Multi-AUV Based Underwater Target Tracking Method via Reinforcement Learning in Dynamic Ocean Environment

Tianxiang Xing*,+, Jingzehua Xu†,+, Jun Du*, Xiangwang Hou*, Tianyu Xing†, Yong Ren*
*Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
†Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, 518055, China
E-mail: hxw21@mails.tsinghua.edu.cn

***Abstract*—Autonomous underwater vehicle (AUV) has gradually been developed to showcase its significant value in performing underwater tasks. Given the current situation that single AUV is constrained with limited abilities in detection, information processing and communication efficiency, multi-AUV system is employed to complete complexed tasks, among which target tracking is one application where multi-AUV system shows superior performance. However, most of the existing studies utilize simplified scenario like constant current velocity, which fails to simulate real dynamic ocean environment including current and obstacles. These factors can affect the overall path planning policies of AUVs or cause mission failure and economic loss. To improve the stability and validity of AUVs in the target tracking task, this paper introduces a multi-AUV based underwater target tracking method via reinforcement learning (RL) in dynamic ocean environment with current and obstacles. The concept of standoff circle is adopted to balance the distance between the AUVs. To be specific, we first model the ocean current and analyze the model of AUV. Then, the target tracking task is modeled as a Markov decision process, while RL is applied to achieve navigation and position control between AUVs and the target. The numerical simulation results reveal the best success rate of the AUVs at 83%, which proves the aforementioned method can perform better tracking accuracy and achieve superior robustness.**

***Index Terms*—AUV, Target tracking, Underwater environment, Reinforcement learning.**

I. Introduction

AUTONOMOUS underwater vehicles (AUVs) have been widely used on various underwater applications due to its excellent mobility, safety and outstanding performance in real-time communication. Compared with single AUV, multi-AUVs can carry more sensors, loads and operation equipment, perform more difficult missions without manual operation in the marine environment, such as underwater engineering, oceanic resources exploration and marine ecological survey [1] where target tracking is an important, frequently applied operation. Unlike target hunting, which simply requires the AUV to reach to the vicinity of the target [2], target tracking tasks demand that the AUV follow the target's path as precisely as possible while dodging any potential collision between AUVs and the target.

Many AUV target tracking methods have been proposed by researchers. Dong *et al.* [3] designed an adaptive target tracking method based on extended Kalman filter to simultaneously estimate the state of an AUV, and built a neural network compensator to correct error. Cao *et al.* [4] proposed an integrated algorithm by combining the Glasius bio-inspired neural network and bio-inspired cascaded tracking control approach to minimize tracking errors. However, most of these target tracking methods neglected the dynamic underwater environment with the obstacle and ocean current. Collision with sea bottom objects brings economic loss and directly lead to mission failure, while ocean currents can deviate the path of AUVs, affecting the overall accuracy of target tracking.

Most of the existing target tracking methods are model-based that requires large amounts of parameters. While for static targets these solutions are relatively straightforward, in a dynamic environment with a mobile target, the stability of such method drops dramatically. To enhance system stability, more and more researchers are using reinforcement learning (RL) approach to address target tracking problems. Fang *et al.* [5] used deep deterministic policy gradient algorithm to control the trajectory of AUV in the underwater horizontal plane and underwater 3-D space. Moon *et al.* [6] propose a novel RL approach for multiple unmanned aerial vehicles (UAVs) target tracking in challenging 3-D environments. The simulation results show that the proposed RL-based UAV controller provides a target tracking method with high accuracy and a very low time consumption. By employing RL, AUVs are able to interact with the environment and present self -decision making abilities. This advantage suggests the possibility of applying RL in dynamic ocean environment target tracking scenario.

Based on the analysis above, this study introduces a multi-AUV based underwater target tracking method via RL in dynamic ocean environment with current and obstacles to improve the stability and validity of AUVs in the target tracking task. We set the AUVs on a standoff circle around the target, which is beneficial for collecting multidimensional information of the target while improving the efficiency and stability of the target tracking task [7]. Also, navigating on standoff circle brings convenience for potential forthcoming entrapment of the target. To be specific, we first introduce the kinematic model of the AUVs, then create a dynamic ocean environment with obstacles and current modeled as multiple

+ These authors contribute equally to this work.

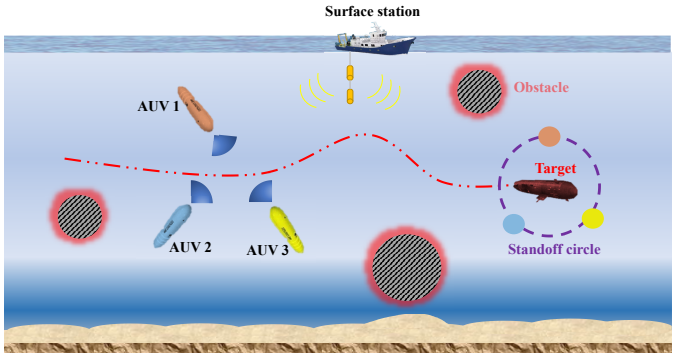

Fig. 1. Illustration of the multi-AUV underwater target tracking system model.

vortexes. Then we establish the multi-AUV underwater target tracking system model. Furthermore, the concept of states, actions and reward functions is introduced to establish the Markov decision process (MDP) modeling. A multi-AUV path planning method based on RL algorithm is developed to solve the problem, with the relative position of the AUVs to the target as the algorithm input, and the velocity value and direction of the AUVs as output. During the tracking process, the distance between AUVs and distance between AUV and target are carefully controlled to eliminate potential collision.

The rest of this paper is structured as follows. In section II, we introduce the system model of underwater target tracking task in detail, including the AUV kinematic model, and the modeling of dynamic ocean environment. In Section III, the problem formulation and methodology are described, which include MDP modeling to accomplish the design of state spaces, action spaces and reward functions of the target tracking task. Then we introduce the multi-agent soft actor-critic (MASAC) algorithm to solve this problem. In Section IV, simulation experiments are conducted to evaluate the performance of the proposed method and its advantage over traditional RL algorithms, followed by the conclusion in Section V.

II. System Model

In this study, we consider a multi-AUV underwater target tracking model, as illustrated in Fig. 1, where AUVs navigate in a dynamic ocean environment with obstacles, and operate the underwater target tracking task while realizing avoidance of dangerous areas. First, L AUVs are denoted as the set $\boldsymbol{A} = \{AUV_1, AUV_2, ..., AUV_L\}$. Then we note the position of AUV i at time slot t as $\boldsymbol{q}_i(t) = (x_i(t), y_i(t))$, where $x_i(t)$ and $y_i(t)$ denote the x, y coordinate of AUV i at time slot t. Similarly, target T is described by position $\boldsymbol{p}_T(t) = (x_T(t), y_T(t))$. Moreover, an obstacle is set with radius r_b and position vector $\boldsymbol{q}_b = (x_b, y_b)$ to label its center. The obstacle remains static throughout the entire process regardless of ocean current. During the execution, an assumption is made where the positions of AUVs and targets can be acquired by underwater sensors, and eventually broadcasted to AUVs. Also, the velocity and yaw angulsr velocity of an AUV can be obtained by its measurement devices. The AUV kinematic model and the dynamic ocean environment model are discussed in the following subsection.

A. AUV Kinematic Model

Without the interference of ocean current, the kinematic model of an AUV can be described as

$$\begin{cases} \dot{\boldsymbol{q}}(t) = \boldsymbol{v}(t), \\ \dot{\boldsymbol{v}}(t) = \boldsymbol{F}(t)/M, \end{cases} \tag{1}$$

where $t \in [0, T]$, $T > 0$, $\boldsymbol{q}(t) \in \mathbb{R}^2$ is the position vector of the AUV in the 2-D plane, $\boldsymbol{v}(t) \in \mathbb{R}^2$ is the 2-D velocity vector, while $\boldsymbol{F}(t) \in \mathbb{R}^2$ is the force vector, and M is the mass of the AUV. In this paper, the velocity value of an AUV is set to a constant value v, and the AUV navigates by changing its yaw angle $\phi(t)$. This is a common modeling method which has been applied to torpedo-shape AUVs, or vehicles that does not use thrusters [8].

B. Dynamic Ocean Environment Modeling

The dynamic characteristics of ocean environment are mainly derived from ocean currents, which are generated by rotation of the earth. Since the common velocity value of an AUV is below 10 m/s [9], the impact of the ocean current remains to be a significant part affecting AUVs' motions. To precisely simulate real world ocean current field, we use 2-D Navier-Stokes equation to model the ocean current into vortex as follows [10]

$$\frac{\partial \omega_c}{\partial t} + (\vec{\boldsymbol{u}}\nabla)\omega_c = v_f \Delta \omega_c, \tag{2}$$

where ω_c represents the vorticity of the vortex, while $\vec{\boldsymbol{u}} = (u_x, u_y)$ denotes the 2-D velocity field vector and v_f represents the current viscosity. Δ and ∇ are Laplacian operator and gradient operator, respectively. However, Eq. (2) requires viscosity of the current, which is very difficult to obtain. To simplify the equation, an approximate manner of Eq. (2) is given as follows

$$u_x(\boldsymbol{q}(t)) = -\frac{\Omega_v \cdot (y - y_0)}{2\pi \left\|\boldsymbol{q}(t) - \boldsymbol{q_v}\right\|_2^2} \cdot \left(1 - e^{-\frac{\|\boldsymbol{q}(t) - \boldsymbol{q_v}\|_2^2}{r_v^2}}\right), \tag{3}$$

$$u_y(\boldsymbol{q}(t)) = \frac{\Omega_v \cdot (x - x_0)}{2\pi \left\|\boldsymbol{q}(t) - \boldsymbol{q_v}\right\|_2^2} \cdot \left(1 - e^{-\frac{\|\boldsymbol{q}(t) - \boldsymbol{q_v}\|_2^2}{r_v^2}}\right), \tag{4}$$

$$\omega_c(\boldsymbol{q}(t)) = \frac{\Omega_v}{\pi r_v^2} \cdot e^{-\frac{\|\boldsymbol{q}(t) - \boldsymbol{q_v}\|_2^2}{r_v^2}}, \tag{5}$$

where $\boldsymbol{q_v}$ denotes the center of Lamb vortex in the 2-D ocean plane. The strength of the vortex is described as Ω_v, and the radius of the vortex is r_v. Besides, Fig. 2 shows the simulation result of the ocean current field with three vortexes.

While navigating in the ocean, underwater vehicles such as the target and AUVs are applied drag force generated by the vortexes. The relationship between the exerted drag force and velocity of the vortex current can be described in the following equation [11]

$$\boldsymbol{F}(\boldsymbol{q}, t) = \frac{1}{2}\rho A C_d \|\boldsymbol{u}(\boldsymbol{q}, t)\|_2^2, \tag{6}$$

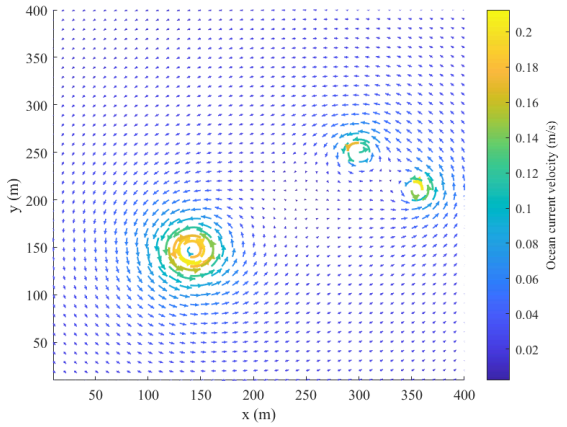

Fig. 2. Three ocean vortexes with radius 4, 10, 4 and strength $8\text{m}^2/\text{s}$, $20\text{m}^2/\text{s}$, $8\text{m}^2/\text{s}$, respectively. Brighter color suggests higher current velocity

where A stands for the cross-sectional area if the vehicle moves along the current direction, C_d denotes the drag coefficient, and ρ is the density of seawater. Then $\boldsymbol{F}(\boldsymbol{q})$ deviates the vehicle by adding a velocity increment $\Delta\boldsymbol{u}(\boldsymbol{q})$ on it with the same direction as $\boldsymbol{F}(\boldsymbol{q})$. Apparently, we have

$$\Delta\boldsymbol{u}(\boldsymbol{q},t) = \sigma_0 \boldsymbol{F}(\boldsymbol{q},t). \tag{7}$$

Furthermore, if the navigation velocity of the vehicle at position $\boldsymbol{q}$ is $(\boldsymbol{v}_0(\boldsymbol{q})$, then the actual velocity vector before it leaves position $\boldsymbol{q}$ can be described as

$$\boldsymbol{v}(\boldsymbol{q},t) = \boldsymbol{v}_0(\boldsymbol{q},t) + \Delta\boldsymbol{u}(\boldsymbol{q},t). \tag{8}$$

III. Problem Formulation and Methodology

A. Markov Decision Process

The process of the multi-AUV underwater target tracking task can be modeled as a MDP, which is an optimal decision-making process based on the Markov decision theory, and is suitable for dynamic stochastic systems and RL, which aims to improve the agent's policy to maximize returns. The MDP M can be expressed by the combination of the state space S, action space A, the state transition function $P^a_{SS'}$ and the reward functions R^a_S, $M = (S, A, P^a_{SS'}, R^a_S)$. The state space, action space and reward functions for the target tracking problem in this paper are designed as follows:

1) State space: The state space should be carefully defined in order to cover every possible state of the AUVs. Also, the state space must avoid redundancy, otherwise it costs large time expense in the later training phase, and even leads to convergence failure. As suggested in Section II-A, one possible way to describe the state of the AUV i at time slot t can be defined as follows

$$S_i(t) = \left[\boldsymbol{q_i}(t), \boldsymbol{v_i}(t), \phi(t)\right]. \tag{9}$$

The overall state space at time slot t is then given as $S(t) = \{S_i(t)\}$. Moreover, to make the simulation process operable, the state values in set $S(t)$ must be discretized.

2) Action space: Given the aforementioned assumption that all AUVs navigate with the same velocity, the action space of AUV i can be easily defined as follows:

$$A_i \in \{\omega_{\min}, \ldots, \omega_{\max}\}, \tag{10}$$

where $\omega_{\min}$ and $\omega_{\max}$ are the minimum and maximum possible yaw angular velocity values of an AUV. And the overall action space of the AUVs is $A = \{A_i\}$.

3) Reward function: The main purpose of introducing the reward function is to realize policy improvement of AUVs via RL training. With carefully pre-designed reward functions, the AUVs can achieve the goal of approaching the target closely and then following its path with a pre-set distance r_0, which is also the radius of the standoff circle. Besides, punishment function is also designed so that an AUV can avoid collision with the obstacle and the other AUVs. Finally, the AUVs are controlled to averagely encircle around the target on the standoff circle. The specific reward functions are designed as follows:

Target Approaching $R^{(1)}(t)$*:* To rapidly approach the target, an AUV should receive a constantly existing reward that is negatively correlated with the distance between AUV i and the target noted as $d_i(t)$. So we have $R^{(1)}(t)$ defined as

$$R^{(1)}(t) = -\sum_i d_i(t). \tag{11}$$

Target Following $R^{(2)}(t)$*:* Target following requires an AUV to maintain an exact distance with the target, thus, we define this reward of AUV i as

$$R_i^{(2)}(t) = \begin{cases} -ad_i(t)^2 + 2ar_0 d_i(t) + c, & d_i(t) < r_0, \\ 0, & d_i(t) \geq r_0, \end{cases} \tag{12}$$

where a and c stand for constant values.

Collision Avoidance $R^{(3)}(t)$*:* $R^{(3)}(t)$ is a contingent negative reward which is only effective when two objects (AUVs or obstacles) collide. Thus, for AUV i, we have

$$R_i^{(3)}(t) = \begin{cases} R_a, & \text{if colliding with other AUVs,} \\ R_b, & \text{if colliding with the obstacle,} \\ 0, & \text{else,} \end{cases} \tag{13}$$

Standoff Circle Formation $R^{(4)}(t)$*:* $R^{(4)}(t)$ should be designed based on the distance between AUV i and AUV j, noted as $d_{ij}(t)$. And the expected position of AUVs can be different according to the number of AUVs. Specifically, the expected position of three AUVs should be an equilateral triangle on the standoff circle with side length $d_0 = \sqrt{3}r_0$. As a result, we have

$$R_{i,j}(t)^{(4)} = \begin{cases} e^{(\mu(d_0 - d_{ij}(t))-1)}, & d_{ij}(t) \leq d_0, \\ e^{(\mu(d_{ij}(t) - d_0)-1)}, & d_{ij}(t) \geq d_0, \end{cases} \tag{14}$$

where μ is a coefficient based on experiments.

Combining equations Eq. (11) $\sim$ Eq. (14), the reward of the multi-AUV underwater target tracking task can be obtained by

$$R(t) = R^{(1)}(t) + \sum_i R_i^{(2)}(t) + \sum_i R_i^{(3)}(t) + \sum_{i,j} R_{i,j}^{(4)}(t). \tag{15}$$

Algorithm 1 MASAC

Input: initial policy parameter ϑ, Q-function parameters φ_1, φ_2, empty replay buffer B.

Set target parameters to main parameters:

$$\varphi_{\mathrm{t},1} \leftarrow \varphi_1, \varphi_{\mathrm{t},2} \leftarrow \varphi_2$$

while not convergence:

Observe state s and choose an action $a \sim \pi_\theta(\cdot|s)$.

Take a step in the environment, then observe the next state s', acquire reward r, and update signal d to decide whether s' is the termination state.

Add (s, a, r, s', d) to B.

If s' is the termination state, reset the environment.

for m **in range** (number of updates) **do**

Sample a batch $B = \{(s, a, r, s', d)\}$ from B.

Compute the target network for the Q-functions:

$$y(r,s',d) = r + \beta(1-d)\left(\min_{j=1,2} Q_{\varphi_{\mathrm{t},j}}(s',\widetilde{a}') - \alpha \log \pi_\vartheta\left(\widetilde{a}'|s'\right)\right),$$

where $\widetilde{a}' \sim \pi_\vartheta\left(\cdot \mid s'\right)$

Update Q-functions:

$$\nabla_{\varphi_j} \frac{1}{|B|} \sum_{(s,a,r,s',d)\in B} \left(Q_{\varphi_j}(s,a) - y\left(r,s',d\right)\right)^2 \quad \text{for } j = 1,2$$

Update the policy:

$$\nabla_\vartheta \frac{1}{|B|} \sum_{s\in B} \left(\min_{j=1,2} Q_{\varphi_j}\left(s, \tilde{a}_\vartheta(s)\right) - \alpha \log \pi_\vartheta\left(\tilde{a}_\vartheta(s) \mid s\right)\right),$$

where $\widetilde{a}_\vartheta(s)$ is a sample from $\pi_\vartheta\left(\cdot \mid s'\right)$.

Update the target network:

$$\varphi_{\mathrm{t},j} \leftarrow \sigma\varphi_{\mathrm{t},j} + (1-\sigma)\varphi_j \quad \text{for } j = 1,2$$

end for

B. Multi-Agent Soft Actor-Critic Algorithm

To achieve multi-AUV underwater target tracking with high accuracy, an efficient RL algorithm is employed. This paper mainly focuses on MASAC algorithm, which is suitable for continuous action space and achieves better learning efficiency than traditional policy gradient algorithms.

MASAC is an off-policy actor-critic RL algorithm. Compared with other actor-critic algorithms, MASAC does not require meticulous parameter tuning and can process large samples. Meanwhile, MASAC possesses promising convergence properties [12]. By training a stochastic policy $\pi(a|s)$, it aims to maximize the reward and entropy at the same time. In another word, MASAC has the ability to successfully achieve the expected goal while exploring as randomly as possible. The details of MASAC are shown in Algorithm 1.

IV. Simulation settings and results

For the convenience and validity of simulation, we employ three AUVs in this experiment with the safety radius (collision occurs within this range) r. The experiment uses OpenAI Gym environment which is automatically wrapped in a compatible layer.

TABLE I
Parameter Settings.

Name	Symbol	Value
Environment:		
Radius of the obstacle	r_b	1.85m
Radius of AUV	r	0.05m
Vortex 1-3 position	$\boldsymbol{q}_{v1}$	$(15.7, 1)$
	$\boldsymbol{q}_{v2}$	$(-5.9, -5.3)$
	$\boldsymbol{q}_{v3}$	$(10, 5.2)$
Radius of vortex 1-3	r_{v1}	0.4m
	r_{v2}	1m
	r_{v3}	0.4m
Strength of vortex 1-3	Ω_{v1}	0.02m^2/s
	Ω_{v2}	0.05 m^2/s
	Ω_{v3}	0.02 m^2/s
Action set:		
Yaw angular velocity (min)	$\omega_{\min}$	$-\pi/3$ s^{-1}
Yaw angular velocity (max)	$\omega_{\max}$	$\pi/3$ s^{-1}
Reward function:		
Coefficient a	a	-6
Coefficient c	c	-25
Collision punishment a	R_a	-5
Collision punishment b	R_b	-9
Designed coefficient $\boldsymbol{\mu}$	μ	0.1
Learning period:		
Simulation step	ΔT	0.1s
Discount factor	β	0.99
Target network update speed	τ	0.005
Learning rate	L	$3e^{-4}$
Buffer size	S_1	$1e^6$
Batch size	S_2	256
Maximum episode length	T	200s
Number of episodes	N	500

In the simulation, the target's initial position is placed on the origin of a 40m $\times$ 40m square area. We use x-y coordinate system to describe the position of an object, in which one unit length represents 1m. The target is on (0,0), and the obstacle's center is placed on (7,8). Three AUVs are placed at (16,16), (16,15), (16,14), respectively. The target has an original speed at 0.01m/s and initial yaw angle 0.9π. Three AUVs have the same velocity value at $v = 0.1$m/s, while their initial yaw angles are randomly selected from $[\pi, 3\pi/2]$. Other necessary environment settings and training parameters are listed in Table I.

The entire process of multi-AUV target tracking under dynamic ocean environment is presented in Fig. 3, which is the result of using MASAC to train the AUVs after 500 episodes. The green thick curve is the trajectory of target, while the thinner curves are the trajectories of AUVs. As shown in Fig. 3, the target trajectory is not a straight line due to the obstruction of vortexes. The result demonstrates that all AUVs can successfully track the target by following target path and maintaining a safe distance at 2m. The AUVs also learn to bypass the obstacle while navigating. Besides, they are equally distributed around the standoff circle and form an equilateral triangle. This behavior pattern of the AUVs proves the feasibility of using MASAC to accomplish multi-AUV cooperative tasks with excellent performance.

To further validate the above conclusions, we compare

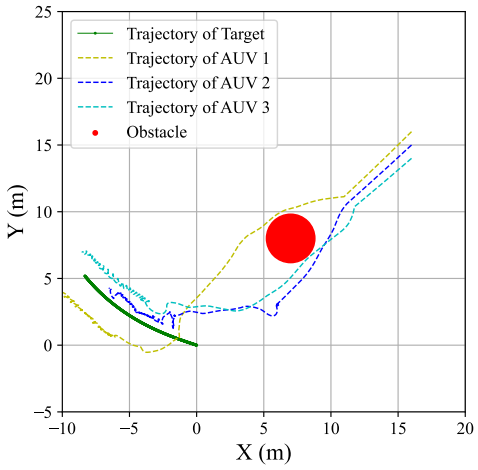

Fig. 3. The scenario of multi-AUVs tracking target while avoiding obstacle with the assistance of the MASAC algorithm.

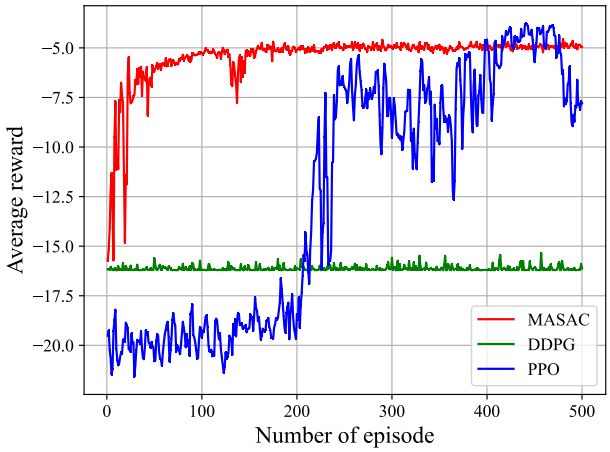

Fig. 4. Training curves of MASAC, DDPG and PPO after 500 episodes.

the convergence ability and tracking success rate between MASAC and other traditional RL algorithms including DDPG and PPO in Fig. 4 and Fig. 5. In Fig. 4, MASAC-employed AUVs learn to avoid the obstacle after approximately 50 episodes. and converges after about 150 episodes. Nevertheless, DDPG and PPO-employed AUVs reveal obvious disability in convergence and have much less average reward. Another important feature is the tracking success rate, defined as the ratio of the number of successful steps to the number of steps after an AUV makes initial contact with to the standoff circle. A step of AUV i is regarded successful only if both the following two criteria are satisfied:

(a) No collision occurs throughout the navigating process of AUV i.

(b) AUV i remains approximately on the standoff circle with $d_i \in (1.8, 2.2)$.

Under criteria *(a)*, *(b)*, the success rates of all AUVs are displayed separately in Fig. 5. The result shows that all the MASAC-employed AUVs have tracking success rate over 75%, while the general success rate of the system (the

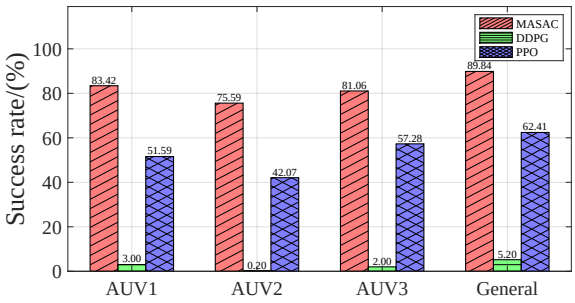

Fig. 5. Tracking success rate of single AUV and multi-AUVs under different algorithms including MASAC, DDPG and PPO.

possibility of at least 1 AUV successfully tracking the target) at 89.84%, which proves multi-AUV system to be more reliable than single AUV system. Besides, Fig. 5 illustrates the outstanding performance of MASAC, as all AUVs trained with MASAC receive remarkable success rate improvement than with other two algorithms.

V. Conclusions

This paper introduced a flexible and robust multi-AUV cooperation method to achieve target tracking task in complexed ocean environment including obstacle and vortexes. Meanwhile, we borrow the concept of standoff circle, encircling the AUVs around the target to ensure efficiency and stability of detection. Simulation results validate that RL-based MASAC approach can converge rapidly and achieve promisingly competitive performance. The experiment also implicates multi-AUV system to be much more advanced than single AUV system, and has the potential to operate more complicated cooperative tasks.

References

[1] C. Wang, J. Du, J. Wang, and Y. Ren, "AUV path following control using deep reinforcement learning under the influence of ocean currents," in *Proceedings of the 2021 5th International Conference on Digital Signal Processing*, 2021, pp. 225–231.

[2] Z. Yang, J. Du, Z. Xia, C. Jiang, A. Benslimane, and Y. Ren, "Secure and cooperative target tracking via AUV swarm: A reinforcement learning approach," in *2021 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2021, pp. 1–6.

[3] L. Dong, H. Xu, X. Feng, X. Han, and C. Yu, "An adaptive target tracking algorithm based on EKF for AUV with unknown non-gaussian process noise," *Applied Sciences*, vol. 10, no. 10, p. 3413, 2020.

[4] X. Cao, H. Sun, and G. E. Jan, "Multi-AUV cooperative target search and tracking in unknown underwater environment," *Ocean Engineering*, vol. 150, pp. 1–11, 2018.

[5] Y. Fang, Z. Huang, J. Pu, and J. Zhang, "AUV position tracking and trajectory control based on fast-deployed deep reinforcement learning method," *Ocean Engineering*, vol. 245, p. 110452, 2022.

[6] J. Moon, S. Papaioannou, C. Laoudias, P. Kolios, and S. Kim, "Deep reinforcement learning multi-UAV trajectory control for target tracking," *IEEE Internet of Things Journal*, vol. 8, no. 20, pp. 15 441–15 455, 2021.

[7] S. Lim, Y. Kim, D. Lee, and H. Bang, "Standoff target tracking using a vector field for multiple unmanned aircrafts," *Journal of Intelligent & Robotic Systems*, vol. 69, pp. 347–360, 2013.

[8] I. Masmitja, M. Martin, K. Katija, S. Gomariz, and J. Navarro, "A reinforcement learning path planning approach for range-only underwater target localization with autonomous vehicles," in *2022 IEEE 18th International Conference on Automation Science and Engineering (CASE)*. IEEE, 2022, pp. 675–682.

[9] Z. Zeng, K. Sammut, A. Lammas, F. He, and Y. Tang, "Efficient path re-planning for AUVs operating in spatiotemporal currents," *Journal of Intelligent & Robotic Systems*, vol. 79, pp. 135–153, 2015.
[10] Z. Fang, J. Wang, J. Du, X. Hou, Y. Ren, and Z. Han, "Stochastic optimization-aided energy-efficient information collection in internet of underwater things networks," *IEEE Internet of Things Journal*, vol. 9, no. 3, pp. 1775–1789, 2021.
[11] K. M. Tan, T.-F. Lu, and A. Anvar, "Drag coefficient estimation model to simulate dynamic control of autonomous underwater vehicle (AUV) motion," in *20th international congress on modelling and simulation*, Adelaide, SA, Australia, Dec. 2013, pp. 963–969.
[12] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *International conference on machine learning*. PMLR, 2018, pp. 1861–1870.