39}
$$

Therefore, $V_E(\boldsymbol{s}(t))$ is a Lyapunov function for $\boldsymbol{s}$, and the system described in (8) is asymptotically stable in the mean with the final state satisfying $\boldsymbol{s}_f=\epsilon$ (ϵ locates in the neighborhood of 0).

B. Nash Equilibrium

Therefore, along with the condition $\boldsymbol{s}_f=\epsilon$ (ϵ locates in the neighborhood of 0), the payoff function in (14) can be rewritten as follows:

$$
\begin{aligned}
&P_E(\boldsymbol{p},\boldsymbol{q},\boldsymbol{s}_0)\\
&=\frac{1}{2}\sum_{i=1}^{M}\int_0^{T_h}\Big\{\boldsymbol{p}_i^{\mathrm{T}}\big[\alpha_i^d g_i^d+\beta_i^c g_i^c\big]\boldsymbol{p}_i-\frac{\boldsymbol{q}^{\mathrm{T}}\boldsymbol{q}}{\|\boldsymbol{U}_i-\boldsymbol{T}_l\|^2}\Big\}\mathrm{d}t\\
&\quad+V(\boldsymbol{s}_0)+\int_0^{T_h}\dot{V}\mathrm{d}t\\
&=\frac{1}{2}\sum_{i=1}^{M}\int_0^{T_h}\Big\{\boldsymbol{p}_i^{\mathrm{T}}\big[\alpha_i^d g_i^d+\beta_i^c g_i^c\big]\boldsymbol{p}_i-\frac{\boldsymbol{q}^{\mathrm{T}}\boldsymbol{q}}{\|\boldsymbol{U}_i-\boldsymbol{T}_l\|^2}\Big\}\mathrm{d}t\\
&\quad+V(\boldsymbol{s}_0)+\int_0^{T_h}\nabla V_E\bigg(\boldsymbol{F}_s\boldsymbol{s}+\sum_{i=1}^{M}\boldsymbol{G}_{12i}\boldsymbol{p}_i+\boldsymbol{G}_{21}\boldsymbol{q}\bigg)\mathrm{d}t.
\end{aligned}
\tag{40}
$$

By combining Lemma 1 and (20), the payoff function can be further expressed as follows:

$$
\begin{aligned}
&P_E(\boldsymbol{p},\boldsymbol{q},\boldsymbol{s}_0)\\
&=V(\boldsymbol{s}_0)\\
&\quad+\int_0^{T_h}\Bigg\{\frac{\sqrt{2}}{2}\sum_{i=1}^{M}\big[\alpha_i^d g_i^d+\beta_i^c g_i^c\big]\big(\boldsymbol{p}_i-\boldsymbol{p}_i^*\big)^{\mathrm{T}}\big(\boldsymbol{p}_i-\boldsymbol{p}_i^*\big)\\
&\qquad-\frac{\sqrt{2}}{2}\sum_{i=1}^{M}\|\boldsymbol{U}_i-\boldsymbol{T}_l\|^{-2}\big(\boldsymbol{q}^*+\boldsymbol{q}\big)^{\mathrm{T}}\big(\boldsymbol{q}-\boldsymbol{q}^*\big)\\
&\qquad-\frac{1}{2}\sum_{i=1}^{M}\Big(\big[\alpha_i^d g_i^d+\beta_i^c g_i^c\big]\boldsymbol{p}_i^{*\mathrm{T}}\boldsymbol{p}_i\\
&\qquad\qquad-\|\boldsymbol{U}_i-\boldsymbol{T}_l\|^{-2}\boldsymbol{q}^{*\mathrm{T}}\boldsymbol{q}\Big)\\
&\qquad+\frac{1}{2}\sum_{i=1}^{M}\Big(\big[\alpha_i^d g_i^d+\beta_i^c g_i^c\big]\boldsymbol{p}_i^{\mathrm{T}}\boldsymbol{p}_i^*\\
&\qquad\qquad-\|\boldsymbol{U}_i-\boldsymbol{T}_l\|^{-2}\boldsymbol{q}^{\mathrm{T}}\boldsymbol{q}^*\Big)\Bigg\}\mathrm{d}t
\end{aligned}
\tag{41}
$$

$$
\begin{aligned}
&P_E(\boldsymbol{p}^*, \boldsymbol{q}, \boldsymbol{s}_0) \\
&= V(\boldsymbol{s}_0) - \int_0^{T_h} \Bigg\{ \frac{\sqrt{2}}{2} \sum_{i=1}^{M} \|\boldsymbol{U}_i - \boldsymbol{T}_l\|^{-2} \big(\boldsymbol{q}^* + \boldsymbol{q}\big)^{\mathrm{T}} \big(\boldsymbol{q} - \boldsymbol{q}^*\big) \\
&\qquad - \frac{1}{2} \sum_{i=1}^{M} \|\boldsymbol{U}_i - \boldsymbol{T}_l\|^{-2} (\boldsymbol{q}^{*\mathrm{T}} \boldsymbol{q} - \boldsymbol{q}^{\mathrm{T}} \boldsymbol{q}^*) \Bigg\} \mathrm{d}t \\
&= V(\boldsymbol{s}_0) - \int_0^{T_h} \sum_{i=1}^{M} \|\boldsymbol{U}_i - \boldsymbol{T}_l\|^{-2} \\
&\quad \times \Bigg\{ \frac{\sqrt{2}}{2} \big(- \boldsymbol{q}^{*\mathrm{T}} \boldsymbol{q}^* + \boldsymbol{q}^{\mathrm{T}} \boldsymbol{q} \big) \\
&\qquad + \left(\frac{\sqrt{2}}{2} - \frac{1}{2} \right) (\boldsymbol{q}^{*\mathrm{T}} \boldsymbol{q} - \boldsymbol{q}^{\mathrm{T}} \boldsymbol{q}^*) \Bigg\} \mathrm{d}t \\
&= V(\boldsymbol{s}_0) - \int_0^{T_h} \sum_{i=1}^{M} \|\boldsymbol{U}_i - \boldsymbol{T}_l\|^{-2} \Bigg\{ \frac{\sqrt{2}}{2} \big(- \boldsymbol{q}^{*\mathrm{T}} \boldsymbol{q}^* + \boldsymbol{q}^{\mathrm{T}} \boldsymbol{q} \big) \Bigg\} \mathrm{d}t
\end{aligned} \tag{42}
$$

$$
\begin{aligned}
&P_E(\boldsymbol{p}, \boldsymbol{q}^*, \boldsymbol{s}_0) \\
&= V(\boldsymbol{s}_0) \\
&\quad + \int_0^{T_h} \Bigg\{ \frac{\sqrt{2}}{2} \sum_{i=1}^{M} \big[\alpha_i^d g_i^d + \beta_i^c g_i^c\big] \big(\boldsymbol{p}_i - \boldsymbol{p}_i^*\big)^{\mathrm{T}} \big(\boldsymbol{p}_i - \boldsymbol{p}_i^*\big) \\
&\qquad - \frac{1}{2} \sum_{i=1}^{M} \Big(\big[\alpha_i^d g_i^d + \beta_i^c g_i^c\big] \boldsymbol{p}_i^{*\mathrm{T}} \boldsymbol{p}_i \\
&\qquad\quad - \big[\alpha_i^d g_i^d + \beta_i^c g_i^c\big] \boldsymbol{p}_i^{\mathrm{T}} \boldsymbol{p}_i^* \Big) \Bigg\} \mathrm{d}t \\
&= \frac{\sqrt{2}}{2} \sum_{i=1}^{M} \int_0^{T_h} \big[\alpha_i^d g_i^d + \beta_i^c g_i^c\big] \big(\boldsymbol{p}_i - \boldsymbol{p}_i^*\big)^{\mathrm{T}} \big(\boldsymbol{p}_i - \boldsymbol{p}_i^*\big) \mathrm{d}t \\
&\quad + V(\boldsymbol{s}_0).
\end{aligned} \tag{43}
$$

It can be seen from (41) that $P_E(\boldsymbol{p}^*, \boldsymbol{q}, \boldsymbol{s}_0) \leq P_E(\boldsymbol{p}^*, \boldsymbol{q}^*, \boldsymbol{s}_0) = V(\boldsymbol{s}_0) \leq P_E(\boldsymbol{p}, \boldsymbol{q}^*, \boldsymbol{s}_0)$, and thus, the Nash equilibrium is obtained.

REFERENCES

[1] S. Guan, J. Wang, C. Jiang, R. Duan, Y. Ren, and T. Q. S. Quek, "MagicNet: The maritime giant cellular network," *IEEE Commun. Mag.*, vol. 59, no. 3, pp. 117–123, Mar. 2021.

[2] Z. Fang, J. Wang, J. Du, X. Hou, Y. Ren, and Z. Han, "Stochastic optimization-aided energy-efficient information collection in Internet of Underwater Things networks," *IEEE Internet Things J.*, vol. 9, no. 3, pp. 1775–1789, Feb. 2022.

[3] Z. Fang, J. Wang, C. Jiang, Q. Zhang, and Y. Ren, "AoI-inspired collaborative information collection for AUV-assisted Internet of Underwater Things," *IEEE Internet Things J.*, vol. 8, no. 19, pp. 14559–14571, Oct. 2021.

[4] T. Li et al., "Applications of multi-agent reinforcement learning in future internet: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 24, no. 2, pp. 1240–1279, 2nd Quart., 2022.

[5] V. G. Lopez, F. L. Lewis, Y. Wan, E. N. Sanchez, and L. Fan, "Solutions for multiagent pursuit-evasion games on communication graphs: Finite-time capture and asymptotic behaviors," *IEEE Trans. Autom. Control*, vol. 65, no. 5, pp. 1911–1923, May 2020.

[6] E. Garcia, D. W. Casbeer, A. Von Moll, and M. Pachter, "Multiple pursuer multiple evader differential games," *IEEE Trans. Autom. Control*, vol. 66, no. 5, pp. 2345–2350, May 2021.

[7] Z. E. Fuchs and P. P. Khargonekar, "Generalized engage or retreat differential game with escort regions," *IEEE Trans. Autom. Control*, vol. 62, no. 2, pp. 668–681, Feb. 2017.

[8] J. Chen, W. Zha, Z. Peng, and D. Gu, "Multi-player pursuit–evasion games with one superior evader," *Automatica*, vol. 71, pp. 24–32, Sep. 2016.

[9] R. Kamalapurkar, J. R. Klotz, P. Walters, and W. E. Dixon, "Model-based reinforcement learning in differential graphical games," *IEEE Trans. Control Netw. Syst.*, vol. 5, no. 1, pp. 423–433, Mar. 2018.

[10] Y.-J. Chen, D.-K. Chang, and C. Zhang, "Autonomous tracking using a swarm of UAVs: A constrained multi-agent reinforcement learning approach," *IEEE Trans. Veh. Technol.*, vol. 69, no. 11, pp. 13702–13717, Nov. 2020.

[11] J. Moon, S. Papaioannou, C. Laoudias, P. Kolios, and S. Kim, "Deep reinforcement learning multi-UAV trajectory control for target tracking," *IEEE Internet Things J.*, vol. 8, no. 20, pp. 15441–15455, Oct. 2021.

[12] Z. Xia et al., "Multi-agent reinforcement learning aided intelligent UAV swarm for target tracking," *IEEE Trans. Veh. Technol.*, vol. 71, no. 1, pp. 931–945, Jan. 2022.

[13] W. Sun, P. Tsiotras, T. Lolla, D. N. Subramani, and P. F. J. Lermusiaux, "Pursuit-evasion games in dynamic flow fields via reachability set analysis," in *Proc. Amer. Control Conf. (ACC)*, Seattle, WA, USA, May 2017, pp. 4595–4600.

[14] J. Ni, L. Yang, L. Wu, and X. Fan, "An improved spinal neural system-based approach for heterogeneous AUVs cooperative hunting," *Int. J. Fuzzy Syst.*, vol. 20, no. 2, pp. 672–686, Feb. 2018.

[15] J. Yan, Z. Guo, X. Yang, X. Luo, and X. Guan, "Finite-time tracking control of autonomous underwater vehicle without velocity measurements," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 52, no. 11, pp. 1–15, Nov. 2021.

[16] X. Yu et al., "Path planning in multiple-AUV systems for difficult target traveling missions: A hybrid metaheuristic approach," *IEEE Trans. Cognit. Develop. Syst.*, vol. 12, no. 3, pp. 561–574, Sep. 2020.

[17] Y. Li, B. Li, W. Yu, S. Zhu, and X. Guan, "Cooperative localization based multi-AUV trajectory planning for target approaching in anchor-free environments," *IEEE Trans. Veh. Technol.*, vol. 71, no. 3, pp. 3092–3107, Mar. 2022.

[18] W. Wei, J. Wang, Z. Fang, J. Chen, Y. Ren, and Y. Dong, "3U: Joint design of UAV-USV-UUV networks for cooperative target hunting," *IEEE Trans. Veh. Technol.*, vol. 72, no. 3, pp. 4085–4090, Mar. 2023.

[19] W. Wei, J. Wang, J. Du, Z. Fang, C. Jiang, and Y. Ren, "Underwater differential game: Finite-time target hunting task with communication delay," in *Proc. IEEE Int. Conf. Commun.*, Seoul, South Korea, May 2022, pp. 3989–3994.

[20] W. Shi, S. Song, C. Wu, and C. L. P. Chen, "Multi pseudo Q-learning-based deterministic policy gradient for tracking control of autonomous underwater vehicles," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 12, pp. 3534–3546, Dec. 2019.

[21] Z. Lin et al., "Efficient parallel split learning over resource-constrained wireless edge networks," 2023, *arXiv:2303.15991*.

[22] T. Mylvaganam, M. Sassano, and A. Astolfi, "A differential game approach to multi-agent collision avoidance," *IEEE Trans. Autom. Control*, vol. 62, no. 8, pp. 4229–4235, Aug. 2017.

[23] C. Ge, J. H. Park, C. Hua, and X. Guan, "Nonfragile consensus of multiagent systems based on memory sampled-data control," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 51, no. 1, pp. 391–399, Jan. 2021.

[24] Y. Fu and T. Chai, "Online solution of two-player zero-sum games for continuous-time nonlinear systems with completely unknown dynamics," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 12, pp. 2577–2587, Dec. 2016.

[25] S. Zhou, Y. Cheng, X. Lei, Q. Peng, J. Wang, and S. Li, "Resource allocation in UAV-assisted networks: A clustering-aided reinforcement learning approach," *IEEE Trans. Veh. Technol.*, vol. 71, no. 11, pp. 12088–12103, Nov. 2022.

[26] R. Lowe, Y. Wu, A. Tamar, A. Harb, P. Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," in *Proc. Adv. Neural Inf. Process. Syst.*, Red Hook, NY, USA, Dec. 2017, pp. 6382–6393.

[27] R. Cui, C. Yang, Y. Li, and S. Sharma, "Adaptive neural network control of AUVs with control input nonlinearities using reinforcement learning," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 47, no. 6, pp. 1019–1029, Jun. 2017.

[28] M. Liu, Y. Wan, F. L. Lewis, and V. G. Lopez, "Adaptive optimal control for stochastic multiplayer differential games using on-policy and off-policy reinforcement learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 12, pp. 5522–5533, Dec. 2020.

Wei Wei (Student Member, IEEE) received the B.S. degree in electronics engineering from Shandong University, Jinan, China, in 2020, and the M.S. degree (Hons.) in electronic and information engineering from Tsinghua University, Beijing, China, in 2023. She is currently pursuing the Ph.D. degree with the Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong.

Her research interests include multiagent reinforcement learning, game theory, heterogeneous system design, resource allocation, edge computing, and vehicular networking.

Jingjing Wang (Senior Member, IEEE) received the B.Sc. degree (Hons.) in electronic information engineering from the Dalian University of Technology, Dalian, Liaoning, China, in 2014, and the Ph.D. degree (Hons.) in information and communication engineering from Tsinghua University, Beijing, China, in 2019.

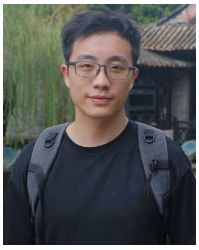
From 2017 to 2018, he visited the Next Generation Wireless Group chaired by Prof. Lajos Hanzo at the University of Southampton, Southampton, U.K. He is currently a Professor at the School of Cyber Science and Technology, Beihang University, Beijing. He has published over 100 IEEE journal/conference papers. His research interests include artificial intelligence (AI)-enhanced next-generation wireless networks, UAV networking, and swarm intelligence.

Dr. Wang was a recipient of the Best Journal Paper Award of IEEE ComSoc Technical Committee on Green Communications and Computing in 2018, and the Best Paper Award of the IEEE International Conference on Communications (ICC) and the IEEE International Wireless Communications and Mobile Computing Conference (IWCNC) in 2019. He has served as a Guest Editor for IEEE INTERNET OF THINGS JOURNAL. He is serving as an Editor for the IEEE WIRELESS COMMUNICATIONS LETTERS and the IEEE OPEN JOURNAL OF THE COMMUNICATIONS SOCIETY.

Jun Du (Senior Member, IEEE) received the B.S. degree in information and communication engineering from the Beijing Institute of Technology, Beijing, China, in 2009, and the M.S. and Ph.D. degrees in information and communication engineering from Tsinghua University, Beijing, in 2014 and 2018, respectively.

From October 2016 to September 2017, she was a sponsored researcher, and she visited the Imperial College London, London, U.K. She is currently an Assistant Professor with the Department of Electrical Engineering, Tsinghua University. Her research interests are mainly in communications, networking, resource allocation, and system security problems of heterogeneous networks and space-based information networks.

Dr. Du was a recipient of the Best Student Paper Award from IEEE Global Conference on Signal and Information Processing (GlobalSIP) in 2015, the Best Paper Award from IEEE International Conference on Communications (ICC) in 2019, and the Best Paper Award from International Wireless Communications and Mobile Computing Conference (IWCMC) in 2020.



Zhengru Fang (Student Member, IEEE) received the B.S. degree (Hons.) in electronics and information engineering from the Huazhong University of Science and Technology (HUST), Wuhan, China, in 2019, and the M.S. degree (Hons.) from Tsinghua University, Beijing, China, in 2022. He is currently pursuing the Ph.D. degree with the Department of Computer Science, City University of Hong Kong, Hong Kong.

His research interests include collaborative perception, vehicle-to-everything (V2X), age of information, and mobile edge computing.

Mr. Fang has been serving as a reviewer for IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS (JSAC), IEEE INTERNET OF THINGS JOURNAL (IoTJ), IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY (TVT), IEEE WIRELESS COMMUNICATIONS LETTERS (WCL), IEEE COMMUNICATIONS LETTERS (CL), and *IEEE Vehicular Technology Magazine* (VTM).

Yong Ren (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees from the Harbin Institute of Technology, Harbin, China, in 1984, 1987, and 1994, respectively, all in electronic engineering.

From 1995 to 1997, he worked as a Post-Doctoral Researcher at the Department of Electrical Engineering, Tsinghua University, Beijing, China, where he is currently a Full Professor with the Department of Electronic Engineering and serves as the Director for the Complexity Engineered Systems Laboratory. He has authored or coauthored more than 400 technical papers in the area of computer network and mobile telecommunication networks. His current research interests include complex system theory and its applications to the optimization of the Internet, the Internet of Things and ubiquitous network, cognitive networks, and cyber-physical systems.

Dr. Ren has served as a reviewer for more than 40 international journals or conferences.

C. L. Philip Chen (Fellow, IEEE) received the M.Sc. degree in electrical engineering from the University of Michigan, Ann Arbor, MI, USA, in 1985, and the Ph.D. degree in electrical engineering from Purdue University, West Lafayette, IN, USA, in 1988.

He was a Chair Professor and the Dean of the Department of Computer and Information Science, Faculty of Science and Technology, University of Macau, Macau, China. He is currently a Chair Professor and the Dean of the College of Computer Science and Engineering, South China University of Technology, Guangzhou, China. His current research interests include systems, cybernetics, and computational intelligence.

Dr. Chen is a member of Academia Europaea (AE), the European Academy of Sciences and Arts (EASA), and the International Academy of Systems and Cybernetics Science (IASCYS). He is also a fellow of the American Association for the Advancement of Science, the International Association of Pattern Recognition (IAPR), the Chinese Association of Automation (CAA), and the Hong Kong Institute of Engineers (HKIE). He received the IEEE Norbert Wiener Award in 2018 for his contribution to systems and cybernetics, and machine learning.