International Journal of Innovative Computing, Information and Control
Volume **21**, Number **1**, February **2025**

ICIC International ©2025 ISSN 1349-4198
pp. **69–84**

OBSTACLE AVOIDANCE AND COMBAT PATH PLANNING ALGORITHM OF UNMANNED SURFACE VEHICLE VIA Q-TRANSFORMER LEARNING

Hongyu Guo*, Hao Gu and Lintao Dou

Technological Innovation Center of Littoral Test
Jiangsu Automation Research Institute
No. 18, Shenghu Avenue, Lianyungang 222000, P. R. China
{ guhao3638; doult99 }@163.com
*Corresponding author: ghy1996@mail.ustc.edu.cn

Received April 2024; revised August 2024

Abstract. *Unmanned surface vehicle (USV) is regarded as a vital force for future naval warfare, as it can replace manned ships in various combat missions. The operational sea area of USVs inevitably encounters various obstacles, which require USV to avoid and plan paths to reach the target area, search for and strike the targets. To address these application requirements, this paper proposes a USV obstacle avoidance and combat path planning (UOACPP) algorithm based on the Transformer architecture and Q-learning. Different reward functions are designed in the proposed UOACPP method according to the stages of USV combat tasks, making it adaptable to various scenarios, such as search and attack. Compared with conventional obstacle avoidance and path planning algorithms, the proposed UOACPP method can simultaneously avoid obstacles and plan paths for search and attack tasks, which achieves the integration of obstacle avoidance and path planning in a unified framework. Moreover, the adopted Transformer architecture and Q-learning method can effectively improve the training efficiency. The experimental results demonstrate the effectiveness and superiority of the proposed UOACPP scheme over existing ones.*

Keywords: Obstacle avoidance, Path planning, USV, Deep reinforcement learning, UOACPP, Q-learning

1. **Introduction.** With the rapid development of technology, unmanned maritime equipment has demonstrated high combat efficiency, strong flexibility, low personnel casualties, and a wide range of combat capabilities, making it an important means of electronic confrontation, precision strikes, and special operations in future intelligent warfare [1-3]. Unmanned surface vehicle (USV) has the advantages of high mobility, low cost, good concealment, and strong weapon-carrying capabilities [4,5]. During combat, they can perform various maritime combat tasks such as area defense, maritime reconnaissance and surveillance, mine clearance, anti-submarine warfare. Hence, it is essential to develop intelligent control algorithms for USV to achieve the striking invading enemy targets, and safeguarding important area [6-8].

When executing various maritime combat tasks, small and medium-sized USVs operate in operational sea areas that are mostly near the coast or islands, which are often obstructed by reefs and other obstacles [9-11]. In addition, the battlefield environment is often complex, where the USVs need to dynamically plan a safe strike path based on environmental status information of the operational area and the location information

DOI: 10.24507/ijicic.21.01.69

of the enemy. Therefore, USVs are required to achieve autonomous navigation and obstacle avoidance within a fixed sea area, ensuring their own navigation safety during the combat process [12-14]. That is, USV can plan a safe and efficient combat path based on the requirements of the combat task and the environmental status information of the maritime battlefield that they have mastered from both a global and local perspective. Some interesting results of obstacle avoidance have been reported in the existing literature [15-17]. To mention a few, the authors in [18] proposed an uninterrupted collision-free path planning system, which can improve the operational performance of multiple UAVs in ocean sampling missions with high robustness. A cooperative trajectory planning algorithm for USV-UAV coupled systems was proposed in [19] to ensure that the USVs execute safe and smooth paths when traversing multi-obstacle maps autonomously. The authors in [20] proposed a path planning method for autonomous surface vehicles to solve the underwater target tracking optimization problem. However, due to the dynamic environment in maritime combat tasks, the USV is also required to avoid being attacked by blue-side opponents to ensure their own safety and completing the established combat mission. The challenge of simultaneous obstacle avoidance and complex combat tasks has made path planning strategies difficult to design, and this research issue has not yet been fully addressed.

In recent years, deep reinforcement learning (DRL) has made breakthroughs in many complex decision-making problems in fields such as video games and navigation path planning [21-23]. Especially in the field of decision-making, it enables intelligent agents to find optimal strategies well in various complex situations, including successful obstacle avoidance, simultaneous facing of multiple targets and tasks. Because the environment of the complex combat tasks is often dynamic and unstable, it is significant to develop the DRL method for USV system to achieve intelligent control decision [24-26]. Many different DRL results of USV systems have been reported in the available literature during the past decades [27-29]. For example, the authors in [30] proposed a capture algorithm based on reinforcement learning and a distributed partially observable multi-target hunting proximal strategy optimization algorithm to solve the problem of multi-target capture for USV system; a neural network structure and a semi-Markov decision process model was established in [31] for the USV collision avoidance problem, where the DRL-based collision avoidance method was proposed; the authors in [32] proposed a cooperative navigation task approach using a hybrid multi-intelligence deep reinforcement learning framework that utilizes a heuristic mechanism to guide USVs in group task learning. However, when USVs execute obstacle avoidance and complex combat tasks, there exist high dynamic environment perception problem and long decision-making processes, which may degrade the performance of traditional DRL methods. Hence, it is crucial to improve the structure of the DRL methods for USV system to tackle the aforementioned issues, which motivates the current work.

Based on the aforementioned issues, this paper proposes a USV obstacle avoidance and combat path planning (UOACPP) algorithm based on the Transformer architecture and combined with Q-learning. The main contributions are listed as follows. (i) The proposed UOACCP method can plan an efficient search and strike navigation route from a global perspective based on the geographical environment of the maritime battlefield, avoiding collisions with obstacles in the operational area. In addition, when USVs detect the status information of blue-side opponents, the developed approach can plan a series of ad hoc combat plans such as attack and evasion of blue-side attacks in a local area. (ii) The proposed method combines deep Q-learning with the Transformer network structure to achieve state-aware enhancement, avoiding excessive forgetting of historical observation information. By designing the encoder and decoder structures with attention mechanisms,

it solves the problem of the LSTM network's sharp decline in accuracy under long-term input, effectively improving training efficiency. (iii) In order to satisfy the requirements of actual combat, different reward functions are designed according to the stages of unmanned boat combat tasks, making it flexible to be applied to various scenarios, such as search and attack.

The remaining parts of the paper are organized as follows. The problem statement is presented in Section 2. Then, the preliminaries are given in Section 3. Section 4 proposes a Q-Transformer-based path planning method, and experimental studies are performed in Section 5. Section 6 concludes this work.

2. **Problem Statement.** In practice, USV executing the task of attacking blue-side opponents in a fixed sea area is a common and typical style of unmanned maritime combat. It is necessary to plan a safe strike path for USV based on the environmental status information of the operational area and the location information of the blue side. A schematic diagram of the path planning scenario for a single unmanned boat searching and attacking blue-side opponents in a fixed sea area is shown in Figure 1. In the absence of interference, the detection and attack range of USV model can be shown in Figure 2. The light blue area represents the detection area of the USV, and the light green area represents the attack range of the USV. Both detection and attack have certain blind spots. The maximum detection range is 25 km, and the maximum missile range is 20 km.

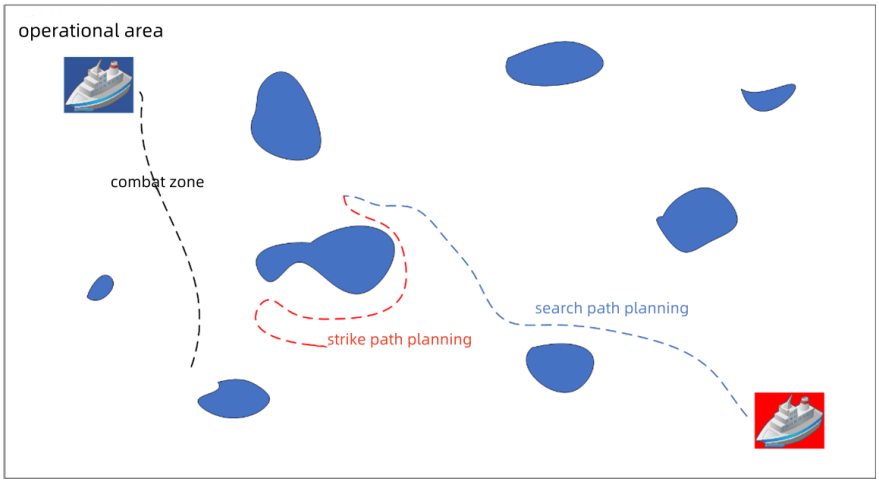


FIGURE 1. USV combat path planning scenario

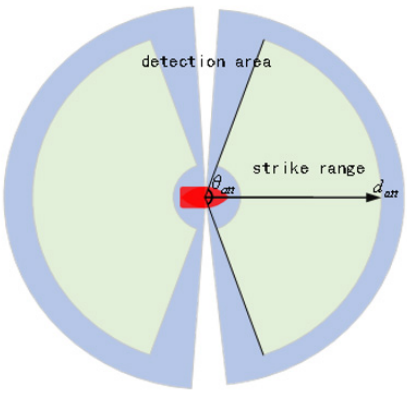


FIGURE 2. USV detection and strike range

Taking the forward attack range of a USV as an example, θ_{att} is the angle of the attack range, and d_{att} denotes the attack distance of the USV. The condition for USV to launch missiles is presented as follows:

$$\begin{cases} \psi_o < \theta_{att} \\ d_o < d_{att} \\ \text{s.t.} \\ \psi_o = \arctan[(y_e - y_o)/(x_e - x_o)] \\ d_o = \sqrt{(y_e - y_o)^2 + (x_e - x_o)^2} \end{cases} \tag{1}$$

where ψ_o is the heading angle of the unmanned boat, and d_o denotes the relative distance between the unmanned boat and its tracked target. x_e is the position coordinate of tracked target in the x-axis direction; y_e is the position coordinate of tracked target in the y-axis direction; x_o is the position coordinate of our unmanned boat in the x-axis direction; y_o is the position coordinate of our unmanned boat in the y-axis direction.

If there are obstacles within the detection range of the USV, the detection ability of the USV for the area behind the obstacles will be interfered with, and the detection accuracy will decrease. Therefore, during the search phase, the USV needs to avoid obstacles while reducing the detection blind zone and efficiently search the operational area. When a blue-side target is detected, a reasonable strike route needs to be planned based on the terrain and strike range. Under the premise of ensuring its own safety, the USV needs to discover and destroy blue-side targets in a timely manner. Due to the long decision-making process, multiple task objectives, and high dynamic environment perception, intelligent algorithms often forget past experiences and are difficult to apply to complex strike planning tasks.

3. Preliminaries.

3.1. Q-learning.

As a reinforcement learning method, Q-learning is used to solve decision-making problems based on environmental and reward signals. It is a model-free learning method that generates the optimal strategy for taking action in different states by iteratively updating the value function.

The core idea of Q-learning is based on the Bellman equation, which uses the relationship between the Q-value of the current state and the maximum Q-value of the next state to update the values in the Q-table. Through continuous interaction with the environment, observing reward signals, and updating the Q-table, Q-learning can gradually optimize the strategy, enabling the agent to obtain the maximum cumulative reward through continuous exploration and exploitation. Specifically, given state s and action a, the Q-value update formula is

$$Q(s,a) = (1-\alpha) * Q(s,a) + \alpha * (r + \gamma * \max(Q(s',a'))) \tag{2}$$

where $Q(s,a)$ is the Q-value of state s and action a; α denotes the learning rate (a parameter between 0 and 1); r represents the immediate reward observed from the environment; γ is the discount factor (a parameter between 0 and 1); s' is the next state, and a' is the optimal action selected in the next state s'.

3.2. Transformer networks.

The core idea of Transformer is to use self-attention mechanism to capture the dependencies in the input sequence. The self-attention mechanism allows the model to focus on different positions in the input sequence when generating output, thus better modeling the long-distance dependencies within the sequence. This enables the Transformer model to perform highly parallel computations, making full use of hardware resources such as GPUs, and accelerating the training and inference speed of USV obstacle avoidance and combat path planning tasks. In addition, the Transformer

structure also solves the problem that RNN models cannot perform parallel computations directly due to their sequential processing nature. The Transformer model consists of an encoder and a decoder. The encoder is responsible for converting the input sequence into a series of high-level feature representations, while the decoder generates the target sequence step by step based on the output of the encoder.

4. **Q-Transformer-Based Path Planning Algorithm.** This section presents the Q-Transformer UOACPP algorithm, which addresses the issues of long decision-making processes and high dynamic environment perception in complex tasks.

4.1. **The Markov decision processes.** Obstacle avoidance and combat path planning for unmanned boats can usually be modeled as Markov decision process (MDP), which describes the process of learning decision-making strategies through interaction with the environment. MDP can be represented by the following equation:

$$M = [S, A, P, \gamma, R] \tag{3}$$

where S denotes the state space sensed by sensors when an unmanned vessel performs a mission, while A represents the action space of the unmanned vessel. R is the reward function, which the USV will use to continuously adjust its behavioral strategy and ultimately achieve obstacle avoidance and attack. γ is the discount factor that determines future rewards and current rewards. P is the transition probability function defined as in Equation (4).

$$P(s_{t+1}|s_t, a_t) = P(s_{t+1}|s_t, a_t, s_{t-1}, a_{t-1}, \ldots) \tag{4}$$

The MDP of the unmanned vessel and the environment is shown in Figure 3. The USV selects an action under the observed environmental state s_t; the environment updates the state s_{t+1} based on the behavior and state transition probability P, and returns the instantaneous reward γ_{t+1} to the USV.

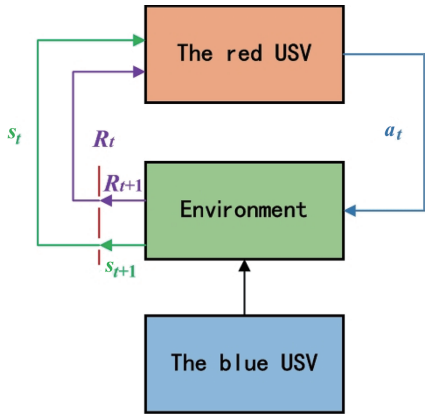


FIGURE 3. MDP of the USVs and environment

4.2. **Enhanced state awareness.** In Markov decision processes, the agent's decisions are based on fully perceived states. However, in complex marine environments, the environmental information perceived by the USV may change with the vessel's actions. As a result, the decisions based on the current perception may forget the information it perceived in the previous moment, thus making it unable to make accurate decision control. To solve this problem, currently use the long short-term memory (LSTM) to increase the storage capacity of previous perception information. LSTM can achieve higher accuracy under short-term input conditions, but its accuracy drops sharply under long-term input conditions, and its operating efficiency is also lower. To address the issues of LSTM,

Transformer network can parallelly process long-term data and has achieved excellent performance in various fields.

Based on the Transformer network, the USV in the red side can enhance its perception ability of the current environmental state. The input of the network is the current and previous perception features, including obstacle positions and hidden spatial features of blue side, and the output is the action that the USV in the red side needs to take at the current moment.

4.3. **UOACPP algorithm.** Based on the above discussions, the UOACPP algorithm is presented in this subsection, which can model the obstacle avoidance and combat path planning of USV during mission execution using the MDP framework. Then, the Q-Transformer method is developed to learn the optimal strategy π, which realizes global path planning and local combat decision planning, as shown in Figure 4.

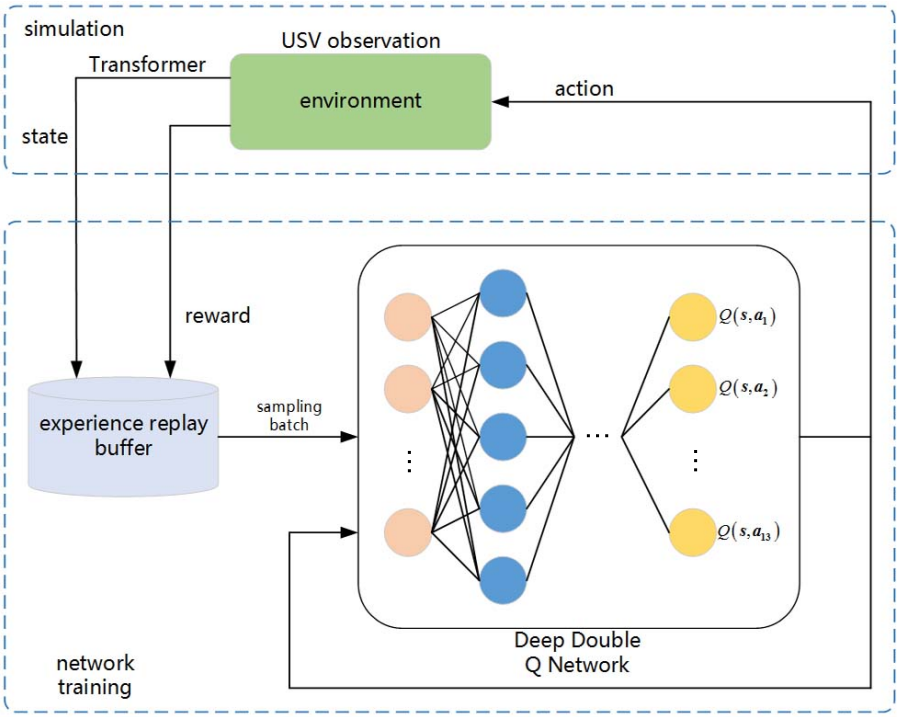


FIGURE 4. UOACPP algorithmic structure diagram

According to the MDP interaction framework, the USV will perform rudder operations (action a) based on the perceived information (state s), and receive rewards r at each step. At the end of each round, the total reward obtained will be calculated. The calculation of the reward is shown in Equation (5).

$$R(h) = r(s_t, a_t) + \gamma * r(s_{t+1}, a_{t+1}) + \cdots = \sum_{t=1}^{T} \gamma^{k-1} r(s_t, a_t) \tag{5}$$

where the total discounted reward $R(h)$ is the sum of the discounted rewards after time t, and γ is the discount factor.

The three signs that indicate the end of the unmanned vessel obstacle avoidance and combat path planning event are as follows.

1) If the USV collides during navigation, then the round of combat ends.
2) During the search phase, if the unmanned vessel fails to detect the USV in the blue side within the specified time, then the round of combat ends.

3) During the combat phase, if the USV in the red side is hit by the USV in the blue side, then the round of combat ends.

The main goal of the UOACPP algorithm is to learn the optimal strategy π from the event. It can be represented as

$$\pi^* = \arg\max_{\pi} E_{p^{\pi}(h)}[R(h)] \tag{6}$$

where $p^{\pi}(h)$ denotes the probability density of the event. The learning of intelligent agent control strategy is achieved by acting state on the value function $Q^{\pi}(s,a)$, then $Q^{\pi}(s,a)$ can also be represented as

$$Q^{\pi}(s,a) = \arg\max_{\pi} E_{p^{\pi}(h)}[R(h)|s_1 = s, a_1 = a] \tag{7}$$

When the red unmanned vessel takes a specific action in the current state s_t, the USV will receive the expected instantaneous reward $r(s_t,a_t)$ that

$$r(s_t,a_t) = E_{p(s_{t+1}|s_t,a_t)}[r(s_t,a_t)] \tag{8}$$

Through recursion, $Q^{\pi}(s,a)$ can be rewritten as

$$Q^{\pi}(s,a) = r(s_t,a_t) + \gamma E_{\pi(a_{t+1}|s_{t+1})p(s_{t+1}|s_t,a_t)}[Q^{\pi}(s_{t+1},a_{t+1})] \tag{9}$$

By utilizing the aforementioned equation, one can obtain the following updated state-action value function:

$$Q_{k+1}(s_t,a_t) = Q_k(s_t,a_t) + \alpha[r(s_t,a_t) + \gamma \max Q_k(s_{t+1},a_{t+1}) - Q(s_t,a_t)] \tag{10}$$

To further address the dynamic obstacle avoidance and evasion problem of USV in the combat process, deep learning is utilized to enhance the expression ability of Q-learning. The state sequence of the USV is taken as input, and then output the Q-values corresponding to all actions. The agent selects an action based on the Q-values and interacts with the environment. As the network parameters are continuously updated, the evaluation of the state-action pairs by the network model will gradually approach the true Q-value. Therefore, the agent can directly select the action with the maximum Q-value to complete the task and achieve the goal of maximizing the RL return.

During the exploration process of the agent in the environment, the algorithm stores the quadruple tuple (s_t,a_t,r_t,s_{t+1}) in the experience buffer sequence. When the memory of the experience buffer sequence reaches a certain amount, the algorithm can sample the experience according to the algorithm, as shown in Equation (11):

$$Q(s_t,a_t)^{DDQN} = r(s_t,a_t) + \gamma \max_{a} Q\left(s_{t+1},a;\theta_t^-\right) \tag{11}$$

According to Equation (11), the Q-value of state s_t at each step of the Q-value update process is taken as the maximum state-action Q-value of the next state. This leads to a serious overestimation of the Q-value during the update process, resulting in instability and poor results during training. That is, the updated Q-value and the evaluated Q-value are both based on the same policy θ, leading to an overcoupling situation. To address this, the UOACPP method uses a double deep Q-learning network to update the Q-value. Network Q_1 estimates the Q-function, while network Q_2 makes decisions based on the Q-function. The network parameters of the two networks are periodically synchronized according to the algorithm, effectively reducing the impact of network parameter updates on the Q-function and increasing the stability of the model. The above Q-value update process can be written as follows:

$$Q(s_t,a_t)^{DDQN} = r(s_t,a_t) + \gamma Q(s_{t+1}, \arg\max Q\left(s_{t+1},a;\theta_t\right), \theta_t^-) \tag{12}$$

By minimizing the loss function sequence that changes at each iteration, the Q-network can be trained as

$$L(\theta_t) = E_{s,a}\left[(y_i - Q(s,a;\theta_t))^2\right] \tag{13}$$

where y_i denotes $Q(s_t,a_t)^{DDQN}$. The optimization goal of the loss function is the parameter θ_1 in $Q_1(s_t,a_t;\theta_1)$. The gradient descent optimization method is used to make the $Q_1(s_t,a_t;\theta_1)$ function approach the evaluation value calculated by $Q_2(s_t,a_t;\theta_2)$. To ensure that network $Q_2(s_t,a_t;\theta_2)$ can accurately estimate the Q-function value, the algorithm needs to update the parameter θ_2 with parameter θ_1 every once in a while.

4.4. **Reward function design.** Effective and reasonable reward function design is one of the necessary conditions for the success of reinforcement learning in solving complex tasks. When planning the obstacle avoidance and combat path of a USV, it is essential to consider various situations that the USV may encounter throughout the entire process and design the reward function accordingly. Based on the experimental scenario, the combat process of the USV in the red side is divided into two stages, and different reward functions are set for each stage as follows.

1) The reward function during the call-to-search phase

At the beginning of the call-to-search phase, the USV in the red side knows the initial orientation of the blue team's USV, but the subsequent reconnaissance route and specific location of the blue team's USV is unknown. The red team's USV needs to take action within a certain time to search for the location of the USV in the blue side. During the search phase, the red team's USV may encounter the following situations:

a) The USV in the red side collides with an obstacle;

b) The USV in the red side locates the blue team's USV;

c) The USV in the red side searches continuously but does not find the blue team's USV.

For the first situation, when the red team's USV collides while sailing, a greater negative feedback $r_{collision}$ can be given. For the second situation, from a global perspective, the straight-line distance between the red team's unmanned boat and the blue team's unmanned boat is calculated. The closer the two are, the more positive the reward will be. Therefore, when the blue team's USV is within the search range of the red team's USV without any collision, the red team's USV can receive a larger positive reward. Conversely, if the red team's USV keeps moving away from the blue team's USV, negative feedback will be given to the reward function, which is recorded by r_{linear_dist}. For the last situation, if the red team's USV sails aimlessly without collision, a smaller reward $r_{aimleaaly}$ will be given, and the reward will gradually increase over time. After a long search, $r_{aimleaaly}$ will accumulate to a large reward value. In summary, the reward function for the red team's USV in the search task can be given by

$$r = r_{collision} + r_{linear\_dist} + r_{aimleaaly} \tag{14}$$

At the beginning of the training, collisions occur frequently, leading to the end of the round quickly. After a long period of training, the red team's USV learns how to avoid collisions, and it will always be zero in the reward function $r_{collision}$.

2) The reward function during the attack phase

At the beginning of the attack phase, the red team's USV and the blue team's USV discover each other, and the red team's USV will take certain actions to use obstacles as cover and approach the blue team's USV to the greatest extent possible. During the combat process, the red team's USV may encounter the following situations:

a) The USV in the red side collides with an obstacle;

b) The USV in the red side loses the location of the blue team's USV;

c) The red team's USV follows the blue team's USV, and implements an attack within an appropriate range.

For the first situation, when the red team's USV collides while sailing, a greater negative feedback $r_{collision}$ can be given. For the second situation, the path distance between the red team's USV and the blue team's USV is calculated based on the sensor data. This path distance is different from the second term of the reward function for the USV in the red side during the search phase, as it considers terrain information. The closer the two boats are to each other and the less obstruction there is between them, the more positive the reward the red team's USV will receive. Conversely, if the red team's USV keeps moving away from the blue team's USV or if there is an obstruction in between that prevents the red team's USV from detecting the blue team's USV, negative feedback will be given to the red team's unmanned boat, which is recorded by r_{path_dist}. For the last situation, if the USV in the red side follows the blue team's USV throughout the process without collision, a smaller reward $r_{aimleaaly}$ will be given to the USV. When the blue team's USV enters the red team's attack range, the reward value will be increased. Additionally, a large reward value r_{end} is added for task success or failure, and the conditions for obtaining the reward are set according to the task success/failure end event in Table 1. In summary, the reward function for the red team's USV during the attack phase is

$$r = r_{collision} + r_{path\_dist} + r_{aimleaaly} + r_{end} \tag{15}$$

5. **Experimental Results.** This article employs the intelligent gaming platform developed by the Jiangsu Automation Research Institute to edit the unmanned boat combat scenario, and uses its embedded USV simulation model to conduct a red-blue USV game confrontation experiment.

5.1. **Scenario design.** Based on intelligence and satellite information, the blue team comprehensively judges that there is a key military node of the red team in a certain range area. To obtain the key combat information of the red team and ensure the efficiency and concealment of the operation, they dispatch one USV to the northwest of the area for reconnaissance and raid. To ensure that the key command node is not detected or attacked and considering that the sea area has many obstacles and large surface ships are inconvenient to pass the red team dispatched USV to conduct regional searches discover and destroy the blue side.

In this experiment, both the red and blue USVs are aware of the location and shape information of all obstacles. The blue USV is conducting reconnaissance activities here and is unaware of the red team's military strength. The red USV is executing a call-to-search mission but only knows the approximate location of the blue team and needs to rely on onboard radar for detection and search. The performance parameters of the red and blue unmanned boats are the same, and the radar detection and weapon attack ranges are 25 km and 20 km, respectively. The red team's objective in this experiment is to discover and destroy the blue team's USV in the shortest possible time while ensuring its own safety. The combat mission description for the red team's USV is shown in Table 1.

5.2. **Experimental conditions.**

5.2.1. *Training parameter configuration.* This paper conducts deep reinforcement learning training by using the sample data generated through the mutual game confrontation between the red team's USV with the UOACPP algorithm and the blue team's USV with the predefined rule.

TABLE 1. The combat mission description of the USV in the red side

Phase	Target	Task success end event	Task failure end event	Description
Call-to-search phase	The red team's USV is required to navigate through a fixed sea area, avoid obstacles, and conduct a rapid search of the area to locate the blue team's USV.	The red team's USV detects the location of the blue team's USV.	1) The red team's USV collides with an obstacle; 2) Within the specified time, the red team's USV fails to detect the blue team's USV.	The red team's USV is aware of the blue team's USV's direction and weapon equipment performance, but is unaware of the blue team's USV's specific location and movement.
Attack phase	The red team's USV must leverage the terrain to plan a strike route that considers the blue team's mobility while ensuring its own safety.	The red team's USV successfully destroyed the blue team's USV.	1) The red team's USV collides with an obstacle; 2) The red team's USV is destroyed; 3) The blue team's USV is not destroyed by the end of the scenario	The red team's USV is aware of the blue team's USV's location, but the detection accuracy is affected by obstacles.

According to the experimental setup, this experiment will simultaneously open 30 simulation environments for 30 rounds of game confrontation based on the size of the expected scenario and the overall training resource configuration. That is, the size of an episode is 30 rounds of game data. Sample data is collected and deep reinforcement learning training is performed using a cluster server composed of 36 CPUs and 1 GPU. The average reward value and average decision steps of each episode during the unmanned boat intelligent agent training process are recorded. The effect data of the intelligent agent training is visualized using TensorBoard, and the parameter values of the neural network are saved every 10 episodes to generate an intelligent USV model, which is convenient for obtaining the required intelligent agent model in a timely manner. In addition, the neural network's hidden layer size is set to 128 layers, because the hidden layers are responsible for learning complex representations and patterns from the input data, enabling the network to make accurate predictions and decisions. The discount factor is set to 0.9, and the learning rate is set to $2e^{-4}$ to improve the learning efficiency.

5.2.2. *Construction of the blue team's rule.* In this experiment, the UOACPP-based USV in the red side will play against the rule-based USV in the blue side to achieve obstacle avoidance and combat path planning by DRL learning training. To prevent the rule-based model from having problems such as less change in action decisions and weak generalization ability, a blue rule-based USV model with multiple decision-making options is established on the simulation platform. According to the environmental state information and weapon information, different combat tasks are triggered when the battlefield situation changes, and a series of decision-making processes such as detection, evasion, and strike strategies are performed for the blue team.

The blue team's rule-based USV is constructed to use the two longest edges of the obstacle area as a rectangle no-sail zone. At the start of the scenario, the blue team first selects a distant target point in the combat area and plans several feasible polyline paths to reach the target point. Then, based on a distance-related probability distribution, a random path is selected to navigate for reconnaissance. During the voyage, if the red team is detected, the blue team will evade in the opposite direction. When the red team's USV enters the attack range, the blue team will immediately launch weapons to strike. If the blue team loses sight of the red team's USV, it will reselect another direction as the target point and plan several polyline paths to reach the target point, and then randomly select a path for searching. The navigation path of the blue team's USV constructed is not fixed, ensuring that the battlefield situation faced by the red team's USV intelligent agent is different in each round of training, thereby improving the intelligent agent's generalization ability.

5.3. **Analysis of experimental results.** The scenario of USV obstacle avoidance and combat path planning can be divided into two stages. The first stage involves obstacle avoidance and search path planning in a fixed sea area, while the second stage involves combat path planning to attack the blue USV. The combat processes and rewards of the two stages are different; thus, the experiment is also divided into two stages. In the first stage, the USV in the red side is required to search for the blue unmanned boat within a specified time. In the second stage, the trained USV model is inherited to continue the red-blue confrontation training.

5.3.1. *Obstacle avoidance search path planning.* The UOACPP method proposed in this paper guides the red USV to search for the blue USV within a fixed sea area in the first stage, achieving the task requirement of quickly discovering the blue USV and improving search efficiency. Before discovering the target, the red USV will estimate the range where

the blue USV is more likely to appear based on its approximate orientation and proceed with the search. During this process, the USV needs to avoid obstacles to prevent collisions that may lead to mission failure.

To compare the training effects of the UOACPP algorithm and verify the advantages of the Transformer structure, ablation experiments are conducted. DDQN algorithm and commonly used LSTM network structure are trained separately. The average reward of every ten rounds of training is taken to plot the reward curve of the ablation experiment, as shown in Figure 5.

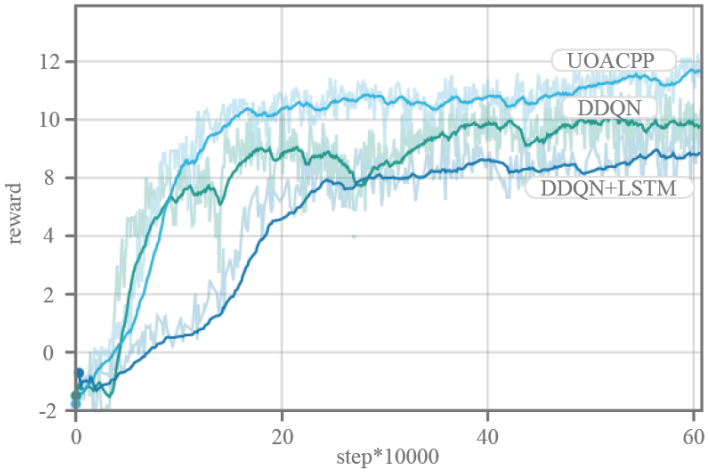


FIGURE 5. Reward curves of different methods under obstacle avoidance task

During the initial training phase, the red USV collides with obstacles due to its random movement pattern, resulting in a penalty feedback. Subsequently, the USV gradually learns to avoid obstacles and is able to detect the blue USV. The reward functions of the three methods gradually converge with different time to detect the blue USV due to different search strategies. The UOACPP algorithm trains a policy that could detect the blue target earlier than the other two methods, with a higher converging reward value, faster convergence speed, and more stable convergence curve.

5.3.2. *Strike mission path planning.* In the second stage of combat training, the USV intelligent agent model established in the search phase is inherited. The USV has already acquired obstacle avoidance capabilities. Ablation experiments are conducted, and the reward variation process is shown in Figure 6. It can be seen that the convergence time

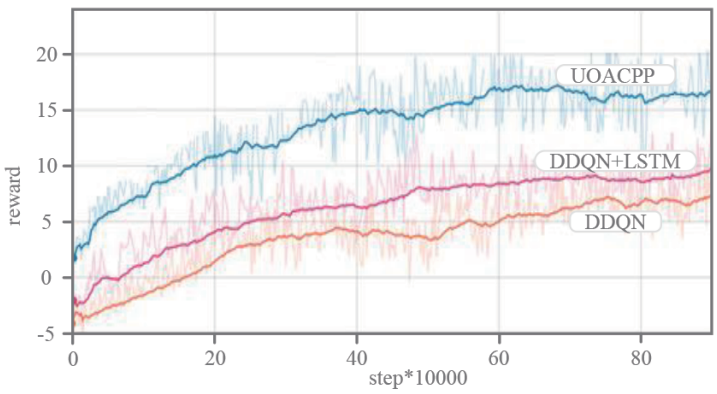


FIGURE 6. The reward curve for different methods in the strike mission

is longer in the combat phase. The convergence time of the three methods is similar, but the convergence value of the UOACPP algorithm is far superior to the other two training methods. The USV intelligent agent obtained has stronger combat capabilities and can destroy blue USVs at a lower cost.

5.3.3. *Operational effect analysis.* Furthermore, three stable USV intelligent agent models trained using different methods are obtained. Their average time, shortest time, and longest time to detect the blue target in 50 rounds of combat are recorded. The simulation acceleration ratio is 1 : 60, and the recorded time is simulation time. The results are shown in Figure 7. The UOACPP algorithm detects the blue USV in an average time of 3.5 minutes, which is 40.7% more efficient than direct DDQN algorithm search, and 23.9% more efficient than LSTM network search. In UOACPP method, the difference between the shortest and longest time is minimal, and the search path is more stable. In addition, the longest time to detect the blue target in the proposed method is still superior to direct DDQN algorithm. The results indicate that the UOACPP algorithm proposed in this paper can plan a safe and efficient obstacle avoidance search path for the red USV search phase.

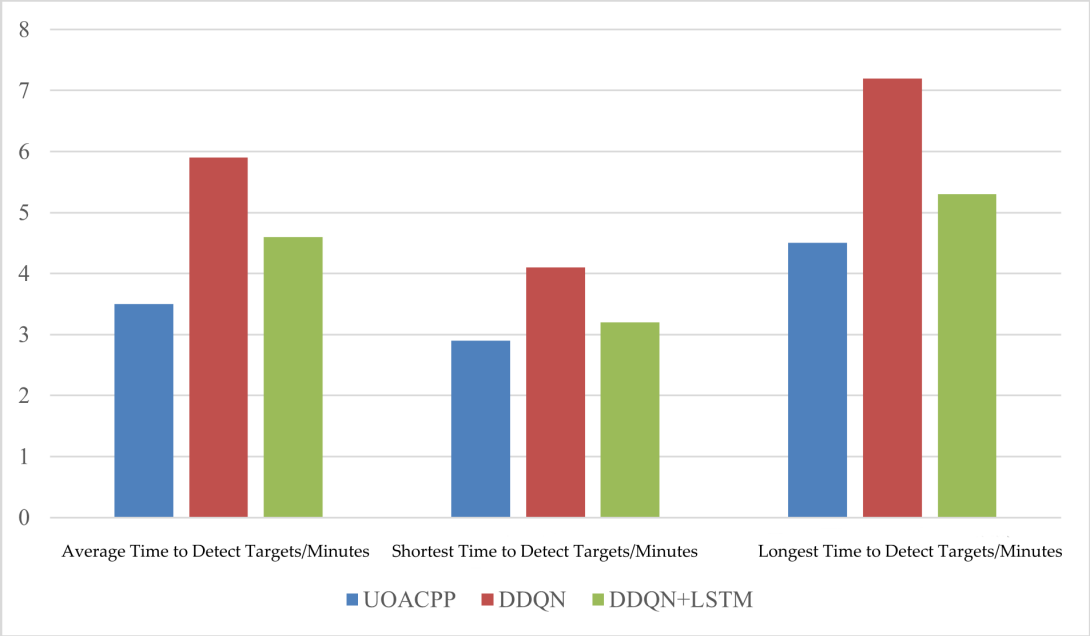
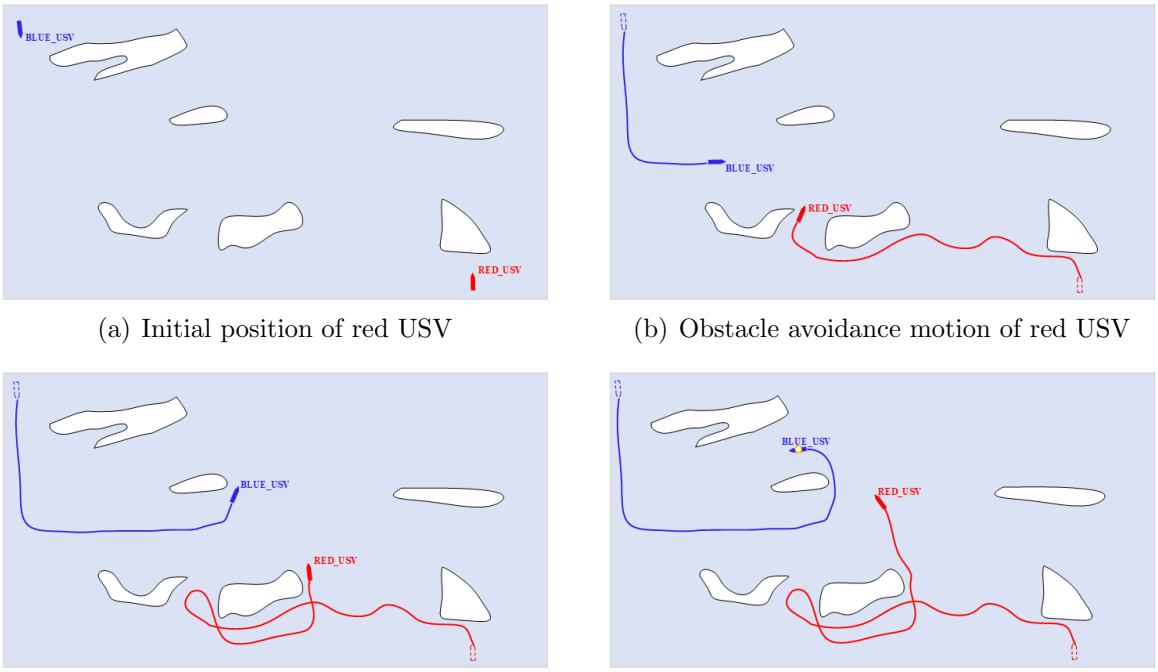


FIGURE 7. Statistical results of different algorithms in search task

Then, we analyze the process of USV searching for the blue target by the UOACPP algorithm. Firstly, the red USV departs from the initial position and encounters an obstacle while searching towards the upper left corner, as shown in Figure 8(a). The red USV moves downward to avoid the obstacle and continues to search for the area where the blue USV may exist after bypassing the island reef obstacle, as shown in Figure 8(b). Then the red USV uses a larger island reef to move along the back of the obstacle to the possible location of the blue USV, as shown in Figure 8(c). The red USV uses the obstacle for evasion, while the blue USV is in an open area and could not find effective cover, ultimately leading to its destruction, as shown in Figure 8(d).

In summary, after two stages of training, the red USV can plan an efficient search path from a global perspective while avoiding static obstacles. When the blue target is detected, the USV can judge whether it is advantageous for the USV to attack based on the surrounding obstacles and the blue USV's location. Under attack, the red USV can



(a) Initial position of red USV

(b) Obstacle avoidance motion of red USV

(c) Red USV's combat path to attack the blue USV

(d) The red USV attacks the blue USV

FIGURE 8. Responding process of USVs in search phase

use obstacles to evade along the shortest path and plan subsequent attack paths using terrain advantages to achieve a one-hit kill. Experiment results show that the UOACPP algorithm proposed in this chapter can achieve automatic obstacle avoidance and combat path planning for USVs, completing military actions such as search and attack against blue USV.

6. **Conclusions.** This paper addresses the problem of USV path planning for attacking blue adversaries in a fixed maritime domain. To cope with challenges such as long decision-making cycles, multiple task objectives, and high dynamic environmental awareness, we propose a Q-Transformer-based USV obstacle avoidance and combat path planning algorithm. First, we design a Transformer-based policy network for the red USV to enhance its perception of the current environmental state. Second, we construct multi-stage rewards for search and strike tasks to improve the learning ability of single complex tasks. Finally, we conduct experiments in a corresponding scenario and verify the effectiveness of our method by comparing it with various baseline algorithms such as DDQN and DDQN+LSTM for obstacle avoidance and path planning. The experiment results show the effectiveness and the superiority of the proposed UOACPP method.

Future research could focus on developing adaptive learning algorithms for dynamic environments, improving multi-agent coordination for USVs, and integrating real-time data fusion to enhance situational awareness and decision-making. Additionally, optimizing reward structures and the Q-Transformer model for practical applications would further enhance USV performance in path planning and obstacle avoidance.

REFERENCES

[1] J. Liu, X. Qin, B. Qi and X. Cui, 3D online path planning of UAV based on improved differential evolution and model predictive control, *International Journal of Innovative Computing, Information and Control*, vol.16, no.1, pp.315-329, 2020.

[2] Y. Yang, J. Wu and W. Zheng, Attitude control for a station keeping airship using feedback linearization and fuzzy sliding mode control, *International Journal of Innovative Computing, Information and Control*, vol.8, no.12, pp.8299-8310, 2012.
[3] S. Zaghi, G. Dubbioso, R. Broglia and R. Muscari, Hydrodynamic characterization of USV vessels with innovative SWATH configuration for coastal monitoring and low environmental impact, *Transportation Research Procedia*, vol.14, pp.1562-1570, 2016.
[4] J. Han and J. Kim, Three-dimensional reconstruction of a marine floating structure with an unmanned surface vessel, *IEEE Journal of Oceanic Engineering*, vol.44, no.4, pp.984-996, 2019.
[5] R. J. Yan, S. Pang, H. B. Sun and Y. J. Pang, Development and missions of unmanned surface vehicle, *Journal of Marine Science and Application*, vol.9, pp.451-457, 2010.
[6] A. Gonzalez-Garcia and H. Castañeda, Guidance and control based on adaptive sliding mode strategy for a USV subject to uncertainties, *IEEE Journal of Oceanic Engineering*, vol.46, no.4, pp.1144-1154, 2021.
[7] Z. Chen, T. Bao, B. Zhang, T. Wu, X. Chu and Z. Zhou, Deep reinforcement learning methods for USV control: A review, *2023 China Automation Congress (CAC)*, Chongqing, China, pp.1526-1531, 2023.
[8] J. Li, G. Zhang, Q. Shan and W. Zhang, A novel cooperative design for USV-UAV systems: 3D mapping guidance and adaptive fuzzy control, *IEEE Transactions on Control of Network Systems*, vol.10, no.2, pp.564-574, 2023.
[9] K. Yu, X. Liang, M. Li, Z. Chen, Y. Yao, X. Li, Z. Zhao and Y. Teng, USV path planning method with velocity variation and global optimization based on AIS service platform, *Ocean Engineering*, vol.236, 109560, 2021.
[10] P. Yao, R. Zhao and Q. Zhu, A hierarchical architecture using biased min-consensus for USV path planning, *IEEE Transactions on Vehicular Technology*, vol.69, no.9, pp.9518-9527, 2020.
[11] H. Niu, Y. Lu, A. Savvaris and A. Tsourdos, An energy-efficient path planning algorithm for unmanned surface vehicles, *Ocean Engineering*, vol.161, pp.308-321, 2018.
[12] Y. Long, S. Liu, D. Qiu, C. Li, X. Guo, B. Shi and M. S. AbouOmar, Local path planning with multiple constraints for USV based on improved bacterial foraging optimization algorithm, *Journal of Marine Science and Engineering*, vol.11, no.3, 489, 2023.
[13] Y. Ma, M. Hu and X. Yan, Multi-objective path planning for unmanned surface vehicle with currents effects, *ISA Transactions*, vol.75, pp.137-156, 2018.
[14] X. Zhu, B. Yan and Y. Yue, Path planning and collision avoidance in unknown environments for USVs based on an improved D* lite, *Applied Sciences*, vol.11, no.17, 7863, 2021.
[15] X. Guo, M. Ji, Z. Zhao, D. Wen and W. Zhang, Global path planning and multi-objective path control for unmanned surface vehicle based on modified particle swarm optimization (PSO) algorithm, *Ocean Engineering*, vol.216, 107693, 2020.
[16] Z. Wang, G. Li and J. Ren, Dynamic path planning for unmanned surface vehicle in complex offshore areas based on hybrid algorithm, *Computer Communications*, vol.166, pp.49-56, 2021.
[17] D. V. Lyridis, An improved ant colony optimization algorithm for unmanned surface vehicle local path planning with multi-modality constraints, *Ocean Engineering*, vol.241, 109890, 2021.
[18] S. MahmoudZadeh, A. Abbasi, A. Yazdani, H. Wang and Y. Liu, Uninterrupted path planning system for Multi-USV sampling mission in a cluttered ocean environment, *Ocean Engineering*, vol.254, 111328, 2022.
[19] T. Huang, Z. Chen, W. Gao, Z. Xue and Y. Liu, A USV-UAV cooperative trajectory planning algorithm with hull dynamic constraints, *Sensors*, vol.23, 1845, 2023.
[20] I. Masmitja et al., Dynamic robotic tracking of underwater targets using reinforcement learning, *Sci. Robot.*, vol.8, no.80, 7811, 2023.
[21] Y. Zhao, X. Qi, Y. Ma, Z. Li, R. Malekian and M. A. Sotelo, Path following optimization for an underactuated USV using smoothly-convergent deep reinforcement learning, *IEEE Transactions on Intelligent Transportation Systems*, vol.22, no.10, pp.6208-6220, 2020.
[22] B. Du, B. Lin, C. Zhang, B. Dong and W. Zhang, Safe deep reinforcement learning-based adaptive control for USV interception mission, *Ocean Engineering*, vol.246, 110477, 2022.
[23] Y. Zhao, Y. Ma and S. Hu, USV formation and path-following control via deep reinforcement learning with random braking, *IEEE Transactions on Neural Networks and Learning Systems*, vol.32, no.12, pp.5468-5478, 2021.
[24] Y. Wang, W. Liu, J. Liu and C. Sun, Cooperative USV-UAV marine search and rescue with visual navigation and reinforcement learning-based control, *ISA Transactions*, vol.137, pp.222-235, 2023.

[25] H. Xu, N. Wang, H. Zhao and Z. Zheng, Deep reinforcement learning-based path planning of underactuated surface vessels, *Cyber-Physical Systems*, vol.5, no.1, pp.1-17, 2019.
[26] S. Wang, F. Ma, X. Yan, P. Wu and Y. Liu, Adaptive and extendable control of unmanned surface vehicle formations using distributed deep reinforcement learning, *Applied Ocean Research*, vol.110, 102590, 2021.
[27] Y. Cheng, Z. Sun, Y. Huang and W. Zhang, Fuzzy categorical deep reinforcement learning of a defensive game for an unmanned surface vessel, *International Journal of Fuzzy Systems*, vol.21, pp.592-606, 2019.
[28] P. Lai, Y. Liu, W. Zhang and H. Xu, Intelligent controller for unmanned surface vehicles by deep reinforcement learning, *Physics of Fluids*, vol.35, no.3, 37111, 2023.
[29] Z. Sun, Y. Fan and G. Wang, An intelligent algorithm for USVs collision avoidance based on deep reinforcement learning approach with navigation characteristics, *Journal of Marine Science and Engineering*, vol.11, no.4, 812, 2023.
[30] J. Xia, Y. Luo, Z. Liu, Y. Zhang, H. Shi and Z. Liu, Cooperative multi-target hunting by unmanned surface vehicles based on multi-agent reinforcement learning, *Defence Technology*, vol.29, pp.80-94, 2023.
[31] J. Woo and N. Kim, Collision avoidance for an unmanned surface vehicle using deep reinforcement learning, *Ocean Engineering*, vol.199, 107001, 2020.
[32] S. Nantogma, S. Zhang, X. Yu, X. An and Y. Xu, Multi-USV dynamic navigation and target capture: A guided multi-agent reinforcement learning approach, *Electronics*, vol.12, 1523, 2023.

Author Biography

Hongyu Guo received the B.S. degree in Central South University in 2018 and M.S. degree in China Ship Research Academy in 2022, China. He is currently pursuing Ph.D. degree in China Ship Research Academy. His current research interests include the application of artificial intelligence in decision-making, computing resource scheduling and large language models.

Hao Gu received the Ph.D. degree in Northwestern Polytechnical University, China, in 2007. He served as the director of the 716th Research Institute of China State Shipbuilding Corporation Limited from 2016 to 2023. He is currently a Professor and Doctoral Supervisor with the Northwestern Polytechnical University and the Jiangsu Automation Research Institute. He is a Fellow of the Chinese Institute of Command and Control. His research interests include shipborne combat command system, fire control and command control systems, and weapon equipment system evaluation.

Lintao Dou received the B.S. degree in Harbin Institute of Technology in 2003 and M.S. degree in China Ship Research Academy in 2006, China. He is currently a Professor and Graduate Supervisor with the Jiangsu Automation Research Institute. His research interests include system simulation, artificial intelligence and pattern recognition.