drones

MDPI

Article

Energy-Efficient Distributed AUV Swarm for Target Tracking via LSTM-Assisted Offline-to-Online Reinforcement Learning

Renbo Li [1,2], Denghui Li [2], Xiangxin Zhang [3] and Weiming Ni [2,*]

[1] Second Institute of China Aerospace Science and Industry Corporation, Beijing 100039, China
[2] Department of Communication Science and Engineering, School of Information Science and Technology, Fudan University, Shanghai 200438, China
[3] Research and Development Department, Shanghai Internet of Things Co., LTD. (SHIOTC), Shanghai 200042, China
* Correspondence: weimingni@fudan.edu.cn

Highlights

What are the main findings?

- An area information entropy and energy-efficient strategy for the AUV swarm are developed, which can decentralize the AUV swarm as much as possible to track the escaping target, thereby enhancing both tracking success rate and energy efficiency.
- A long-short term memory-based offline-to-online multi-agent reinforcement learning scheme is designed, which fully leverages historical data and rapidly generates optimal policies in new scenarios, thereby enhancing task success rates.

What are the implications of the main findings?

- The proposed method demonstrates excellent scalability and portability, which can be used as a foundation for future validation and deployment on actual AUV platforms.
- The proposed scheme can provide valuable guidance for autonomous unmanned systems to track targets collaboratively, while helping to extend the tracking duration of the AUV swarm and reduce deployment costs.

Abstract

In recent years, autonomous underwater vehicles (AUVs) have been increasingly employed for target surveillance and tracking. However, the limited performance and information-processing capability of a single AUV make it difficult to achieve high-precision tracking in practice. To address these challenges, this paper proposes an online-to-offline multi-agent reinforcement learning (MARL) framework that employs offline training on historical data to obtain the expert policy. Then, the optimal policy is generated by online fine-tuning technology, which enhances the training efficiency of reinforcement learning in new scenarios. To expand the surveillance range of AUV swarms, a distributed cooperative strategy based on area information entropy (AIE) is introduced. To reduce energy consumption in complex marine environments containing obstacles and vortices, ocean current and energy consumption models are introduced, together with an energy-efficiency optimization strategy. Furthermore, a long short-term memory (LSTM) network is integrated into the offline-to-online MARL framework to predict time-varying environmental states, thereby improving tracking accuracy and energy efficiency. Experimental results show that the proposed scheme is superior to the baseline schemes in terms of energy consumption, task success rate, and distance between AUVs. In addition, various performance indicators of the extended AUV swarm are also superior to the baseline schemes, demonstrating that the proposed scheme has excellent performance and scalability.

Check for updates

Academic Editors: Jian Chen, Min Xia and Hui Xie

Received: 8 December 2025
Revised: 27 December 2025
Accepted: 30 December 2025
Published: 26 February 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

Keywords: autonomous underwater vehicles (AUVs); offline-to-online multi-agent reinforcement learning; target tracking; area information entropy (AIE); long short-term memory (LSTM)

1. Introduction

Over the past few years, rapid advances in artificial intelligence-based underwater wireless communication have attracted significant attention in the field of autonomous underwater vehicles (AUVs). Owing to their strong maneuverability and excellent stealth capabilities, AUVs have been widely applied in tasks such as ship surveillance [1], underwater rescue [2], and target tracking [3]. The AUV swarm collaborative target-tracking technology addresses the limitations of single-AUV systems—such as limited detection range and low mission success rates—and has attracted growing research interest [4,5]. The AUV swarm enable precise and continuous monitoring and tracking while reducing overall energy consumption in complex underwater environments, though achieving this remains highly challenging.

Early studies approached the problem of tracking highly maneuverable targets by first investigating the capture of stationary or slow-moving targets [6,7]. These studies relied on acoustic channels for both target detection and communication. During tracking, AUVs are required to continuously optimize their strategies based on locally observed environmental information, thereby improving system performance and ensuring stable tracking. AUVs exhibit game-theoretic behavior during tracking, although related research remains in its early stages. Most AUV swarm target-tracking strategies in marine environments are adapted from control schemes originally designed for unmanned aerial and ground vehicles. Traditional approaches include the potential field algorithm [8] and various intelligent algorithms [9,10]. Yu et al. [10] designed a hybrid metaheuristic-based target-tracking framework that employs the shortest-path faster method and the A* algorithm to plan motion paths for multiple AUVs, and further developed a differential evolution–clustering–ant colony optimization algorithm to refine their motion strategies. Liang et al. [11] proposed an immune-inspired target-tracking scheme that generates motion plans for unmanned underwater vehicle (UUV) swarms using both global and local control mechanisms. In this scheme, the global control mechanism enables the UUV swarm to search for underwater targets, while the local control mechanism directs individual vehicles to track the target. Ni et al. [12] studied a security-based target-tracking scheme that integrates a dynamic alliance mechanism with an improved genetic algorithm to assign motion control strategies to multiple AUVs, thereby achieving efficient and stable tracking performance. Although model-based approaches perform well in ground and aerial scenarios, their dependence on extensive prior knowledge limits their scalability in unfamiliar underwater environments. Moreover, the complexity of underwater environments hinders the effectiveness of these approaches in highly dynamic tracking tasks [13].

Ocean current fields in time-varying underwater environments significantly influence the propagation of acoustic signals. Acoustic signals experience severe attenuation in marine environments, limiting the ability of AUVs to perceive environmental information and communicate with each other [14]. Moreover, the limited sensing and information-gathering capabilities of AUVs reduce the success rate of target tracking. Fortunately, a distributed AUV swarm can make decisions based on limited local observations, thereby overcoming the challenges associated with the absence of a centralized node [15]. Liu et al. [16] developed a formation-tracking framework based on distributed model predictive control, which generates virtual AUV state trajectories to achieve target tracking under ocean current disturbances and communication delays. Wei et al. [17] proposed a swarm

https://doi.org/10.3390/drones10030158

hyper-heuristic-based scheme that considers ocean currents and energy consumption to plan stable navigation routes in complex ocean current environments. Fang et al. [18] investigated an energy-oriented trajectory-optimization scheme that employs a particle swarm optimization algorithm to perform vehicle scheduling in ocean current environments, thereby reducing energy consumption. However, most existing studies assume access to global information, which makes their control strategies impractical in real-world scenarios [19]. In addition, complex underwater currents greatly influence the energy consumption of the AUV swarm, thereby shortening the operational duration of the AUV swarm. Fortunately, clustering-based communication protocols [20,21] can reduce the communication energy consumption of AUV swarms.

In recent years, deep reinforcement learning (DRL), benefiting from its strong environmental interaction and feature representation capabilities, has achieved remarkable success in robotic autonomous control and has emerged as a promising solution for multi-agent target tracking [22]. Zhang et al. [23] proposed a deep Q-network (DQN)-based path planning scheme that leverages DQN to enhance the capabilities of monitoring dynamic obstacles and adaptive local navigation, thereby enhancing the autonomy of AUVs in complex environments. Bu et al. [24] designed a DRL-based data collection scheme that plans the AUV's navigation trajectory during data collection tasks by considering the age of information of the nodes and AUV's real-time motion energy consumption. Yu et al. [25] suggested a soft actor-critic (SAC)-based path planning scheme that optimizes the AUV's navigation trajectory by optimizing the reward function, thereby reducing both navigation distance and time. Compared with centralized control schemes, multi-agent reinforcement learning (MARL) approaches can accomplish complex underwater tasks without relying on a complete environmental model [26,27], particularly with respect to mission success rate, energy efficiency, and adaptability to ocean current environments. Zhang et al. [28] formulated the target-tracking problem as a Markov decision process (MDP) and proposed a multi-agent deep deterministic policy gradient (MADDPG)-based scheme that extends the target prediction network and achieves stable tracking performance. Xing et al. [29] proposed a DRL-based energy-saving scheme that employed sensors to perceive the underwater environment and enabled AUVs to cruise autonomously. Xia et al. [30] proposed a multi-agent target-tracking scheme featuring the distributed tracking strategy to improve the tracking success rate. Wang et al. [31] devised a DRL-based cooperative tracking system that enables the AUV swarm to collaboratively perform target-tracking tasks in harsh marine conditions. Song et al. [32] proposed an end-to-end target-tracking framework that employs a DRL-based control method to generate motion policies for multiple vehicles, thereby improving tracking accuracy. Zhu et al. [33] presented an elastic, software-defined MARL-based collaborative tracking framework that enhances the scalability of the AUV swarm and enables high-performance target tracking. Yang et al. [34] designed an end-to-end MARL-based target-tracking framework that adopts the Centralized Training and Distributed Execution (CTDE) paradigm to achieve secure tracking in adversarial scenarios. The MARL framework is based on the CTDE paradigm [35], which can efficiently acquire informative representations through environmental interactions while reducing training time. To obtain more environmental information and time-related feature information, the LSTM network is introduced into MARL, thereby enhancing its decision-making capabilities. Liu et al. [36] designed an MARL-LSTM-based internet of things algorithm that uses LSTM to predict the states and actions of other agents, thereby enhancing the performance of the IoT network. Wu et al. [37] proposed an MADDPG-LSTM-based traffic light control scheme that mitigates environmental instability caused by partial observable states, thereby enhancing system efficiency. Although online reinforcement learning (RL) is adaptable to various underwater scenarios, the online training of RL requires significant time and

computational resources. Therefore, offline-to-online RL effectively addresses this issue by training an optimal policy using historical data, then optimizing the policy through online fine-tuning in new scenarios, thereby enhancing training efficiency. Lee et al. [38] designed an offline-to-online RL-based autonomous driving scheme that trains a unified multi-task policy by using historical data, thereby enhancing the sample efficiency and learning efficiency of RL. Xu et al. [39] studied a offline-to-online MARL-based cooperative tracking scheme that generates expert policy through offline MARL, then employs online MARL to rapidly generate the optimal policy for target-tracking tasks.

Inspired by the aforementioned analysis, this paper presents an offline-to-online MARL-based target-tracking scheme designed to achieve efficient, reliable target tracking for AUV networks, and enhances the training efficiency. The proposed framework introduces an innovative distributed tracking strategy that incorporates cooperative decision-making among the AUV swarm thereby improving the overall tracking efficiency of the AUV swarm. In addition, the framework achieves energy conservation by perceiving complex underwater environments and integrates a long short-term memory (LSTM) network into the offline-to-online MARL framework for enhancing learning efficiency. We contribute the following contributions:

- We incorporate the concept of Area Information Entropy (AIE) into the tracking strategy of the AUV swarm, which improves tracking success rates through distributed collaboration. Furthermore, an energy-efficient strategy is proposed to reduce AUV's energy consumption and extend the endurance of the AUV swarm.
- Since the formulated constrained target-tracking problem is a high-dimensional NP-hard problem, DRL can effectively address it. Therefore, the optimization problem is first modeled as an MDP. Then, we propose a multi-agent twin-delayed deep deterministic policy gradient (MATD3)-based scheme to address this issue. Additionally, we adopt offline training and online fine-tuning to enhance the training efficiency of the proposed scheme.
- A novel LSTM layer is integrated into the neural networks of the online MATD3 algorithm to predict the states of time-varying environments. During each iteration, the proposed algorithm employs the LSTM layer to estimate unobserved states, thereby improving the training efficiency of the framework. Experimental results indicate that the proposed framework greatly exceeds baseline schemes with respect to success rate, energy efficiency, and safety performance.

We organize the remainder of this paper to be the following. Section 2 presents the system model in detail, while Section 3 introduces the design details of the proposed offline-to-online MATD3-LSTM algorithm. Experimental results and analysis are presented in Section 4, and finally, Section 5 draws the conclusions.

2. System Model

We consider a scenario where multiple AUVs are deployed to perform underwater target-tracking tasks on a 2D plane with a depth of d meters. As shown in Figure 1, multiple AUVs navigate within a turbulent ocean current environment, where they employ sonar equipment to detect the escaping target and communicate, thereby maintaining continuous tracking for the target. Let there be I AUVs and the position of AUV i is denoted to be $\boldsymbol{P}_i(t) = [x_i(t), y_i(t)]^{\mathrm{T}}$, $i \in I$. Additionally, the escaping target's position is represented as $\boldsymbol{P}_g(t) = [x_g(t), y_g(t)]^{\mathrm{T}}$. Furthermore, assume there are N vortices and K obstacles in the ocean environment, and the positions of vortex n and obstacle k are represented as $\boldsymbol{P}_n(t) = [x_n(t), y_n(t)]^{\mathrm{T}}$, $n \in N$, and $\boldsymbol{P}_k(t) = [x_k(t), y_k(t)]^{\mathrm{T}}$, $k \in K$.

During the tracking process, each AUV is equipped with ocean current velocity sensors that detect local turbulence information to optimize its trajectory-planning strategy.

The motion model, acoustic communication and detection model, distributed strategy, ocean current model, and energy consumption model will be described in Sections 2.1–2.5, respectively. In addition, for convenience, the important notations are summarized in Table 1.

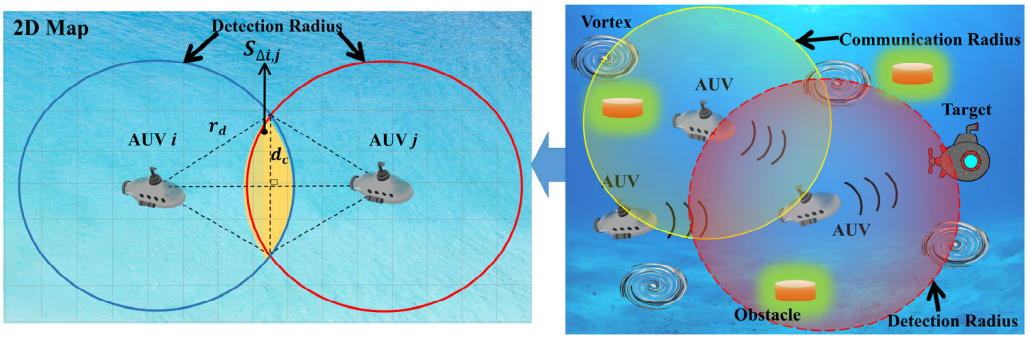

Figure 1. Illustration of the AUVs warm-based target tracking.

Table 1. List of Notations.

Notation	Description
I	The AUV's number
N	The vortice's number
K	The obstacle's number
$\boldsymbol{P}_i$	The position of the AUV i
$\boldsymbol{P}_g$	The target's position
$\boldsymbol{P}_n$	The position of the vortex n
$\boldsymbol{P}_k$	The position of the obstacle k
$\boldsymbol{P}_\zeta$	The position of the cell ζ
ψ_i, $v_{i,1}$, $v_{i,2}$, ω_i	The yaw angular, surge velocity, sway velocity, and yaw angular velocity of the AUV i
$\boldsymbol{\tau}_i$	The control input of the AUV i
TS	The target strength of the reflection area
SL	The emission strength of the information source
NL	The ambient noise level
TL	The attenuation of sound intensity
f	The center acoustic frequency
N_t, N_s, N_w, N_{th}	The turbulent, ship, wind, and thermal noises in underwater communication
Γ	The vorticity of the vortex
ξ	The fluid viscosity
Λ	The vortex strength
ς	The radius of the Lamb vortex
C_1	The AUV's cross-sectional area
C_2	The AUV's dragging coefficient
T	The total time of the tracking task

2.1. Motion Model

Since AUVs perform target tracking on a two-dimensional (2D) plane, a three-degree-of-freedom (3-DOF) dynamic model is introduced, without loss of generality, to describe their motion states. In the world coordinate frame, the state vector of AUV i is $\eta_i = [x_i, y_i, \psi_i]^{\mathrm{T}}$, where ψ_i represents the yaw angle of the AUV i. Additionally, its state matrix is $v_i = [v_{i,1}, v_{i,2}, \omega_i]^{\mathrm{T}}$ in the body coordinate frame, where $v_{i,1}$, $v_{i,2}$, and ω_i are defined as the surge, sway, and yaw angular velocities. Because the sampling interval is relatively short, the AUV motion can be approximated as uniformly accelerated. Accordingly, based on Fossen's motion equations, the 3-DOF dynamic model of AUV i is [40]

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

$$\boldsymbol{M}\dot{\boldsymbol{v}}_i + \boldsymbol{C}(\boldsymbol{v}_i)\boldsymbol{v}_i + \boldsymbol{D}(\boldsymbol{v}_i)\boldsymbol{v}_i + \boldsymbol{G}(\boldsymbol{v}_i) = \boldsymbol{\tau}_i, \tag{2}$$

where $\boldsymbol{M}$ represents the AUV's additional mass matrix, primarily including key parameters such as the device's mass and rotational habitus parameters. $\boldsymbol{C}$ stands for the Coriolis centripetal matrix, and $\boldsymbol{D}$ represents the damping matrix, while $\boldsymbol{G}$ represents the composite matrix formed by the combination of gravity and buoyancy. Additionally, $\boldsymbol{\tau}_i$ is the control input, which is used to control its motion. Meanwhile, the transformation matrix $\boldsymbol{J}$ can be expressed as

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

2.2. Communication and Detection Models

The AUVs detects the environment and communicates with each other by using sonar technology [34,41]. The sonar equation can be expressed as

$$EM = TS + DI - NL - DT - 2TL + SL\,, \tag{4}$$

where TS, DI, NL, and DT indicate the target strength, directivity index of the sonar, ambient noise level, and detection threshold, respectively. EM and SL denote the echo margin of the sonar and the emission strength. Additionally, TL denotes the attenuation of sound intensity after the sound wave with central acoustic frequency f propagates a certain distance d, i.e.,

$$\begin{cases} TL = 20\lg(d) + d \times \alpha(f) \times 10^{-3}, \\ \alpha(f) = \frac{0.11f^2}{1+f^2} + \frac{44f^2}{4400+f^2} + 2.75 \times 10^{-4} f^2 + 0.003, \end{cases} \tag{5}$$

where $\alpha(f)$ represents the attenuation formula of underwater sound waves [42]. Additionally, environmental noise NL consists of turbulent, ship, wind, and thermal noises, denoted as N_t, N_s, N_w, and N_{th} [43]. Furthermore, the noise model is denoted as

$$NL(f) = N_t(f) + N_s(f) + N_w(f) + N_{th}(f). \tag{6}$$

where the noise elements are represented, respectively, as

$$\begin{cases} 10\log N_t(f) = 17 - 30\log f; \\ 10\log N_s(f) = 30 + 20s + \log\left(\frac{f^{26}}{(f+0.03)^{60}}\right); \\ 10\log N_w(f) = 50 + 70w^{1/2} + \log\left(\frac{f}{(f+0.4)^2}\right); \\ 10\log N_{th}(f) = -15 + 20\log f, \end{cases} \tag{7}$$

where w and $s \in [0,1]$ denote wind speed (m/s) and shipping activity factor, respectively. Assuming that there is a fixed f, the AUV's maximum detection distance can be expressed:

$$r_d = arg\max_d \{EM(d) \geq 0\}. \tag{8}$$

Additionally, the maximum communication distance of the AUV can be derived from the sonar equation as d_c.

2.3. Monitoring Redundancy and Distributed Strategies

The scenario of a AUV swarm monitoring underwater targets is shown in Figure 1. The AUVs detect underwater targets using acoustic detection technology. The target discovery condition is defined as $\|(x_i, y_i) - (x_g, y_g)\| \leq r_d$. Thus, the AUVs' target search can be mapped onto an OXY plane within the inertial coordinate system. To classify and label the area information, we rasterize the detecting environment for modeling. The search area, defined with dimensions $L \times W$, is then rasterized as

$$\begin{cases} cell_x = ceil(L/\ell) \\ cell_y = ceil(W/\ell) \end{cases}, \tag{9}$$

where ℓ is the scaling factor of rasterization and $ceil(\cdot)$ is the integer function that outputs the smallest integer not less than the inner formula. In addition, we store lattice information in table $\mathbb{C}(t) = \{c_1(t), \ldots, c_\zeta(t), \ldots, c_Y(t)\}$ at time t, where cell ζ denotes the corresponding rasterized cell, and $Y = cell_x \times cell_y$. Since the rasterized cells are pretty small, we use $\boldsymbol{P}_\zeta = (x_\zeta, y_\zeta)$ to represent the position of cell ζ. We assume that the rasterized cells satisfy the detecting condition $d_{i,\zeta} = \|\boldsymbol{P}_i - \boldsymbol{P}_\zeta\| \leq r_d$ at time t, which indicates the rasterized cell ζ is searched by AUV i, and then the cell ζ is labeled as $c_\zeta(t) = 1$; otherwise $c_\zeta(t) = 0$. The AUVs use acoustic detection techniques to gather accurate information about underwater targets. Thus, a detecting rasterized cell ζ indicates that the AUV swarm acquires complete information about the cell coordinates $\boldsymbol{P}_\zeta = (x_\zeta, y_\zeta)$.

The AUV swarm must first locate the underwater target before initiating the tracking tasks. A larger monitoring area for the AUV swarm increases the success rate of tracking the target. To assess search efficiency, we define the search coverage as

$$R_c = D_s / D_e, \tag{10}$$

where D_s denotes the area searched by the UAV swarm, and D_e represents the total environmental area.

When AUVs operate in close proximity during tracking tasks, monitoring redundancy may occur, thereby reducing monitoring efficiency and overall task success rate. As illustrated in Figure 1, the overlapping regions among AUVs indicate monitoring redundancy. To minimize redundant monitoring regions and promote decentralized coverage, the Area Information Entropy (AIE) metric is employed to guide AUVs in expanding their monitoring range. AIE is designed to quantify the independence of target monitoring among AUVs, which is directly related to the degree of spatial overlap between their sensing regions. The overlapping region between AUV i and AUV j can be computed using spatial geometric relationships.

$$\begin{cases} S_{\Delta i,j} &= 2\left(\pi r_d^2 \frac{2\arcsin(\frac{d_c}{2r_d})}{2\pi} - \frac{d_c}{2}\sqrt{r_d^2 - (\frac{d_c}{2})^2}\right), \\ d_c &= 2\sqrt{r_d^2 - (\frac{d_{i,j}}{2})^2}, \end{cases} \tag{11}$$

where r_d denotes the AUV's maximum detecting distance while $d_{i,j}$ denotes the distance between AUV i and AUV j. Thus, we use the ratio between the overlapping area and the monitoring area to express the probability of monitoring redundancy

$$p(x_i) = (\sum_{i \neq j} S_{\Delta i,j})/(\pi r_d^2). \tag{12}$$

Based on entropy theory, AIE is calculated to be

$$AIE = -\sum_{i=1}^{I} p(x_i) \log\left(p(x_i)\right) \tag{13}$$

AIE is positively correlated with the distance between AUVs, allowing AUVs to use AIE to increase their separation and reduce monitoring redundancy.

2.4. Turbulent Ocean Environments Model

Due to the effect of the Earth's rotation, the strength of ocean currents in the horizontal plane is significantly greater than that in the vertical plane. Therefore, the horizontal component of ocean currents plays a dominant role in AUV motion. Based on this assumption, the two-dimensional (2D) Navier–Stokes equations [44] are employed to model the ocean current field as

$$\frac{\partial \Gamma}{\partial t} + (\boldsymbol{V}_t \nabla)\Gamma = \xi \Delta \Gamma \tag{14}$$

where $\boldsymbol{V}_t = (v_{tx}, v_{ty})$ denotes the field of ocean velocity, and v_{tx} and v_{ty} represent velocity components along the $\boldsymbol{X}$ and $\boldsymbol{Y}$ axes. In addition, ξ represents the vorticity of the vortex and the fluid viscosity, respectively. Moreover, ∇ is the gradient operator, and Δ represents the Laplacian operator. In practical applications, AUVs have horizontal acoustic Doppler current profilers (H-ADCPs) to detect and analyze ocean current characteristics. By interpreting flow-field data, these instruments estimate the number, position, and rotational direction of Lamb vortices, thereby providing an initial approximation of each vortex's radius and intensity [45]. To simplify the analytical process, the approximate solution of the two-dimensional (2D) Navier–Stokes equations is constructed as a superposition of single-point vortex solutions, as follows:

$$\begin{cases} v_{tx}(\boldsymbol{P}_i(t)) = & -\frac{\Lambda \cdot (y_i(t) - y_0)}{2\pi \|\boldsymbol{P}_i(t) - \boldsymbol{P}_0\|^2} \left(1 - e^{-\frac{\|\boldsymbol{P}_i(t) - \boldsymbol{P}_0\|^2}{\varsigma^2}}\right), \\ v_{ty}(\boldsymbol{P}_i(t)) = & \frac{\Lambda \cdot (x_i(t) - x_0)}{2\pi \|\boldsymbol{P}_i(t) - \boldsymbol{P}_0\|^2} \left(1 - e^{-\frac{\|\boldsymbol{P}_i(t) - \boldsymbol{P}_0\|^2}{\varsigma^2}}\right), \end{cases} \tag{15}$$

and

$$\Gamma(\boldsymbol{P}_i(t)) = \frac{\Lambda}{\pi \varsigma^2} e^{-\frac{\|\boldsymbol{P}_i(t) - \boldsymbol{P}_0\|^2}{\varsigma^2}}, \tag{16}$$

where Λ and ς indicate the vortex strength and radius of the Lamb vortex, respectively. In addition, $\boldsymbol{P}_0 = (x_0, y_0)$ denotes the coordinate of Lamb vortex center.

2.5. Energy Consumption Model

The energy consumption of each AUV during target tracking primarily arises from overcoming hydrodynamic resistance caused by ocean currents. Since this study mainly focuses on target-tracking tasks and the energy consumption associated with communication is negligible, the system energy consumption can be considered to be the motion energy

consumption of all AUVs. During the target-tracking process, the relative velocity of AUV i with respect to ocean current is expressed as [46]

$$\boldsymbol{v}_{r_i}(\boldsymbol{P}_i) = \boldsymbol{V}_i - \boldsymbol{V}_t(\boldsymbol{P}_i), \tag{17}$$

where $\boldsymbol{V}_i$ denotes the velocity of AUV i. Based on the computational fluid dynamics method [47], the value of moving resistance of the AUV i is

$$F_i = \frac{1}{2}\zeta C_1 C_2 \|\boldsymbol{v}_{r_i}(\boldsymbol{P}_i)\|^2, \forall i \in I, \tag{18}$$

where ζ denotes the seawater density, C_1 and C_2 are the AUV's cross-sectional area along the current direction and the AUV's dragging coefficient. Furthermore, according to Equations (17) and (18), the motion power of the AUV i is

$$W_i = F_i \|\boldsymbol{v}_{r_i}(\boldsymbol{P}_i)\|. \tag{19}$$

Thus, in the target-tracking task, the energy consumption of the AUV swarm is

$$E_{total} = \sum_{i=1}^{I} \int_0^T W_i \mathrm{d}t. \tag{20}$$

where T is the total time of the target-tracking task.

3. Problem Formulation and Algorithm Design

We first model the target-tracking task as an MDP, then formulate a constrained optimization problem, and design a corresponding reward function. Finally, we develop an algorithm to solve this issue.

3.1. Markov Decision Process Modeling

Tracking escaping targets in oceanic environments while maintaining AUV safety is a complex NP-hard problem. Furthermore, considering the interactive characteristics between multi-agent systems and the underwater environment, thus we model it as an MDP. An MDP can be expressed as a five-tuple [48]

$$\Phi = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma), \tag{21}$$

where $\mathcal{P}$ and $\mathcal{R}$ represent the state transition probability distribution and the reward function, respectively. $\gamma \in [0, 1]$ stands for the discount factor. In addition, the state space $\mathcal{S}$ and the action space $\mathcal{A}$ of all vehicles are expressed as follows.

1. **State Space**

During the target-tracking process, each AUV possesses acoustic communication and detection equipment to acquire environmental information. Due to the signal attenuation during underwater propagation, the observation of each AUV includes its own state, target state, and other AUV states within its communication radius. Therefore, the observation space of AUV i, where $i \in I$, is represented as

$$\boldsymbol{o}_i(t) = \left\{\boldsymbol{S}_i(t), \boldsymbol{S}_{j'}(t), \boldsymbol{P}_k(t), \boldsymbol{P}_g(t)\right\}, \tag{22}$$

where $\boldsymbol{S}_i(t) = \{\boldsymbol{P}_i(t), \boldsymbol{V}_i(t), \theta_i(t)\}$ is the state of the AUV i. In addition, $\boldsymbol{S}_{j'}(t) = \{\boldsymbol{P}_{j'}(t), \boldsymbol{V}_{j'}(t), \theta_{j'}(t)\}$ is the state of the vehicle j', where AUV j' is the other vehicle within

the communication radius of AUV i. $\boldsymbol{P}_k(t)$ denotes the position of the obstacle k. $\boldsymbol{P}_g(t)$ denotes the underwater target position state.

2. **Action Space**

In target-tracking tasks, AUVs determine their subsequent actions based on observed environmental states and corresponding feedback. The action space of AUV i, $i \in I$, is expressed as

$$\boldsymbol{a}_i(t) = \{a_{i,b}(t), \omega_i(t)\}, \tag{23}$$

where $a_{i,b}(t)$ and $\omega_i(t)$ are the acceleration and angular velocity of AUV i, respectively. Additionally, $\|a_{i,b}(t)\| \leq a_{max}$ and $\|\omega_i(t)\| \leq \omega_{max}$ represent the movement restrictions of AUV i.

3. **Reward Function**

In reinforcement learning, agents continuously optimize their policies based on the rewards received from their actions, which directly affects the overall performance of the AUV swarm. Therefore, the design of an appropriate reward function is crucial for achieving effective learning. Given the complexity of the target-tracking task, the details of the reward function are elaborated.

3.2. Optimization Problem Formulation

In the target-tracking task, each AUV executes action $a(\mathrm{t})$ and obtains reward $R(\mathrm{t})$ based on observation $o(\mathrm{t})$ and policy π to maximize the total discounted reward $J(\phi)$. During task execution, the AUV swarm must satisfy a series of constraints. Therefore, the constrained optimization problem is formulated as

$$\max_{\pi} J(\phi) = \max_{\pi} \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t R(t)\right] \tag{24a}$$

$$\text{s.t. } d_{i,j}(t) \geq r_s, i \neq j, \forall i, j \in I, \tag{24b}$$

$$d_{i,g}(t) < r_d, \tag{24c}$$

$$d_{i,k}(t) > r_k + r_s, \tag{24d}$$

$$\|a_{i,b}(t)\| \leq a_{\max}, \|\omega_i(t)\| \leq \omega_{\max}, \tag{24e}$$

where Equation (24a) denotes the optimization objective, which aims to maximize discount reward in the target-tracking task. Equation (24b) indicates that AUV i should maintain a safe distance r_s from AUV j. Moreover, Equation (24c) indicates that the target should be monitored by AUV i to prevent the task from failing. Additionally, Equation (24d) indicates that AUVs should keep safety, and Equation (24e) indicates that AUVs should meet the vehicle dynamic requirements when performing target-tracking tasks.

3.3. Reward Function Design

The AUV swarm should simultaneously consider target approach, communication maintenance, target monitoring, energy conservation, distributed coordination, and safety assurance. Therefore, a corresponding reward function is designed for the AUV swarm to optimize AUV trajectories and enable stable and efficient target tracking.

To avoid losing information about the tracked target, the AUV i must approach the target consistently to complete the target-tracking task. Therefore, when the AUV i approaches the target, it will receive a tracking reward

$$r_i^{(1)}(t) = d_{i,g}(t-1) - d_{i,g}(t), \tag{25}$$

where $d_{i,g}(t-1)$ and $d_{i,g}(t)$ represent the distances between AUV i and the target at time $t-1$ and time t, respectively.

To ensure AUV i avoids collisions with obstacles, other AUVs and the target while performing the tracking task, we design a dangerous area penalty for AUV i

$$r_i^{(2)}(t) = \begin{cases} 0, & d_{i,j}(t) > r_s \text{ and } d_{i,g}(t) < r_d \text{ and } d_{i,k}(t) > r_k + r_s, \\ -p_1, & otherwise, \end{cases} \tag{26}$$

where p_1 represents a normal number.

To ensure that the AUV i operates within the simulation area, we designed a boundary penalty for AUV i.

$$r_i^{(3)}(t) = \begin{cases} -p_2, & \text{AUV } i \text{ reaches boundary}, \\ 0, & \text{otherwise}, \end{cases} \tag{27}$$

where p_2 is a normal number.

To ensure communication connectivity between AUVs, we design a communication reward for AUV i.

$$r_i^{(4)}(t) = \begin{cases} r_c - d_{i,j}, & d_{i,g}(t) > r_d, \\ -p_3, & \text{otherwise}, \end{cases} \tag{28}$$

where p_3 is a normal number and r_c denotes the maximum acoustic communication distance.

To lower energy consumption of each AUV during target-tracking tasks, we design an energy penalty for the AUV i.

$$r_i^{(5)}(t) = -W_i(t). \tag{29}$$

To encourage AUVs to decentralize as much as possible when performing target-tracking tasks, we consider the AIE as the decentralization reward for AUV i.

$$r_i^{(6)}(t) = -\sum_{i=1}^{I} p(x_i) \log\left(p(x_i)\right). \tag{30}$$

Thus, the reward function of the AUV i can be constructed as

$$r_i(t) = \sum_{n=1}^{6} \chi_i^{(n)} r_i^{(n)}(t), \tag{31}$$

where $\chi_i^{(n)}$ is the weighting coefficient.

3.4. Offline-to-Online MATD3-LSTM Framework

Figure 2 shows the framework of the proposed algorithm. Specifically, this framework consists of two stages: the offline training stage and the online fine-tuning stage. The offline training stage utilizes historical data to train and obtain an expert policy, while the online fine-tuning stage applies this expert policy to new scenarios, performing online fine-tuning to enhance the training efficiency of MARL. At each stage, each AUV is controlled by a TD3 algorithm. The AUV swarm adopts the MATD3 algorithm within the CTDE framework [49], and assigns an individual controller to each AUV. Each AUV has its own LSTM network [50], which helps it to enhance feature representation capability. During centralized training, each AUV accesses the observations and actions of all AUVs and trains its own networks using this information to evaluate the actions generated by its actor network. During distributed execution, each AUV selects actions based on its own observations and policy, and receives rewards through interactions with the environment.

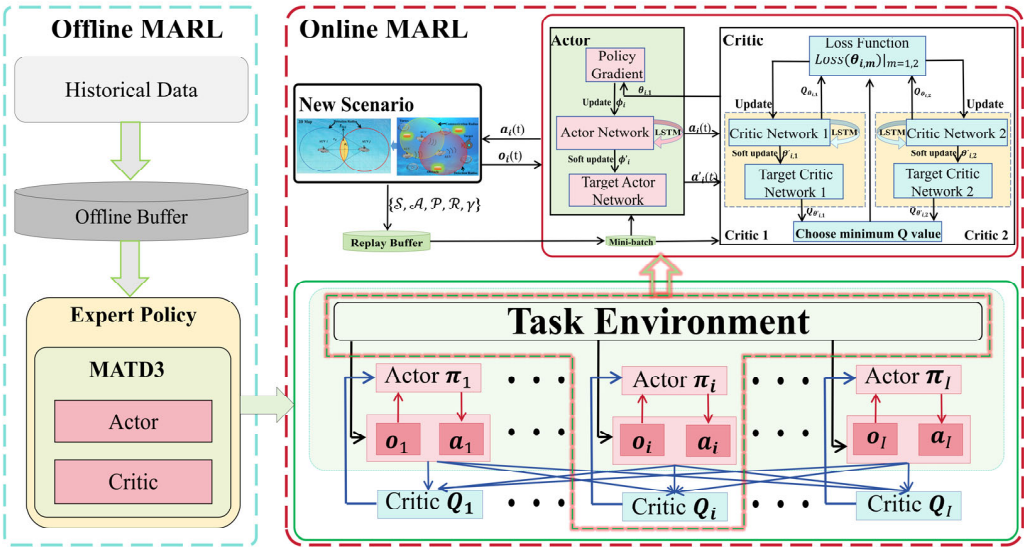

Figure 2. The framework of the proposed offline-to-online MATD3-LSTM algorithm in target-tracking task.

During the training stage, the data $(o, a, r, \hat{o})$ of a mini-batch with size B is randomly sampled from the experience pool Ψ, where $o = \{o_1, \ldots, o_i, \ldots, o_I\}$, a $= \{a_1, \ldots, a_i, \ldots, a_I\}$, $r = \{r_1, \ldots, r_i, \ldots, r_I\}$, and $\hat{o} = \{\hat{o}_1, \ldots, \hat{o}_i, \ldots, \hat{o}_I\}$ represent the observations, actions, rewards and the next observations of all AUVs, respectively. Then, for the AUV i, the update of critic networks is denoted as

$$Loss(\theta_{i,m}) = \frac{1}{B}\sum_{b=1}^{B}[(y_i^b - Q_{\theta_{i,m}}(\boldsymbol{o}^b, \boldsymbol{a}^b))^2], m = 1, 2, \tag{32}$$

where $\theta_{i,m}$ is the parameter of m-th critic network, $\boldsymbol{o}^b$ and $\boldsymbol{a}^b$ are the observations and actions of all AUVs. Additionally, $Q_{\theta_{i,m}}$ denotes the output of m-th critic network. The two target critic networks are employed to obtain the target value

$$y_i^b = r_i^b + \gamma \min_{m=1,2} Q_{\theta'_{i,m}}(\hat{\boldsymbol{o}}^b, \hat{\boldsymbol{a}}^b) \mid_{\hat{a}_i = \pi_{\phi_i}(\hat{o}_i^b) + \epsilon}, \tag{33}$$

where $\epsilon \sim clip(\mathcal{N}(0, \sigma^2), -c, c)$, σ^2 and c denote the Gaussian noise variance and the clipped threshold, respectively. $Q_{\theta'_{i,m}}$ denotes the m-th target critic network, $\theta'_{i,m}$ is the parameter of the m-th target critic network. In addition, $\hat{\boldsymbol{o}}^b$ and $\hat{\boldsymbol{a}}^b$ denote the next observations and next actions of all AUVs in the b-th sample. γ denotes the discount factor.

Then, for the AUV i, the update of the actor network uses policy gradient $\nabla_{\phi_i} J(\phi_i)$, being expressed as

$$\nabla_{\phi_i} J(\phi_i) = \frac{1}{B}\sum_{b=1}^{B}[\nabla_{a_i} Q_{\theta_{i,1}}(\boldsymbol{o}^b, \boldsymbol{a}^b) \mid_{a_i^b = \pi_{\phi_i}(o_i^b)} \times \nabla_{\phi_i} \pi_{\phi_i}(o_i^b)]. \tag{34}$$

Furthermore, for the AUV i, the parameters of the target networks are updated with a fixed update rate $\tau \in [0, 1]$, and they are respectively represented as

$$\phi'_i = \tau\phi_i + (1 - \tau)\phi'_i, \tag{35}$$

$$\theta'_{i,m} = \tau\theta_{i,m} + (1 - \tau)\theta'_{i,m}, m = 1, 2. \tag{36}$$

During the offline phase, we train the agent by using historical datasets to obtain the expert policy. According to [51], we employ the MATD3 algorithm to obtain the expert policy. Additionally, the historical data originates from other target-tracking scenarios, in

which case the expert policy can be smoothly fine-tuned in the new scenario. Specifically, we randomly sample B' from the offline datasets. The critic network updates using a gradient method, which can be expressed as

$$\nabla_{\theta_{i,m'}} L(\theta_{i,m'}) = \frac{1}{B'} \sum_{b'=1}^{B'} [(y_i^{b'} - Q_{\theta_{i,m'}}(\boldsymbol{o}^{b'}, \boldsymbol{a}^{b'})) \nabla_{\theta_{i,m'}} Q'_{\theta_{i,m'}}(\boldsymbol{o}^{b'}, \boldsymbol{a}^{b'})], m' = 1, 2. \tag{37}$$

The goal is to update the actor policy by maximizing discount rewards:

$$J(\phi_i) = \frac{1}{B'} \sum_{b'=1}^{B'} [Q_{\theta_{i,1}}(\boldsymbol{o}^{b'}, \boldsymbol{a}^{b'})]. \tag{38}$$

The process of the proposed algorithm is summarized in Algorithm 1.

Algorithm 1 The offline-to-online MATD3-LSTM Algorithm

```
Initialize the network parameters of all AUVs and replay buffer Ψ'.
for each offline training episode do
Update θ_{i,m'} by Equation (37);
    if t mod t_d ==0:
        Update φ_i by Equation (38);
    end if
end for
Initialize the replay buffer Ψ'
for each online training episode do
  Reset the training environment;
  for t = 0 to T do
      for i ∈ I do
          Based on observation o_i, AUV i gets action a_i, obtains reward r_i, and new
            observation ô_i;
      end for
      Store data (o, a, r, ô) into Ψ;
      Update observation o ← ô;
      for i ∈ I do
        Randomly sample a mini-batch B from Ψ;
        Calculate the target value by Equation (33);
        Update critic networks by Equation (32);
        if t mod t_d ==0:
            Update actor network by Equation (34);
          Updates on target networks φ'_i and θ'_{i,m} by Equations (35) and (36);
        end if
    end for
  end for
end for
```

4. Simulation Results and Discussion

This section first describes the simulation setup, including both environmental and algorithmic parameters. The effectiveness of the proposed algorithm in target-tracking tasks is then verified through extensive simulations. Furthermore, by conducting experiments under diverse scenarios, comparing the results with baseline methods, and analyzing the outcomes, the superiority of the proposed framework is demonstrated.

In the simulation, we select a 120 m × 120 m experimental scenario for modeling, with a depth of −100 m. We set the grid resolution to 0.5 m, the time interval to 1 s, and the total task time to 100 s. Then, the number of vortices and obstacles are set to 4 and 5, with their positions randomly assigned and r_k being 5 m. In addition, the positions of the AUVs and the target are randomly assigned. Additionally, the parameters of the sonar equation are set as follows: emission strength of the information source SL = 120 dB, detection threshold DT = 20 dB, and directivity index DI = 3 dB. The dynamic parameters of the underwater vehicle are set as follows: maximum velocity $v_{\max}$ is 2 m/s, maximum acceleration $a_{\max}$ is 1 m/s^2, and maximum angular velocity $\omega_{\max}$ is π rad/s. Additionally, the main parameters of the proposed algorithm are set as follows: training steps are 1500, discount factor is 0.99, soft update rate is 0.005, and batch size is 64. Detailed parameters are presented in Table 2. Additionally, the target is controlled by reinforcement learning, exhibiting the gaming feature with the AUV swarm. Furthermore, to demonstrate the superiority of the proposed scheme, two baseline schemes are selected to evaluate the performance. The details of the baseline schemes are shown as follows:

Table 2. Simulation Parameters.

	Parameters	**Value**
System parameters	The detection threshold DT	20 dB
	The directivity index DI	3 dB
	The sound frequency f of underwater communication	20 kHz
	The emission strength of the information source SL	120 dB
	The maximum velocity $v_{\max}$ of AUVs	2 m/s
	The maximum acceleration $a_{\max}$	1 m/s^2
	The maximum angular velocity $\omega_{\max}$	π rad/s
	The scaling factor ℓ	0.5
	The Strength Λ of the vortex	1
	The radius ς of the vortex	40 m
	The AUV's cross-sectional area C_1	1.0
	The AUV's dragging coefficient C_2	0.6
Proposed algorithm parameters	The total time of the tracking task	100 s
	Online training episodes	1500
	Offline training episodes	1500
	Discount factor	0.99
	Soft update rate	0.005
	Batch size	64
	Memory capacity	1,000,000

Baseline 1: As a compared scheme, it considers tracking reward, communication reward, monitoring reward, and safety reward. Additionally, in the target-tracking task, the AUV swarm employs the online MATD3 algorithm.

Baseline 2: Compared to our scheme, this scheme only considers tracking reward, communication reward, and monitoring reward. Additionally, the AUV swarm adopts the online MADDPG algorithm [52] to perform target-tracking tasks.

First, the average rewards of different schemes during training are compared, where the average reward is calculated using a moving-average window method. The average rewards of the three schemes are illustrated in Figure 3. All three schemes eventually converge, while the proposed framework exceeds the baseline methods with respect to total reward. Furthermore, the proposed framework exhibits higher average rewards with smaller fluctuations, indicating significantly better convergence stability and performance than the baseline methods.

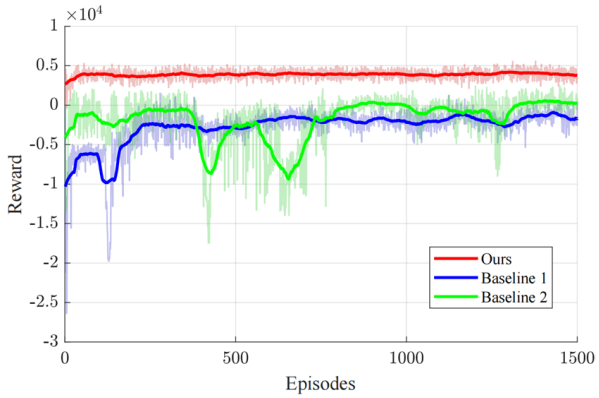

Figure 3. The average reward comparison of different schemes.

Figure 4 illustrates the trajectories of the AUV swarm and the target under the proposed framework, where the target's maximum velocity is 1 m/s. The target randomly appears within the monitoring area of the AUV swarm and then begins to escape while being tracked by the swarm. As shown in Figure 4, the AUVs consistently maintain an appropriate distance from the target while preserving sufficient spacing among themselves. This behavior results from the inclusion of a safety reward, together with an AIE reward that encourages dispersion among AUVs. Furthermore, the AUV swarm achieves stable target tracking owing to the combined effects of the monitoring and tracking rewards. These rewards collectively drive the swarm to maintain consistent monitoring and reliable tracking of the target.

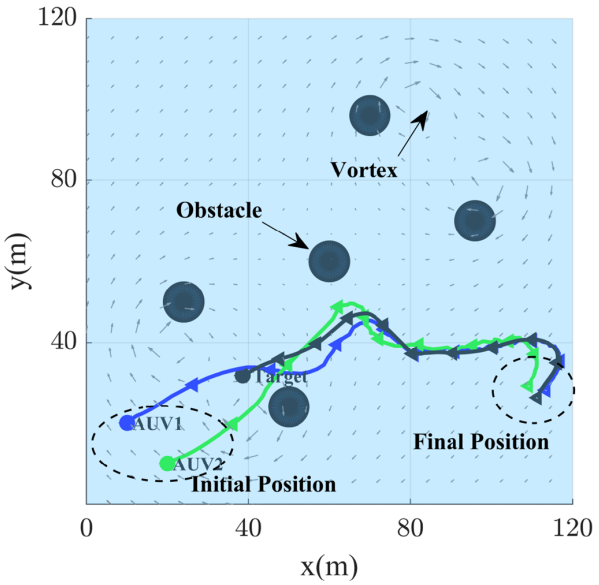

Figure 4. Trajectories of the AUV swarm and the target in a complex marine environment (scenario with two AUVs and one target).

Figure 5 presents the total energy consumption of different schemes under various target maximum velocities. As shown in Figure 5, the total energy consumption gradually increases with the target's maximum velocity. This trend occurs because a higher target velocity increases maneuvering demands, forcing the AUV swarm to expend more energy to enhance its own agility for effective target tracking. Furthermore, compared with the baseline methods, the proposed framework achieves significantly lower energy consumption. This improvement is attributed to the inclusion of an energy-consumption reward, which effectively reduces the total energy usage.

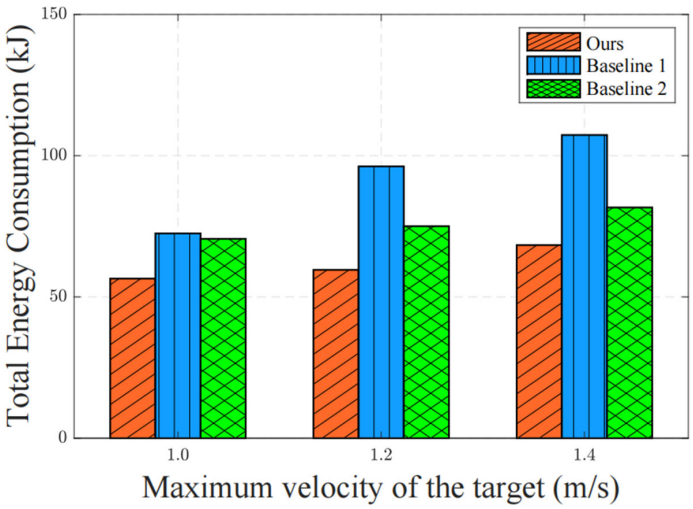

Figure 5. Total energy consumption of different schemes under different maximum velocities of the target.

Figure 6 presents the mission success rates of the three schemes under various target maximum velocities. The target-tracking mission is considered unsuccessful when communication within the AUV swarm is interrupted or collisions occur. As the target's maximum velocity increases, the success rate of the proposed framework remains stable and significantly surpasses that of the baseline methods. This improvement results from the AIE reward, which further promotes decentralized and cooperative target tracking among AUVs. In addition, the proposed framework incorporates a safety reward that enhances the collision avoidance capability of the AUVs.

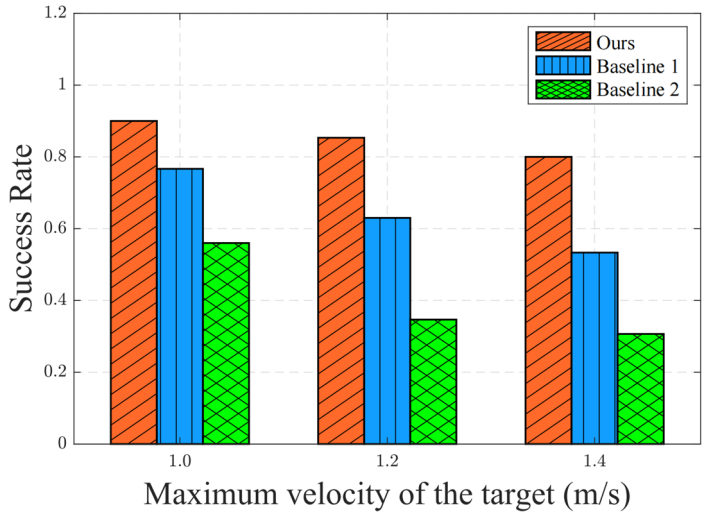

Figure 6. Success rate of different schemes under different maximum velocities of the target.

To show the safety performance, the inter-AUV distances over the task duration are plotted in Figure 7. The results indicate that the proposed framework significantly outperforms the baseline methods in terms of maintaining larger minimum inter-AUV distances. This improvement can be attributed to the AIE reward, which promotes decentralized coordination within the AUV swarm, thereby indirectly increasing inter-AUV distances and enhancing the overall safety. Furthermore, compared with the baseline methods, the inter-AUV distances in the proposed framework tend to stabilize over time, further highlighting its superior stability and safety performance.

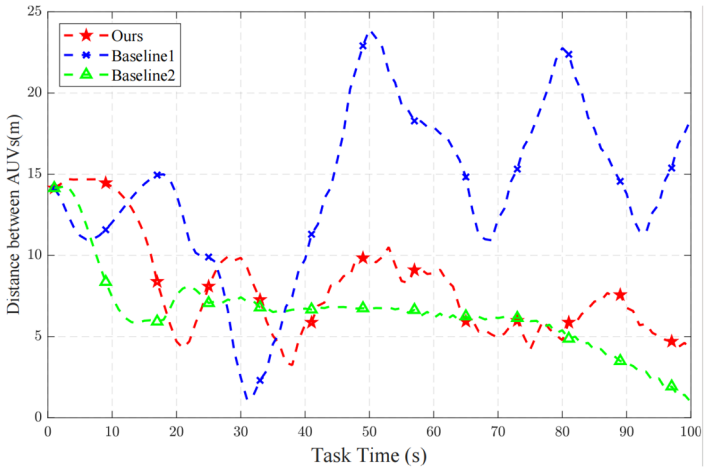

Figure 7. Distances between AUVs versus task time of different schemes (scenario with two AUVs and one target).

To evaluate the stability and safety of different schemes during the target-tracking process, the distances between AUVs and the target are compared in Figure 8. The AUVs rapidly approach the target and maintain stable tracking, demonstrating superior tracking stability compared with the baseline methods. Moreover, the proposed framework achieves a larger minimum distance between the AUVs and the target than the baseline methods. This improvement is attributed to the integration of an LSTM layer, which enhances the stability of the MATD3 algorithm and enables the AUVs to capture richer environmental information. Furthermore, the proposed framework incorporates a safety reward that enables the AUVs to maintain an appropriate safe distance from the target.

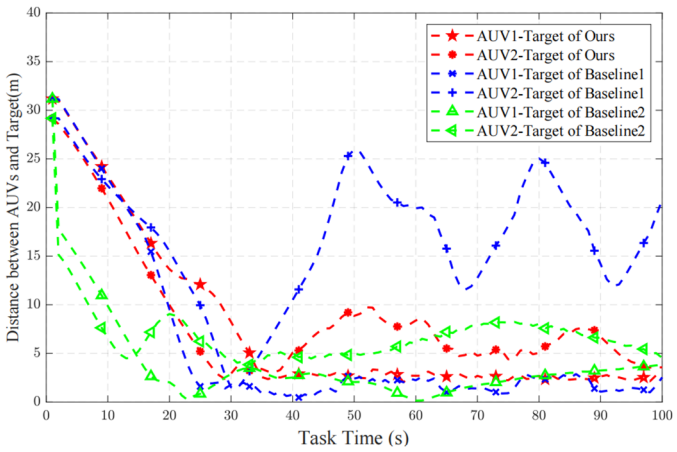

Figure 8. Distances between AUVs and target versus task time of different schemes (scenario with two AUVs and one target).

To evaluate the scalability of the proposed framework, an additional experiment is conducted in which three AUVs are employed to track an escaping target. The target randomly appears within the monitoring area and then begins to escape while being tracked by the AUV swarm. As shown in Figure 9, the AUVs consistently maintain safe distances from one another and from the target. This behavior results from the inclusion of both a safety reward and an AIE reward, which encourage the AUVs to track the target in a safe and decentralized manner. In addition, the monitoring and tracking rewards enable the AUVs to maintain stable and continuous target tracking.

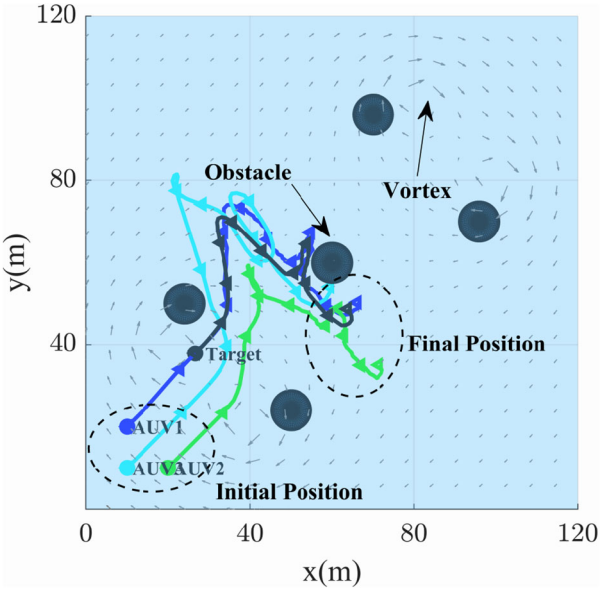

Figure 9. Trajectories of the AUV swarm and the target in a complex marine environment (scenario with three AUVs and one target).

In addition, the minimum distances among different AUVs over the task duration are compared, as shown in Figure 10. The results indicate that the proposed framework outperforms the baseline methods in terms of maintaining larger minimum inter-AUV distances. This improvement is attributed to the AIE reward, which promotes decentralized coordination among AUVs. Furthermore, the integration of an LSTM layer enhances the MATD3 algorithm, enabling the proposed framework to capture richer environmental information and achieve superior control performance.

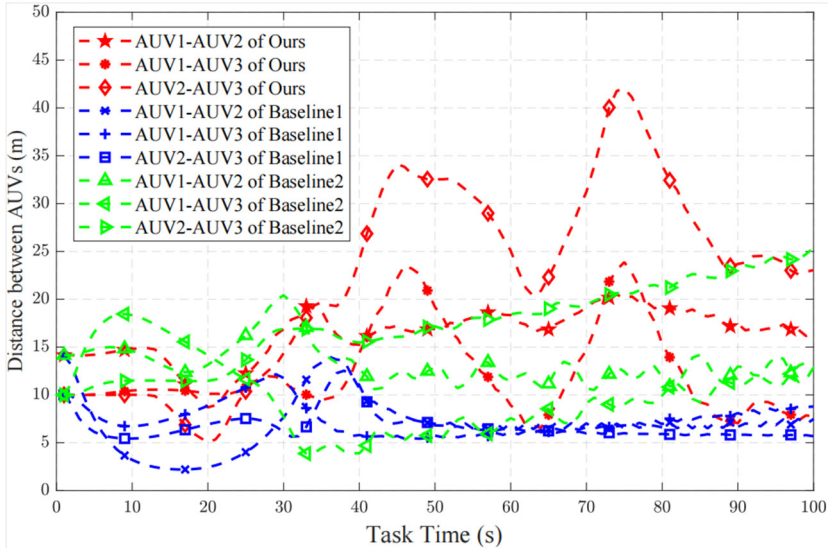

Figure 10. Distances between AUVs versus task time of different schemes (scenario with three AUVs and one target).

To evaluate the stability and safety of different schemes, the distances between AUVs and the target are compared in Figure 11. The AUVs gradually approach the target and then stabilize their tracking, with the proposed framework outperforming the baseline methods. This improvement is attributed to the inclusion of a safety reward, which enables the AUVs to maintain an appropriate safe distance from the target throughout the tracking process.

Furthermore, the integration of an LSTM layer enhances the MATD3 algorithm, thereby improving the AUVs' feature extraction capability and overall tracking performance.

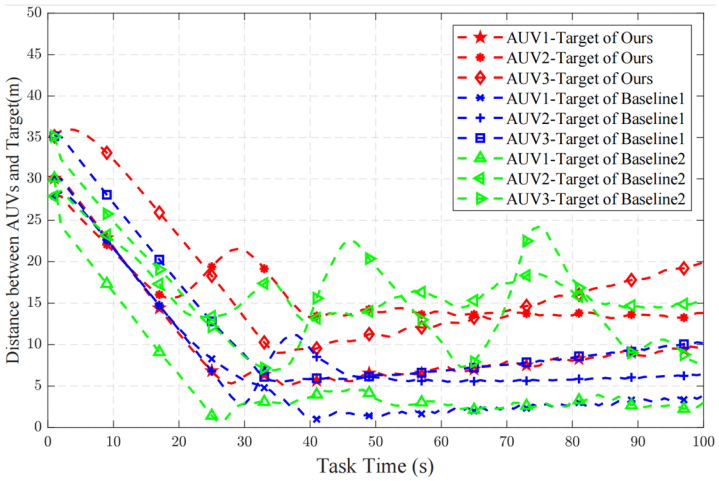

Figure 11. Distances between AUVs and target versus task time of different schemes (scenario with three AUVs and one target).

5. Conclusions

This paper proposes an energy-efficient distributed AUV swarm target-tracking scheme, which balances energy conservation and coverage range. We integrated the AIE mechanism into the distributed strategy of the AUV swarm, which significantly enhance the AUVs' decentralization capabilities during the target-tracking process, thereby indirectly improving mission success rates. In addition, we introduce real-time energy consumption into the AUV's control strategy, which helps AUVs reduce total energy consumption during target tracking. Moreover, we combine the LSTM layer into the neural network of MARL, which assists MARL in extracting more time-related feature information, thereby enhancing the control performance of the AUV swarm. Furthermore, we employ an offline-to-online RL method to improve sample efficiency and training performance. Experimental results demonstrate that the proposed scheme outperforms the baseline schemes in terms of success rate, energy consumption, and the coverage range, while also exhibiting excellent scalability. Future physical experiments will be applied to verify the algorithm under realistic environmental conditions, and establish a stronger foundation for practical deployment in the AUV swarm collaborative tasks.

Author Contributions: Conceptualization, R.L.; methodology, R.L. and D.L.; software, R.L. and D.L.; validation, R.L., D.L. and X.Z.; formal analysis, R.L., D.L. and X.Z.; investigation, R.L. and D.L.; resources, X.Z. and W.N.; data curation, R.L. and W.N.; writing—original draft preparation, R.L. and D.L.; writing—review and editing, R.L., D.L., X.Z. and W.N.; visualization, R.L. and D.L.; supervision, X.Z. and W.N.; project administration, X.Z.; funding acquisition, X.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Ministry of Science and Technology of the People's Republic of China (Grant No.: Ys2021YFC200328).

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Conflicts of Interest: Author Renbo Li was employed by the Second Institute of China Aerospace Science and Industry Corporation. Author Xiangxin Zhang was employed by the Research and Development Department, Shanghai Internet of Things Co., LTD. (SHIOTC). The remaining authors

declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Sahoo, A.; Dwivedy, S.K.; Robi, P. Advancements in the field of autonomous underwater vehicle. *Ocean Eng.* **2019**, *181*, 145–160. [CrossRef]
2. Wang, H.; Tong, S.; Wang, A.; Zhang, W.; Hu, Z.; Peng, Z. Heterogeneous cross domain coordinated control of ASV-AUV system for maritime search and rescue. *Ocean Eng.* **2024**, *306*, 117950. [CrossRef]
3. Zou, T.; Situ, W.; Yang, W.; Zeng, W.; Wang, Y. A method for long-term target anti-interference tracking combining deep learning and CKF for LARS tracking and capturing. *Remote Sens.* **2023**, *15*, 748. [CrossRef]
4. Cai, W.; Zhang, M.; Zheng, Y.R. Task Assignment and Path Planning for Multiple Autonomous Underwater Vehicles Using 3D Dubins Curves †. *Sensors* **2017**, *17*, 1607. [CrossRef]
5. Qiao, Q.; Kong, X.; Wu, S.; Liu, G.; Zhang, G.; Yang, H.; Zhang, W.; Yang, Y.; Jia, L.; He, C. A bio-inspired MEMS wake detector for AUV tracking and coordinated formation. *Remote Sens.* **2023**, *15*, 2949. [CrossRef]
6. Deghat, M.; Shames, I.; Anderson, B.D.; Yu, C. Localization and circumnavigation of a slowly moving target using bearing measurements. *IEEE Trans. Autom. Control* **2014**, *59*, 2182–2188. [CrossRef]
7. Zhang, C.; Cheng, P.; Du, B.; Dong, B.; Zhang, W. AUV path tracking with real-time obstacle avoidance via reinforcement learning under adaptive constraints. *Ocean Eng.* **2022**, *256*, 111453. [CrossRef]
8. Pang, W.; Zhu, D.; Chen, M.; Xu, W. RG-SAPF: A Scheme for Cooperative Escorting of Underwater Moving Target by Multi-AUV Formation Systems Based on Rigidity Graph and Safe Artificial Potential Field. *Sensors* **2025**, *25*, 6823. [CrossRef]
9. Hao, K.; Zhao, J.; Li, Z.; Liu, Y.; Zhao, L. Dynamic path planning of a three-dimensional underwater AUV based on an adaptive genetic algorithm. *Ocean Eng.* **2022**, *263*, 112421. [CrossRef]
10. Yu, X.; Chen, W.-N.; Hu, X.-M.; Gu, T.; Yuan, H.; Zhou, Y.; Zhang, J. Path planning in multiple-AUV systems for difficult target traveling missions: A hybrid metaheuristic approach. *IEEE Trans. Cogn. Dev. Syst.* **2019**, *12*, 561–574. [CrossRef]
11. Liang, H.; Fu, Y.; Kang, F.; Gao, J.; Qiang, N. A behavior-driven coordination control framework for target hunting by UUV intelligent swarm. *IEEE Access* **2020**, *8*, 4838–4859. [CrossRef]
12. Ni, J.; Yang, L.; Wu, L.; Fan, X. An improved spinal neural system-based approach for heterogeneous AUVs cooperative hunting. *Int. J. Fuzzy Syst.* **2018**, *20*, 672–686. [CrossRef]
13. Li, Y.; He, X.; Lu, Z.; Jing, P.; Su, Y. Comprehensive ocean information-enabled auv motion planning based on reinforcement learning. *Remote Sens.* **2023**, *15*, 3077. [CrossRef]
14. Jiang, W.; Yang, X.; Tong, F.; Yang, Y.; Zhou, T. A low-complexity underwater acoustic coherent communication system for small AUV. *Remote Sens.* **2022**, *14*, 3405. [CrossRef]
15. Cao, X.; Zhu, D. A survey of cooperative hunting control algorithms for multi-AUV systems. In Proceedings of the 32nd Chinese Control Conference, Xi'an, China, 26–28 July 2013; pp. 5791–5795.
16. Liu, Q.; Min, B.; Gao, J.; Wang, M.; Xu, G. Hierarchical practical formation tracking control for AUVs with predictive motion planning. *Ocean Eng.* **2025**, *321*, 120381. [CrossRef]
17. Wei, D.; Wang, F.; Ma, H. Autonomous Path Planning of AUV in Large-Scale Complex Marine Environment Based on Swarm Hyper-Heuristic Algorithm. *Appl. Sci.* **2019**, *9*, 2654. [CrossRef]
18. Fang, Z.; Wang, J.; Du, J.; Hou, X.; Ren, Y.; Han, Z. Stochastic optimization-aided energy-efficient information collection in Internet of Underwater Things networks. *IEEE Internet Things J.* **2021**, *9*, 1775–1789. [CrossRef]
19. Kulkarni, I.S.; Pompili, D. Task allocation for networked autonomous underwater vehicles in critical missions. *IEEE J. Sel. Areas Commun.* **2010**, *28*, 716–727. [CrossRef]
20. Juwaied, A.; Jackowska-Strumillo, L.; Majchrowicz, M. Enhanced Distributed Energy-Efficient Clustering (DEEC) Protocol for Wireless Sensor Networks: A Modular Implementation and Performance Analysis. *Sensors* **2025**, *25*, 4015. [CrossRef]
21. Juwaied, A.; Jackowska-Strumillo, L. DL-HEED: A Deep Learning Approach to Energy-Efficient Clustering in Heterogeneous Wireless Sensor Networks. *Appl. Sci.* **2025**, *15*, 8996. [CrossRef]
22. Zhang, Y.; Mou, Z.; Gao, F.; Jiang, J.; Ding, R.; Han, Z. UAV-enabled secure communications by multi-agent deep reinforcement learning. *IEEE Trans. Veh. Technol.* **2020**, *69*, 11599–11611. [CrossRef]
23. Zhang, K.; Ye, Y.; Chen, K.; Li, Z.; Li, K. Enhanced AUV Autonomy Through Fused Energy-Optimized Path Planning and Deep Reinforcement Learning for Integrated Navigation and Dynamic Obstacle Detection. *J. Mar. Sci. Eng.* **2025**, *13*, 1294. [CrossRef]
24. Bu, F.; Luo, H.; Ma, S.; Li, X.; Ruby, R.; Han, G. AUV-Aided Optical—Acoustic Hybrid Data Collection Based on Deep Reinforcement Learning. *Sensors* **2023**, *23*, 578. [CrossRef]
25. Yu, X.; Sun, Y.; Wang, X.; Zhang, G. End-to-end AUV motion planning method based on soft actor-critic. *Sensors* **2021**, *21*, 5893. [CrossRef]

26. Zhou, C.; Wu, W.; He, H.; Yang, P.; Lyu, F.; Cheng, N.; Shen, X. Deep reinforcement learning for delay-oriented IoT task scheduling in SAGIN. *IEEE Trans. Wirel. Commun.* **2020**, *20*, 911–925. [CrossRef]
27. Wang, Y.; Wang, W.; Chen, D. Knowledge-Guided Reinforcement Learning with Artificial Potential Field-Based Demonstrations for Multi-Autonomous Underwater Vehicle Cooperative Hunting. *J. Mar. Sci. Eng.* **2025**, *13*, 423. [CrossRef]
28. Zhang, M.; Chen, H.; Cai, W. Hunting task allocation for heterogeneous multi-AUV formation target hunting in IoUT: A game theoretic approach. *IEEE Internet Things J.* **2023**, *11*, 9142–9152. [CrossRef]
29. Xing, T.; Wang, X.; Ding, K.; Ni, K.; Zhou, Q. A multi-source-data-assisted AUV for path cruising: An energy-efficient DDPG approach. *Remote Sens.* **2023**, *15*, 5607. [CrossRef]
30. Xia, Z.; Du, J.; Wang, J.; Jiang, C.; Ren, Y.; Li, G.; Han, Z. Multi-agent reinforcement learning aided intelligent UAV swarm for target tracking. *IEEE Trans. Veh. Technol.* **2021**, *71*, 931–945. [CrossRef]
31. Wang, C.-C.; Wang, Y.-L.; Shi, P.; Wang, F. Scalable-MADDPG-based cooperative target invasion for a multi-USV system. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, *35*, 17867–17877. [CrossRef]
32. Song, D.; Gan, W.; Yao, P.; Zang, W.; Zhang, Z.; Qu, X. Guidance and control of autonomous surface underwater vehicles for target tracking in ocean environment by deep reinforcement learning. *Ocean Eng.* **2022**, *250*, 110947. [CrossRef]
33. Zhu, S.; Han, G.; Lin, C.; Zhang, F. Underwater Multiple AUV Cooperative Target Tracking Based on Minimal Reward Participation-Embedded MARL. *IEEE Trans. Mob. Comput.* **2025**, *24*, 4169–4182. [CrossRef]
34. Yang, Z.; Du, J.; Xia, Z.; Jiang, C.; Benslimane, A.; Ren, Y. Secure and cooperative target tracking via AUV swarm: A reinforcement learning approach. In Proceedings of the 2021 IEEE Global Communications Conference (GLOBECOM), Madrid, Spain, 7–11 December 2021; pp. 1–6.
35. Zhou, C.; Wang, C.; Bao, L.; Gao, X.; Gong, J.; Tan, M. Frequency Diversity Array Radar and Jammer Intelligent Frequency Domain Power Countermeasures Based on Multi-Agent Reinforcement Learning. *Remote Sens.* **2024**, *16*, 2127. [CrossRef]
36. Liu, P.; An, K.; Lei, J.; Sun, Y.; Liu, W.; Chatzinotas, S. Computation rate maximization for SCMA-aided edge computing in IoT networks: A multi-agent reinforcement learning approach. *IEEE Trans. Wirel. Commun.* **2024**, *23*, 10414–10429. [CrossRef]
37. Wu, T.; Zhou, P.; Liu, K.; Yuan, Y.; Wang, X.; Huang, H.; Wu, D.O. Multi-agent deep reinforcement learning for urban traffic light control in vehicular networks. *IEEE Trans. Veh. Technol.* **2020**, *69*, 8243–8256. [CrossRef]
38. Lee, D.; Kwon, M. Scenario-free autonomous driving with multi-task offline-to-online reinforcement learning. *IEEE Trans. Intell. Transp. Syst.* **2025**, *26*, 13317–13330. [CrossRef]
39. Xu, J.; Zhang, Z.; Wang, J.; Han, Z.; Ren, Y. Multi-AUV pursuit-evasion game in the Internet of Underwater Things: An efficient training framework via offline reinforcement learning. *IEEE Internet Things J.* **2024**, *11*, 31273–31286. [CrossRef]
40. Xing, T.; Wang, X.; Ding, K.; Ni, K.; Zhou, Q. Improved Artificial Potential Field Algorithm Assisted by Multisource Data for AUV Path Planning. *Sensors* **2023**, *23*, 6680. [CrossRef]
41. Yang, S.; Liu, X.; Su, Y. A Traffic-Aware Fair MAC Protocol for Layered Data Collection Oriented Underwater Acoustic Sensor Networks. *Remote Sens.* **2023**, *15*, 1501. [CrossRef]
42. Wang, H.; Huang, Y.; Luo, F.; Yang, L. Multi-Node Joint Power Allocation Algorithm Based on Hierarchical Game Learning in Underwater Acoustic Sensor Networks. *Remote Sens.* **2022**, *14*, 6215. [CrossRef]
43. Stojanovic, M. On the relationship between capacity and distance in an underwater acoustic communication channel. *ACM SIGMOBILE Mob. Comput. Commun. Rev.* **2007**, *11*, 34–43. [CrossRef]
44. Garau, B.; Alvarez, A.; Oliver, G. AUV navigation through turbulent ocean environments supported by onboard H-ADCP. In Proceedings of the 2006 IEEE International Conference on Robotics and Automation, Orlando, FL, USA, 15–19 May 2006; pp. 3556–3561.
45. Zeng, Z.; Sammut, K.; Lammas, A.; He, F.; Tang, Y. Efficient path re-planning for AUVs operating in spatiotemporal currents. *J. Intell. Robot. Syst.* **2015**, *79*, 135–153. [CrossRef]
46. Gong, H.; Er, M.J.; Liu, Y.; Ma, C. Three-dimensional optimal trajectory tracking control of underactuated AUVs with uncertain dynamics and input saturation. *Ocean Eng.* **2024**, *298*, 116757. [CrossRef]
47. Tan, K.M.; Lu, T.-F.; Anvar, A. Drag coefficient estimation model to simulate dynamic control of autonomous underwater vehicle (AUV) motion. In Proceedings of the 20th International Congress on Modelling and Simulation, Adelaide, Australia, 1–6 December 2013; pp. 1–6.
48. Lv, Y.; Lei, J.; Yi, P. A Local Information Aggregation-Based Multiagent Reinforcement Learning for Robot Swarm Dynamic Task Allocation. *IEEE Trans. Neural Netw. Learn. Syst.* **2025**, *36*, 10437–10449. [CrossRef]
49. Zhou, C.; Li, J.; Shi, Y.; Lin, Z. Research on Multi-Robot Formation Control Based on MATD3 Algorithm. *Appl. Sci.* **2023**, *13*, 1874. [CrossRef]
50. Zheng, J.; Li, K.; Mhaisen, N.; Ni, W.; Tovar, E.; Guizani, M. Exploring deep-reinforcement-learning-assisted federated learning for online resource allocation in privacy-preserving EdgeIoT. *IEEE Internet Things J.* **2022**, *9*, 21099–21110. [CrossRef]

51. Chen, J.; Luo, N. An offline-to-online reinforcement learning framework with trajectory-guided exploration for industrial process control. *J. Process Control* **2025**, *154*, 103535. [CrossRef]
52. Lowe, R.; Wu, Y.I.; Tamar, A.; Harb, J.; Pieter Abbeel, O.; Mordatch, I. Multi-agent actor-critic for mixed cooperative-competitive environments. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 6382–6393.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.