Hierarchical Heterogeneous Multi-Agent Cross-Domain Search Method Based on Deep Reinforcement Learning

Shangqun Dong, Meiqin Liu, *Senior Member, IEEE*, Shanling Dong, *Member, IEEE*, Ronghao Zheng, *Member, IEEE*, and Ping Wei, *Senior Member, IEEE*

***Abstract*— Marine target searching is a complex task due to large search areas, unique signal propagation characteristics, and limited visibility, posing significant challenges for single-agent or homogeneous multi-agent systems. In response, we propose a novel hierarchical heterogeneous multi-agent (HHMA) framework designed for underwater search scenarios. This framework integrates three types of vehicles moving in different domains—unmanned aerial, surface, and underwater vehicles, effectively overcoming the limitations of single or double-agent configurations. We begin by elucidating the advantages of the HHMA system in target searching, providing the kinematic modeling, while also transforming sonar detecting data and defining the search problem. The mission is decomposed to three human-comprehensible subtasks that are adaptive to both environmental conditions and equipment capabilities: moving, target estimating and trajectory planning. The target estimating subtask is effectively modeled as a Markov Decision Process, retaining its memory capability. Additionally, we extend multi-agent reinforcement learning to multi-policy reinforcement learning, facilitating the training of interdependent policies. The efficacy of our approach is demonstrated through simulations, comparing it with rule-based methods. Simulation results underscore the significance of the HHMA system and validate the proposed training methodology.**

***Index Terms*— Hierarchical heterogeneous multi-agent, cross-domain, multi-policy reinforcement learning, target searching.**

I. Introduction

OCEANS offer humanity an abundance of resources, but our exploration and exploitation are constrained by the formidable challenges of the sea. Unmanned vehicles have emerged as valuable substitutes for humans, playing a pivotal role in diverse tasks such as search operations [1], [2], [3], [4], monitoring [3], [4], [5], [6], and target hunting [7].

Unmanned underwater vehicles (UUVs) stand out as popular tools for oceanic exploration, capable of navigating three-dimensional space in the ocean and leveraging sensors to gather valuable underwater information. However, with the escalating demand for oceanic missions, single agents are proved to be inadequate to meet human requirements [7], [8]. This has spurred a transition towards multi-agent systems, garnering significant attention from researchers.

While multi-UUVs effectively extend exploration capabilities [9], they face challenges such as energy shortages and the necessity for periodic surfacing to update positions and transmit signals. These limitations are inherent to the characteristics of UUVs and cannot be addressed merely by increasing their numbers [7]. Consequently, the evolution in this field spans from homogeneous agents [6] to heterogeneous agents [2], [3], [4], [7], [8], [10], and from single-domain [3] to cross-domain [2], [4], [7], [8], [10].

The first system to be considered is the system consisting of an unmanned surface vehicle (USV) and multiple UUVs [8]. In this configuration, the USV serves as a crucial signal relay between the UUVs and the extra-marine environment. UUVs can operate without the need for surfacing to update positions and transmit signals. Additionally, the USV is equipped to search for targets both on the sea surface and above it. However, it is important to note that the USV's search range is relatively limited, and its movement on the sea surface is constrained due to the necessity of maintaining communication with the UUV—a key participant in underwater target searches. Therefore, an unmanned aerial vehicle (UAV) is incorporated to address this limitation, which enhances the system's search range significantly, effectively expanding the capability to locate both surface and air targets [2], [3], [4], [7].

Here we present the establishment of a HHMA system, using a typical underwater target searching mission as a case study. Unlike many previous works that treat equipment as agents, in light of the aforementioned vehicles requirements and the need for ease of human comprehension, our approach decomposes the underwater target searching into three distinct

Manuscript received 18 February 2024; revised 25 April 2024; accepted 23 May 2024. This work was supported in part by the National Natural Science Foundation of China under Grant U23B2060, in part by the Joint Fund of Ministry of Education for Pre-Research of Equipment under Grant 8091B042220, and in part by the Fundamental Research Funds for Xi'an Jiaotong University under Grant xtr072022001. The Associate Editor for this article was J. Yan. *(Corresponding author: Meiqin Liu.)*

Shangqun Dong and Ping Wei are with the National Key Laboratory of Human-Machine Hybrid Augmented Intelligence, the National Engineering Research Center for Visual Information and Applications, and the Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: dongshangqun@stu.xjtu.edu.cn; pingwei@xjtu.edu.cn).

Meiqin Liu is with the National Key Laboratory of Human-Machine Hybrid Augmented Intelligence, Xi'an Jiaotong University, Xi'an 710049, China, and also with the College of Electrical Engineering, Zhejiang University, Hangzhou 310027, China (e-mail: liumeiqin@zju.edu.cn).

Shanling Dong and Ronghao Zheng are with the College of Electrical Engineering, Zhejiang University, Hangzhou 310027, China (e-mail: shanlingdong28@zju.edu.cn; rzheng@zju.edu.cn).

Digital Object Identifier 10.1109/TITS.2024.3417698

subtasks: moving, target estimating, and trajectory planning. Corresponding policies are trained for each of these subtasks. These subtasks align with human intuition, and their inputs and outputs are comprehensible for humans. By doing so, people can grasp the reasoning behind the intelligent system's decisions, eliminating the mystery of a complete black-box scenario. Besides, with this approach, the UUVs only need to undertake the simplest moving subtask, which will significantly reduce the UUVs' energy consumption for computation, thereby increasing its endurance. Meanwhile, the other complex subtasks are managed by the USV, aligning with its greater load capacity and extending the overall system's working time.

The moving policy commands UUVs to move towards the moving target position (MTP) with constraints. Though USVs move in two-dimensional surface and UAVs usually maintain a fixed altitude [4], their moving policies can be obtained with same reinforcement learning (RL) methods. The target estimating policy works to estimate the target area with the assistance of UUVs' underwater exploration, facilitating the system in rapidly locating the target. Simultaneously, the trajectory planning policy is responsible for determining the UUVs' MTPs to explore the underwater environment and approach the target for precise positioning.

As the moving subtask operates independently, we initially train this policy using the proximal policy optimization (PPO) [11] algorithm. Subsequently, the moving policy becomes an integral part of the training environment for other policies. The output of the target estimating policy initially adjusts environmental states, influencing the input for the trajectory planning policy. In turn, the output of the trajectory planning policy also interacts with the environment, further adjusting the input for the target estimating policy. Recognizing the interdependence between the target estimating and trajectory planning policies, we propose a multi-policy reinforcement learning (MPRL) method based on the multi-agent reinforcement learning (MARL) approach to simultaneously train the two interdependent policies that are serially coupled, resulting in excellent performance. Simulation results validate the effectiveness of the proposed method in solving underwater searching missions. The main contributions of this paper include:

1) A hierarchical heterogeneous multi-agent system is proposed with consideration for cross-domain connectivity and actual constraints.
2) A hierarchical mission decomposition method is introduced, which not only addresses the search mission within realistic constraints but also enhances human comprehension of the system.
3) The MPRL, a training method for serial coupled interdependent policies based on MARL, is proposed and can successfully train these policies simultaneously.

To ensure a coherent structure, the remainder of this paper is organized as follows: Section II provides a brief review of related research. Section III models the system to delineate the role of each equipment and elucidates the workflow of the entire system. The mission division process and the RL based solution to this problem is detailed in Section IV. In Section V, the simulation setting is introduced, and relevant simulation results are presented. Finally, Section VI concludes the study and suggests areas for future research.

II. Related Works

In the realm of underwater target searching missions, cross-domain heterogeneous multi-agent research has gained substantial attention due to its demonstrated superior system performance [2], even though it is still in its early stages. In April 2016, research institutions and companies in Norway and Portugal conducted an experiment to assess the capabilities of heterogeneous vehicles, incorporating state-of-the-art vehicle systems [12]. The study successfully demonstrated the superior performance of heterogeneous vehicle systems in executing oceanic missions.

The typical solution for a 3U system involves the UUV taking on the role of search, with the USV serving as a relay [4], [7], [10] or central control capable of processing cross-domain information [8]. Additionally, USV can assist UUV in achieving long-term high-precision positioning, a challenging task for UUV [4] alone in GPS-denied underwater environments [5]. Due to the fact that the detection and communication methods used in air cannot be directly applied underwater, UAVs are unable to contribute directly to underwater search missions. Consequently, they hover in the air, primarily detecting targets on the sea surface [1], integrating and transmitting system information to human.

Control methods for multiple vehicles are commonly classified into two types: centralized methods and distributed methods [13]. Centralized methods operate under the assumption that a central station is available and capable of controlling the entire group of vehicles. Conversely, distributed methods eliminate the need for a central station but come with the trade-off of a more complex system and control algorithm. Due to the increasing demand for agents' independence, distributed methods are considered more promising for homogeneous vehicles. However, in 3U system, the USV serves as a suitable central station for UUVs due to its communication relay role. By this way, the USV can obtain comprehensive information to make more informed decisions. Additionally, UUVs have data processing capabilities, albeit weaker, allowing them to execute small calculation missions, similar by UAVs. Therefore, the 3U system exhibits characteristics of centralized and distributed methods.

To establish the communicating framework of UUV-based underwater system, the software-defined networking (SDN) architecture is employed for UUV-based underwater wireless network, synchronizing network information and executing network operations [2], [8]. Various methods, such as rule-based approaches [1], [10], artificial potential field methods [8], particle swarm optimization (PSO) [3], [4], and RL [5], [6], [8], [9], [14], [15] are utilized for trajectory planning.

Most studies train policies to map sensor output and the system's current state to motion commands [7], [9], with objectives such as minimizing time [4], energy consumption [7] or tracking error [5], among others, offering a straightforward approach. However, the trained policy or

policy decision remains a black box for humans. The lack of transparency in the trained policy does not enhance human confidence in the system, which is crucial for potential system users.

Contrary to merely assigning missions to vehicles, our approach involves decomposing the overall task into multiple subtasks directly relevant to humans. Additionally, this approach is designed independently of the vehicle's attributes to make the policies independent and suitable for different vehicles. Each subtask is designed to be understandable for humans, significantly aiding in their comprehension of the mission.

However, the interdependence of some policies presents a challenge in their training. Hierarchical reinforcement learning (HRL) aims to simplify long-horizon RL tasks by breaking them down into a hierarchy of subtasks, where a higher-level policy learns to perform the task by selecting optimal subtasks as actions. Each subtask is itself a RL problem, with a lower-level policy learning to solve it [16], [17]. While HRL can simplify complex tasks, it is not specifically designed to address the training of interdependent policies, which is a critical aspect of our proposed method.

While earlier proposals laid the groundwork for collaborative efforts among multi-heterogeneous agents, they often encountered shortcomings during implementation in simulations. To optimize the simulation environment, numerous studies simplified the motion process by overlooking the vehicles' kinetics and opting for a discrete moving approach [8], [9]. Furthermore, some research constrained the movement of UUVs to a two-dimensional (2D) surface [5], [7], [8]. These simplifications deviate from the actual environmental conditions, impacting the deployment of methods in real-world scenarios, albeit at the cost of mitigating the complexity of policy training. Consequently, the same policy training methods may become inadequate in more intricate and realistic environments.

III. System Model and Problem Formulation

A. System Model

In this paper, we delve into a hierarchical multi-agent cross-domain system featuring a UAV and several basic formations. Each basic formation consists of a USV and three UUVs. Fig. 1 illustrates one UAV and two basic formations, providing a clearer representation of the system discussed in this paper.

The UUV serves as the primary executor for underwater target searching. Equipped with sonar-related instruments, UUVs utilize sonar signals to detect underwater targets by exploring the underwater environment.

The USV primarily functions as a communication relay between UUVs and the UAV. Additionally, it serves as the primary calculator for various types of information, benefiting from its superior loading capacity and electrical capabilities compared to the UUV and UAV.

The UAV serves as the air-based monitor, integrating all pre-processed information from the basic formations and sending it to the human command center (HCC) if necessary. Moreover, it provides a broader detection range in the air.

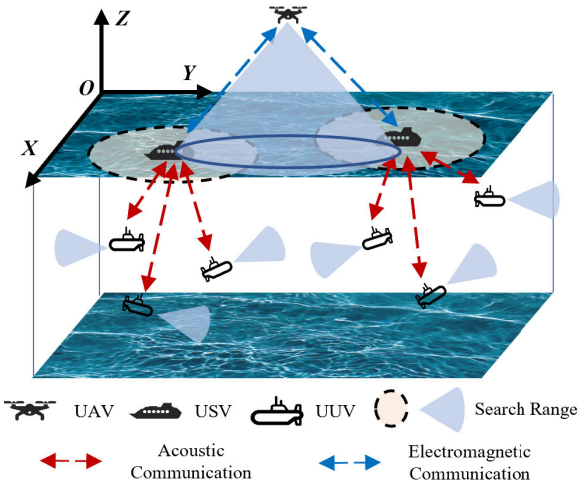


Fig. 1. Hierarchical multi-agent cross-domain system for underwater target searching.

The link between the UUV and USV relies on an acoustic channel, while an electromagnetic channel is utilized for signal transmission between the USV and UAV. All vehicles are abstracted as particles with directional navigation in three-dimensional space. In the ground frame, the sea surface level is defined as the OXY plane, with the Z-axis oriented vertically upward following the right-hand rule.

1) UUV: The 5-degree-of-freedom (DOF) mathematical model of an under-actuated UUV is established [18]. Generally, we make use of Fig. 2 to describe the motion coordinate system and its detected range of an UUV. $O-XYZ$ is earth-fixed coordinate system and $O'-X'Y'Z'$ is the UUV's body-fixed coordinate system. Fig. 2 describes the rotation coordinate transformation of fully actuated vehicles to adapt to all vehicles. In time slot t, the UUV rotates p^U, q^U, and r^U around the X'_t, Y'_t, and Z'_t axes in sequence and the body-fixed coordinate system changed to $O'_{t+1}-X'_{t+1}Y'_{t+1}Z'_{t+1}$, p^U, q^U, r^U refer to pitch, yaw, roll angular velocity in the body-fixed coordinate system with their positive direction following the right-hand rule. Then $\eta^U=\left[x^U, y^U, z^U\right]$ denotes the position of the UUV in the earth-fixed coordinate system, $\zeta^U=\left[\phi^U, \theta^U, \psi^U\right]$ represents pitch, yaw and roll angle in the earth-fixed coordinate system, with the positive direction following the right-hand rule. u^U, v^U, w^U are the velocities in surge (X of the body-axis coordinate system), sway (Y of the body-axis coordinate system), and heave (Z of the body-axis coordinate system),

Consider 6-DOF kinematics equation of an fully actuated UUV as

$$\dot{\epsilon}^U = T^U\left(\phi^U, \theta^U, \psi^U\right)\nu^U \tag{1}$$

where $\epsilon^U=\left[\eta^U, \zeta^U\right]^T \in R^6$ is the posture vector, and $\nu^U=\left[u^U, v^U, w^U, p^U, q^U, r^U\right]^T \in R^6$ is the velocity vector. The transformation matrix between $\dot{\epsilon}^U$ and ν^U can be written as [19]

$$\dot{\epsilon}^U=\begin{bmatrix} J_1^U & O_{3\times3} \\ O_{3\times3} & J_2^U \end{bmatrix}\nu^U \tag{2a}$$

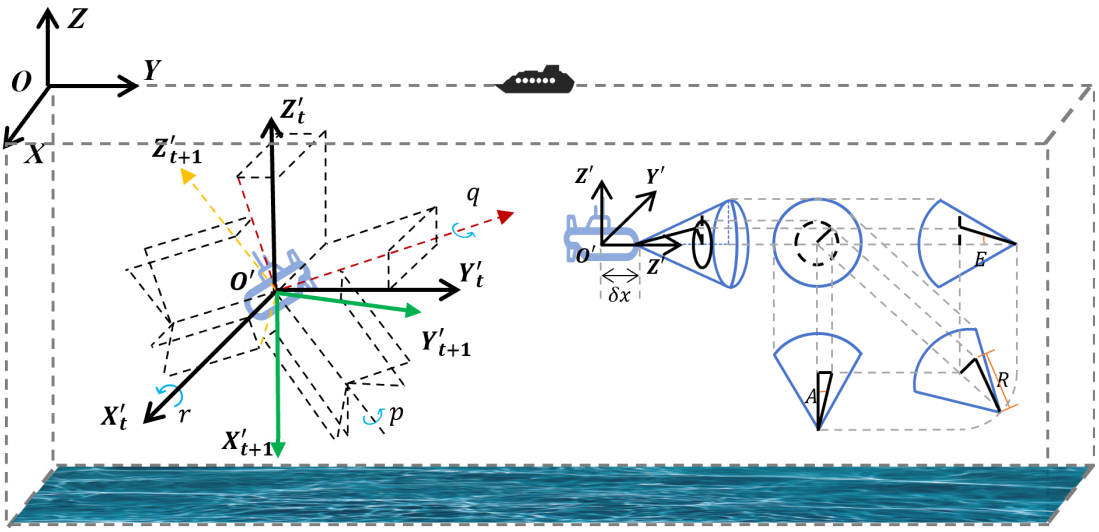


Fig. 2. Motion coordinated system of UUV and FLS sketch map.

$$J_1^U = \begin{bmatrix} c\psi c\theta & -s\psi c\phi + c\psi s\theta s\phi & s\psi s\phi + c\psi c\phi s\theta \\ s\psi c\theta & c\psi c\phi + s\phi s\theta s\psi & -c\psi s\phi + s\theta s\psi c\phi \\ -s\theta & s\theta s\phi & c\theta c\phi \end{bmatrix} \tag{2b}$$

$$J_2^U = \begin{bmatrix} 1 & s\phi t\theta & c\phi t\theta \\ 0 & c\phi & -s\phi \\ 0 & s\phi/c\theta & c\phi/c\theta \end{bmatrix} \tag{2c}$$

where $s\cdot := sin(\cdot)$, $c\cdot := cos(\cdot)$, $t\cdot := tan(\cdot)$, $\psi := \psi^U$, $\theta := \theta^U$, $\phi := \phi^U$.

As for most UUVs are under-actuated, we only discuss the 5-DOF UUVs in this paper, where $\phi^U = 0$ and $p^U = 0$. The kinematics equation is changed to

$$\dot{\epsilon}^U = T^U(\theta^U, \psi^U)\nu^U \tag{3a}$$

$$T^U = \begin{bmatrix} c\psi^U c\theta^U & -s\psi^U & c\phi^U s\theta^U & 0 & 0 \\ s\psi^U c\theta^U & c\psi^U & s\theta^U s\psi^U & 0 & 0 \\ -s\theta^U & 0 & c\theta^U & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1/c\theta^U \end{bmatrix} \tag{3b}$$

where $\epsilon^U = \left[x^U, y^U, z^U, \theta^U, \psi^U\right]^T$, $\nu^U = \left[u^U, v^U, w^U, q^U, r^U\right]^T$.

The UUVs navigate within a three-dimensional (3D) underwater environment, equipped with forward looking sonar (FLS) positioned at the front of the UUV [20], as depicted in Fig. 2. Without loss of generality, we assume that the detection range of FLS forms a rotational sector with a detected signal represented as $sig = [A, E, R]$, as illustrated in Fig. 2. Here, $A \in [-\theta_H, \theta_H]$ denotes the azimuth angle, $E \in [-\theta_V, \theta_V]$ represents the elevation angle, and $R \in [0, R_{max}]$ is the detected range. The parameters θ_H, θ_V, R_{max} represent the maximum azimuth angle, elevation angle, and detected range, respectively. The sonar signal [5], which is typically influenced by noise, is modeled as follows:

$$sig = sig_1 + sig_2 \tag{4a}$$

$$\frac{sig_2}{sig_1} = N(0,1) \times \max\left(\frac{abs(A)}{\theta_H}, \frac{abs(E)}{\theta_V}, \frac{R}{R_{max}}\right) \tag{4b}$$

where sig_1 is the accurate signal, and sig_2 is the noise signal related to the relative position. $N(0, 1)$ is the Gaussian distribution with a mean of 0 and a variance of 1. In this configuration, the closer the proximity to the periphery of the detected area, the higher the level of noise. The variable sig is defined with respect to the body-fixed coordinate system and can be transformed into the earth-fixed coordinate system to provide a clearer representation of the detected position through the following equations:

$$\eta_{sig} = J_1^U \left[x_{sig}, y_{sig}, z_{sig}\right]^T + \eta^U \tag{5a}$$

$$x_{sig} = \frac{R}{\sqrt{1 + \tan^2(A) + \sin^2(E)}} + \delta x \tag{5b}$$

$$y_{sig} = \sqrt{R^2 - 1 - \tan^2(A)} \tag{5c}$$

$$z_{sig} = \sqrt{R^2 - 1 - \sin^2(E)} \tag{5d}$$

where η_{sig} represents the signal position relative to the earth-fixed coordinate, $\left[x_{sig}, y_{sig}, z_{sig}\right]$ denotes the coordinates of the signal position in the body-fixed coordinate, and the sonar's origin is shifted by δx along the axis X' of body-fixed coordinate.

Finally, the signal is set as a 4D vector $\kappa = [\beta, \eta_{sig}]$, $\beta = 1$ if the target is within the sonar search range, else $\beta = 0$ and $\kappa = [0, 0, 0, 0]$.

2) USV: UUVs can establish signal communication with a specific USV within the same basic formation since a portion of the USV is submerged underwater. Subsequently, the USV can relay the UUVs' information to the external environment. This configuration enables UUVs to efficiently fulfill their missions, allowing them to focus more attentively on their tasks. Without assistance from the USV, UUVs are required to surface periodically to establish communication with the external environment and update their positions independently. The USV operates on the sea surface, having three degrees of freedom (DOF), namely ϕ^S, ψ^S, z^S, w^S, p^S, $r^S = 0$, so its

kinetic equation is expressed as:

$$\dot{\epsilon}^S = T^S\left(\theta^S\right)\nu^S \tag{6a}$$

$$T^S = \begin{bmatrix} c\theta^S & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{6b}$$

where $\epsilon^S = \left[x^S,\ y^S,\ \theta^S\right]^T$, $\nu^S = \left[u^S,\ v^S,\ q^S\right]^T$.

In this paper's context, USVs also play the role of surface and air target searchers, but they cannot move as the search mission demand, and their main function is to act as relays between UUVs and the external environment. Thus, their moving rule is defined as follows:

$$\min \sum (d^{S2U}) \tag{7a}$$

$$s.t.\ \max(d^{S2U}) < d^{S2U}_{max} \tag{7b}$$

where d^{S2U} is the distance between the USV and UUV. d^{S2U}_{max} is the maximum distance for establishing a reliable communication connection between the USV and UUV. Eq. 7a is used for minimum the total distance between USV and each UUVs in the same basic formation to minimum the communication energy consumption, and Eq. 7b is used to ensure USV can communicate with all UUVs in the same basic formation.

As the USV integrates information from all UUVs in the same basic formation, it can estimate the target's position and integrate it with the UAV's state in the same basic formation to issue position commands to these vehicles. The USV undertakes the most crucial and numerous tasks due to its capacity of heavy loads. The UAV and UUV themselves have relative limited energy, and to ensure more consistent task performance, it is necessary to minimize computational complexity of UAV and UUV as much as possible.

3) UAV: Given the distinct characteristics of UAV air mobility, it broadens the scope for environment observation, effectively compensating for the limitations in surface and air observation capabilities within the UUV-USV system. Concurrently, in the presence of multiple basic formations, the UAV can act as a central component for aggregating all formation information, facilitating comprehensive communication with the HCC. As UAV's energy load is consistently lower than that of USVs, and UAV is always visible to all USVs [7]. It can also designate a specific USV to communicate with the HCC after aggregating all formation information.

Given that the UAV converage radius can reach up to 2.5km when hovering at an altitude of 2km [15], there is no constraints on the distance between the UAV and the USVs in this paper, allowing the UAV to move as desired. Since there is only one UAV in the system, as shown in Fig. 1, the UAV can just hover in certain position or follow a predefined trajectory, such as the four fundamental search types proposed by the International Aeronautical and Maritime Search and Rescue [1], [21]. These search types encompass sector search missions, extended square search missions, parallel searches, and air-sea coordinated search missions, selected based on the mission requirements. Additionally, various extended search paths, such as extended sector search paths and coverage paths [1], are available for the UAV to follow, depending of the specific demands of different missions.

B. Problem Formulation

The main difficulties lies in the movement and control of UUVs, as USVs and UAV each have their own distinct motion rules. For UUVs, their movement is guided by commands transmitted by the USV, adhering to the same basic formation to explore the underwater environment and determine the underwater target's position using the equipped FLS. The underwater search mission in this paper is defined as follows:

$$\min \quad d^{U2T}_{min} \tag{8a}$$

$$\text{s.t.} \quad d^{U2T}_{min} = \min_{i\in\{0,1,\ldots,n\}} \left\| \eta^U_i(t) - \eta^T(t) \right\| \tag{8b}$$

$$\eta^U, \eta^T \in \left[\eta^U_{min},\ \eta^U_{max}\right] \tag{8c}$$

$$\zeta^U \in \left[\zeta^U_{min},\ \zeta^U_{max}\right] \tag{8d}$$

$$\nu^U \in \left[-\nu^U_b,\ \nu^U_b\right] \tag{8e}$$

$$a^U \in \left[-a^U_b,\ a^U_b\right] \tag{8f}$$

$$\max(d^{S2U}) < d^{S2U}_{max} \tag{8g}$$

where d^{U2T}_{min} is the minimum distance between USV and UUVs, $\eta^U_i = \left[x^U_i,\ y^U_i,\ z^U_i\right]$ is the ith UUVs' position, $i \in \{0,\ 1,\ \ldots,\ n\}$ is the index of UUV in the basic formation, η^T is the target position, η^U_{min} and η^U_{max} are the vector representing the minimum and maximum positions of UUVs and the target, ζ^U_{min} and ζ^U_{max} are the minimum and maximum angles of UUVs. a^U represents the UUV's acceleration, which serves as the control command in this paper. ν^U_b, a^U_b denote the boundaries of velocities and acceleration. The subscripts and time annotations are omitted in η^U, ν^U, a^U to simplify expressions.

The constraints encompass limitations on the position, angle, velocity, and acceleration of UUVs, defined by mission requirements or inherent attributes. Additionally, to ensure a reliable communication connection with the USV and external environment, we impose restrictions on the distance between the USV and UUVs within the same basic formation.

IV. Mission Division and Its RL-Based Solution

In accordance with cross-domain environmental attributes, we divide the underwater searching mission into three subtasks: moving, target estimating, and trajectory planning. The USV is responsible for the target estimating and trajectory planning subtasks, while the UUV handles the moving subtask.

The entire process of this mission is illustrated in Fig. 3, which also indicates the information required for each subtask. Initially, the simulation environment randomly resets relevant parameters within their boundaries. Subsequently, the USV estimates the target area based to related information, such as the posture and signal of UUVs. The USV plans the MTPs of UUVs and communicates them to the UUVs. Finally, the UUVs move to the their respective positions, and the entire

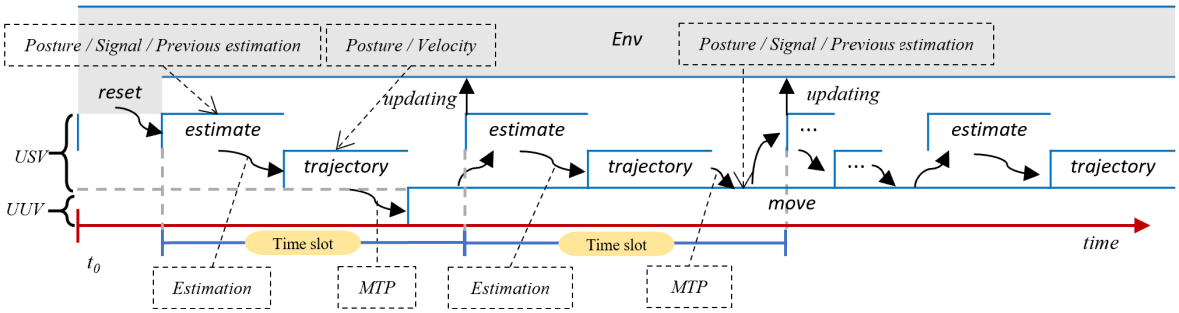


Fig. 3. Process of underwater searching mission.

process repeats until the mission is completed for information collection.

While integrating the target estimating policy and trajectory planning policy into a single policy is feasible, this approach will result in a lack of transparency in the agents' intentions. Such opacity could adversely affect the system's effectiveness, as human operators may interrupt or even prohibit the system. Additionally, under the current configuration, the inputs for all three policies consist of information comprehensible to humans. This design choice establishes an interface that allows human intervention in the mission execution if the system performs poorly. Human intelligence can sometimes outperforms artificial intelligence, and this interface ensures that humans can contribute their insights or propose more suitable solutions when needed.

A. Moving Subtask

The USV controls the UUV's motion by providing MTP to it, and the underwater traffic control problem can be modeled as an Markov decision process (MDP) $\{S^M, A^M, P^M, R^M, \gamma^M\}$.

- *State* S^M: S^M is the state space of moving policy, and we denote the state of UUV in the time slot t by $s^M(t) = \{\epsilon^U(t), v^U(t), \eta^{MTP}(t) - \eta^U(t)\}$, which consists of the posture and velocity of UUV, and the vector from UUV's current position to MTP in time slot t.
- *Action* A^M: the action space is defined as the acceleration of UUV which is denoted as $a^M(t) = a^U(t)$.
- *State Transition* P^M: The state transition function P^M models the probabilistic evolution of the state over time in response to actions taken. In this context, it governs the kinetic of the UUV as it moves. Given the current state $s^M(t)$ and the action taken $a^M(t)$, the subsequent state $s^M(t+1)$ is stochastically determined by the transition function. This probabilistic process can be expressed as:

$$s^M(t+1) \sim P^M(\cdot|s^M(t), a^M(t)) \tag{9}$$

  The precise form of the probability distribution depends on the specific modeling considerations and the underwater traffic control scenario.
- *Reward* R^M: after taking $a^M(t)$, the state transforms from $s^M(t)$ to $s^M(t+1)$, and the environment will give a reward $r^M(t)$ to judge the quality of moving policy $\pi^M(a^M(t)|s^M(t))$. The reward can prompt π^M to evolve to approach UUV to the MTP, which is defined as followed:

$$R^M(t) = 0.9R_1^M(t) + 0.1R_2^M(t) + R_3^M(t) \tag{10a}$$
$$R_1^M(t) = (d^{U2M}(t-1) - d^{U2M}(t))/u_{max}^U \tag{10b}$$
$$R_2^M(t) = -(\omega/\pi)^2 \tag{10c}$$
$$R_3^M(t) = \begin{cases} 10, & \text{if } d^{U2M}(t) < 1 \\ 0, & \text{if } d^{U2M}(t) \geq 1 \end{cases} \tag{10d}$$

  where $R_1^M(t)$ is the reward related to the distance difference of time slot t and $t-1$ between MTP and current position, $d^{U2M}(t) = \|\eta^U(t) - \eta^{MTP}(t)\|$ denotes the distance between UUV and MTP, $\|\cdot\|$ is the norm of vector, u_{max}^U is the UUV's maximum velocity in surge used for $R_1^M(t)$ normalization, and $R_2^M(t)$ is the reward corresponds to the angle ω between the current movement direction of the UUV and the direction of the line connecting the current position and MTP. $R_3^M(t)$ is the terminal reward when the distance between MTP and current position is less than 1.
- *Discount Factor* γ^M: The discount factor γ^M plays a crucial role in influencing the agent's decision-making by determining the importance of future rewards. In the context of the underwater traffic control problem, it represents the degree to which the UUV values long-term objectives in its motion policy. A higher γ^M assigns greater significance to future rewards, encouraging the UUV to consider the impact of its actions on the overall mission success over time. Conversely, a lower γ^M prioritizes more immediate rewards, potentially leading to a more myopic decision-making approach. The choice of an appropriate discount factor depends on the specific requirements and characteristics of the underwater environment, balancing the trade-off between short-term and long-term objectives in the UUV's motion control strategy.

The moving policy π^M is executed by UUV, and its training data collecting overview is shown in Fig. 4.

We employ proximal policy optimization (PPO) algorithm [11] to train policy π^M, and PPO is a RL algorithm employed for training artificial neural networks, particularly well-suited for problems with continuous action spaces. PPO aims to address the instability issues present in traditional policy gradient methods by introducing a technique known as "proximal policy clipping", ensuring

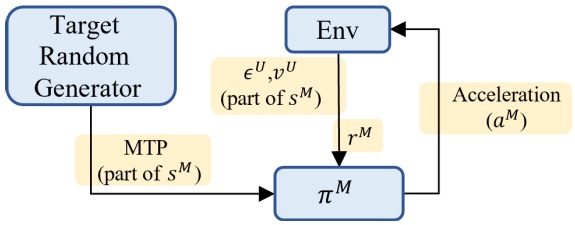


Fig. 4. Overview of training data collecting process for π^M.

relative conservatism in updating the policy with respect to the previous one.

As its core, PPO iteratively optimizes the policy while attempting to avoid introducing excessively large changes during updates to maintain stability throughout training. The algorithm achieves this by adjusting the policy through maximizing the objective of expected cumulative rewards, while concurrently controlling the magnitude of policy updates to overcome the instability issues inherent in policy gradient methods.

B. Target Estimating Subtask

By aggregating related states of all UUVs within the same basic formation and the estimation last time slot, the USV can estimate the target region by target estimating policy π^E. This estimate aids in planning the trajectories of UUVs, accelerating the search process, and enhancing the operator's understanding of the search pace.

While it is evident that the target estimating task benefits from historical data, a balance must be struck to alleviate data storage demands and reduce computational costs. Therefore, we propose a method for this estimating task by just using an uni-directional feedforward neural network (FNN). To utilize historical data more effectively, we incorporate the principles of recurrent neural network (RNN) [22], which has the ability to use internal state (memory) to process arbitrary sequences of inputs. We initialize the internal state with the estimating region from the previous time slot, which serves as part of the input for the next time slot's estimation. Additionally, the confidence parameter helps model the relationship between consecutive estimations. This approach allows FNN to retain a form of memory, providing some functionalities akin to RNNs but with a more streamlined network architecture.

To represent the target region, we model it as a globe, denoted as $C(t): (x-x^E(t))^2+(y-y^E(t))^2+(z-z^E(t))^2 \leq (r^E(t))^2$. Here, $O^E(t)=\left[x^E(t), y^E(t), z^E(t)\right]$ represents the estimating globe center coordinates of the target region, and $r^E(t)$ denotes the radius of the globe in time slot t.

Specifically, we use $O^E(t-1)$ and $r^E(t-1)$ as part of the input to the policy network. The action of the policy network is represented as $a^E(t)=\left\{O'^E(t),\ r'^E(t),\ \lambda(t)\right\}$, where $\lambda(t) \in [0,1]$ signifies the confidence associated with $[O'^E(t),\ r'^E(t)]$. By this way, the environment can adaptively use the historical experience to estimate the target position. Consequently, the last estimating result is given by $[O^E(t),\ r^E(t)] = (1-\lambda(t))\left[O^E(t-1),\ r^E(t-1)\right]+\lambda(t)\left[O'^E(t),\ r'^E(t)\right]$.

By adopting this configuration, only the last time slot's estimating results need to be saved and can be seen as a part of the current state. The estimation is independent on the past information, the estimating mission has the non-aftereffect property [23]. The target estimating task can therefore be modeled as an MDP denoted as $\left\{S^E, A^E, P^E, R^E, \gamma^T\right\}$, as outlined below:

- *State* S^E: The state encapsulates the posture and signal of UUVs within the same basic formation, along with the target position estimation of the previous time slot. It can be represented as $s^E(t) = \left\{\cup(\epsilon^U(t),\ \kappa(t)),\ O^E(t-1),\ r^E(t-1)\right\}$, where $O^E(0)$ and $r^E(0)$ are randomly initialized within predefined boundaries, $\cup(\epsilon^U(t),\ \kappa(t))$ means the aggregation of all UUVs' postures and signals at time t.
- *Action* A^E: The action is represented as the estimating value $a^E(t)=\left\{O'^E(t),\ r'^E(t), \lambda(t)\right\}$, with $\lambda(t) \in [0,1]$ indicating the estimating confidence in time slot t. The estimation in time slot t is calculated as follows:

$$[O^E(t), r^E(t)] = (1-\lambda(t))[O^E(t-1), r^E(t-1)] + \lambda(t)[O'^E(t), r'^E(t)] \tag{11}$$

- *State Transition* P^E: The state transition function P^E governs the probabilistic evolution of the state over time in response to actions taken. Notably, as the trajectory planning policy varies, P^E also changes over the course of training. Mathematically, this can be expressed as:

$$s^E(t+1) \sim P^E(t)(\cdot|s^E(t), a^E(t)) \tag{12}$$

- *Reward* R^E: The reward is associated with the distance between the true target position and the estimating target position. The reward is defined as follows:

$$R^E(t) = 0.7R_1^E(t) + 0.3R_2^E(t) + R_3^E(t) \tag{13a}$$

$$R_1^E(t) = d^{E2T}/d_{max}^{E2T} \tag{13b}$$

$$R_2^E(t) = \begin{cases} -0.1, & r^E(t) - d^{E2T} > 0, \\ 1, & 0 \leq r^E(t) - d^{E2T} < 5, \\ 0, & r^E(t) - d^{E2T} \geq 5 \end{cases} \tag{13c}$$

$$R_3^E(t) = \begin{cases} 5, & d^{E2T} \leq 10 \ \& \ r^E(t) - d^{E2T} > 0, \\ 0, & \text{others} \end{cases} \tag{13d}$$

Here, $R_1^E(t)$ is the reward associated with the distance between the true target position and the estimating target position, denoted as $d^{E2T} = \|O^E(t) - \eta^T(t)\|$. $R_2^E(t)$ penalizes the action if $\eta^T(t)$ is not within the estimating region, rewards it if $\eta^T(t)$ is within the estimating region, and avoids giving a reward if the estimated region is excessively large, as it may not positively impact the search mission. $R_3^E(t)$ is the terminal reward when the distance d^{E2T} is less than 10 and the true target position is within the estimating region.
- *Discount Factor* γ^E: The discount factor γ^E determines the importance of future rewards in the target estimating task.

C. Trajectory Planning Subtask

After the USV estimates the target region, the next step is to guide the UUVs to determine the target location. The USV takes the policy π^P to plan trajectories for UUVs. This process also can be modeled as an MDP denoted as $\{S^P, A^P, P^P, R^P, \gamma^P\}$ to address the problem.

- *State* S^P: the state comprises the posture and velocity of all UUVs and the estimating region, denoted as $s^P(t) = \{\cup(\epsilon^P(t),\ v^P(t)),\ O^E(t),\ r^E(t)\}$.
- *Action* A^P: the action is the MTPs of each UUV in the same basic formation which would be the next time slot input of the moving policy.
- *State Transition* P^P: The state transition function P^P governs the probabilistic evolution of the state over time in response to actions taken. Notably, as the target estimating policy varies, P^P also changes over the course of training. Mathematically, this can be expressed as:

$$s^P(t+1) \sim P^P(t)(\cdot|s^P(t), a^P(t)) \tag{14}$$

- *Reward* R^P: The HHMA system aims to swiftly search the target by the trajectory planning policy. Therefore, the reward is related to the distance between target and each UUV, as well as the distance between each pair of UUVs. The reward is defined as follows:

$$R^P(t) = 0.9R_1^P(t) + 0.1\ R_2^P(t) + R_3^P(t) \tag{15a}$$

$$R_1^P(t) = -\min(d^{U2E}(t)) \tag{15b}$$

$$R_2^P(t) = \sum(d^{U2U}(t)) \tag{15c}$$

$$R_3^P(t) = \begin{cases} 10, & \min(d^{U2E}(t)) < 5 \\ 0, & others \end{cases} \tag{15d}$$

where $R_1^P(t)$ is associated with the minimum distance between the estimating target position and each UUV ($d^{U2E}(t)$), aiming to encourage the nearest UUV to assess whether the estimating position is correct. $R_2^P(t)$ is the sum of distances between each pair of UUVs $d^{U2U}(t)$. This encourages UUVs to thoroughly explore the environment, aiding the target estimating policy π^E in accurately estimating the target position. R_3^P represents terminal reward, triggered when the minimum $d^{U2E}(t)$ is less than 5.

- *Discount Factor* γ^P: The discount factor γ^P determines the importance of future rewards in the target estimating task.

D. Multi-Policy Reinforcement Learning

As the output of target position estimating policy will change the environment before the trajectory planning policy been executed as shown in Fig. 5, the two serial coupled policies cannot be trained solely as the moving subtask can be trained rely on the target random generator as shown in Section. IV-A.

In the context of the target estimating task, the probabilistic nature of the state transition function P^E is influenced by changes in the environment when π^P undergoes variations during RL training. Moreover, the presence of γ^E introduces

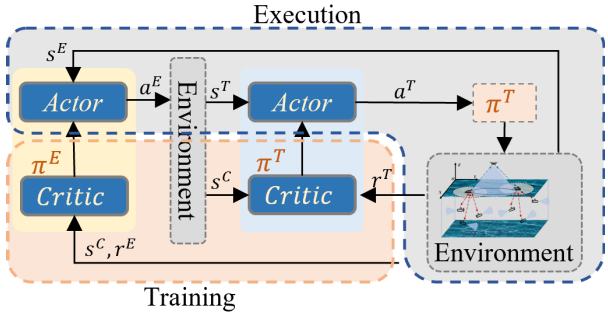


Fig. 5. Overview of the MPRL.

an additional layer of complexity, as different policies π^P not only affect their own future rewards but also impact the overall environmental dynamics. The same considerations extend to π^P.

Given the interdependence between these two policies, we introduce a MPRL approach, drawing inspiration from the principles of MARL. The training paradigm for both policies adheres to the centralized training and decentralized execution (CTDE) methodology, similar to MARL. Compared with the MARL, MPRL have following differences:

1) Asynchronous policy execution introduces a higher degree of dependency.
2) A different reward is given in real time as each policy is executed, while a team reward is given after all agents have executed their policies simultaneously for MARL.
3) The inputs for the two policies and their corresponding critic networks are obtained at different times. In contrast, in MARL, the inputs for all agents, along with those for their corresponding critic networks, if each has one, or for the shared critic, if present, are obtained simultaneously.

Recognizing that the previous policy can influence the state transition function of another policy, we strive to minimize the impact of prior policies on current updates. To achieve this, we employ an on-policy algorithm, specifically IPPO [24], which builds upon the PPO algorithm, for the simultaneous update of π^E and π^P.

The training procedure is graphically represented in Fig. 5. In this framework, we employ an actor-critic architecture, with the input for the critic networks denoted as $s_C^E(t) = \{\epsilon(t),\ v(t),\ sig(t),\ O^E(t-1),\ r^E(t-1),\ \eta^T\}$ and $s_C^P(t) = \{\epsilon(t),\ v(t),\ sig(t),\ O^E(t),\ r^E(t),\ \eta^T\}$. Notably, it is crucial to highlight that the input O^E and r^E differ between the critic networks of π^E and π^P within the same time slot, underscoring the sequential nature inherent in the operation of these two policies.

To be clearly, the pseudocode for the MPRL is presented in Algorithm 1. the pseudocode comprehensively outlines the processes involving the three policies and specifies the condition for pushing data to the buffer, detailing both the "how" and "when" aspects of this operation.

V. Simulation Results

A. Setting

All the settings are the same along the whole simulation.

Algorithm 1 Multi-Policy Reinforcement Learning

Initialize experience buffer $\mathcal{D}$:
- Critic: $\{s_C^E, ns_C^E, s_C^P, ns_C^P\}$
- Estimating: $\{s_A^E, a^E, ns_A^E, r^E\}$
- Planning: $\{s_A^P, a^P, ns_A^P, r^P\}$

for $e = 1$ to max_epoch **do**
  Reset environment
  Set step $\leftarrow 0$
  Get s_A^E and s_C^E
  Execute $a^E \leftarrow \pi^E(s_A^E)$
  Get s_A^P and s_C^P
  Execute $a^P \leftarrow \pi^P(s^P)$
  Get r^E, r^P, done
  if not done and step $<$ max_steps **then**
    step $\leftarrow$ step $+ 1$
    Get ns_A^E and ns_C^E
    store $\{s_A^E, a^E, ns_A^E, r^E\}$ to $\mathcal{D}$[Estimating]
    Update $s^E \leftarrow ns_A^E$
    Execute $a^E \leftarrow \pi^E(s^E)$
    Get ns_A^P and ns_C^P
    store $\{s_A^P, a^P, ns_A^P, r^P\}$ to $\mathcal{D}$[Planning]
    store $\{s_C^E, ns_C^E, s_C^P, ns_C^P\}$ to $\mathcal{D}$[Critic]
    Update $s_A^P \leftarrow ns_A^P$, $s_C^E \leftarrow ns_C^E$, $s_C^P \leftarrow ns_C^P$
    Execute $a^P \leftarrow \pi^P(s^P)$
    Get r^E, r^P, done
  end if
  Update policies with $\mathcal{D}$
  Clear $\mathcal{D}$
end for

In simulations, UUVs are located in a 200m $\times$ 200 $\times$ 200m cube region, USV is located in the sea surface, and UAV is located in the air. All equipment and the target are randomly located in their moving region with constraints. As for the speed of UUVs, we set $v_b = [2, 2, 2, \pi/6, \pi/6]$ and $a_b = [0.4, 0.4, 0.4, \pi/30, \pi/30]$, the unit of linear velocity and angular velocity is m/s and rad/s, the unit of linear acceleration and angular acceleration is m/s^2 and rad/s^2, and the time interval is set as 1s. And some parameters of the system and algorithms are shown in Table I.

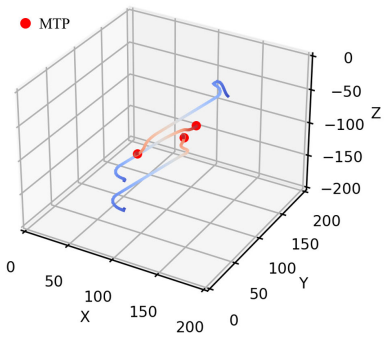
If certain UUVs detect the target and the distance from these UUVs to the target is less than 10m, the search mission is considered successfully completed. Furthermore, to guarantee the efficacy of the moving policy, the termination condition for the moving policy is defined as reaching a terminal distance of 1m.

B. Results

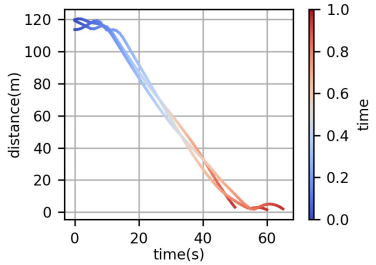
1) Moving Subtask: After training, the agent smoothly navigates to the desired MTP as per the moving policy. The resulting moving trajectory, and distance to the target are illustrated in Fig. 6, showcasing outcomes from three distinct tests. Remarkably, the agent successfully reaches MTP in 99.3% of the 1000 tests. Due to kinetic constraints, UUVs undergo directional adjustments to align with their MTP, as depicted in Fig. 6 (a). Throughout this reorientation process, there may be an initial increase in distance. Once the UUV attains a

TABLE I
PARAMETERS' SETTING

	Parameters	Values
System Parameter	UUV max velocity	$[2, 2, 2, \pi/6, \pi/6]$
	UUV max acceleration	$[0.4, 0.4, 0.4, \pi/30, \pi/30]$
	UUV freedom	5
	time interval	1s
	sonar parameter	$[60, \pi/3]$
	target region of x and y	$[50, 150]$
	target region of z	$[-150, -50]$
PPO Parameters	learning rate	[1e-5, 1e-5]
	clip parameter	0.2
	max grad norm	0.5
	batch size	64
	buffer size (max step)	640
	γ	0.95
	update epoch	10
	advantage normalization	True
Moving Parameters	terminal distance	1m
	state dim	13
	action dim	5
Multi-policy Parameter	terminal distance	10m
	Estimating state dim	31
	Estimating action dim	5
	Planning state dim	34
	Planning action dim	9
	critic input's dim	49
Rule-based Parameter	initial heights (z)	-20m / -80m / -160m
	initial x and y	$[0, 0]$ or $[100, 100]$
	height difference	20m
	trajectories' distance	20m

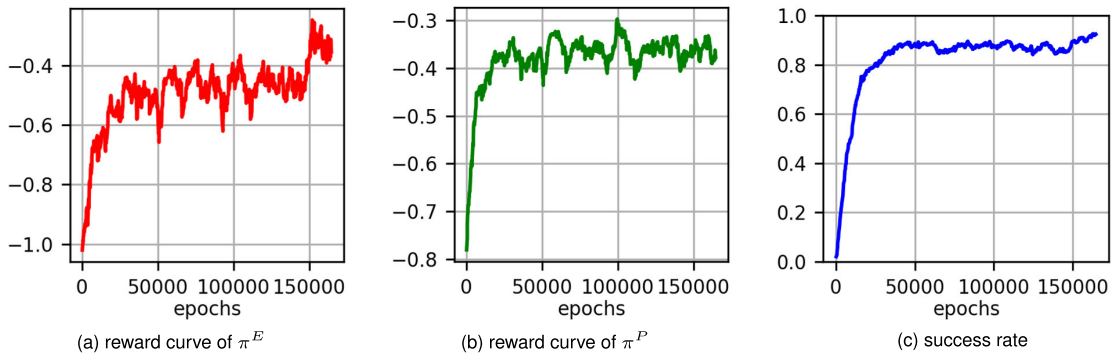


(a) Moving trajectory



(b) Distance to the target

Fig. 6. Moving trajectory and the distance to the target.



(a) reward curve of π^E (b) reward curve of π^P (c) success rate

Fig. 7. Training results of MPRL.

suitable orientation, the distance exhibits a linear decrease until it approaches the target. Fine-tuning of the UUV's position occurs to meet the completion requirement, a process visually represented in the last time slots of Fig. 6 (b).

2) Target Position Estimating and Trajectory Planning Subtask: During the training process, the corresponding reward and the success rates of the search missions are illustrated in Fig. 7.

As depicted in Fig. 7 (a) and Fig. 7 (b), both policies undergo simultaneous optimization throughout the training process, and their reward curves ascend as anticipated.

Furthermore, as training progresses, the success rate of the underwater searching mission steadily rises, as illustrated in Fig. 7 (c). This notable achievement signifies the effectiveness and applicability of the proposed algorithm, which consistently achieves successful outcomes in underwater search operations.

3) The Searching Mission: During the test of the search mission, the success rate is 97% in 1000 tests. Moreover, the average time consumed in the successful tests is 113.72s.

The UUVs' moving trajectory is presented in Fig. 8 (a), showcasing the effeciency of the proposed framework in solving the problem. This is further supported by the analysis of Fig. 8 (b) and Fig. 8 (c). In Fig. 8 (b), the distance between the estimated position and the target position is depicted, providing evidence of the effectiveness of the target estimating policy. Fig. 8 (c) illustrates the minimum distance between the UUVs and the target. The Fig. 8 demonstrates the success of the search process, wherein a UUV approaches the target and accurately determines its position with the assistance of other equipment and policies.

C. Comparison Experiment

In the same simulation environment, we conducted a comparison experiment to assess the efficacy of the method proposed in this paper. Due to the absence of various applied scenarios, we employ rule-based methods [1] for the comparative experiment. Taking into account the the applied scenarios, we select parallel and extended square search strategies as shown in Fig. 9.

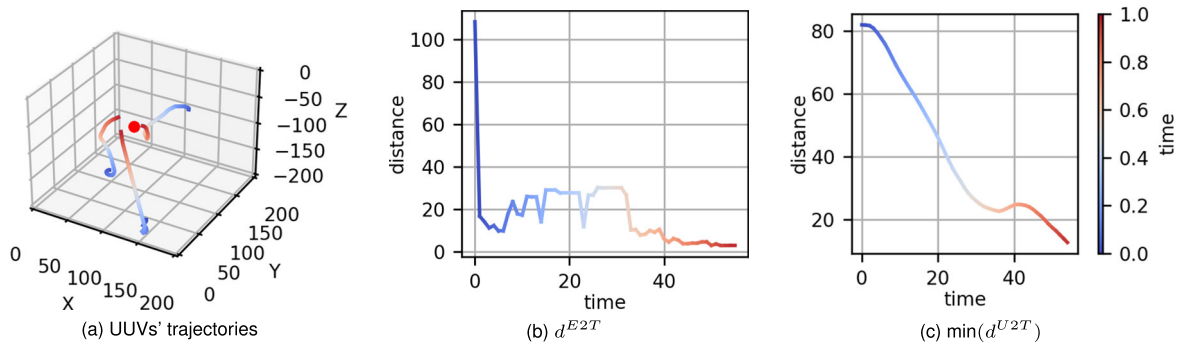
After completing the search on each surface, UUVs move to a deeper layer with a height difference of 20m, maintaining consistency with the distance between close trajectories as determined by the terminal condition defined in Section V-A.

TABLE II
COMPARISON EXPERIMENTAL RESULTS

Method	Success Rate	Success Rate Decrease (%)	Ave. Search Time (s)	Ave. Search Time Increment (%)
Our Method	97.0	-	113.72	-
Parallel	78.37	18.63	1505.26	1223.65
Dense Parallel	100.0	-2.92	2580.30	2168.99
Extended Square	77.45	19.55	1643.93	1345.60
Dense Extended Square	99.92	-3.09	3004.74	2542.23

The experiments are conducted without considering kinematic constraints, allowing the UUVs to move at their maximum velocity of 2m/s with a uniform information updating time interval of 1s. Additionally, the success condition is relaxed to only require that d^{U2T} should be less than 10m. With this setup, the search success rate and the average successful search time for parallel and extended square searching methods are 77.6%, 78.4%, 1505.26s, and 1643.93s, respectively. These times are 1223.65% and 1345.60% higher than those achieved using our method. Moreover, the success rates are 19.21% and 20.15% lower than our method.The low success rate is due to the circular searching path section cannot cover all area as shown in Fig. 9(a-b), therefore, we also take the corresponding dense searching path as shown in Fig. 9(c-d), where the trajectories' distance is reduced to $10\sqrt{2}$m. The dense search results are detailed in Table II, which explores the search space comprehensively. From this, we observe that a 100% success rate is achieved by covering all areas. However, the search time has increased significantly, highlighting the low efficiency of the dense search approach. For enhanced clarity in presenting the experimental outcomes, we encapsulate them in Table II, providing a concise summary of the success rates and average successful search times for both parallel and extended square searching methods, alongside our proposed approach.

Although these searching methods are conducted under somewhat ideal conditions, the performance of rule-based methods is still notably inferior to our proposed method.



(a) UUVs' trajectories (b) d^{E2T} (c) $\min(d^{U2T})$

Fig. 8. Searching results.

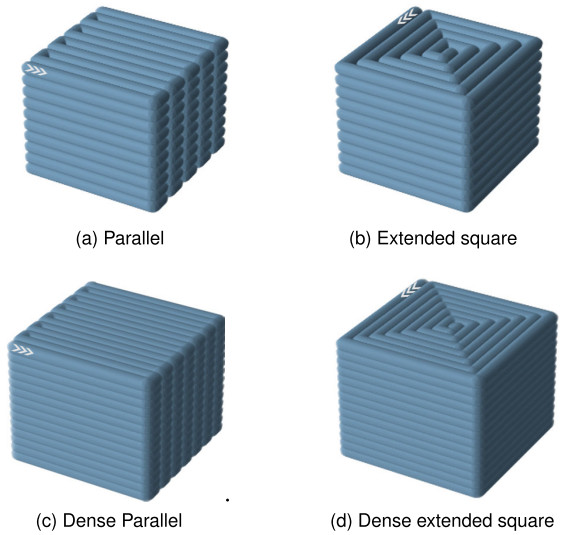


Fig. 9. Rule-based trajectories.

VI. Conclusion

In this paper, we propose a novel hierarchical framework adept at handling underwater target searching missions with cross-domain heterogeneous multi-vehicles. This framework incorporates three types of vehicles, namely UAVs, USVs, and UUVs, to collectively overcome the limitations of single or dual-agent approaches. We decompose the underwater search mission into three human-comprehensible subtasks designed to be adaptive to both environmental conditions and equipment capabilities: moving, target estimating, and trajectory planning. To address the training of interdependent policies, we propose the MPRL framework, which is based on MARL and CTDE, and implement it with the aid of a modified IPPO algorithm. The efficacy of our approach is demonstrated through simulations and comparisons with rule-based methods. In summary, our method enhances the flexibility and precision of underwater operations, offering a promising solution for complex search and rescue missions.

References

[1] J. Li, G. Zhang, C. Jiang, and W. Zhang, "A survey of maritime unmanned search system: Theory, applications and future directions," *Ocean Eng.*, vol. 285, Oct. 2023, Art. no. 115359.

[2] C. Lin, G. Han, M. Guizani, Y. Bi, J. Du, and L. Shu, "An SDN architecture for AUV-based underwater wireless networks to enable cooperative underwater search," *IEEE Wireless Commun.*, vol. 27, no. 3, pp. 132–139, Jun. 2020.

[3] Y. Wu, K. H. Low, and C. Lv, "Cooperative path planning for heterogeneous unmanned vehicles in a search-and-track mission aiming at an underwater target," *IEEE Trans. Veh. Technol.*, vol. 69, no. 6, pp. 6782–6787, Jun. 2020.

[4] C. Ke and H. Chen, "Cooperative path planning for air–sea heterogeneous unmanned vehicles using search-and-tracking mission," *Ocean Eng.*, vol. 262, Oct. 2022, Art. no. 112020.

[5] I. Masmitja et al., "Dynamic robotic tracking of underwater targets using reinforcement learning," *Sci. Robot.*, vol. 8, no. 80, Jul. 2023, Art. no. eade7811.

[6] D. Song, W. Gan, P. Yao, W. Zang, Z. Zhang, and X. Qu, "Guidance and control of autonomous surface underwater vehicles for target tracking in ocean environment by deep reinforcement learning," *Ocean Eng.*, vol. 250, Apr. 2022, Art. no. 110947.

[7] 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.

[8] C. Lin, G. Han, T. Zhang, S. B. H. Shah, and Y. Peng, "Smart underwater pollution detection based on graph-based multi-agent reinforcement learning towards AUV-based network ITS," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 7, pp. 7494–7505, Jul. 2023, doi: 10.1109/TITS.2022.3162850.

[9] Y. Hou, G. Han, F. Zhang, C. Lin, J. Peng, and L. Liu, "Distributional soft actor-critic-based multi-AUV cooperative pursuit for maritime security protection," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 6, pp. 6049–6060, Jun. 2024.

[10] J. Lindsay et al., "Collaboration of heterogeneous marine robots toward multidomain sensing and situational awareness on partially submerged targets," *IEEE J. Ocean. Eng.*, vol. 47, no. 4, pp. 880–894, Oct. 2022, doi: 10.1109/JOE.2022.3156631.

[11] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," 2017, *arXiv:1707.06347*.

[12] M. Ludvigsen et al., "Network of heterogeneous autonomous vehicles for marine research and management," in *Proc. OCEANS MTS/IEEE Monterey*. Monterey, CA, USA: IEEE, Sep. 2016, pp. 1–7.

[13] Y. Cao, W. Yu, W. Ren, and G. Chen, "An overview of recent progress in the study of distributed multi-agent coordination," *IEEE Trans. Ind. Informat.*, vol. 9, no. 1, pp. 427–438, Feb. 2013.

[14] C. Zhao, J. Liu, M. Sheng, W. Teng, Y. Zheng, and J. Li, "Multi-UAV trajectory planning for energy-efficient content coverage: A decentralized learning-based approach," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 10, pp. 3193–3207, Oct. 2021.

[15] H. Kang, X. Chang, J. Mišić, V. B. Mišić, J. Fan, and Y. Liu, "Cooperative UAV resource allocation and task offloading in hierarchical aerial computing systems: A MAPPO based approach," *IEEE Internet Things J.*, vol. 10, no. 12, pp. 10497–10509, Jun. 2023.

[16] S. Pateria, B. Subagdja, A.-H. Tan, and C. Quek, "Hierarchical reinforcement learning: A comprehensive survey," *ACM Comput. Surv.*, vol. 54, no. 5, pp. 1–35, Jun. 2022.

[17] Z. Gu et al., "Safe-state enhancement method for autonomous driving via direct hierarchical reinforcement learning," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 9, pp. 1–18, Sep. 2023.

[18] J. Xu, F. Huang, D. Wu, Y. Cui, Z. Yan, and X. Du, "A learning method for AUV collision avoidance through deep reinforcement learning," *Ocean Eng.*, vol. 260, Sep. 2022, Art. no. 112038.

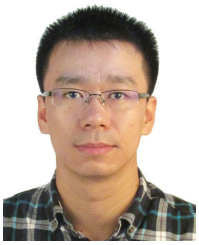
[19] T. I. Fossen, *Guidance and Control of Ocean Vehicles*. Hoboken, NJ, USA: Wiley, 1999.

[20] S. T. Havenstrøm, A. Rasheed, and O. San, "Deep reinforcement learning controller for 3D path following and collision avoidance by autonomous underwater vehicles," *Frontiers Robot. AI*, vol. 7, Jan. 2021, Art. no. 566037.
[21] *IAMSAR Manual: International Aeronautical and Maritime Search and Rescue Manual*, International Maritime Organization (IMO), London, U.K., Jul. 2024. [Online]. Available: https://www.imo.org/en/
[22] O. I. Abiodun, A. Jantan, A. E. Omolara, K. V. Dada, N. A. Mohamed, and H. Arshad, "State-of-the-art in artificial neural network applications: A survey," *Heliyon*, vol. 4, no. 11, Nov. 2018, Art. no. e00938.
[23] Z. Ling, Y. Zhang, and X. Chen, "A deep reinforcement learning based real-time solution policy for the traveling salesman problem," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 6, pp. 1–12, Jun. 2023.
[24] C. S. de Witt et al., "Is independent learning all you need in the StarCraft multi-agent challenge?" 2020, *arXiv:2011.09533*.

Shangqun Dong received the B.E. degree from Northwestern Polytechnical University (NWPU), Xi'an, China, in 2019, and the M.D. degree from the University of Chinese Academy of Sciences (UCAS), Beijing, China, in 2022. He is currently pursuing the Ph.D. degree with the College of Artificial Intelligence, Xi'an Jiaotong University (XJTU). His current research interests include reinforcement learning, multi-agent systems, human–machine hybrid intelligence.

Meiqin Liu (Senior Member, IEEE) received the B.E. and Ph.D. degrees in control theory and control engineering from Central South University, Changsha, China, in 1994 and 1999, respectively. She was a Post-Doctoral Research Fellow with Huazhong University of Science and Technology, Wuhan, China, from 1999 to 2001. She was a Professor with the College of Electrical Engineering, Zhejiang University, Hangzhou, China, from 2001 to 2021. She was a Visiting Scholar with The University of New Orleans, New Orleans, LA, USA, from 2008 to 2009. She is currently a Professor with the National Key Laboratory of Human-Machine Hybrid Augmented Intelligence, Xi'an Jiaotong University, Xi'an, China, and also with the College of Electrical Engineering, Zhejiang University. She has authored more than 200 papers in major journals and international conferences. She has led 16 national or provincial or ministerial projects in the last five years, including nine projects funded by the National Natural Science Foundation of China (NSFC). Her work was supported by Zhejiang Provincial Natural Science Fund for Distinguished Young Scholars in 2010 and by the National Science Fund for Excellent Young Scholars of China in 2012. Her current research interests include the theory and application of artificial intelligence, multi-sensor networks, information fusion, and nonlinear systems. She won the second prize of the Science and Technology Award of Zhejiang Province in 2013 and the first prize of the Natural Science Award of Chinese Association of Automation in 2019.

Shanling Dong (Member, IEEE) received the B.S. degree in automation from Xidian University, Xi'an, China, in 2014, and the Ph.D. degree in control science and engineering from Zhejiang University, Hangzhou, Zhejiang, China, in 2019. Supported by China Scholarship Council, she was a Visiting Graduate Student with the University of California at Riverside, USA, from September 2017 to September 2018. She was a Research Associate and a Post-Doctoral Research Fellow with the City University of Hong Kong, Hong Kong, SAR, China, from December 2019 to December 2020. She is currently with the College of Electrical Engineering, Zhejiang University. Her research interests include Markov jump systems, fuzzy systems, multi-agent systems, and ocean robots.



Ronghao Zheng (Member, IEEE) received the B.S. degree in electrical engineering and the M.S. degree in control theory and control engineering from Zhejiang University, Hangzhou, China, and the Ph.D. degree in mechanical and biomedical engineering from the City University of Hong Kong. He is currently with the College of Electrical Engineering, Zhejiang University. His research interests include distributed algorithms and control, especially the coordination of networked mobile robot teams with applications in automated systems and security.

Ping Wei (Senior Member, IEEE) received the B.E. and Ph.D. degrees from Xi'an Jiaotong University, Xi'an, China. He has been a Post-Doctoral Researcher with the Center for Vision, Cognition, Learning, and Autonomy (VCLA), University of California at Los Angeles (UCLA). He is currently a Professor with the Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University. His research interests include computer vision, machine learning, and computational cognition.