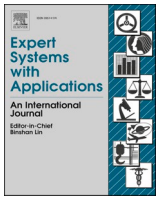
Expert Systems With Applications 256 (2024) 124906

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



USV formation navigation decision-making through hybrid deep reinforcement learning using self-attention mechanism

Zhewen Cui [1], Wei Guan [*,1], Xianku Zhang

Navigation College, Dalian Maritime University, Dalian 116026, China

ARTICLE INFO

Keywords:
USV
Formation
Collision avoidance
Decision-making
Hybrid deep reinforcement learning

ABSTRACT

To address the challenging of balancing Unmanned Surface Vessel (USV) autonomous collision avoidance and formation maintenance in uncertain environments, a formation construction and navigation decision-making strategy based on Hybrid Deep Reinforcement Learning (HDRL) is proposed in this study. The novelty of this study is that: (1) A HDRL training approach is proposed, incorporating diverse DRL for virtual leader and followers, thereby significantly enhancing the decision-making adaptability of USVs formation. (2) The multi-head attention mechanism of decentralized Critic strategy is used to enhance the attentional focus of the HDRL algorithm towards different agents, thereby effectively improving convergence speed. (3) The method employs a meticulously designed hybrid reward function and incorporates the Optimal Reciprocal Collision Avoidance (ORCA) for speed selection, thereby providing optimal speed recommendations based on the current situation. It is worth emphasizing that the method exhibits exceptional performance in terms of success rate, navigation time, average reward value, and other relevant indicators within the simulation. Finally, the method is validated through the utilization of real-time navigation data in the Panama Canal, thereby substantiating the potential engineering applicability.

1. Introduction

The continuous advancement of maritime strategic resources has garnered significant attention towards the study of ship formation in relation to maritime combat, search and rescue operations, and escort missions (Park & Yoo, 2019). In maritime combat missions, carrier formations can effectively accomplish intricate strategic tasks by means of inter-formation communication and cooperation, thereby enhancing overall operational safety. This capability often surpasses that of a single ship. In addition, the multi-agent system plays a pivotal role in satellite communication (He et al., 2022), network security (Zhang, Sun, & Hu, 2020), UAV cooperative control (Su, Bhowmick, & Lanzon, 2023), and other domains, offering a novel solution for the investigation of ship formation.

The Unmanned Surface Vessel (USV) formation problem poses the following technical challenges: (1) The formation of the USV fleet is required to be established at the commencement of the mission and maintained throughout the entire voyage; (2) To ensure the safety of USV formation navigation, it is imperative to consider collision avoidance. Specifically, USV formations need to strike a balance between collision avoidance and maintaining formation; (3) The generalization capability of the formation strategy is a crucial aspect in the method. The USV formation must achieve a high success rate of collision avoidance in diverse environments.

In the existing literature, several well-established formation strategies have been identified, including behavioral control of giving (Cheng, Zhang, & Jiang, 2023), virtual structure (Benzerrouk, Adouane, & Martinet, 2014; Morris, Kumar, Biswas, & Mohan, 2024), and leader–follower strategy (Consolini, Morbidi, Prattichizzo, & Tosques, 2008). Currently, the leader–follower method is widely adopted among various approaches, which fundamental concept is to assign all members of a formation as either leaders or followers. The leader navigates along a predetermined or temporary path, exerting control over the movement trajectory of the entire formation. The followers, guided by distance and orientation information relative to the leader, achieve formation control by following their leader.

In (Dai, He, Chen, & Jin, 2020), the formation tracking control of non-holonomic mobile robots is achieved while considering

* Corresponding author.
E-mail addresses: cuizhewen123@dlmu.edu.cn (Z. Cui), gwwtxdy@dlmu.edu.cn (W. Guan), zhangxk@dlmu.edu.cn (X. Zhang).
[1] Zhewen Cui and Wei Guan have contributed equally to this work and share first authorship.

https://doi.org/10.1016/j.eswa.2024.124906
Received 13 April 2024; Received in revised form 3 June 2024; Accepted 25 July 2024
Available online 29 July 2024

communication constraints. In (Cai & Hu, 2017), the author employs a leader–follower multi-agent system with distributed control law, utilizing solely local information. Ghommam *et al.* (Ghommam & Saad, 2018) designed a simplified method for USV formation planning by integrating LOS guidance with leader–follower. The elastic tracking problem (Rezaee, Parisini, & Polycarpou, 2021) is addressed by adopting a formation method that comprises of a leader, a group of healthy agents, and a group of malicious agents under attack. In (Thuyen, Thanh, & Anh, 2023), a synovial double ring control-based algorithm for Autonomous Underwater Vehicle (AUV) formation control was proposed, which employs radial basis function neural network in conjunction with adaptive law to address unknown dynamic parameters and external interference.

The virtual structure strategy considers the entire formation as a unified entity, necessitating the determination of kinematic and dynamic characteristics of the virtual structure, followed by deduction of corresponding characteristics for the virtual target points on said structure (Yuan, He, & Wang, 2019). The robot is capable of achieving formation control by accurately tracking the corresponding virtual target point through the controller. In (Li et al., 2023), an Underwater Glide (UG) formation algorithm based on virtual hinge construction of multi-body system is described. The integration of virtual structure and Artificial Potential Field (APF) algorithm (Zhen, Wan, Li, & Jiang, 2022) presents a novel approach for collision avoidance in AUV formation. In (Benzerrouk et al., 2014), the decision algorithm is modified to conform to the motion characteristics by integrating the robot kinematics constraint with the virtual structure. Zhang *et al.* (Zhang, Yu, Li, & Zhang, 2021) proposed a USV formation control method based on event triggering, ensuring the stability of the formation even in the presence of wind and wave interference.

Although the aforementioned approaches have yielded a series of findings, they remain inadequate for addressing the challenge of dynamic collision avoidance in unfamiliar environments. The conventional control formation methods exhibit certain limitations in facilitating multi-agent cooperative communication (Mehdifar, Bechlioulis, Hendrickx, & Dimarogonas, 2023); Although the leader–follower method exhibits a straightforward control structure and is easily implementable, enabling the followers to achieve formation control by maintaining a predetermined position offset from the leader, this approach excessively relies on the decision-making capability of the leader (Khodamipour, Khorashadizadeh, & Farshad, 2023); The virtual structure strategy necessitates the perpetual maintenance of a rigid formation by its members, thereby compromising flexibility and adaptability (Ning et al., 2023). Furthermore, stringent formation constraints can lead to frequent control commands, heightened energy consumption, and even actuator saturation.

The research on autonomous collision avoidance for ships can be broadly categorized into two domains: (1) conventional planning methods, such as A-star (Auh et al., 2024), Reciprocal Velocity Obstacle (RVO) (Alonso-Mora, Breitenmoser, Beardsley, Siegwart, & Ieee, 2012; Han et al., 2022), APF (Lyu & Yin, 2019; Sang, You, Sun, Zhou, & Liu, 2021), Dynamic Window Approach (DWA) algorithm (Guan & Wang, 2023), etc.; (2) artificial intelligence algorithms, including Ant Colony Optimization (ACO) (Wu & Gao, 2023), Particle Swarm Optimization (PSO) algorithm (Yang et al., 2018), etc. The former exhibits inherent limitations in its application: conventional global path planning methods such as the A-star, RRT, and related algorithms experience significant decision-making challenges when confronted with unknown and dynamic obstacle environments (Cui, Li, & Yan, 2016); the RVO algorithm lacks sufficient foresight capabilities to make reasonable predictions about future scenarios; The APF algorithm necessitates meticulous parameter adjustment of the repulsive field function; the DWA algorithm exhibits limited global awareness in decision making, resulting in subpar dynamic obstacle avoidance performance and susceptibility to local optima (Han, Wang, Wang, & He, 2022).

The emergence of the DRL algorithm presents a novel solution for decision-making in multi-agent formation. Shen *et al.* (Shen et al., 2019) applied DNN models trained using the Deep Q Learning (DQN) algorithm to three test ships, demonstrating the practicality of this approach through ship experiments and establishing a crucial foundation for theoretical research on employing DRL algorithms for ship collision avoidance decision-making. In addition, the DRL algorithm is extensively employed for distributed formation control (Sui, Pu, Yi, & Wu, 2021). In (Pu, Zhang, Ai, Qiu, & Yi, 2023), an online training scheme was be proposed, which combines the advantages of both model-driven and data-driven approaches, demonstrating remarkable performance in terms of robot formation travel time and generalization capability. However, overcoming the issue of overestimation poses a formidable challenge in value learning, while non-uniform noise can lead to an excessive amplification of action value. Even with the incorporation of dueling networks (Wu et al., 2020), decision-making strategies may still exhibit errors when selecting actions within intricate environments.

The strategy-based learning can effectively address problems with high-dimension continuous action spaces. In (Zhao, Ma, & Hu, 2021), a formation-following strategy is developed, demonstrating seamless switching and following of 5 USVs in formation. The Optimal Reciprocal Collision Avoidance (ORCA) algorithm, in addition, generates expert experience, thereby effectively enhancing the training efficiency and success rate of the DRL algorithm for robot formation (Sui et al., 2021); The implementation of a collaborative strategy (de Souza et al., 2021) effectively addresses the issue of Unmanned Aerial Vehicle (UAV) formation pursuit; The integration of the Obstacle Zone Target (OZT) with the action space was considered by Sawada et al. (Sawada, Sato, & Majima, 2021) to address the issue of collision risk prediction. In (Chang, Shan, Zhang, & Dai, 2023), a robot formation navigation method based on the Deep Deterministic Policy Gradient (DDPG) algorithm was proposed, which aims to optimize the collaborative among multiple agents.

Building upon the aforementioned research findings, this study proposes a formation navigation decision-making strategy for USVs using multi-agent deep reinforcement learning. The primary contributions and innovations of this study are outlined as follows:

1) A novel formation strategy is proposed, wherein all members of the formation act as followers and employ Multi-agent Deep Reinforcement Learning (MADRL) training to interact with the environment, acquiring the ability to effectively balance collision avoidance and formation maintenance. Notably, the virtual leader is trained using Single-agent Deep Reinforcement Learning (SADRL) independently from any followers within the formation.
2) The HDRL algorithm incorporates a Multi-head self-attention (MHA) mechanism, which partitions the observation space based on correlation strength. This enables different heads to capture multiple feature representations, thereby mitigating potential biases arising from a single attention mechanism.
3) The virtual leader and followers are equipped with a compound reward and punishment mechanism, tailored to the characteristics of different tasks. Additionally, ORCA is incorporated into speed selection to provide optimal speed recommendations based on the current formation conditions.
4) The proposed strategy exhibits superior performance in terms of success rate, sailing time, average reward value, and other relevant metrics when compared to the traditional formation method. The engineering application potential of this study is validated using practical navigation data from the Panama Canal.

The remainder of this paper is organized as follows: The ship motion mathematics and DRL are given in Section II. The Section IV introduces the formation strategy and the construction of neural network in detail. Finally, the experimental part is arranged in Section V and the conclusion is proposed in Section VI.

2. Preliminaries and problem formulation

2.1. Ship motion mathematical model

In this study, all USVs are assumed to possess identical structure and motion models. To account for the ship's horizontal plane motion state and the intricate influence of hydrodynamic derivatives on training the USV Formation Navigation Decision-making (UFND) model, a simplified 3-DOF motion model (Fossen, 2011) is employed as depicted in Fig. 1, with its corresponding equation presented as follows:

$$\begin{aligned} \dot{\eta} &= R(\psi)\nu \\ M\dot{\nu} &= \tau - C(\nu)\nu - D(\nu)\nu - g(\nu) + \tau_w \end{aligned} \tag{1}$$

where $\eta = [x, y, \psi]^T$ signifies the positional coordinates and course angle, $\nu = [u, v, r]^T$ represents the velocity vector, $g(\nu) = [g_u, g_v, g_r]^T$ denotes the unmodeled dynamic model, τ_w is the sum of the external forces of environmental interference, and $R(\psi)$ corresponds to the rotation matrix. The system inertia matrix M is composed of the mass matrix representing the additional mass in fluid mechanics and the rigid body mass matrix. The $C(\nu)$ matrix comprises a centripetal matrix and a hydrodynamic Coriolis matrix. The $D(\nu)$ is a matrix that can be represented by the sum of linearly damped and nonlinear damped superpositions. The moment of three degrees of freedom is denoted by $\tau = [X_\delta\delta, Y_\delta\delta, N_\delta\delta]^T$.

Both virtual leader and followers must rely on the mathematical model of ship motion during UFND model training to determine action output and ensure that the trained model adhere ship motion characteristics.

2.2. Deep reinforcement learning

Reinforcement learning (RL) is a distinct form of machine learning that diverges from both supervised and unsupervised learning paradigms (Ladosz, Weng, Kim, & Oh, 2022). This approach hinges upon the agent's continual interaction with its environment to acquire rewards, thereby iteratively optimizing its own action strategy to maximize long-term returns. The Markov Decision Process (MDP) as the ideal form of RL in mathematics consists of four parts: $\langle S, A, P, r\rangle$. The state space is denoted by S, the action space by A, the distribution of state transition probability by $P{:}S \times A \times S \to \mathrm{R}$, which is employed to describe the dynamic characteristics of the environment. $r{:}S \times A \to \mathrm{R}$ represents the reward function, which can also be interpreted as the k - step return $\sum_{t=0}^{\infty}\gamma^t r(s_{t+k}, a_{t+k})$ starting from t.$\gamma \in (0,1)$ denotes the discount factor. The mapping of each state s in the policy can be denoted as $\pi(\cdot|s)$, while the stochastic policy provides the probability $\pi(a|s)$ of selecting action a given state s. $\varphi(\pi) = \mathrm{E}_\pi[\sum_{t=0}^{\infty}\gamma^t r(s_t, a_t)]$ is employed to assess the efficacy of the strategy.

Nevertheless, the intricate nature of the observation space and task gives rise to the issue of dimensionality explosion in RL. This challenge has been progressively addressed with the emergence of deep learning techniques. The incorporation of DNN not only endows the DRL algorithm with the capability to effectively approximate the optimal value function and strategy, but also indirectly enhances decision-making performance (Hasselt, Guez, & Silver, 2015; Lillicrap et al., 2015; Schulman, Moritz, Levine, Jordan, & Abbeel, 2015). Therefore, the DRL plays a pivotal and indispensable role in various frontier technology fields such as autonomous driving (Li, Zhao, Zhang, & Chen, 2019), traffic planning (Li et al., 2023), strategy games (Vinyals et al., 2019) and Natural Language Processing (NLP) (Banino, Badia, Walker, Scholtes, & Blundell, 2021).

3. Proposed approach

3.1. Formation model

The virtual leader strategy is employed to USVs formation model for the tasks of maintaining formation and making navigation decisions (Zhao et al., 2021). As illustrated in Fig. 2, USV^V represents the virtual leader whose actual position coordinates can be denoted as $p^{V_t}(x^{V_t}, y^{V_t})$; USV_i^F, USV_j^F, USV_k^F denotes the followers, while $p_*^{F_t}(x_*^{F_t}, y_*^{F_t})$ and $p_*^{F_p}(x_*^{F_p}, y_*^{F_p})$ represent the real positions and their expected arrival positions respectively, among them $(*) = [i, j, k]$; e_d signifies the distance error within the virtual leader formation. Taking USV_i^F as an example, e_{fi} represents the error in formation distance, which denotes the deviation between the actual position and its position at the formation distance. d_i and θ_i represent the differences in distance and angle between the followers and virtual leader after forming a stable formation. The expected position of USV_i^F within the formation can be expressed as follows:

$$x_i^{F_p}|\theta = \theta_0 = x^{V_t}|\theta = \theta_0 + d_i\cos\left(\arctan\left(\frac{dy}{dx}|\theta = \theta_0\right) + \theta_i\right) \tag{2}$$

$$y_i^{F_p}|\theta = \theta_0 = y^{V_t}|\theta = \theta_0 + d_i sin\left(\arctan\left(\frac{dy}{dx}|\theta = \theta_0\right) + \theta_i\right) \tag{3}$$

The error in the formation distance of USV_i^F among followers can be expressed as:

$$e_{fi} = \sqrt{\left(x_i^{F_p} - x_i^{F_t}\right)^2 + \left(y_i^{F_p} - y_i^{F_t}\right)^2} \tag{4}$$

Assuming a formation of m USVs, the collective error in formation

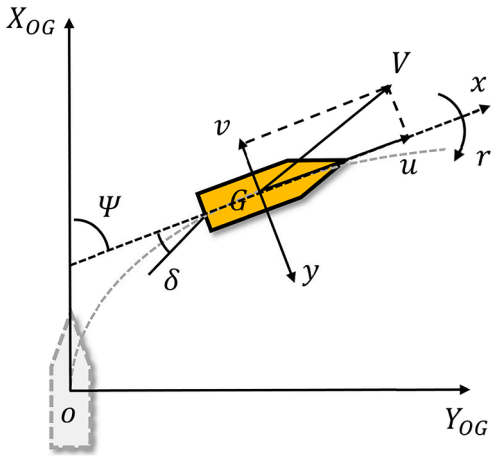


Fig. 1. 3-DOF ship motion mathematical model.

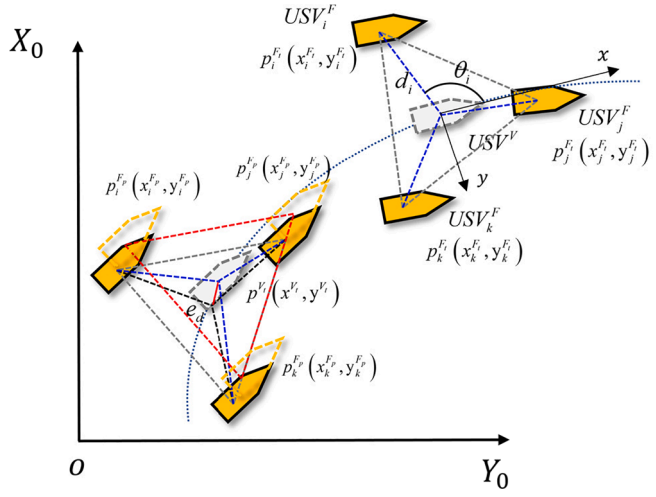


Fig. 2. The USV formation model based on virtual leader.

distance can be represented as a weighted average of multiple individual errors:

$$e_f = \frac{\sum_i^{m+i} \sqrt{\left(x_i^{F_p} - x_i^{F_t}\right)^2 + \left(y_i^{F_p} - y_i^{F_t}\right)^2}}{m} \tag{5}$$

Furthermore, the objective of USV formation construction and maintaining can be defined as achieving satisfaction of condition $\lim_{t\to\infty} e_f = 0$ within a sufficiently short time t interval.

Remark 1. *This study does not take into account the leader's path error when designing the virtual leader formation strategy. This is because the primary focus of this study is on formation navigation, which is achieved through MADRL algorithm. Unlike the formation tracking control, the virtual leader is independent of followers and does not need to rely on the path planning algorithm to plan the expected arrival position of the virtual leader. The real-time collision avoidance during movement is entirely dependent on SADRL. This formation strategy, which is independent of the leader's path error, significantly simplifies the subsequent reward function design.*

3.2. Formation strategy

The formation strategy framework by HDRL is illustrated in Fig. 3. It is imperative for the USVs formation navigation strategy to preserve the individual members' capacity to autonomously make collision avoidance decisions, thereby enhancing the adaptability of the formation. In this study, the virtual leader and the follower are acted as distinct entities with separate roles and strategies to improve the flexibility of the USVs formation. The virtual leader assumes responsibility for providing coordinates indicating the expected arrival location of the follower formation, while followers are assigned with maintaining the formation, and the distinct DRL strategies are employed for each role. Specifically, the virtual leader utilizes the more generalizing SADRL algorithm to provide coordinates of the expected arrival position for followers. Meanwhile, the followers employ the MADRL algorithm to ensure communication among formation members. Furthermore, both of them are required to avoid collision with unknown and dynamic obstacles. Consequently, the designed HDRL algorithm must strike a balance between formation maintenance and collision avoidance.

3.2.1. Virtual leader decision-making algorithm

The virtual leader is solely responsible for providing the expected arrival position coordinates to followers and making collision avoidance, without considering the motion state and position coordinates of followers. Therefore, we employ the Proximal Policy Optimization (PPO) algorithm as a navigation strategy for the virtual leader. The PPO algorithm enhances policy learning by synchronously updating two network parameters within the Actor-Critic network structure. An Actor network is defined as a policy approximation function $\pi(a|s)$ that generates action a. The Critic network is employed to approximate the action value function $Q_\pi(s,a)$ and assess its quality based on scores. The method requires the identification of an optimal value within the current network parameters, such that the updated objective function surpasses its current counterpart.

The PPO algorithm constrains the objective function through clipping method, thereby minimizing the discrepancy between new and old parameters. Consequently, the loss function of the PPO algorithm is formulated as follows:

$$L^{PPO}(\theta) = E_{\tau\sim\pi}\left[\sum_{t=0}^{T}\left[\min(\eta_\theta^t \widehat{G}^t, clip(\eta_t(\theta), 1-\varepsilon, 1+\varepsilon)\widehat{A}^t)\right]\right] \tag{6}$$

$$\eta_\theta^t = \frac{\pi(a_t|s_t)}{\pi_{old}(a_t|s_t)} \tag{7}$$

where η_θ^t represents the ratio between old and new strategies,$clip(*)$ denotes the clipping function, and ε signifies the clipping factor. In comparison to other policy-based algorithms, the PPO algorithm exhibits notable advantages such as consistent performance, robust adaptability to diverse network structures, and simultaneous handling of problems involving discrete continuous action spaces.

3.2.2. Followers decision-making algorithm

The followers of formation decision-making algorithm employ a multi-agent PPO via decentralized Critic strategy with a multi-head self-attention mechanism (MHA-MAPPO). The loss update of the traditional MAPPO algorithm is not discussed in this paper, but can be referred to the literature (Yu, Velu, Vinitsky, Wang, & Wu, 2021).

The formation task is defined as a fully cooperative scenario, wherein all followers maintain a cooperative relationship. Despite adopting different action strategies, their cumulative returns remain identical. We employ a hybrid communication framework, characterized by centralized training and decentralized execution. The followers transmit its own state observation data to the central controller, which acquires the

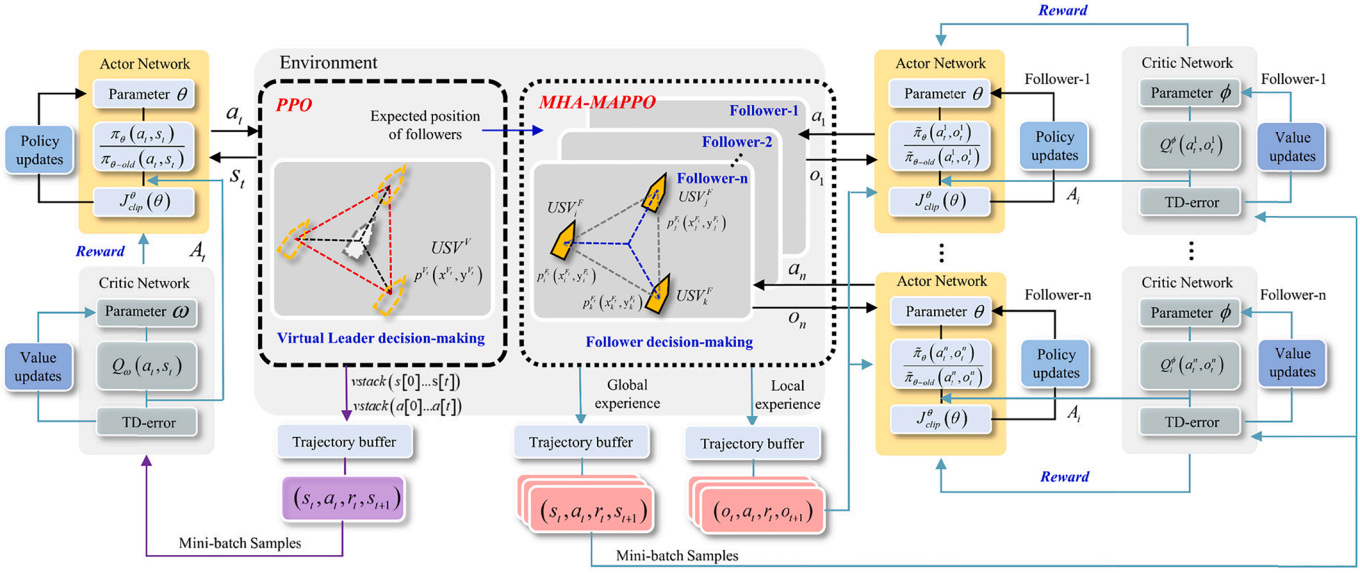


Fig. 3. The USV formation sailing decision-making strategy based on hybrid training framework.

global state and conducts model training, subsequently providing the optimal decentralized control strategy to followers. The trained followers are no longer able to establish communication with the central controller, and thus relies on its own local observation state and action strategy function to generate the optimal action.

We incorporate a decentralized Critic strategy with a multi-head self-attention mechanism (Iqbal & Sha, 2018). In contrast to the conventional approach in the field of NLP, our method facilitates information exchange among followers by querying each other regarding their observations and operations. This acquired information is then integrated into an estimation of the value function, without any assumptions about temporal or spatial positioning of the input. The collective contribution Q function for agent i can be formulated as follows:

$$Q_i^{\phi}(o,a) = f_i(g_i(o_i,a_i),x_i) \tag{8}$$

$$x_i = \sum_{j\neq i} a_j v_j = \sum_{j\neq i} a_j h(V g_j(o_j,a_j)) \tag{9}$$

where $o=(o_1,\ldots,o_N)$ denotes the observation space of the Critic strategy, while $a=(a_1,\ldots,a_N)$ represents the action, $i\in\{1,\ldots,N\}$, and f_i refers to a two-layer fully connected, g_i denotes a single-layer fully connected, x_i represents the weighted sum of values from other followers, and v_j signifies the embedded function used for coding. The multi-head mechanism partitions the observation information of different followers based on their correlation, thereby enabling diverse heads to capture multiple feature representations and mitigate potential biases arising from a single attention mechanism. The MHA can be formulated as follows:

$$Head_i = Attention(QW_i^q, KW_i^k, VW_i^v) \tag{10}$$

$$MultiHead(Q,K,V) = concat([Head_i]_{i=1}^{h})W_i^h \tag{11}$$

where, Q represents the query matrix, K represents the key matrix, V represents the value matrix, W_i^q, W_i^k, W_i^v represents the weight of Q,K,V, and h represents the number of heads. The α_j is regarded as the weight of MHA and is used to measure $g_i(o_i,a_i)$. This process is shown in Fig. 4, which can be expressed as:

$$\alpha_j \propto \exp(g_j(o_j,a_j)W_i^{kT}W_i^{qT}g_i(o_i,a_i)) \tag{12}$$

Similar to the previous and current strategies employed in the PPO algorithm, the loss function of the Critic network in MHA-MAPPO algorithm can be defined as follows:

$$L(\phi) = \frac{1}{Bn}\sum_{i=1}^{B}\sum_{j=1}^{n}\left[\max\left[(Q_i^{\phi}(o,a)-y_i)^2\right], (clip(Q_i^{\phi}(o,a), Q_{i-old}^{\phi}(o,a)-\varepsilon, Q_{i-old}^{\phi}(o,a)+\varepsilon)-y_i)^2\right] \tag{13}$$

$$y_i = r_i + \gamma E_{a'\sim\pi_{\bar{\theta}}}[Q_i^{\bar{\phi}}(o',a') - \alpha\log(\pi_{\bar{\theta}}(a_i|o_i))] \tag{14}$$

The loss function of the policy network is shown as follows:

$$L(\theta) = \frac{1}{Bn}\sum_{i=1}^{B}\sum_{j=1}^{n}\min\left[\eta_{\theta,i}^{(j)}A_i(o,a), clip(\eta_{\theta,i}^{(j)}, 1-\varepsilon, 1+\varepsilon)(o_i^{(j)})\right] A_i(o,a)] + \alpha\frac{1}{Bn}\sum_{i=1}^{B}\sum_{j=1}^{n}S_{(\theta)}\left[\pi_{\bar{\theta}}\right] \tag{15}$$

$$\eta_{\theta,i}^{(j)} = \frac{\pi_{\theta}(a_i^{(j)}|o_i^{(j)})}{\pi_{\theta-old}(a_i^{(j)}|o_i^{(j)})} \tag{16}$$

where, $S_{(\theta)}$ represents the policy entropy function, α denotes the regulatory factor of the entropy function, B signifies the scale of batch size, n indicates the number of agents, and y_i stands for the discount reward.

Remark 2. *In most prior studies, although the leader and followers are treated as an integral entity, they are both implemented using identical algorithms (Zhao et al., 2021). Considering the distinct nature of the two tasks, it is imperative to ensure that the virtual leader exhibits a superior collision avoidance success rate across diverse environments. Evidently, the SADRL framework training yields enhanced model generalization effects compared to MADRL. In addition, the followers require the MADRL algorithm to serve as the decision module, enabling interaction and cooperation through the communication framework.*

3.3. State and action space design

To enhance the generalization capability of the UFDN model, we leverage lidar sensors to enable independent exploration of the virtual leader and followers in the environment for acquiring observation data, thereby eliminating their reliance on positioning and global mapping, as depicted in Fig. 5. The observation information acquired through virtual leader detection is $o_v=[o_1,o_2,o_3\ldots o_{24}]$. Furthermore, this study assumes that virtual leader possesses the capability to obtain its own position $p^{V_t}(x^{V_t},y^{V_t})$ and expected position $p^d(x^d,y^d)$. Consequently, the state space of the virtual leader is defined as $S_{USV}^{VL}=[o_v,x^{V_t},y^{V_t},x^d,y^d]$. The followers are required to gather the observation space encompassing all members of the formation as input, which can be represented as $o_F=[(o_1^1,\ldots o_{24}^1),\ldots,(o_1^n,\ldots o_{24}^n)]$, n denotes the number of followers. Additionally, assuming that both the follower's own position and expected arrival positions can be utilized for hybrid communication, its state space can be represented as: $S_{USV}^F=[o_F,x_*^{F_t},y_*^{F_t},x_*^{F_p},y_*^{F_p}]$, $(*)\in\{1,\ldots,n\}$.

The action space represents the agent's feedback operation on the environmental impact in each time step. We explore how to adjust both

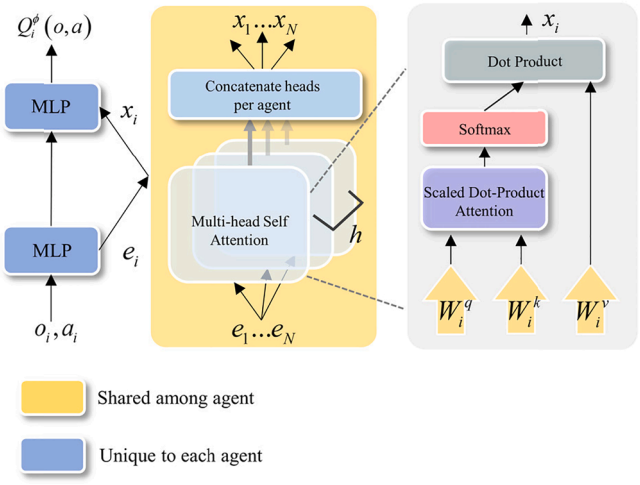


Fig. 4. Application of MHA mechanism in decentralized Critic structure.

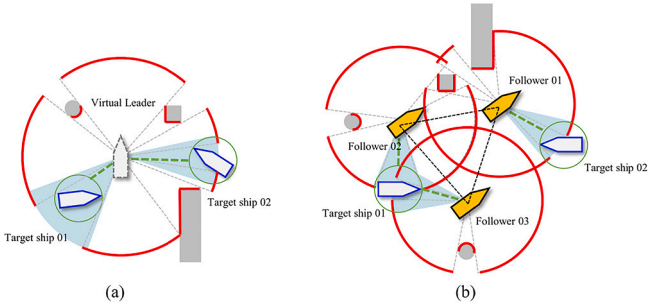


Fig. 5. Lidar based observation space for virtual leader and follower.

the rudder angle and speed to achieve autonomous collision avoidance between the virtual leader and followers. Specifically, we define the action space as a continuous range of rudder angles [−20°, 20°] and a continuous range of speed changes [0 m/s, 5 m/s].

3.4. Network structure design

Considering the absence of moving obstacles in the state space and the lidar sensor's detection process for unknown environments, it is imperative to design a neural network capable of quantifying the impact of different moving obstacles on the USV based on observed state information. Therefore, this study incorporates a Gate Recurrent Unit (GRU) network structure.

The virtual leader exhibits an identical network structure to that of followers, as illustrated in Fig. 6. The GRU can be perceived as an enhanced variant of LSTM, effectively mitigating the issue of gradient vanishing commonly encountered in Recurrent Neural Network (RNN). Specifically, the GRU memorizes long-term information through update gates and reset gates. The reset gate determines the manner in which novel input information is integrated with the preceding memory, while the update gate governs the extent to which the prior memory is retained for the current time step. During the process of formation collision avoidance, the update gate is responsible for encoding and retaining the historical states of mobile obstacles that pose security threats until the subsequent time step. The reset gate integrates the historical state of obstacles preserved by the update gate with newly observed state information, and encodes and organizes the state information based on their priority of importance.

3.5. Reward Function

3.5.1. Reward function for virtual leader

The virtual leader shall effectively execute autonomous collision avoidance and provide accurate expected arrival position for the formation follower members. Throughout this process, the virtual leader is considered an independent entity, not constrained by the formation. Therefore, we have incorporated the distance reward, yaw angle reward, and collision reward based on previous researches (Cui, Guan, Luo, & Zhang, 2023; Z. W. Cui, Guan, Zhang, & Zhang, 2023), while reconfigured their weight parameters. Specifically, the distance reward can effectively guide the virtual leader towards the target; The yaw angle reward facilitates USV course maintenance and enhances its ability to overcome disturbances caused by wind and waves; The collision reward serves as a significant incentive for collision avoidance. Then, the safety incentive has been redesigned to incentivize virtual leader to explore uncharted environments, with the specific formula as follows:

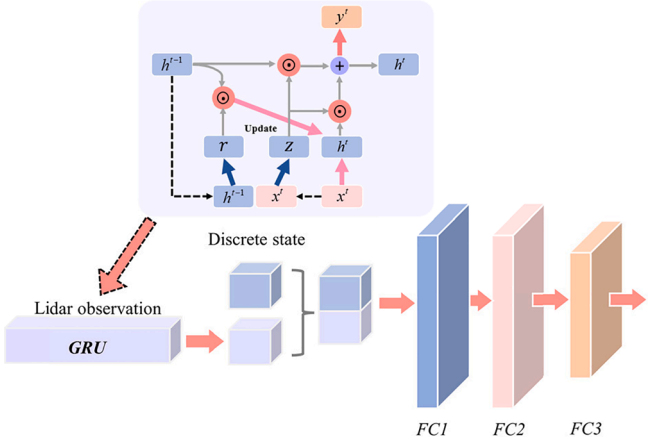


Fig. 6. The neural network design based on GRU structure.

$$R_S^V = \begin{cases} 0, & d_m^V > R_{radar} \\ -\lambda_{s1} \cdot d_m^V, & d_m^V < R_{radar}, d_m^V > S_1 \\ \lambda_{s2} \cdot d_m^V, & d_m^V < R_{radar}, d_m^V < S_1 \end{cases} \tag{17}$$

where, λ_{s1} and λ_{s2} are the weight parameters, d_m^V represents the nearest distance between the virtual leader and the obstacle, which is detected by lidar, R_{radar} represents the detection distance of lidar, and S_1 represents the safe distance. The compound reward of virtual leader is weighted by the above four rewards.

3.5.2. Reward function for followers

The followers must strike a delicate balance between avoiding collisions and maintaining formation, thus necessitating the utilization of three distinct reward functions to accomplish this objective: formation error reward, speed reward, and collision avoidance reward.

The formation error reward is defined as a function of the discrepancy between the expected and actual positions of the follower, with a penalty that proportionally increases in magnitude as the error grows. When the error is kept within a controllable range, a continuous positive reward will be obtained to encourage followers to maintain the formation. The formula is as follows:

$$R_e^F = \begin{cases} 3, & 0 \leqslant e_f \leqslant d_e^1, d_m^F > S_2 \\ -\lambda_e^F \cdot e_f, & d_e^1 < e_f < d_e^2, d_m^F > S_2 \\ -10, & e_f > d_e^2, d_m^F > S_2 \\ 0, & d_m^F < S_2 \end{cases} \tag{18}$$

where, λ_e^F is the weight parameter, d_e^1 is the tolerance error limit, d_e^2 is the maximum error limit, e_f is the follower formation error, and d_m^F represents the minimum distance between the follower and the obstacle.

In formation maintaining, it is imperative to employ speed rewards to maintain followers within an acceptable speed range. To achieve this, we have incorporated the ORCA concept into our speed reward. For each follower and any agent present in the environment, it becomes possible to calculate the relative $ORCA_{F|A}$ of the follower with respect to that particular agent. The set of velocities for collision avoidance of one follower relative to all agents in the environment can be defined by the following formula:

$$ORCA_F^\tau = D(0, v_{\max}^i) \cap \left(\bigcap_{A \neq F} ORCA_{F|A}^\tau \right) \tag{19}$$

where $D(0, v_{\max}^i) = \{v \| \|v - 0\| < v_{\max}^i \}$ represents the set of allowable speeds. To ensure the safety of the formation in collision avoidance, we define v_F^{pref} as the preferred speed, then the expected speed can be obtained:

$$v_F^{new} = \underset{v \in ORCA_F^\tau}{\operatorname{argmin}} \left\| v - v_F^{\text{pref}} \right\| \tag{20}$$

When the distance between the obstacle and any follower falls below the emergency threshold S_2, it indicates a need to suspend formation maintenance and initiate autonomous collision avoidance. The speed limit for followers is lifted. Speed rewards can be quantified as follows:

$$R_v^F = \begin{cases} 0, & d_m^F < S_2 \\ 1, & d_m^F \geqslant S_2, v_v^1 < v_F^{new} < v_v^2 \\ -\lambda_{v1}^F \cdot |v_F^{new} - v| & d_m^F \geqslant S_2, v_F^{new} > v_v^2 \\ -\lambda_{v2}^F \cdot |v_F^{new} - v| & d_m^F \geqslant S_2, v_F^{new} < v_v^1 \end{cases} \tag{21}$$

where, v_v^1 is the speed tolerance limit, v_v^2 is the maximum speed limit, these two values are within the range of actual speed ±0.5 m/s, λ_v^F

represents the speed reward weight.

The collision avoidance reward effectively mitigates the risk of followers colliding with obstacles. When $d_m^F < S_2$, the calculation incorporates the collision avoidance reward. At this point, both the speed reward and formation error reward are set to 0, while the penalty increases as the obstacle gets closer to followers. This can be expressed as follows:

$$R_c^F = \begin{cases} -\lambda_c^F \cdot d_m^F, & 0 < d_m^F < S_2 \\ -1000, & collision \end{cases} \tag{22}$$

where, λ_c^F represents the collision avoidance reward weight. Finally, the follower's compound reward can be expressed as:

$$R^F = R_e^F + R_v^F + R_c^F \tag{23}$$

Remark 3. *The mainline and auxiliary rewards are both incorporated into the follower reward function. In this study, the collision avoidance reward is defined as the mainline reward. Also, to avoid sparse returns, we also introduce auxiliary reward functions, including formation error reward and speed reward. This means that it is imperative for USVs formation to acquire the collision avoidance ability prior to advancing their skills in formation maintenance. After a certain number of training episodes, the USVs gradually gained a clearer understanding of its mainline task (maintain formation and reach target position safely). However, relying solely on mainline rewards for accomplishing the mainline task yields a relatively low success rate and increases the risk of getting trapped in local optima. Consequently, the USV will explore new strategies to enhance the likelihood of achieving the mainline task through learning auxiliary rewards.*

Remark 4. *The formation error reward is defined as a function of the discrepancy between the expected and actual positions of followers, with a penalty that proportionally increases in magnitude as the error grows. When the error is kept within an acceptable range, a continuous positive reward will be obtained to encourage followers to maintain the formation. It is implied that when the actual formation error falls within the range of d_e^1 and d_e^2, it will consistently receive a positive reward of $+3$; otherwise, it will incur a penalty of -10. To optimize the reward value, it is imperative to ensure a stable formation of the USVs within predefined boundaries, thereby ensuring sustained formation stability. Similarly, we have introduced the upper and lower speed threshold into the speed reward function design to ensure that the USVs formation maintains a reasonable velocity range. To maintain formation stability and allow for greater flexibility in adjusting each member's speed, we strive to minimize fluctuations in the formation error during the normal sailing, resulting in a relatively modest continuous positive reward value.*

4. Experiment

4.1. Design of simulation

The experimental platform is configured as follows: Intel Core i9-11900H CPU, Nvidia RTX3090 GPU, ubuntu 20.04, ROS version noetic. Most DRL-based approaches are primarily developed and evaluated on simulation platforms, rather than undergoing real-world testing, to mitigate the exorbitant costs associated with trial and error. Therefore, we employed the Gazebo platform (Meng, Liu, Bucknall, Guo, & Ji, 2022; Zhou et al., 2019), which is based on a physics engine, to validate the aforementioned research. The neural network training process on this platform typically requires approximately 10–12 h for 10,000 episodes. The algorithm parameters are shown in Table 1, and the specific parameters of USV ship model are shown in Table 2. During the UFND model training process, we employed a static water environment to ensure the robustness of the trained model. In the verification simulation, the parameters of wind and wave interference are set as follows: 35° for wave and wind direction; 0.18 m for significant wave height; 8 s for wave period; 0.33 m/s for mean wind speed. The detailed wave and wind model settings can be found in (Bingham, Aguero, Mccarrin, Klamo, & Waqar, 2019).

4.2. UFND model training

In this study, the average reward value, navigation time and training success rate were used to analyze the training results of the UFND model. Among them, the algorithm's learning ability is stronger when the average reward curve converges faster; a shorter navigation time indicates a better decision path for the algorithm; the training success rate represents the model's probability of achieving the task during training.

The training results of the virtual leader are illustrated in Fig. 7. The training comparative experiment encompasses conventional strategy learning algorithms (SAC and DDPG). The DDPG algorithm evidently failed to converge after 10,000 episodes, resulting in subpar performance in terms of navigation time and training success rate. The SAC algorithm addresses the challenge of solving the Boltzmann distribution of value function in continuous space by employing a maximum entropy strategy, achieving a success rate approaching 100 % after training with 6000 episodes. The average reward value of G-PPO (PPO trained with GRU network) converges the fastest, and the peak value is higher than that of SAC algorithm. The GRU model enhances the memory of historical obstacle status in the presence of security threats, enabling the PPO algorithm to acquire collision avoidance skills more efficiently. Consequently, it exhibits superior performance in terms of navigation time and training success rate compared to the preceding two methods.

The training results for formations with 2, 3, 4, and 6 followers are

Table 2
Principal dimensions of the USV.

Parameters	Value
Length (m)	4.85
Beam (m)	2.44
Draft (m)	6.5
Height (m)	1.27
Full Displacement Weight (kg)	374
Maximum speed (kn)	9.72
Rudder area (m^2)	0.83
Max rudder angle (deg)	25
Propulsion (kw)	4.3
Aspect ratio of rudder	1.8452

Table 1
Experimental parameters & settings.

Parameters of MHA-MAPPO Algorithm and reward function								
Parameters	**Symbol**	**Value**	**Parameters**	**Symbol**	**Value**	**Parameters**	**Symbol**	**Value**
Critic learning rate	c_r	0.0003	Batch size	B	2000	Lidar distance	R_{lidar}	60.0 m
Actor learning rate	a_r	0.0003	GRU hidden state	h_{GRU}	128	Safe distance	S_1	50.4 m
Discounted rate	γ	0.95	Formation error weight	λ_e^F	1.88	Emergency distance	S_2	20.7 m
Multi head number	h	4	tolerance error limit	d_e^1	0.5	collision weight	λ_c^F	1.04
Replay memory size	N_R	50,000	maximum error limit	d_e^2	2	speed weight (1)	λ_{v1}^F	0.77
Safety weight (1)	λ_{s1}	1.04	Safety weight (2)	λ_{s2}	1.62	speed weight (2)	λ_{v2}^F	1.25

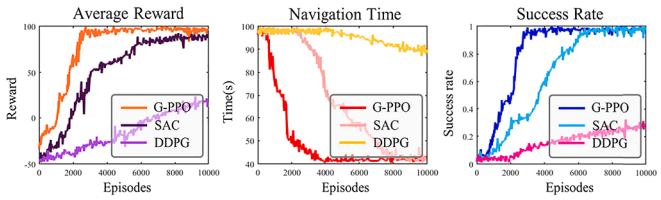


Fig. 7. Comparison of follower training results based on G-PPO, SAC, and DDPG algorithms. The metrics include success rate (higher is better), navigation time (lower is better), and average reward (higher is better).

respectively depicted in Fig. 8 (a), (b), (c), and (d). The training difficulty gradually increases and the convergence speed of the curve decreases as the number of followers grows. In the evaluation of various indicators during 10,000 episodes, it is evident that the MHA-MAPPO outperforms MAPPO. Particularly in complex training scenarios Fig. 8 (d), the UFND model trained by MHA-MAPPO is capable of maintaining a success rate that approaches 100 %. This demonstrates the significant performance of the decentralized Critic's MHA mechanism in training the UFND model. Each follower can independently query observation information from other followers, enabling them to consider the status of their peers when making decision and facilitating better cooperation among multiple agents.

4.3. Training result analysis

The virtual leader strategy experiments involving generalization and comparison encompass both conventional DRL (SAC and DDPG) and classical collision avoidance algorithms (APF and DWA). To ensure the test's fairness, all DRL algorithms are trained in an identical environment until model convergence. Moreover, algorithm parameters undergo individualized processing to maximize their decision-making efficacy. The success rate and sailing time of different algorithms are depicted in Fig. 9(a).

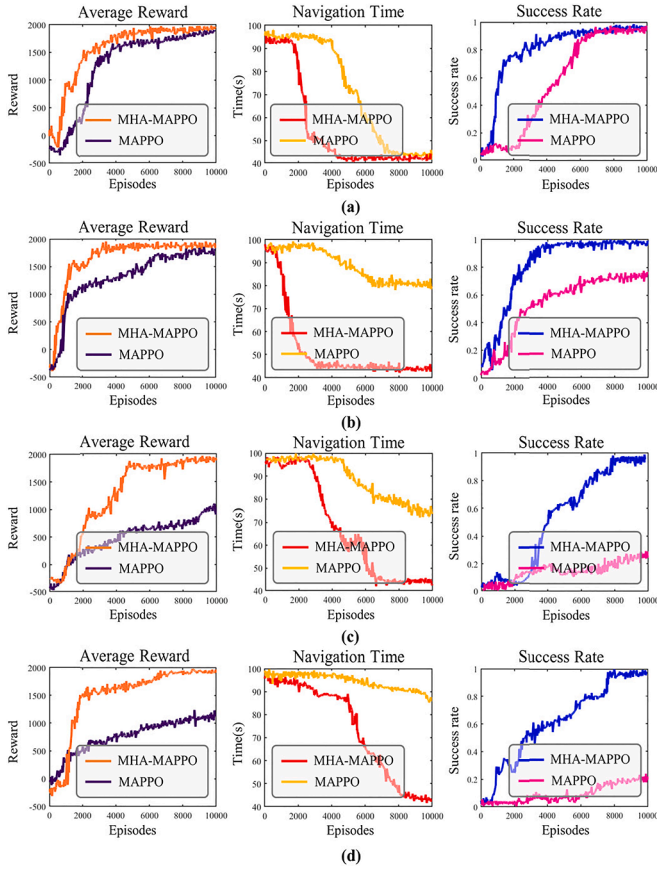


Fig. 8. Comparison of follower training results based on MHA-MAPPO and MAPPO algorithms. The training results for formations with 2, 3, 4, and 6 followers are respectively depicted in (a), (b), (c), and (d). The metrics include success rate (higher is better), navigation time (lower is better), and average reward (higher is better).

The followers of formation generalization experiment comprise the UFDN model trained under the four conditions illustrated in Fig. 8, conducted in an environment with randomly dynamic obstacles ranging from 1 to 10, as well as the experiment involving various algorithms when there are 6 followers. Similar to the former, there are no restrictions on the maximum number of episodes in the MADRL. Among them, F-ORCA and F-APF employ the leader–follower formation control to maintain a desired formation while utilizing the ORCA and APF collision avoidance algorithms respectively to ensure collision-free navigation in the presence of obstacles. The actual output speed of F-ORCA is determined by utilizing the preferred speed derived from formulas 19 and 20. About the F-APF algorithm, it is necessary to consider the influence of obstacles or other agents on the repulsive force generated by the current agent, and design different control models for the leader and the followers:

$$R_j(k) = \sum_{l=1}^{M} \alpha(x_j(k) - x_{ob}^l) \tag{24}$$

$$\begin{cases} u_N(k) = m + kD(k) + \sum_{i \in N_i} a_{N_i} r_{N_i}(k) + \beta \cdot R_N(k) \\ u_i(k) = \varepsilon \sum_{j \in N_i} a_{ij}(x_j(k) - x_i(k) - r_{ij}(k)) + \beta \cdot R_i(k) \end{cases} \tag{25}$$

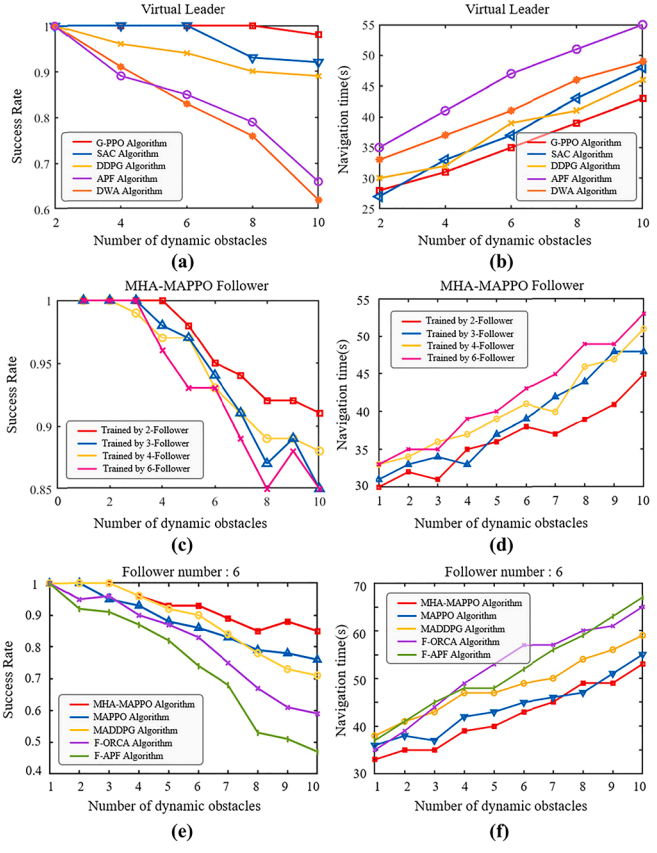


Fig. 9. Comparison of different method with the trained UFND model. (a, b) shown the comparison of collision avoidance ability of virtual leader; (c, d) shown the generalization comparison of four groups of DNN models trained in Fig. 8; (e, f) shown the comparison of different follower method.

This feature enables the agent to effectively avoid obstacles while maintaining the original formation to its fullest extent, thereby ensuring collision avoidance safety. The traditional algorithm for formation is not addressed in this study. The comparison between the success rate of the follower algorithm and the navigation time is illustrated in Fig. 9(b) and Fig. 9(c). Based on data analysis, the following conclusions can be inferred:

1) Regardless of whether it is the virtual leader or follower strategy, the increased number of obstacles within the environment leads to a decline in generalization success rate and an increased navigation time consumption. This phenomenon is commonly observed across all algorithms.
2) The G-PPO algorithm enables virtual leader to maintain a success rate of 98 % in the presence of 10 dynamic obstacles (Fig. 9a). Despite a slight decline, this method still outperforms other algorithms. While its navigation time may not be optimal in environments with fewer obstacles, it progressively demonstrates its efficacy as the number of obstacles increases (Fig. 9b).
3) When comparing the four groups of UFND models trained in Fig. 8, it is evident that an increased formation members lead to a gradual decline in generalization success rate (Fig. 9c), accompanied by an increased navigation time (Fig. 9d).
4) When the number of formation members is fixed, the MHA-MAPPO algorithm exhibits excellent performance across all indicators (Fig. 9e, Fig. 9f). However, when the number of obstacles in the F-APF algorithm exceeds five, the success rate noticeably decreases due to its susceptibility to "zero potential energy points" and local optima in complex environments. Moreover, the collision avoidance effectiveness of F-ORCA is unsatisfactory due to limitations in its observation space.

4.4. Verification experiment

In this section, the reliability of the proposed UFND model for USV formation construction and navigation decision making is verified through three sets of simulations and one set of real water experiments.

Fig. 10 illustrates the process of constructing a hexagonal formation with six USVs. The six USVs, starting from diverse initial positions and directions, formed a hexagonal configuration with the virtual leader at its center as depicted in Fig. 10(a). The followers in different positions can effectively navigate through static obstacles in the environment, enabling them to successfully reach their expected positions. The speed and rudder angle changes of six USVs are depicted in Fig. 10(b) and, Fig. 10(c), representing the output of the MHA-MAPPO algorithm. The USVs exhibit a proactive deceleration behavior in anticipation of their arrival, facilitated by the speed reward and formation error reward mechanisms. Consequently, each USV gradually reduces its velocity to 0 m/s upon reaching the expected positions. The variation process of formation error, depicted in Fig. 10(d), corresponds to the temporal evolution of the distance between different USVs and the expected position. Upon reaching the desired position, the formation error is reduced to 0 m.

The configuration of 3-USV in quiescent waters and their maintained navigation trajectory are illustrated in Fig. 11(a). The 3-USV initiated their trajectories from distinct initial positions and established a geometrically precise triangular formation around T=6.5(s), subsequently commencing deceleration. As depicted in Fig. 11(b), when the leader accelerates, followers adjust their strategy accordingly by accelerating and maintaining a speed of 4 m/s to uphold the formation. At T=17.6(s), Follower-03 initially detects the obstacle ahead and promptly opts to deviate from the formation constraint by executing a port alteration to evade the obstacle. Subsequently, Follower-02 and Follower-01 adopt the strategies of accelerating to execute a port alteration and decelerating to perform a starboard alteration respectively to evade the obstacle, as illustrated in Fig. 11(c). The three USVs

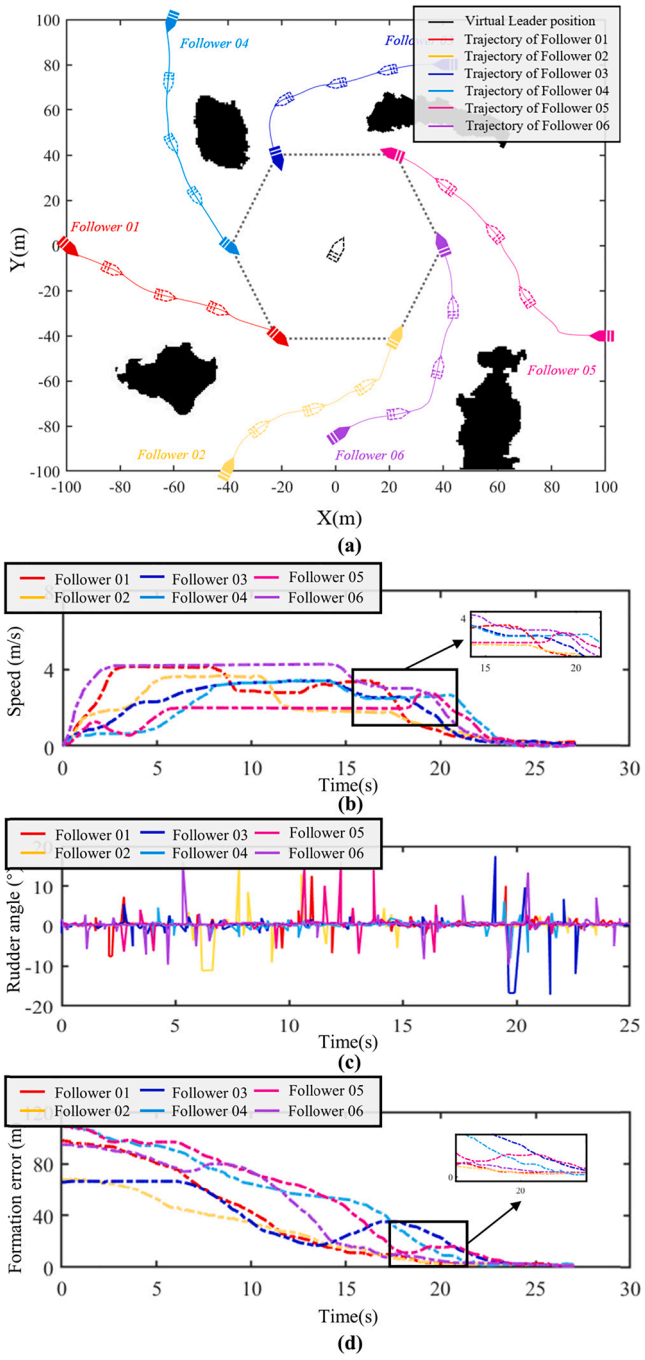


Fig. 10. Hexagonal formation with six USVs. (a) Formation construction trajectory of six USVs. (b) Speeds variation of formation. (c) Rudder angle variation of formation. (d) Follower error variation of formation.

were released from the obstruction and reassembled at T=38.3(s). The final formation decelerates to 0 m at T=45(s) and successfully reaches the destination. The formation error during navigation is shown in Fig. 11(d).

The collision avoidance effectiveness and sailing trajectory of a 6-USV formation in a dynamic environment are depicted in Fig. 12(a). In this setting, there exist 5 non-autonomous vessels that follow predetermined paths at constant velocities without the ability to avoid collisions. During the initial phase of collision avoidance, the 6 followers gradually accelerate and adjust their speed to maintain a stable formation at 4 m/s, as depicted in Fig. 12(b). At T=16.6(s), Follower-02 and Follower-01 detect TS-02 via lidar and opt to disassemble the formation

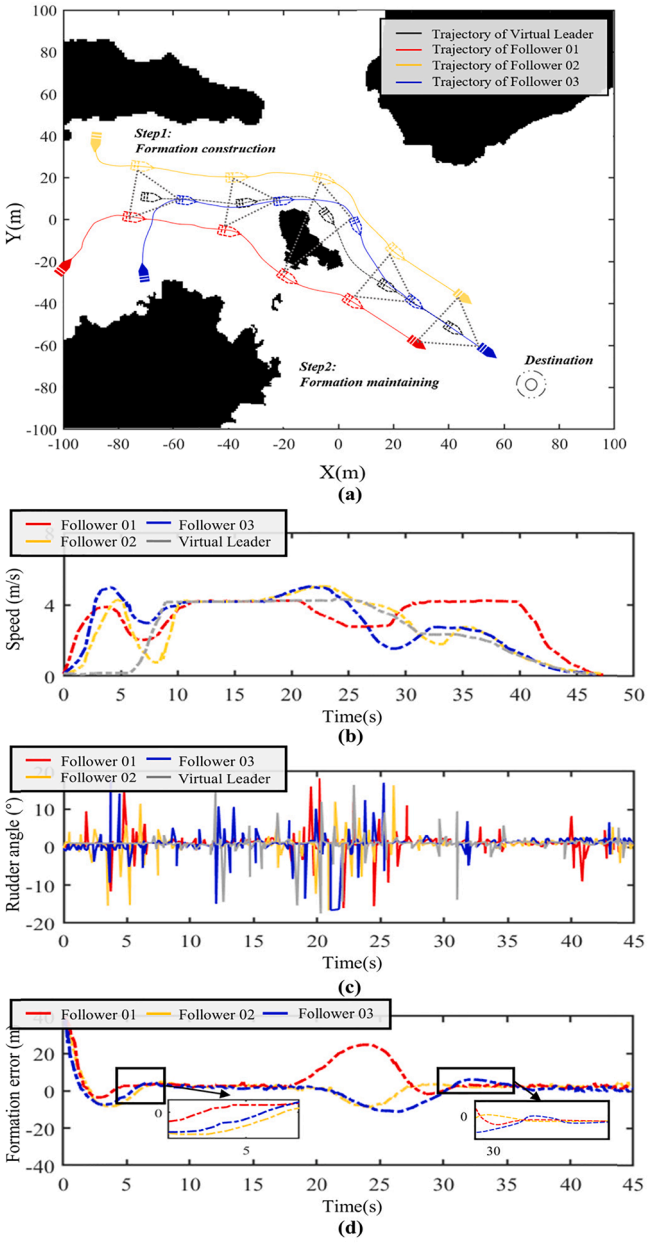


Fig. 11. Formation construction and static obstacle avoidance with three USVs (a) Formation navigation decision-making trajectory of three USVs. (b) Speeds variation of formation. (c) Rudder angle variation of formation. (d) Follower error variation of formation.

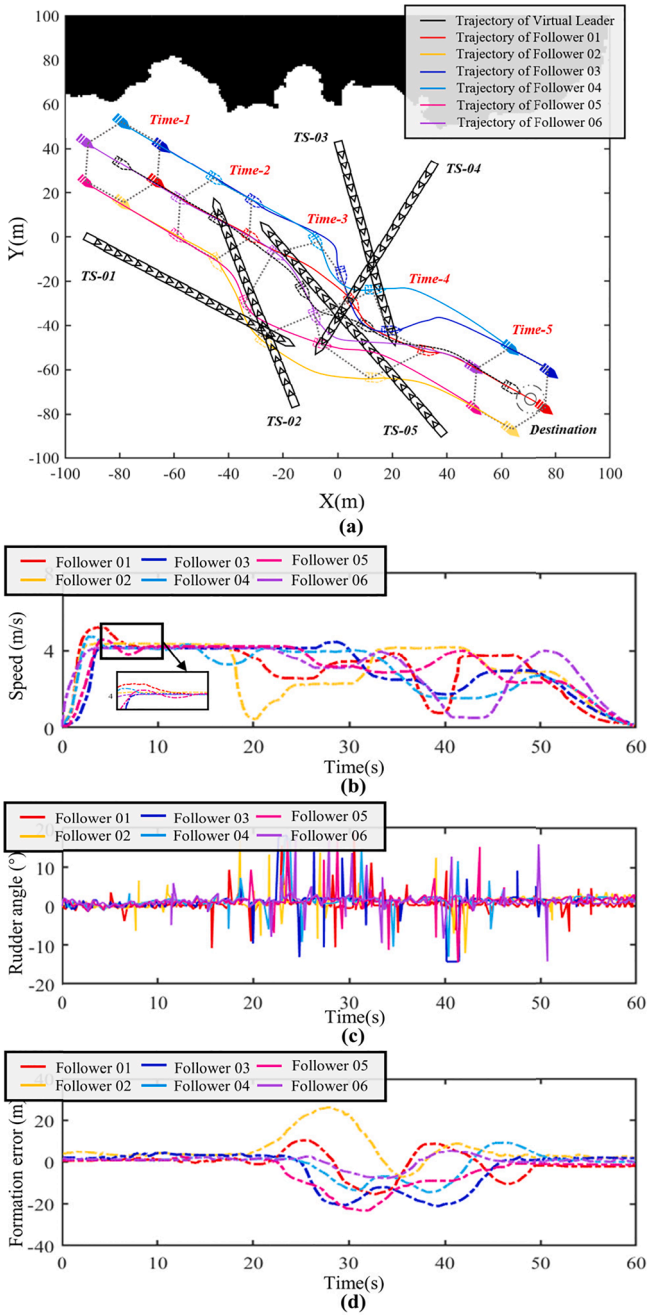


Fig. 12. Formation dynamic collision avoidance with six USVs (a) Formation collision avoidance decision-making trajectory of six USVs. (b) Speeds variation of formation. (c) Rudder angle variation of formation. (d) Follower error variation of formation.

constraint, subsequently executing a starboard alteration for evasion. At this juncture, the formation members are no longer bound by the formation and autonomously make collision avoidance decision-making with target ships. The variation in rudder angle during this process is depicted in Fig. 12(c). Finally, at T=52.4(s), the 6-USV formation will converge and re-establish its configuration. The formation error during navigation is shown in Fig. 12(d).

Finally, the UFND model's efficacy in collision avoidance was evaluated by employing a dataset of authentic sailing data obtained from the Panama Canal on March 2, 2024. Among them, the trajectories of TS-01, TS-02, TS-03, TS-04, and TS-05 represent actual ship sailing paths, with detailed ship information provided in Table 3. Task assumption: The formation starts from the position in Fig. 13(a), and the expected destination in Fig. 13(h); The maximum speed remains 5 m/s (9.72kn). During the experiment, the triangular formation consisting of 3-USV effectively avoids obstacles encountered with real vessels and successfully reaches the designated target. Notably, the UFND model successfully eliminated the USV formation restrictions at the entrance to the Panama Canal and reestablished formation after navigating through narrow waters (Fig. 13e - Fig. 13h). The detailed data of formation collision avoidance are shown in Table 4.

The above experimental results objectively demonstrate that the UFND model exhibits excellent performance in addressing challenges related to USVs formation construction and navigation decision-making, thereby offering a reliable theoretical foundation for future marine formation auxiliary decision-making systems.

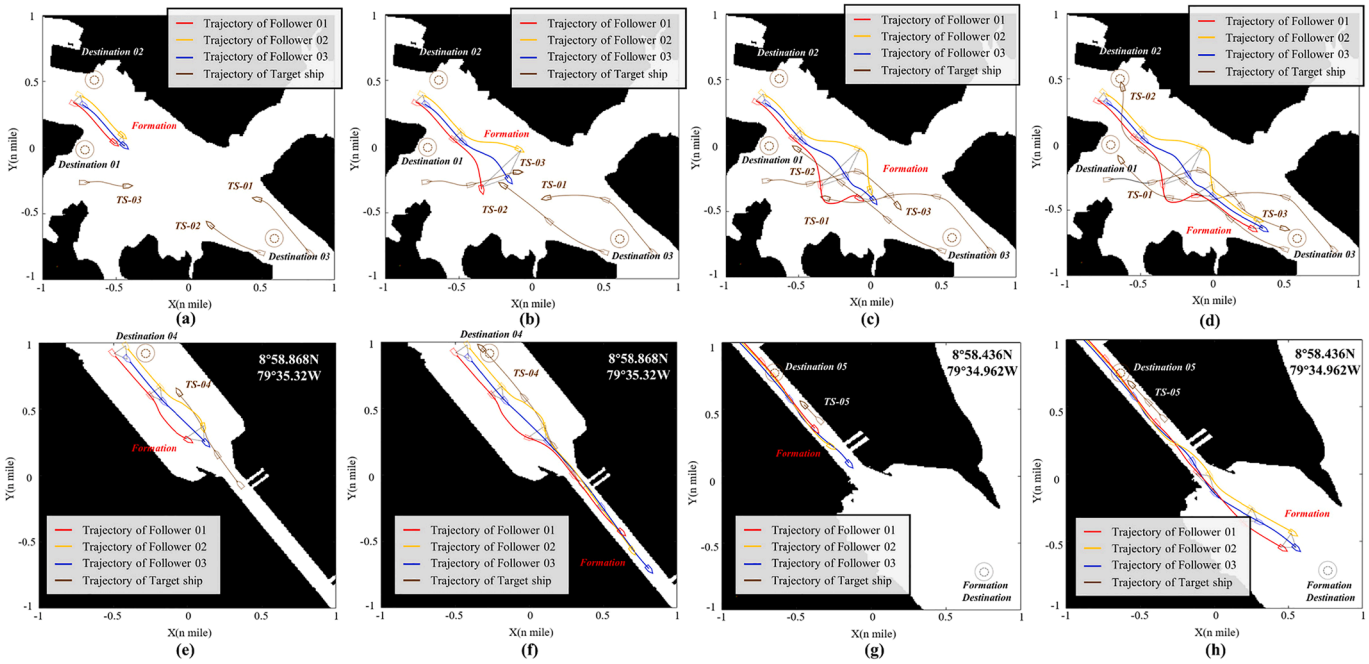


Fig. 13. 3-USV formation navigation decision-making process in the Panama Canal.

Table 3
Ship information in the Panama Canal.

Time	Ship	call sign	MMSI	IMO	category	destination	Length (m)	Beam (m)	Draft (m)	Speed (kn)	Initial course (°)
1916–1924	TS-01	LAUTAU	372****00	–	Pilot vessel	E WA	15.4	5.08	–	5.61	298.1
	TS-02	TUGCS	351****00	–	Tug	CRISTOBAL, PA	26.4	13.0	5.7	6.26	83.5
	TS-03	TUGPD	357****00	996**00	Harbor tender	MIRAFLORES	27.1	14.0	6.2	7.39	276.7
1929–1936	TS-04	9HA3462	229****00	960**79	Container ship	PAPTY	300.4	48.3	14.2	4.63	317.2
	TS-05	TUGCP	372****00	996**86	Tug	COLONA	28.8	14.6	6.0	1.31	318.4

Table 4
Formation collision avoidance decision-making details date.

	minimum distance (n mile)					Encounter angle (°)					Maximum speed (kn)	Minimum speed (kn)
	TS-01	TS-02	TS-03	TS-04	Ts-05	TS-01	TS-02	TS-03	TS-04	TS-05		
Follower-01	0.32	0.37	0.43	0.27	0.09	332.5	294.7	104.1	1.7	0.9	7.79	0.32
Follower-02	0.45	0.43	0.21	0.11	0.10	278.0	3.9	121.4	358.2	358.3	7.78	0.41
Follower-03	0.29	0.13	0.26	0.17	0.07	311.9	2.8	116.5	355.4	359.7	7.79	0.37

5. Conclusion

This study proposes a USV formation navigation decision-making method through HDRL using self-attention mechanism. This method takes into full consideration the motion characteristics of USVs and the constraints imposed by formation. The UFND model incorporates HDRL architectures to independently design the virtual leader and followers, thereby guiding the controlled objects towards accomplishing the target task through a compound reward function. The algorithm design incorporates a decentralized Critic strategy with an MHA mechanism to enhance the convergence speed of the MAPPO algorithm, while introducing a GRU network structure to bolster the agent's assimilation of historical data. The proposed UFND model demonstrates exceptional performance in terms of success rate, navigation time, and convergence speed. Furthermore, when subjected to real sea state of the Panama Canal, formations constructed using this method exhibit remarkable flexibility in collision avoidance decisions and formation reconstruction.

In future work, it is imperative to devise a novel interactive approach for the practical implementation of the UFND model trained by this algorithm on actual vessels, ensuring collision avoidance with real obstacles. This will furnish substantial theoretical underpinning for the investigation of intelligent USVs formation-assisted navigation.

CRediT authorship contribution statement

Zhewen Cui: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing – original draft, Writing – review & editing. **Wei Guan:** Conceptualization, Methodology, Validation, Writing – review & editing, Supervision, Project administration, Funding acquisition. **Xianku Zhang:** Validation, Supervision, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgement

The paper is partially supported by National Natural Science Foundation of China (NO.51409033, NO.52171342), Dalian lnnovation Team Support Plan in the Key Research Field (NO. 2020RT08), 2023 DMU navigation college first-class interdisciplinary research project (NO. 2023XA03), and the Fundamental Research Funds for the Central Universities (NO. 3132023502). The authors would like to thank the anonymous reviews for their valuable comments.

References

Alonso-Mora, J., Breitenmoser, A., Beardsley, P., Siegwart, R., & Ieee. (2012, May 14-18). *Reciprocal Collision Avoidance for Multiple Car-like Robots.* Paper presented at the IEEE International Conference on Robotics and Automation (ICRA), St Paul, MN.

Auh, E., Kim, J., Joo, Y., Park, J., Lee, G., Oh, I., & Moon, H. (2024). Unloading sequence planning for autonomous robotic container-unloading system using A-star search algorithm. *Engineering Science and Technology-an International Journal-Jestech, 50*. https://doi.org/10.1016/j.jestch.2023.101610

Banino, A., Badia, A. P., Walker, J., Scholtes, T., & Blundell, C. (2021). CoBERL: Contrastive BERT for Reinforcement Learning.

Benzerrouk, A., Adouane, L., & Martinet, P. (2014). Stable navigation in formation for a multi-robot system based on a constrained virtual structure. *Robotics and Autonomous Systems, 62*(12), 1806–1815. https://doi.org/10.1016/j.robot.2014.07.004

Bingham, B., Aguero, C., Mccarrin, M., Klamo, J., & Waqar, R. (2019). Toward Maritime Robotic Simulation in Gazebo. *Paper presented at the OCEANS 2019 MTS/IEEE SEATTLE*.

Cai, H., & Hu, G. Q. (2017). Distributed Tracking Control of an Interconnected Leader-Follower Multiagent System. *IEEE Transactions on Automatic Control, 62*(7), 3494–3501. https://doi.org/10.1109/tac.2017.2660298

Chang, L., Shan, L., Zhang, W. L., & Dai, Y. W. (2023). Hierarchical multi-robot navigation and formation in unknown environments via deep reinforcement learning and distributed optimization. *Robotics and Computer-Integrated Manufacturing, 83*. https://doi.org/10.1016/j.rcim.2023.102570

Cheng, W. L., Zhang, K., & Jiang, B. (2023). Fixed-Time Fault-Tolerant Formation Control for a Cooperative Heterogeneous Multiagent System With Prescribed Performance. *Ieee Transactions on Systems Man Cybernetics-Systems, 53*(1), 462–474. https://doi.org/10.1109/tsmc.2022.3186382

Consolini, L., Morbidi, F., Prattichizzo, D., & Tosques, M. (2008). Leader-follower formation control of nonholonomic mobile robots with input constraints. *Automatica, 44*(5), 1343–1349. https://doi.org/10.1016/j.automatica.2007.09.019

Cui, R., Li, Y., & Yan, W. (2016). Mutual Information-Based Multi-AUV Path Planning for Scalar Field Sampling Using Multidimensional RRT*. *Ieee Transactions on Systems Man Cybernetics-Systems, 46*(7), 993–1004. https://doi.org/10.1109/tsmc.2015.2500027

Cui, Z., Guan, W., Luo, W., & Zhang, X. (2023). Intelligent navigation method for multiple marine autonomous surface ships based on improved PPO algorithm. *Ocean Engineering, 287*, Article 115783. https://doi.org/10.1016/j.oceaneng.2023.115783

Cui, Z. W., Guan, W., Zhang, X. K., & Zhang, C. (2023). Autonomous Navigation Decision-Making Method for a Smart Marine Surface Vessel Based on an Improved Soft Actor-Critic Algorithm. *Journal of Marine Science and Engineering, 11*(8). https://doi.org/10.3390/jmse11081554

Dai, S. L., He, S. D., Chen, X., & Jin, X. (2020). Adaptive Leader-Follower Formation Control of Nonholonomic Mobile Robots With Prescribed Transient and Steady-State Performance. *Ieee Transactions on Industrial Informatics, 16*(6), 3662–3671. https://doi.org/10.1109/tii.2019.2939263

de Souza, C., Newbury, R., Cosgun, A., Castillo, P., Vidolov, B., & Kulic, D. (2021). Decentralized Multi-Agent Pursuit Using Deep Reinforcement Learning. *Ieee Robotics and Automation Letters, 6*(3), 4552–4559. https://doi.org/10.1109/lra.2021.3068952

Fossen, T. I. (2011). *Handbook of Marine Craft Hydrodynamics and Motion Control*: Handbook of Marine Craft Hydrodynamics and Motion Control.

Ghommam, J., & Saad, M. (2018). Adaptive Leader-Follower Formation Control of Underactuated Surface Vessels Under Asymmetric Range and Bearing Constraints. *IEEE Transactions on Vehicular Technology, 67*(2), 852–865. https://doi.org/10.1109/tvt.2017.2760367

Guan, W., & Wang, K. (2023). Autonomous Collision Avoidance of Unmanned Surface Vehicles Based on Improved A-Star and Dynamic Window Approach Algorithms. *Ieee Intelligent Transportation Systems Magazine*. https://doi.org/10.1109/mits.2022.3229109

Han, R. H., Chen, S. D., Wang, S. J., Zhang, Z. Q., Gao, R., Hao, Q., & Pan, J. (2022). Reinforcement Learned Distributed Multi-Robot Navigation With Reciprocal Velocity Obstacle Shaped Rewards. *Ieee Robotics and Automation Letters, 7*(3), 5896–5903. https://doi.org/10.1109/lra.2022.3161699

Han, S., Wang, L., Wang, Y. T., & He, H. C. (2022). A dynamically hybrid path planning for unmanned surface vehicles based on non-uniform Theta* and improved dynamic windows approach. *Ocean Engineering, 257*. https://doi.org/10.1016/j.oceaneng.2022.111655

Hasselt, H. V., Guez, A., & Silver, D. (2015). *Deep Reinforcement Learning with Double Q-learning.* Computer Science.

He, Y., Wang, Y. H., Yu, F. R., Lin, Q. Z., Li, J. Q., & Leung, V. C. M. (2022). Efficient Resource Allocation for Multi-Beam Satellite-Terrestrial Vehicular Networks: A Multi-Agent Actor-Critic Method With Attention Mechanism. *IEEE Transactions on Intelligent Transportation Systems, 23*(3), 2727–2738. https://doi.org/10.1109/tits.2021.3128209

Iqbal, S., & Sha, F. (2018). Actor-Attention-Critic for Multi-Agent Reinforcement Learning.

Khodamipour, G., Khorashadizadeh, S., & Farshad, M. (2023). Adaptive formation control of leader-follower mobile robots using reinforcement learning and the Fourier series expansion. *ISA Transactions, 138*, 63–73. https://doi.org/10.1016/j.isatra.2023.03.009

Ladosz, P., Weng, L. L., Kim, M., & Oh, H. (2022). Exploration in deep reinforcement learning: A survey. *Information Fusion, 85*, 1–22. https://doi.org/10.1016/j.inffus.2022.03.003

Li, D., Zhao, D. B., Zhang, Q. C., & Chen, Y. R. (2019). Reinforcement Learning and Deep Learning Based Lateral Control for Autonomous Driving. *Ieee Computational Intelligence Magazine, 14*(2), 83–98. https://doi.org/10.1109/mci.2019.2901089

Li, G. F., Qiu, Y. F., Yang, Y. F., Li, Z. N., Li, S., Chu, W. B., & Li, S. E. (2023). Lane Change Strategies for Autonomous Vehicles: A Deep Reinforcement Learning Approach Based on Transformer. *Ieee Transactions on Intelligent Vehicles, 8*(3), 2197–2211. https://doi.org/10.1109/tiv.2022.3227921

Li, S., Wang, Y. H., Ma, X. J., Yang, M., Yang, S. Q., & Niu, W. D. (2023). A method based on virtual hinges for multi-underwater glider formation. *Ocean Engineering, 286*. https://doi.org/10.1016/j.oceaneng.2023.115565

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., . . . Wierstra, D. (2015). Continuous control with deep reinforcement learning. *Computer Science.*.

Lyu, H., & Yin, Y. (2019). COLREGS-Constrained Real-time Path Planning for Autonomous Ships Using Modified Artificial Potential Fields. *Journal of Navigation, 72*(3), 588–608. https://doi.org/10.1017/s0373463318000796

Mehdifar, F., Bechlioulis, C. P., Hendrickx, J. M., & Dimarogonas, D. V. (2023). 2-D Directed Formation Control Based on Bipolar Coordinates. *Ieee Transactions on Automatic Control, 68*(7), 4175–4190. https://doi.org/10.1109/tac.2022.3206603

Meng, J. W., Liu, Y. C., Bucknall, R., Guo, W. H., & Ji, Z. (2022). Anisotropic GPMP2: A Fast Continuous-Time Gaussian Processes Based Motion Planner for Unmanned Surface Vehicles in Environments With Ocean Currents. *IEEE Transactions on Automation Science and Engineering, 19*(4), 3914–3931. https://doi.org/10.1109/tase.2021.3139163

Morris, S. D., Kumar, V. A., Biswas, R., & Mohan, C. G. (2024). Identification of a Staphylococcus aureus amidase catalytic domain inhibitor to prevent biofilm formation by sequential virtual screening, molecular dynamics simulation and biological evaluation. *International Journal of Biological Macromolecules, 254*. https://doi.org/10.1016/j.ijbiomac.2023.127842

Ning, Z., Ou, D. X., Xie, C., Zhang, L., Gao, B. W., & He, J. F. (2023). Optimal convoy composition for virtual coupling trains at junctions: A coalition formation game approach. *Transportation Research Part C-Emerging Technologies, 154*. https://doi.org/10.1016/j.trc.2023.104277

Park, B. S., & Yoo, S. J. (2019). Adaptive-observer-based formation tracking of networked uncertain underactuated surface vessels with connectivity preservation and collision avoidance. *Journal of the Franklin Institute-Engineering and Applied Mathematics, 356*(15), 7947–7966. https://doi.org/10.1016/j.jfranklin.2019.04.017

Pu, Z. Q., Zhang, T. L., Ai, X. L., Qiu, T. H., & Yi, J. Q. (2023). A Deep Reinforcement Learning Approach Combined With Model-Based Paradigms for Multiagent Formation Control With Collision Avoidance. *IEEE Transactions on Systems Man Cybernetics-Systems, 53*(7), 4189–4204. https://doi.org/10.1109/tsmc.2023.3241337

Rezaee, H., Parisini, T., & Polycarpou, M. M. (2021). Resiliency in dynamic leader-follower multiagent systems. *Automatica, 125*. https://doi.org/10.1016/j.automatica.2020.109384

Sang, H. Q., You, Y. S., Sun, X. J., Zhou, Y., & Liu, F. (2021). The hybrid path planning algorithm based on improved A* and artificial potential field for unmanned surface vehicle formations. *Ocean Engineering, 223*. https://doi.org/10.1016/j.oceaneng.2021.108709

Sawada, R., Sato, K., & Majima, T. (2021). Automatic ship collision avoidance using deep reinforcement learning with LSTM in continuous action spaces. *Journal of Marine Science and Technology, 26*(2), 509–524. https://doi.org/10.1007/s00773-020-00755-0

Schulman, J., Moritz, P., Levine, S., Jordan, M., & Abbeel, P. (2015). High-Dimensional Continuous Control Using Generalized Advantage Estimation. *Computer Science.*.

Shen, H., Hashimoto, H., Matsuda, A., Taniguchi, Y., Terada, D., & Guo, C. (2019). Automatic collision avoidance of multiple ships based on deep Q-learning. *Applied Ocean Research, 86*, 268–288. https://doi.org/10.1016/j.apor.2019.02.020

Su, Y. H., Bhowmick, P., & Lanzon, A. (2023). A robust adaptive formation control methodology for networked multi-UAV systems with applications to cooperative payload transportation. *Control Engineering Practice, 138*. https://doi.org/10.1016/j.conengprac.2023.105608

Sui, Z. Z., Pu, Z. Q., Yi, J. Q., & Wu, S. G. (2021). Formation Control With Collision Avoidance Through Deep Reinforcement Learning Using Model-Guided Demonstration. *IEEE Transactions on Neural Networks and Learning Systems, 32*(6), 2358–2372. https://doi.org/10.1109/tnnls.2020.3004893

Thuyen, N. A., Thanh, P. N. N., & Anh, H. P. H. (2023). Adaptive finite-time leader-follower formation control for multiple AUVs regarding uncertain dynamics and

disturbances. *Ocean Engineering, 269*. https://doi.org/10.1016/j.oceaneng.2022.113503

Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., . . . Georgiev, P. (2019). Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature* (7782).

Wu, H. G., & Gao, Y. L. (2023). An ant colony optimization based on local search for the vehicle routing problem with simultaneous pickup-delivery and time window. *Applied Soft Computing, 139*. https://doi.org/10.1016/j.asoc.2023.110203

Wu, X., Chen, H. L., Chen, C. G., Zhong, M. Y., Xie, S. R., Guo, Y. K., & Fujita, H. (2020). The autonomous navigation and obstacle avoidance for USVs with ANOA deep reinforcement learning method. *Knowledge-Based Systems, 196*. https://doi.org/10.1016/j.knosys.2019.105201

Yang, G., Zhou, F. R., Ma, Y., Yu, Z. Q., Zhang, Y. J., & He, J. L. (2018). Identifying Lightning Channel-Base Current Function Parameters by Powell Particle Swarm Optimization Method. *IEEE Transactions on Electromagnetic Compatibility, 60*(1), 182–187. https://doi.org/10.1109/temc.2017.2705485

Yu, C., Velu, A., Vinitsky, E., Wang, Y., & Wu, Y. (2021). The Surprising Effectiveness of MAPPO in Cooperative, Multi-Agent Games.

Yuan, C. Z., He, H. B., & Wang, C. (2019). Cooperative Deterministic Learning-Based Formation Control for a Group of Nonlinear Uncertain Mechanical Systems. *IEEE Transactions on Industrial Informatics, 15*(1), 319–333. https://doi.org/10.1109/tii.2018.2792455

Zhang, G. Q., Yu, W., Li, J. Q., & Zhang, X. K. (2021). A novel event-triggered robust neural formation control for USVs with the optimized leader-follower structure. *Ocean Engineering, 235*. https://doi.org/10.1016/j.oceaneng.2021.109390

Zhang, Y., Sun, L. C., & Hu, G. Q. (2020). Distributed Consensus-Based Multitarget Filtering and Its Application in Formation-Containment Control. *IEEE Transactions on Control of Network Systems, 7*(1), 503–515. https://doi.org/10.1109/tcns.2019.2926281

Zhao, Y., Ma, Y., & Hu, S. (2021). USV Formation and Path-Following Control via Deep Reinforcement Learning With Random Braking. *IEEE Transactions on Neural Networks and Learning Systems, 32*(12), 5468–5478. https://doi.org/10.1109/tnnls.2021.3068762

Zhao, Y. J., Qi, X., Ma, Y., Li, Z. X., Malekian, R., & Sotelo, M. A. (2021). Path Following Optimization for an Underactuated USV Using Smoothly-Convergent Deep Reinforcement Learning. *IEEE Transactions on Intelligent Transportation Systems, 22*(10), 6208–6220. https://doi.org/10.1109/tits.2020.2989352

Zhen, Q. Z., Wan, L., Li, Y. L., & Jiang, D. P. (2022). Formation control of a multi-AUVs system based on virtual structure and artificial potential field on SE(3). *Ocean Engineering, 253*. https://doi.org/10.1016/j.oceaneng.2022.111148

Zhou, G., Mou, N., Fan, Y., Pi, Q., Bian, W., Zhou, C., . . . Gai, K. (2019). *Deep Interest Evolution Network for Click-Through Rate Prediction.* Paper presented at the National Conference on Artificial Intelligence.