Journal of
Marine Science and Engineering

MDPI

Article

USV Collision Avoidance Decision-Making Based on the Improved PPO Algorithm in Restricted Waters

Shuhui Hao, Wei Guan *, Zhewen Cui and Junwen Lu

Navigation College, Dalian Maritime University, Dalian 116026, China; hjkhsh@163.com (S.H.); cuizhewen123@dlmu.edu.cn (Z.C.); chuandianljw@dlmu.edu.cn (J.L.)
* Correspondence: gwwtxdy@dlmu.edu.cn

Abstract: The study presents an optimized Unmanned Surface Vehicle (USV) collision avoidance decision-making strategy in restricted waters based on the improved Proximal Policy Optimization (PPO) algorithm. This approach effectively integrates the ship domain, the action area of restricted waters, and the International Regulations for Preventing Collisions at Sea (COLREGs), while constructing an autonomous decision-making system. A novel set of reward functions are devised to incentivize USVs to strictly adhere to COLREGs during autonomous decision-making. Also, to enhance convergence performance, this study incorporates the Gated Recurrent Unit (GRU), which is demonstrated to significantly improve algorithmic efficacy compared to both the Long Short-Term Memory (LSTM) network and traditional fully connected network structures. Finally, extensive testing in various constrained environments, such as narrow channels and complex waters with multiple ships, validates the effectiveness and reliability of the proposed strategy.

Keywords: USV; collision avoidance decision-making; PPO; restricted waters; COLREGs

Citation: Hao, S.; Guan, W.; Cui, Z.; Lu, J. USV Collision Avoidance Decision-Making Based on the Improved PPO Algorithm in Restricted Waters. *J. Mar. Sci. Eng.* **2024**, *12*, 1428. https://doi.org/10.3390/jmse12081428

Academic Editor: Sergei Chernyi

Received: 25 July 2024
Revised: 6 August 2024
Accepted: 14 August 2024
Published: 19 August 2024

1. Introduction

In recent years, with the acceleration of global economic integration, the significance of ships as a crucial carrier for international trade has become increasingly prominent. However, this development is also accompanied by a substantial rise in the proportion of maritime accidents [1]. According to accident statistics from the European Maritime Safety Agency regarding EU-flagged ships between 2014 and 2022, navigational accidents accounted for 43% of all incidents [2]. To some extent, Unmanned Surface Vehicles (USVs) can mitigate ship collisions and effectively circumvent human error [3].

USV autonomous navigation relies on accurate and timely decision-making regarding dynamic changes in the marine environment. Factors such as ship maneuverability, perception capabilities, collision avoidance reliability, and the International Regulations for Preventing Collisions at Sea (COLREGs) are considered [1].

1.1. Related Works

In existing research, solutions to USV autonomous decision-making problems are primarily categorized into traditional intelligent algorithms, mathematical model-based and heuristic algorithms, and Deep Reinforcement Learning (DRL)-based algorithms [4].

Traditional intelligent algorithms have been extensively employed in USV autonomous decision-making. Inan et al. proposed an anti-collision decision support system that is capable of predicting the forward position of ships and calculating a collision avoidance route using the Cuckoo Search–Genetic Algorithm (CSGA) [5]. Cheng et al. introduced the Ship Dynamic Collision Avoidance Space Model (SDCASM), incorporating the Genetic Annealing algorithm (GAA) as an enhanced optimization technique for collision avoidance trajectory optimization based on SDCASM [6]. Li et al. proposed a fusion algorithm based on State-tracking Collision Detection and the Simulated Annealing Potential Field

(SCD-SAPF) to address the challenges of obstacle avoidance for Autonomous Underwater Vehicles (AUVs) in dynamic environments [7].

Despite the widespread utilization of traditional intelligent algorithms in the domain of USV autonomous decision-making, they encounter the challenge of adapting to intricate, uncertain, and dynamically evolving environments due to the escalating data volume and problem complexity.

Mathematical model-based and heuristic algorithms utilize mathematical analysis and heuristics to explore solutions. The algorithm proposed by Charalambopoulos et al. introduces an enhanced Probabilistic Roadmaps (PRMs) algorithm for addressing the ship weather routing (SWR) problem. Simulation results in the Aegean Islands and the Mediterranean Sea demonstrate the efficiency of the developed algorithm [8]. Zaccone et al. proposed a collision avoidance algorithm for ships navigating the open sea based on the Rapidly-exploring Random Tree star (RRT-star) algorithm. This algorithm enables the efficient management of multiple dynamic obstacles with varying speeds and trajectories [9]. However, it does not consider the COLREGs. He et al. proposed a Dynamic Anti-collision A-star (DAA-star) algorithm that complies with COLREGs, to address complex multi-ship encounter scenarios. The authors have developed a dynamic search mechanism based on DAA-star that incorporates temporal considerations, thereby achieving efficient dynamic collision avoidance [10]. Yuan et al. proposed an improved Dynamic Window Approach (DWA) algorithm called Utility DWA (UDWA), which takes into account the COLREGs. Also, the velocity sampling area is enhanced through a prioritization mechanism, while improving the velocity function within the objective function to account for wind and wave effects on USVs [11]. Arul et al. presented the V-RVO algorithm, a distributed collision avoidance technique based on Buffering Voronoi Cells (BVCs) and Reciprocal Velocity Obstacles (RVOs), which demonstrates its effectiveness in complex scenarios [12].

Compared to intelligent algorithms, mathematical model-based and heuristic algorithms exhibit lower data quality requirements and demonstrate superior adaptability to complex problems [4]. However, the construction of accurate mathematical models necessitates rigorous assumptions, potentially resulting in excessively impractical models. Additionally, numerous heuristic algorithms may encounter local optimality issues and possess limited generalization capabilities, thereby impeding their ability to meet the demands of future ship autonomous decision-making [13,14].

DRL serves as a decision-making approach and an end-to-end automated driving mechanism for collision avoidance at sea, enabling adaptability to complex, unknown, or dynamic marine environments. Chen et al. proposed a path planning and manipulating approach based on a Q-learning algorithm, which can drive a cargo ship by itself without requiring any human input [15]. However, the Q-learning algorithm manifests the state input in a discrete manner. Zhao et al. proposed a smoothly-convergent DRL (SCDRL) method based on the deep Q network (DQN) and reinforcement learning to solve the path-following problem for an underactuated USV [16]. Nevertheless, it should be noted that the output of DQN algorithm remains discrete. Du et al. proposed an optimized path planning method for coastal ships based on the Deep Deterministic Policy Gradient (DDPG) algorithm and Douglas–Peucker (DP) algorithms, with the incorporation of Long Short-Term Memory (LSTM) to enhance the network structure of DDPG [17]. However, this approach does not consider COLREGs and dynamic obstacles. Chun et al. proposed a collision risk assessment method based on the ship domain and the Closest Point of Approach (CPA), incorporating considerations of ship maneuverability and compliance to COLREGs [18]. Zheng et al. proposed a Rule-guided Visual Supervised Learning (RGVSL) approach to address the limitations of DRL in feature extraction for adaptive collision avoidance decision-making in autonomous complex encounter scenarios, thereby improving vision-supervised learning [19]. However, this approach does not provide speed control for the ship. In contrast, Chun et al. introduced a collision avoidance method based on DRL, which enables ship path and speed control but fails to consider immediate danger situations [20]. Guan et al. proposed a policy optimization navigational

method that incorporates the PRM and Proximal Policy Optimization (PPO) algorithm, while considering immediate danger situations in the reward functions [21]. Meyer E et al. proposed an obstacle avoidance method for autonomous ships based on the PPO algorithm, ensuring they follow the required trajectory without running aground. Trained models are evaluated in real-world scenarios using high-fidelity altitude and AIS tracking data from Trondheim Fjord [22]. However, it should be noted that employing radar lines as inputs may impose certain constraints on state perception. Wang et al. investigated the ships' collision avoidance system in complex collision scenarios, employed the images of ships' encounter situations as inputs for the DRL model and improved algorithm convergence speed through an adaptive parameter sharing method [1]. However, the computation load of the network will be increased due to the inherent pixel dimension and complexity of the image.

1.2. Problems and Contributions

While the DRL algorithm is considered as an advantageous approach for addressing the USV collision avoidance decision-making problem, the following challenges persist:

(1) DRL algorithms often employ a traditional Multilayer Perceptron (MLP) network structure, which partially limits their ability to retain long-term memory of past experiences.
(2) Previous research primarily focused on open sea environments, and there has been insufficient algorithm performance testing and validation in diverse scenarios, such as restricted waters and narrow waterways.

Therefore, addressing the aforementioned issues, this study successfully achieves autonomous collision avoidance decision-making for USVs in restricted waters based on the improved PPO algorithm while adhering to COLREGs. The specific contributions are outlined as follows:

(1) Considering problem 1, we substitute the traditional MLP network structure with the GRU network structure in the PPO algorithm. This modification facilitates information processing flow through a gating mechanism, thereby enabling a better capture of long-term dependencies and enhancing algorithm convergence.
(2) Considering problem 2, the proposed approach takes into account both the ship domain model and ship action area model within restricted waters. Additionally, a novel set of reward functions have been devised to incentivize USVs to comply with the COLREGs departure clause under immediate danger situations. Finally, a series of performance tests were conducted in real marine environments, encompassing narrow channels and complex waters, to comprehensively evaluate the proposed methods.

The remainder of this study is organized as follows: Section 2 focuses on the model and decision-making system construction. Section 3 delves into the improved PPO algorithm, covering the PPO algorithm, the GRU network implementation, the state space and action space, and the reward functions. Section 4 entails experimental verification through the simulation design, the network training, and the simulation validation. Section 5 encompasses the conclusion and a description of future works.

2. Model and Decision-Making System

To achieve the autonomous collision avoidance decision-making of USVs in compliance with COLREGs, it is essential to establish a comprehensive model and system. This chapter presents the ship domain model, the ship action area model, the ship motion mathematical model, the COLREGs model, and the ship autonomous decision-making system employed in this study.

2.1. Ship Domain Model

This study focuses on the autonomous decision-making of USV in restricted waters and narrow channels, utilizing the ship domain model proposed by Wang et al. [23], which

can accurately assess the potential immediate danger faced by USVs. As illustrated in Figure 1, the model employs an asymmetric polygon design, with the size of the model determined by the radial distance from the ship's center to the vertices of the polygon. The calculation formula is shown in Equation (1):

$$r_{\theta_i} = \alpha_{\theta_i} \cdot L \cdot g_{\theta_i}(v) \tag{1}$$

where α_{θ_i} is the normalised radial distance of the domain when the USV is stationary at the polar angle θ_i, L is the length of the USV, $g_{\theta_i}(v)$ is the velocity function at a given polar angle θ_i.

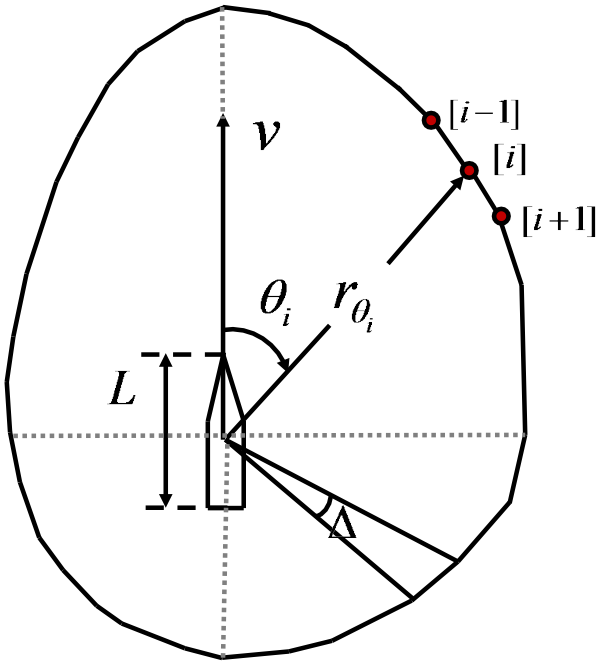


Figure 1. The restricted ship domain model of a USV.

Remark: Δ is the angular interval. The study employed 37 radar lines encircling the USV to ascertain potential violations of the ship domain; therefore, $\Delta = 360^\circ/37$, and the ship domain is represented by a 37-sided polygon. Subsequently, an assessment was carried out on the relationship between r_{θ_i} and length of radar line in θ_i to determine if any infringements occurred with respect to the USV.

2.2. Ship Action Area Model

In this study, the concept of "action area" proposed by Dinh et al. is adopted to define the timing of collision avoidance behavior for a USV operating in restricted waters. In contrast, other collision risk assessment systems can detect the collision within a range of 5 nm [24]. This is suitable for the open sea, but in restricted waters, premature action makes little sense. The concept of "action area" is illustrated in Figure 2; the radius of the area can be calculated by Equation (2) [24]:

$$r_{\mathrm{a}} = D_{\mathrm{f}} + 0.167 \times V_{\mathrm{relative}} \tag{2}$$

where r_{a} is the radius of the action area, D_{f} is the dangerous forward distance, and V_{relative} is the relative speed.

2.3. Ship Motion Mathematical Model

A sufficiently precise mathematical model for the USV motion is indispensable for both designing the controller and conducting simulation research [25]. In order to prioritize

the decision-making problem, the six-degrees of freedom ship motion model can be simplified by only considering the surge, sway, and yaw motions of the USV, as illustrated in Figure 3 [26]. We ultilize the simplified 3DOF ship motion mathematical model proposed by Fossen [27], as shown in Equation (3).

$$\begin{cases} \dot{\eta} = R(\psi)v \\ M\dot{v} = \tau - C(v)v - D(v)v - g(v) + \tau_w \end{cases} \tag{3}$$

where $\eta = [x, y, \psi]^T$ is the position coordinate and heading angle of the USV, $R(\psi)$ is the rotation matrix of the USV, $v = [u, v, r]^T$ is the velocity vector in the direction of the three degrees of freedom of the USV, M is the inertia matrix of the system, τ is the control force; $C(v)$ matrix consists of the centripetal matrix and the hydro-dynamic Coriolis matrix, $D(v)$ is the nonlinear damping matrix. $g(v) = [g_u, g_v, g_r]^T$ represents the unmodeled dynamic model and τ_w is the sum of the external forces of environmental interference. The details of the ship motion mathematical model can be viewed in other relevant papers [27]. The ship parameters are shown in Table 1, which were universally applied to all TSs and the OS in subsequent simulation experiments.

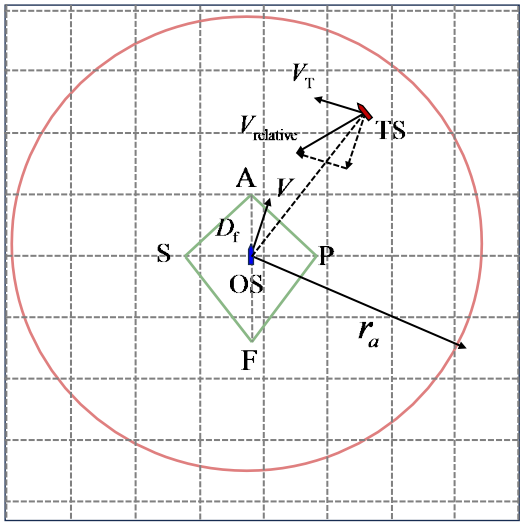


Figure 2. The restricted action area model of a USV.

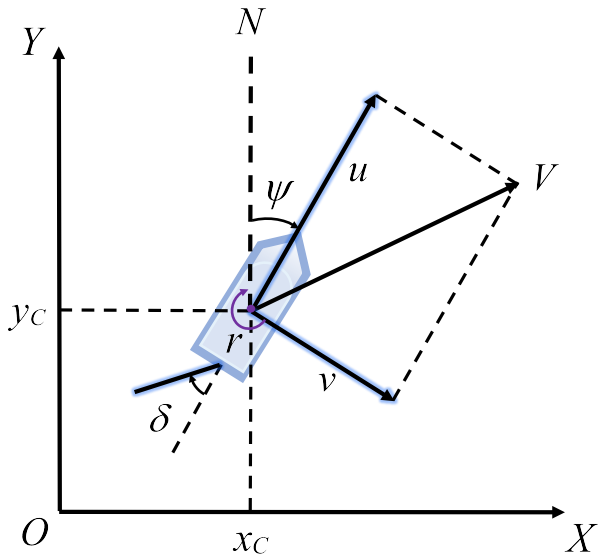


Figure 3. The USV motion mathematical model.

Table 1. Parameters of the USV.

Parameters	Value
Length (m)	7
Beam (m)	3.66
Draft (m)	0.56
Weight (kg)	544
Maximum speed (knot)	20
Max rudder angle (deg)	25
Propulsion (kw)	29.828
Payload (kg)	270

2.4. Ship Autonomous Decision-Making System

When developing the USV autonomous decision-making strategy, it is crucial to establish a comprehensive end-to-end system that includes environment perception and steering control functions. The system, as shown in Figure 4, consists of the sensing module, decision-making module, and control module. Each of these modules plays a vital role in ensuring efficient and accurate USV autonomous decision-making.

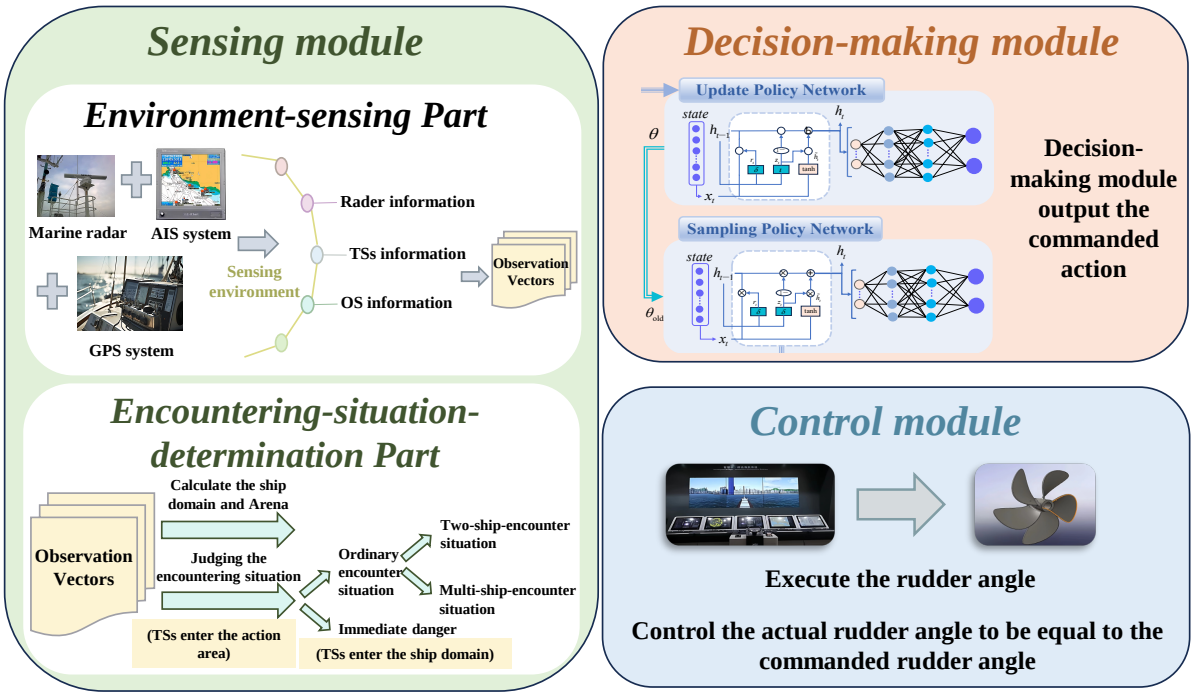


Figure 4. USV autonomous decision-making system.

First, the sensing module, serving as the initial stage of information acquisition, is divided into two sub-parts. The environment-sensing part senses the environmental information and generates a set of detailed observation vectors by the USV's Automatic Identification System (AIS), marine radar, Global Positioning System (GPS) and other hardware devices. Then, based on these observation vectors, the encountering-situation-determination part dynamically calculates the USV ship domain R_i and action area radius r_a, where $R_i = [r_{\theta_1}, r_{\theta_2}, \ldots, r_{\theta_{37}}]$.

After the target ship (TS) enters the action area of the USV, the encountering-situation-determination part judges the encounter situation that the USV faces. In this regard, if the TS does not invade the ship domain of the USV, the encounter situation can be categorized into a two-ship encounter situation and a multi-ship encounter situation within the action area. When the TS invades the ship domain of the USV, it poses an immediate danger.

Subsequently, the decision-making module plays a crucial role in the system by using the DRL algorithm to extract important information from the observation vectors as inputs. It accurately translates the perceptual data into action commands during the navigation, adapting to diverse and complex environments while ensuring safety.

Finally, the control module executes instructions from the decision-making module, ensuring the precise regulation of the USV's heading and propulsion through ship control devices like the rudder and propeller system. This facilitates efficient and safe navigation for the USV by enabling autonomous decision-making.

2.5. COLREGs Model

COLREGs provide the criteria for assessing the encounter situation between two ships, encompassing the rules pertaining to USV autonomous decision-making. The related clauses in this study are as outlined:

(1) The Article 13–17 of Chapter II of the COLREGs specify the collision avoidance actions that the own ship (OS) and target ship (TS) should undertake when two ships encounter each other [28], as illustrated in Figure 5.

(2) In the event of multi-ship encounters, COLREGs does not specify the guidelines that should be followed by the USV. In such cases, different encounter situations may arise between the USV and various TSs, and the USV should strive to comply with the COLREGs guidelines as closely as possible.

(3) In the event of the TS entering the ship domain of the USV, an immediate danger arises. The USV should comply with the provisions outlined in Article 2 of Chapter II and Article 8 of Chapter II of the COLREGs by considering a significant alteration in course and/or speed.

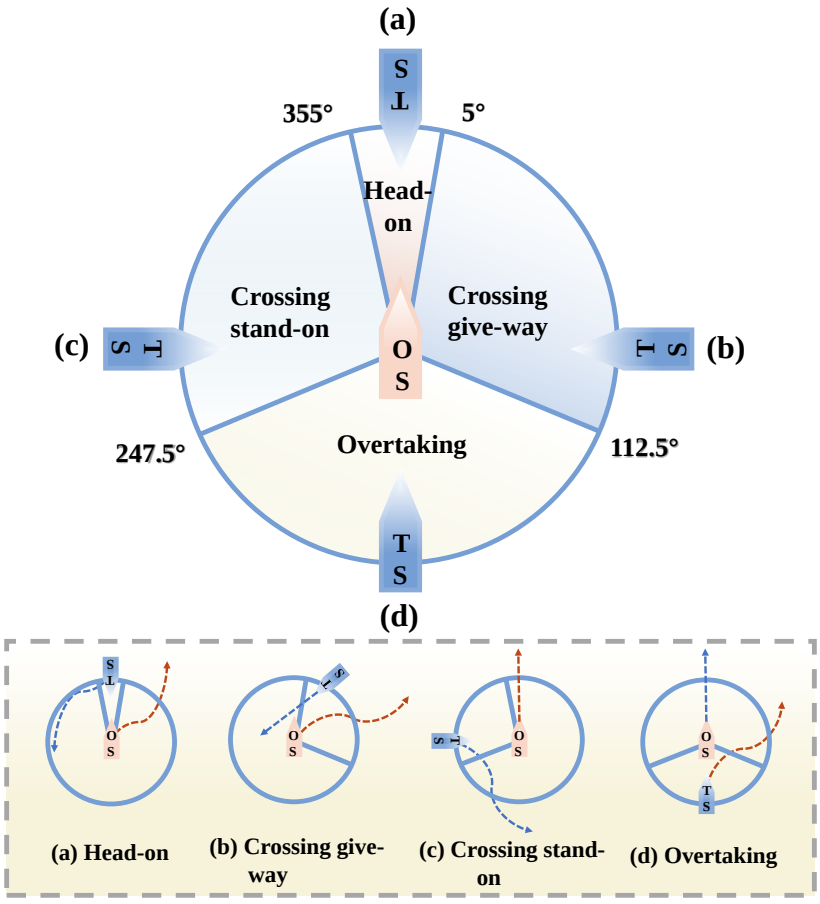


Figure 5. COLREGs model.

3. Improved PPO Algorithm

The current chapter focuses on the main methodology, providing a detailed exposition of the PPO algorithm and the GRU network. Additionally, it elaborates on the state space, the action space, and the reward functions configuration.

3.1. PPO Algorithm

The PPO algorithm is selected as the core decision-making algorithm in this study, representing a significant improvement over the Policy Gradient (PG) algorithm. Incorporating a clipping mechanism and importance sampling, the PPO algorithm significantly enhances data utilization efficiency.

The optimization of strategies in reinforcement learning aims to maximize the accumulation of long-term rewards. Mathematically, this process usually involves solving for the mean gradient $\nabla \overline{R}_\theta$ of R_θ, as shown in Equation (4).

$$\nabla \overline{R}_\theta = E_{\tau \sim P_{\theta(\tau)}}[R(\tau) \cdot \nabla \log P_{\theta(\tau)}] \tag{4}$$

where $R(\tau)$ represents the cumulative reward in an epoch of τ, and $P_{\theta(\tau)}$ represents the occurrence probability of τ.

To ensure a rational allocation of weights for different actions and to address the issue of decreasing the probability caused by unselected actions, the PPO algorithm employs an advantage function $A^\theta(s_t, a_t)$ to replace the original cumulative reward $R(\tau)$. This advantage function quantifies the benefit of taking an action compared to the average action in a given state, enabling a more accurate evaluation of different actions and subsequent updates of strategy parameters. $\nabla \overline{R}_\theta$ can be determined by Equation (5) as follows:

$$\nabla \overline{R}_\theta = E_{\tau \sim P_{\theta(\tau)}}[A^\theta(s_t, a_t) \cdot \nabla \log P_{\theta(\tau)}] \tag{5}$$

The PPO algorithm employs Generalized Advantage Estimation (GAE) for advantage estimation, which introduces parameters to flexibly combine short-term and long-term return information, thereby achieving an effective trade-off between bias and variance. The advantage function integrated with GAE is shown in Equation (9).

$$\begin{aligned}
A^{\mathrm{GAE}(r,\lambda)} &= \sum_{n=1}^{\infty} (1-\lambda)\lambda^{n-1} \hat{A}_t^{(n)} \\
&= (1-\lambda)(\hat{A}_t^{(1)} + \lambda \hat{A}_t^{(2)} + \lambda^2 \hat{A}_t^{(3)} + \dots) \\
&= (1-\lambda)(\delta_t^\nu + \lambda(\delta_t^\nu + \gamma \delta_{t+1}^\nu) + \lambda^2(\delta_t^\nu + \gamma \delta_{t+1}^\nu + \gamma^2 \delta_{t+2}^\nu) + \dots) \\
&= (1-\lambda)(\delta_t^\nu(1 + \lambda + \lambda^2 + \dots) + \gamma \delta_{t+1}^\nu(\lambda + \lambda^2 + \lambda^3 + \dots) + \gamma^2 \delta_{t+2}^\nu(\lambda^2 + \lambda^3 + \lambda^4 + \dots) + \dots) \\
&= (1-\lambda)(\delta_t^\nu \frac{1}{1-\lambda} + \gamma \delta_{t+1}^\nu \frac{\lambda}{1-\lambda} + \gamma^2 \delta_{t+2}^\nu \frac{\lambda^2}{1-\lambda} + \dots) \\
&= \sum_{l=0}^{\infty} (\gamma\lambda)^l \delta_{t+l}^\nu
\end{aligned} \tag{6}$$

where $\lambda \in [0, 1]$ is a hyperparameter. When $\lambda = 0$, we can focus on the advantages of the single-step difference. When $\lambda = 1$, we can obtain a perfectly averaged dominant difference acquired at each step.

Combined with GAE as an advantage function, the PPO algorithm can achieve multi-step update within a single iteration through the utilization of importance sampling (IS). $\nabla \overline{R}_\theta$ is shown in Equation (7), and objective function is shown in Equation (8).

$$\nabla \overline{R}_\theta = E_{(s_t, a_t) \sim \pi_{\theta_{\mathrm{old}}}} \nabla [\frac{P_\theta(a_t \mid s_t)}{P_{\theta_{\mathrm{old}}}(a_t \mid s_t)} A^{\mathrm{GAE}}(s_t, a_t)] \tag{7}$$

$$J^{\theta_{\mathrm{old}}}(\theta) = E_{(s_t, a_t) \sim \pi_{\theta_{\mathrm{old}}}} [\frac{P_\theta(a_t \mid s_t)}{P_{\theta_{\mathrm{old}}}(a_t \mid s_t)} A^{\mathrm{GAE}}(s_t, a_t)] \tag{8}$$

where $J^{\theta_{\text{old}}}(\theta)$ is the objective function, θ_{old} is the network parameter of the sampling policy, θ is the network parameter of the update policy. $P_\theta(a_t \mid s_t)$ and $P_{\theta_{\text{old}}}(a_t \mid s_t)$ represent the probability that s_t selects a_t in the update policy network and sampling policy network, respectively.

In Equation (8), by incorporating the clip clipping strategy instead of solely relying on KL divergence to constrain policy updates, the optimization problem is effectively streamlined, and $J^{\theta_{\text{old}}}(\theta)$ can be obtained by Equation (9).

$$J^{\theta_{\text{old}}}(\theta) \approx \sum_{(s_t,a_t)} \min\left(\frac{P_\theta(a_t \mid s_t)}{P_{\theta_{\text{old}}}(a_t \mid s_t)} A^{\text{GAE}}(s_t,a_t), clip\left(\frac{P_\theta(a_t \mid s_t)}{P_{\theta_{\text{old}}}(a_t \mid s_t)}, 1-\varepsilon, 1+\varepsilon\right) \cdot A^{\text{GAE}}(s_t,a_t)\right) \tag{9}$$

where $clip(a,b,c)$ means limiting a to $[b,c]$. ϵ is a hyper parameter which indicates the range of clipping.

3.2. GRU Network

The operation process of the algorithm is illustrated in Figure 6, where the PPO algorithm updates the strategy network through gradient descent by utilizing objective function (8). Furthermore, the agents interact with the environment to generate experience sequences, which are randomly sampled by the PPO algorithm to construct training samples for both the policy and value networks. During the training, the loss functions of these networks are computed and their parameters are iteratively updated until convergnce is achieved.

It is worth noting that the conventional PPO algorithm employs a MLP as both the strategy network and value network structure, which fails to capture the temporal dependencies at each time step. Moreover, traditional Recurrent Neural Networks (RNNs) are susceptible to gradient vanishing or exploding problems. To address these issues, this study proposes a GRU-MLP network structure as an alternative to MLP. As a simplified version of LSTM network, GRU simplifies the three gating units into two gating units (update gates and reset gates), reducing parameter complexity and overfitting risks while improving training speed.

Specifically, the GRU network acts as a preprocessing component of the input information x_t, generating information about the current moment, including x_t, h_{t-1}, and all previous moments.

Once x_t is entered, the reset gate determines how to combine the new input information with the previous memory h_{t-1}. The reset gate formula is represented by Equation (10)

$$r_t = \sigma(W_r[h_{t-1}, x_t] + b_r) \tag{10}$$

where r_t is used in the candidate hidden state formula to determine the relevance extent between h_{t-1} and the current input information x_t, as shown in Equation (11).

$$\tilde{h}_t = \tanh(W[r_t * h_{t-1}, x_t]) \tag{11}$$

where $\tilde{h}_t$ is the candidate hidden state, which is used to update the hidden status of the current time step.

The update gate is used to help the model decide how much past information should be passed on to the future, that is, updating the memory. The update gate formula is shown in Equation (12), and the formula of updating the memory is shown in Equation (13).

$$z_t = \sigma(W_z[h_{t-1}, x_t]) \tag{12}$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t \tag{13}$$

where the closer z_t is to 1, the more information GRU remembers. The closer z_t is to 0, the more information GRU forgets.

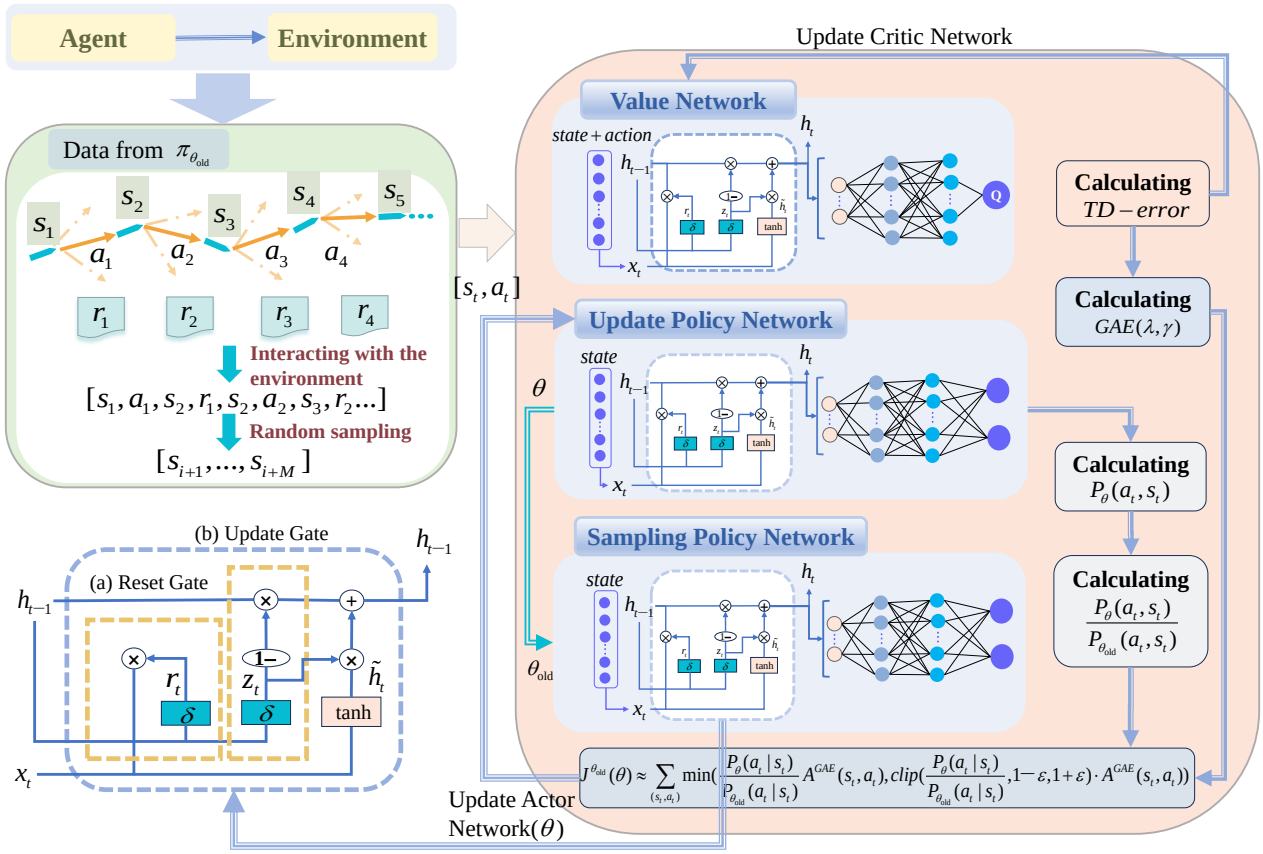


Figure 6. Interaction process and network structure of the improved PPO algorithm.

In the value network, after x_t is preprocessed by GRU network, MLP network outputs the action value. Similarly, in the policy network, after x_t is preprocessed by the GRU network, the MLP network outputs the mean and variance of the selected action in the current state. Also, the thoughtful utilization of reset and update gates in a GRU network can effectively manage and exploit long-term dependencies inherent in time-series data.

3.3. State Space and Action Space

3.3.1. State Space

The sensing module generates a set of observation vectors, which serve as the foundation for constructing the state space. The state space aims to comprehensively represent both the surrounding environmental information and the USV's internal state information. Consequently, the state space is defined as follows as Equation (14):

$$S_t = [S_{\mathrm{ENV}}, S_{\mathrm{OS}}] \tag{14}$$

where S_{ENV} represents the surrounding environmental information. To balance the input complexity and the exploratory potential of the surrounding environment, a configuration of 37 radar lines emitting from the USV center is utilized as the perceptual mode for gathering surrounding information. S_{OS} represents the USV's internal state information.

Hence,

$$S_{\mathrm{ENV}} = [\xi_1, \xi_2, \ldots, \xi_{37}] \tag{15}$$

$$S_{\mathrm{OS}} = [\delta_t, v_t, \psi_t, d_t, D_t] \tag{16}$$

where $[\xi_1, \xi_2, \ldots, \xi_{37}]$ represent the detection lengths of 37 radar lines, in n miles. δ_t represents the command steering rudder angle of the USV. v_t represents the command

speed of the USV. ψ_t represents the difference between the current course and the azimuth angle at the goal. d_t represents the distance between the current position and the goal. D_t is a Boolean value, signifies the occurrence of a collision at the present instant.

3.3.2. Action Space

The action space is designed to enable USV autonomous collision avoidance and navigation in compliance with COLREGs. According to the provisions of Article 8 of Chapter II of the COLREGs, the action space defined in this study is represented in Equation (17).

$$A = \left[\delta_t, u_t\right] \tag{17}$$

where δ_t represents the commanded rudder angle, and its value is continuous within the interval $[-25°, 25°]$. u_t is the commanded speed and its value is continuous within the interval $[0 \text{ knot}, 20 \text{ knots}]$.

It should be noted that, in order to fulfil the actual navigation requirements, a predetermined commanded speed of 14 knots is set when the USV operates beyond the area of the action area. Once the action area of USV is invaded, the speed adjustments are authorized to prevent potential collisions.

3.4. Reward Functions

It is crucial to establish well-designed reward functions in the implementation of USV autonomous decision-making, which should take into account the relationship between mainline reward and auxiliary rewards [29]. The mainline reward aims to guide the agent's behavior towards the successful avoidance of collisions and arriving at the goal, while auxiliary rewards are intended to balance the overall reward functions, provide sufficient feedback and encouragement during learning, and maintain stability in learning even when the mainline reward is unclear. Specifically, the mainline reward is set as the ending reward, while the auxiliary rewards are set as the guided reward, the COLREGs reward, the ship domain reward and the path optimization reward.

3.4.1. Ending Reward

During the voyages, the mission of the USV is terminated upon collision or reaching a intended goal. The ending reward is designed to impose penalties for collision incidents and to provide rewards for safe arrival at the goal, thereby incentivizing the USV to successfully complete tasks without collisions. The reward is defined as

$$R_\alpha^t = \begin{cases} r_{\text{collision}}, & d_{p_t}^{p_{\text{obs}}} \le d_{\text{collision}} \\ r_{\text{goal}}, & d_{p_t}^{p_g} \le d_{\text{goal}} \end{cases} \tag{18}$$

where $d_{p_t}^{p_{\text{obs}}}$ represents the distance from USV to the obstacles, $d_{p_t}^{p_g}$ represents the distance from USV to the goal, $d_{\text{collision}}$ represents the collision distance, and d_{goal} represents the distance of arriving at the goal. When $d_{p_t}^{p_{\text{obs}}}$ is less than or equal to $d_{\text{collision}}$, a collision penalty occurs. When $d_{p_t}^{p_g}$ is less than or equal to d_{goal}, a reward for reaching the goal is generated.

3.4.2. Guided Reward

Relying solely on the end reward function may result in a lack of reward density. Consequently, the guided reward function is specifically designed to provide a continuous stream of rewards to the USV as it approaches the goal. When the USV reaches the goal, positive guidance rewards are provided, whereas negative guidance rewards are assigned when the USV is away from the target. The guided reward is defined as

$$R_\beta^t = \beta(\sqrt{(p_{x_t} - p_{x_g})^2 + (p_{y_t} - p_{y_g})^2} - \sqrt{(p_{x_{t-1}} - p_{x_g})^2 + (p_{y_{t-1}} - p_{y_g})^2}) \tag{19}$$

where β represents the weight of guided reward, (p_{x_t}, p_{y_t}) represents the current coordinate of the USV, $(p_{x_{t-1}}, p_{y_{t-1}})$ represents the coordinate of the USV in the previous moment, (p_{x_g}, p_{y_g}) represents the coordinate of the goal.

3.4.3. COLREGs Reward

When TSs invade the action area of the USV, both the own ship (OS) and the TSs are required to execute collision avoidance manoeuvers in accordance with the COLREGs. Upon TS entry into the action area of USV, the sensing module assesses the encounter situation based on observation vectors and relays the observed results to the decision-making module. If USV behavior adheres to COLREGs, a positive reward is assigned proportionate to distance. Conversely, if USV behavior violates COLREGs, a negative reward correlated with distance is imposed. When TSs invade the domain of USV, a reward of 0 indicates that the USV should execute collision avoidance strategies by complying with the Article 2 departure clause of the COLREGs in an immediate danger situation.

$$R_\gamma^t = \begin{cases} \gamma_h, & d_T \geq r_a \\ \gamma\sqrt{\left(p_{x_t} - p_{x_g}\right)^2 + \left(p_{y_t} - p_{y_g}\right)^2}, & (r_{\theta_i}^T < d_T < r_a) \wedge (contrary\ to\ COLREGs) \\ -\gamma\sqrt{\left(p_{x_t} - p_{x_g}\right)^2 + \left(p_{y_t} - p_{y_g}\right)^2}, & (r_{\theta_i}^T < d_T < r_a) \wedge (departure\ from\ COLREGs) \\ 0, & d_T \leq r_{\theta_i}^T \end{cases} \tag{20}$$

where γ_h represents a hyperparameter that indicates a reward value, γ represents the weight of the COLREGs reward, d_T represents the distance from the USV to the TS, r_a represents the radius of the USV action area, and $r_{\theta_i}^T$ represents the length of the ship domain in the direction of the TS.

3.4.4. Ship Domain Reward

According to the COLREGs, it is advisable for the USV to maintain an inviolable presence within her ship domain during her voyage. In cases where violations occur within the USV ship domain, a penalized reward is assigned, proportional to the cumulative length of violation across all directions. More severe violations yield smaller rewards. Conversely, less severe violations will incur a greater reward.

$$R_\zeta^t = \begin{cases} \zeta e^{\left(\sqrt{(p_{x_t} - p_{x_g})^2 + (p_{y_t} - p_{y_g})^2} \cdot \frac{\sum_{i=1}^{M} (r_{\theta_i} - d_{\theta_i})^2}{4}\right)}, & d_{\theta_i} \leq r_{\theta_i} \\ 0, & d_{\theta_i} > r_{\theta_i} \end{cases} \tag{21}$$

where ζ represents the weight of the ship domain reward, d_{θ_i} represents the length of the radar line in the direction of θ_i.

3.4.5. Path Optimization Reward

The path optimization reward is a negative value correlated with the number of USV navigation steps, which serves to minimize the USV navigation path, and reduce the overall navigation duration. The path optimization reward can be defined as

$$R_\eta^t = \begin{cases} -\eta(m/\eta_h), & A = \text{False} \\ 0, & A = \text{True} \end{cases} \tag{22}$$

where η represents the weight of path optimization reward, m represents the number of USV navigation steps before arriving at the goal, η_h represents a hyperparameter, and A represents a Boolean value. When A is equal to True, it indicates that the USV has arrived at the goal. Conversely, when A is equal to False, it implies that the USV has not yet arrived at the goal.

Therefore, the total reward function is expressed as a sum of the five parts above, as shown in Equation (23).

$$R_{\mathrm{T}}^{t} = R_{\alpha}^{t} + R_{\beta}^{t} + R_{\gamma}^{t} + R_{\xi}^{t} + R_{\eta}^{t} \tag{23}$$

4. Experiment

4.1. Design of Simulation

The simulation experiment consists of two primary stages: network training and trained model verification. Particularly, the simulation verification experiment includes four challenging restricted waters to evaluate the feasibility of the trained USV autonomous decision-making model under different levels of complex sea conditions. In the neural network training, we utilized a stable water environment to enhance the resilience of the trained model [30]. In the verification simulation, the parameters of wind and wave interference are shown in Table 2 [31]. We used Python 3.7.0 as the simulation software and OpenAI Gym as the simulation platform. The computer configuration settings are as follows: Intel Core i7 13700KF CPU, NVIDIA GeForce RTX 4070 Ti.

Table 2. Parameters of the improved PPO algorithm and experiment settings.

Parameter	Symbol	Value	Parameter	Symbol	Value
Discounted rate	γ_D	0.96	R_{β}^{t} weight	β	−0.35
Lamba	λ	0.99	R_{γ}^{t} weight	γ	6
Clipping hyperparameter	ϵ	0.20	R_{ξ}^{t} weight	ξ	−3.5
Learning rate	l_{r}	0.0003	R_{η}^{t} weight	η	0.008
Max steps	$m_{\max}$	1000	Collision penalty	$r_{\mathrm{collision}}$	−2000
GRU hidden layer units number	n_{u}	128	Goal reward	r_{goal}	5000
Collision distance (n mile)	$d_{\mathrm{collision}}$	0.136	R_{γ}^{t} hyperparameter	γ_{h}	2
Goal distance (n mile)	d_{goal}	0.12	R_{η}^{t} hyperparameter	η_{h}	5
wave and wind direction (°)	d	35	significant wave height (m)	w_{h}	0.18
wave period (s)	w_{p}	8	mean wind speed (m/s)	v_{w}	0.33

4.2. Network Training

The parameters of the improved PPO algorithm and the values of the reward functions parameters during the training process are presented in Table 2. Upon initializing the starting point for the USV, the goal point is randomly generated within a predefined scope. In event of a collision between the USV and an obstacle or TS, this training round is immediately terminated. When the USV successfully reaches the goal, the current round continues and a new randomly generated goal distinct from that of the previous round will be established.

The training environment is shown in Figure 7, which is a section of restricted water with static obstacles and TSs. During the initial stages of training, the USV explores the environment where rewards are typically low. As the training progresses, the USV gradually refines her decision-making strategy through continuous interaction with the training environment, leading to a gradual increased reward value. When the training rounds are sufficiently large, the improved PPO algorithm converges, leading to a stable autonomous decision-making model for USV. Based on the model, USV can successfully reach the goal without any collisions in each round until surpassing the maximum step limit.

Figure 8 illustrates the reward value changes throughout the training process. Also, this study conducted a comparative analysis among the improved PPO algorithm based on LSTM, the improved PPO algorithm based on GRU, the PPO algorithm, and the DDPG algorithm. Consequently, the following conclusions are drawn:

(1) The PPO algorithm, the improved PPO algorithm based on LSTM (PPO-LSTM), and the improved PPO algorithm based on GRU (PPO-GRU) all exhibit convergence after approximately 1400 iterations, 1100 iterations, and 600 iterations, respectively. In the 2000 iterations, the DDPG algorithm exhibits suboptimal convergence.

(2) Compared to the PPO algorithm, the PPO-LSTM algorithm incorporates a gating mechanism for regulating information flow, facilitating the effective preservation and dissemination of crucial information, and enhancing network convergence in lengthy sequences. In contrast, the PPO-GRU algorithm offers a streamlined gating structure with fewer parameters than the PPO-LSTM algorithm, resulting in an improved computational efficiency and faster training speed under certain circumstances.

(3) The DDPG algorithm exhibits sensitivity to hyperparameters, and its sample utilization is suboptimal, which probably impedes the convergence of training [32].

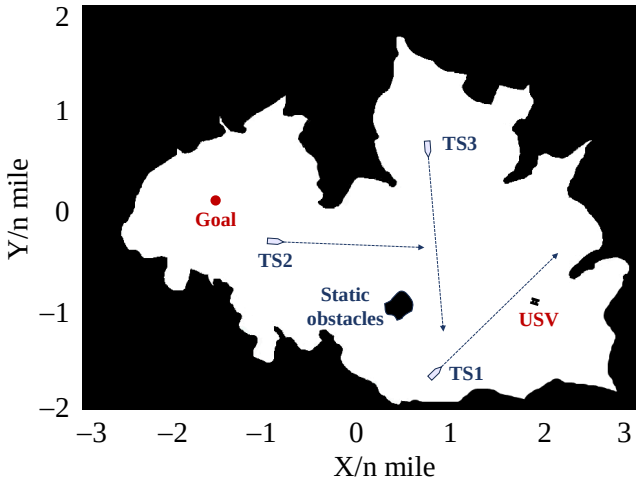


Figure 7. The training environment.

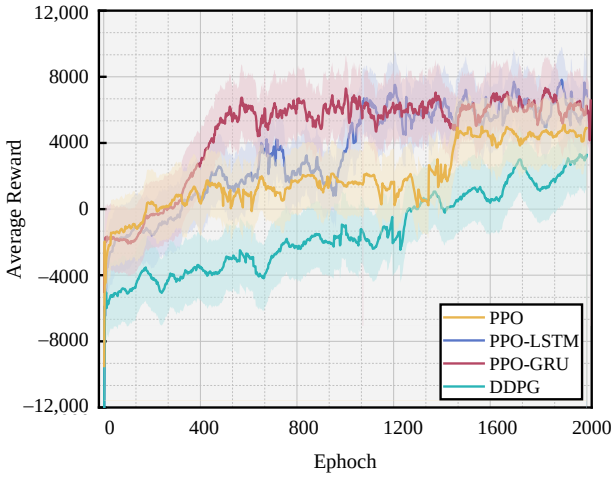


Figure 8. The average reward values of the four algorithms.

4.3. Trained Model Verification

In this section, we verified the network model trained in the real marine environment [33]. To comprehensively evaluate the effectiveness and validity of the trained model, this study designed four sets of experiments to mirror real-world maritime conditions, including the COLREGs compliance verification experiments, the five-ship encounter situation experiment in a narrow channel, the six-ship encounter situation experiment in complex waters and the seven-ship encounter situation experiment in complex waters. These experiments aimed to simulate the potential complications that the USV might encounter during actual voyages, providing a comprehensive validation of the trained model's performance and reliability across various restricted waters.

4.3.1. COLREGs Compliance Verification Experiments

In the COLREGs compliance verification experiment, cthis study focuses on verifying the specific clauses of Article 13–17 of Chapter II of the COLREGs, aiming to simulate four different ship encounter situations. The setting of the USV navigation conditions is presented in Table 3. The experimental results are shown in Figure 9, including the path trajectory of the USV and the TSs, as well as the USV rudder angle and speed during the collision avoidance process.

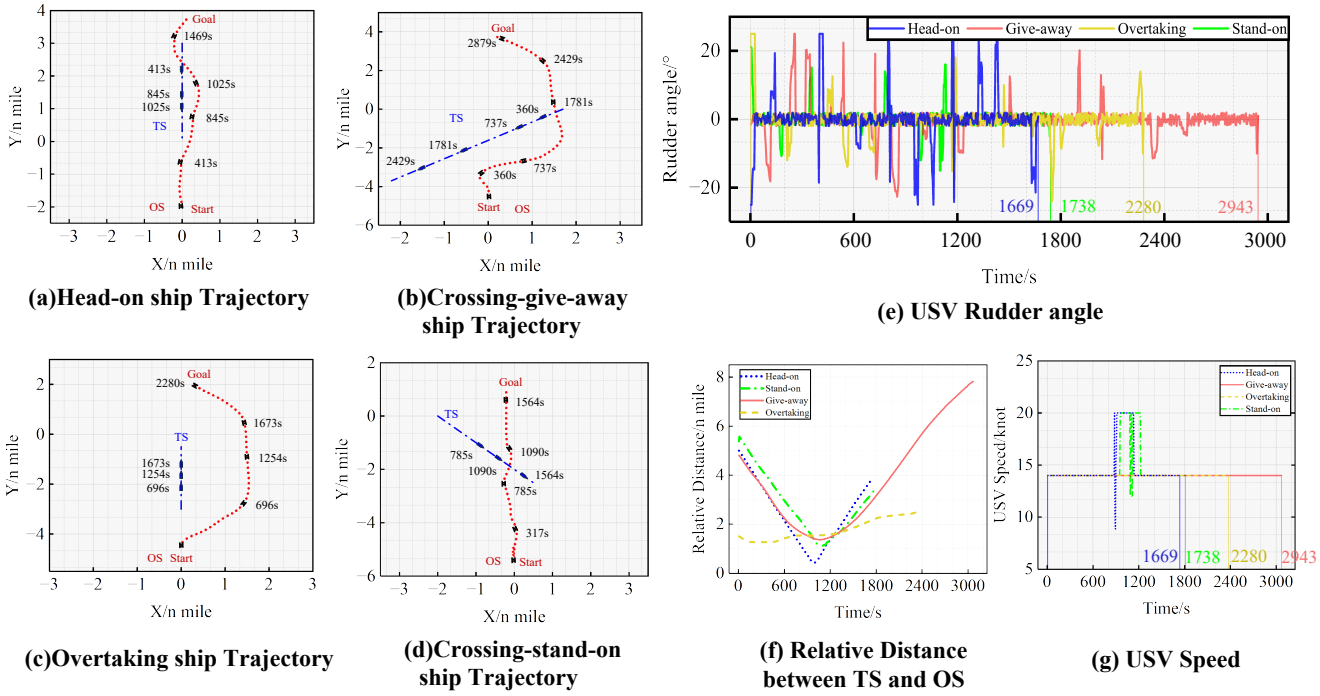


Figure 9. The results of the COLREGs compliance verification experiments.

Table 3. Initial settings of COLREGs compliance verification experiments.

Encounter Situation	OS Start Position (n Mile)	OS Goal Position (n Mile)	TS Start Position (n Mile)	TS Speed (Knot)
Head-on	(0, −2)	(0.16, 3.83)	(0, 3)	7.71
Crossing give-way	(0, −4.5)	(0.16, 3.78)	(1.7, 0)	7.88
Overtaking	(0, −4.5)	(0.25, 2)	(0, −3)	4.74
Crossing stand-on	(0, −5)	(−0.25, 1.04)	(−2, 0)	9.12

In Figure 9a, when the distance between the OS and the TS was 2.94 n miles at 345 s, the TS entered the action area of the OS. The encountering-situation-determination system of the sensing module would judge this encounter situation as a two-ship head-on situation. Consequently, according to the COLREGs, the OS steered with a rudder angle of 25° to alter her course, and sailed along the port side of the TS, successfully avoiding the collision at 855 s.

In the Figure 9b experiment, when the distance between the OS and the TS was 3.2 n mile at 300 s and the TS entered the action area of the OS, the encountering-situation-determination system of the sensing module judged the situation as a two-ship crossing give-way situation. Consequently, the OS steered with a rudder angle of 25° to alter her course to the starboard side, successfully avoiding the collision at 1485 s.

In Figure 9c, when the distance between the OS and the TS was 1.5 nm at 0 s, the TS entered the action area of the OS and the speed of the OS was greater than the TS, so encountering-situation-determination system of the sensing module judged the situation

as an overtaking situation. Consequently, the OS steered with a rudder angle of 25° to alter her course to the starboard side, successfully avoiding the collision at 1395 s.

In Figure 9d, after the TS entered the action area of the OS, the encountering-situation-determination system assessed the situation as a two-ship crossing stand-on situation. In accordance with the COLREGs, the OS was advised to maintain her course and speed and exercise due diligence in awaiting the TS's collision avoidance manoeuvers. Given that the TS was a manned vessel unable to manipulate the rudder, the OS steered with a rudder angle of 14° at 655 s to prevent the immediate collision danger in response to the TS's invasion into the ship domain of the OS. Ultimately, at 910 s, the OS successfully avoided the collision and sailed to the intended goal.

In summary, the trained model enables autonomous collision avoidance decision-making for the USV in four types of two-ship encounter scenarios while ensuring compliance with Article 13–17 in Chapter II of COLREGs. Additionally, when another ship departs from the COLREGs and enters the USV's ship domain, the USV can effectively respond to immediate dangerous situations by promptly adjusting rudder angle and speed.

4.3.2. Five-Ship Encounter Situation Experiment in Narrow Channel

In this section, the channel from Nanjing City to Zhenjiang City is selected as the experimental environment for evaluating the USV's narrow channel navigation capability. The setting of the USV navigation conditions is presented in Table 4, while the corresponding experimental results are illustrated in Figure 10. During the navigation of a narrow channel, it is common for USV to encounter situations where her ship domain is violated. In such cases, the departure clause of Article 2 in Chapter I of the COLREGs becomes imperative for governing the USV's behavior, and the successful execution of a collision avoidance operation should serve as the primary decision-making criterion.

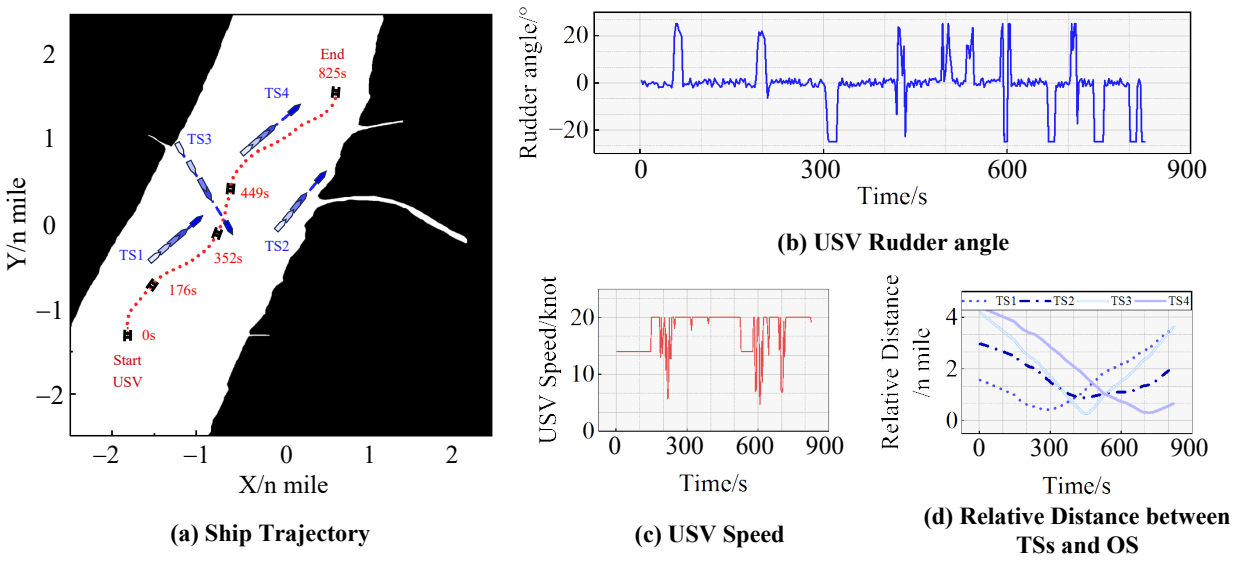


Figure 10. The results of the five-ship encounter situation experiment in a narrow channel.

Table 4. Initial settings of the five-ship encounter situation experiment in a narrow channel.

Ship	Start Position (n Mile)	Goal Position (n Mile)	Speed (Knot)
USV	(−1.82, −1.31)	(0.64, 1.15)	0–20
TS1	(−1.5, −0.4)	(−1, 0.05)	2.93
TS2	(0, 0)	(0.5, 0.6)	3.41
TS3	(−1.2, 0.9)	(−0.6, −0.05)	4.90
TS4	(−0.4, 0.9)	(0.2, 1.4)	3.41

After the initialization of the experiment, TS1, TS2, TS3, TS4 and the USV formed a five-ship encounter situation, and the USV steered with a 25° rudder angle to alter her course to starboard side. Then, at 235 s, TS1 and the USV represented an immediate danger, and the USV subsequently steered with a 22° rudder angle to alter her course, sailing along the starboard side of TS1. At 470 s, the USV steered with a −25° rudder angle to sail between TS2 and TS3. At 600 s, TS4 and the USV again represented an immediate danger, encouraging the USV to steer with a 25° rudder angle to alter her course to starboard side to avoid TS4, ultimately reaching the goal.

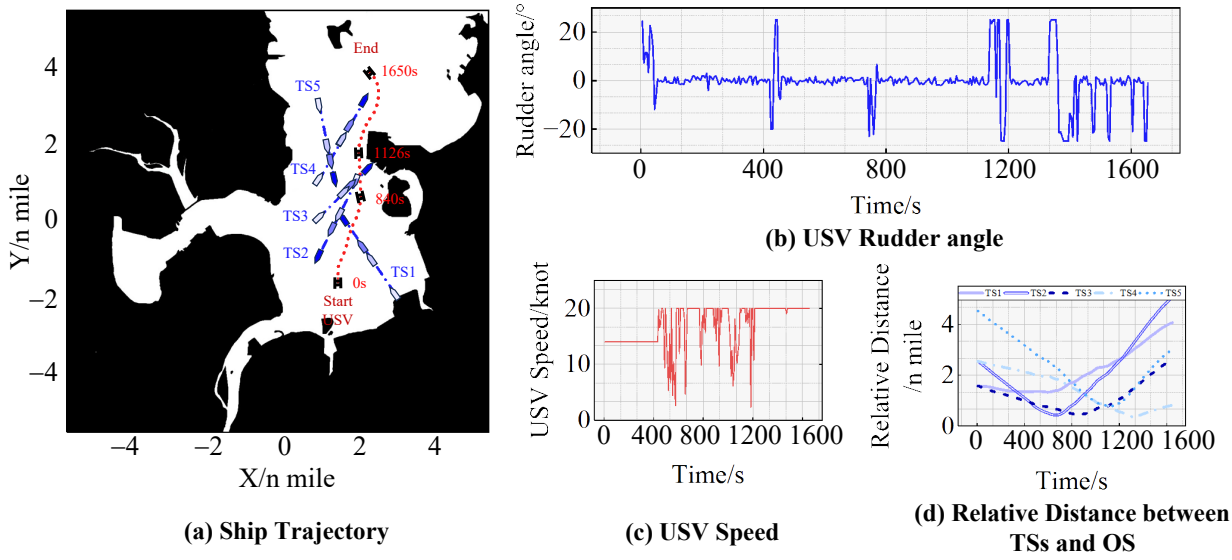
The experimental results demonstrate that:

(1) The trained model enables the USV to safely navigate during the passing through the narrow channel and adjust her speed to avoid collisions in immediate danger situations.

(2) The trained model empowers the USV to analyze the sailing conditions of multiple ships during encounters, make independent decision-makings, and adhere as closely as possible to the COLREGs. In this experiment, when faced with a encounter situation involving four ships, the USV autonomously chose to alter her course to the starboard side of TS1 and TS4, instead of passing TS1 on the port side and TS3 astern, thereby demonstrating compliance with the COLREGs in narrow channel situations.

4.3.3. Six-Ship Encounter Situation Experiment in Complex Waters

In this part, the coastal area of Quanzhou City is chosen as the experimental environment for assessing the navigating capability of USV within a six-ship encounter situation under the complex sea environment. The initialization of the USV navigation conditions is presented in Table 5, while the corresponding experimental results are illustrated in Figure 11.



(a) Ship Trajectory **(b) USV Rudder angle** **(c) USV Speed** **(d) Relative Distance between TSs and OS**

Figure 11. The results of the six-ship encounter situation experiment in complex waters.

Table 5. Initial settings of the six-ship encounter situation experiment in complex waters.

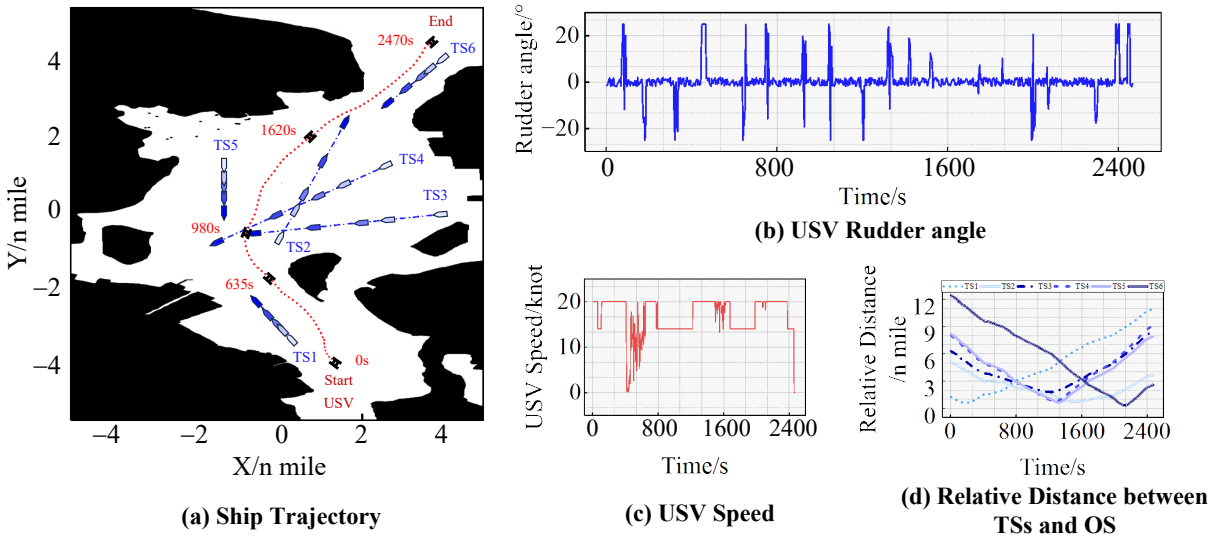
Ship	Start Position (n Mile)	Goal Position (n Mile)	Speed (Knot)
USV	(1.5, −1.7)	(2.3, 3.9)	0–20
TS1	(3, −2)	(1.7, −0.1)	5.02
TS2	(1, 0)	(2.3, 1.3)	4.01
TS3	(1, 1)	(2.2, 3.1)	5.28
TS4	(2, 1)	(1, −1)	4.88
TS5	(1, 3)	(1.4, 1)	4.45

Following the condition initialization, TS1, TS2, TS3 and the USV formed a four-ship encounter situation. Firstly, the USV steered with a 25° rudder angle to alter her course to the starboard side for avoiding TS2 and TS3. When the collision with TS2 and TS3 became clear, the USV, along with TS4 and TS5, formed a three-ship-encountering situation, steering with a 25° rudder angle to alter her course to the starboard side for the purpose of avoiding TS4 at 710 s, ultimately reaching the intended goal.

In this experiment, TS1 and TS2 formed a crossing give-way encounter situation with the USV; the USV passed on the port side of TS2 but on the bow side of TS1. This is because altering her course to starboard side to pass behind TS1 would elevate the risk of collision due to TS1's location at the starboard aft of the USV. Consequently, the USV chose a more cautious decision-making behavior. Moreover, an overtaking situation was formed between the USV and TS4, while a crossing stand-on situation was formed between the USV and TS5. To prevent collision with TS4 and TS5, the USV altered her course to starboard side. These actions exemplify the USV's commitment to compliance with COLREGs as closely as possible in a six-ship encounter situation.

4.3.4. Seven-Ship Encounter Situation Experiment in Complex Waters

In this study, the coastal sea area of Zhanjiang City is selected as the experimental environment for a seven-ship encounter situation experiment in a complex sea environment. The initialization of the USV's navigation conditions is presented in Table 6, while the corresponding experimental results are illustrated in Figure 12.



(a) Ship Trajectory **(b) USV Rudder angle** **(c) USV Speed** **(d) Relative Distance between TSs and OS**

Figure 12. The results of the seven-ship encounter situation experiment in complex waters.

Table 6. Initial settings of the seven-ship encounter situation experiment in complex waters.

Ship	Start Position (n Mile)	Goal Position (n Mile)	Speed (Knot)
USV	(1.32, −4)	(3.64, 4.57)	0–20
TS1	(0.27, −3.33)	(−0.63, −2.32)	1.97
TS2	(0, −0.63)	(1.58, 2.48)	5.08
TS3	(4, 0)	(−0.67, −0.55)	6.85
TS4	(2.65, 1.34)	(−1.59, −0.81)	6.93
TS5	(−1.33, 1.34)	(−1.33, 0)	1.95
TS6	(4, 4.22)	(2.66, 3)	2.64

Following the initialization, TS1 and the USV represented an immediate danger, encouraging the USV to steer with a 25° rudder angle to alter her course and sail along the

starboard side of TS1. Subsequently, at 800 s, the USV was successfully out of immediate danger. At 980 s, a five-ship encounter situation was formed, involving TS2, TS3, TS4, TS5 and the USV. During this encounter, the USV first opted to pass between TS4 and TS5 while sailing along the port side of TS2. Ultimately, the USV approached the starboard side of TS6 to avoid the collision and continued towards its intended goal.

It is worth noting that the USV chose to navigate through the middle of TS4 and TS5 rather than the starboard side of TS2. This decision was based on the relatively close distance between TS3 and TS4 when located on the starboard side of the USV, which would increase the risk of collision caused by an instruction of starboard rudder operation. Consequently, the USV chose a more cautious decision-making behavior. This is evidence that the USV autonomous decision-making very closely complies with the COLREGs in a seven-ship-encountering situation.

4.4. Comparative Experiment

In this chapter, to validate the superiority of the proposed approach, we selected DWA [34] and APF [35] as representative traditional autonomous decision-making methods for the USV. A comparative experiment on success rates was conducted using the DWA algorithm, the APF algorithm, and models trained by the DDPG algorithm, the PPO algorithm, the LSTM-PPO algorithm, and the GRU-PPO algorithm. The experiments were performed in fifty diverse restricted waters. Within each restricted water scenario, three test scenarios were established to assess the algorithms' decision-making capability in multi-ship situations involving encounters, including two, four, and six TSs, respectively. Success was defined as reaching the goal while failure was determined by collisions or local optimality traps. The success rate of each algorithm was individually calculated and presented in Table 7 and Figure 13.

Table 7. Comparative experiments of success rate.

TS Number	Number of Successes/Success Rate					
	DRL Algorithms				Traditional Algorithms	
	DDPG	PPO	PPO-LSTM	PPO-GRU	DWA	APF
2	45/0.90	49/0.98	50/1	50/1	41/0.82	40/0.80
4	34/0.68	40/0.80	45/0.90	49/0.98	21/0.42	17/0.34
6	22/0.44	37/0.74	41/0.82	45/0.90	8/0.16	11/0.22

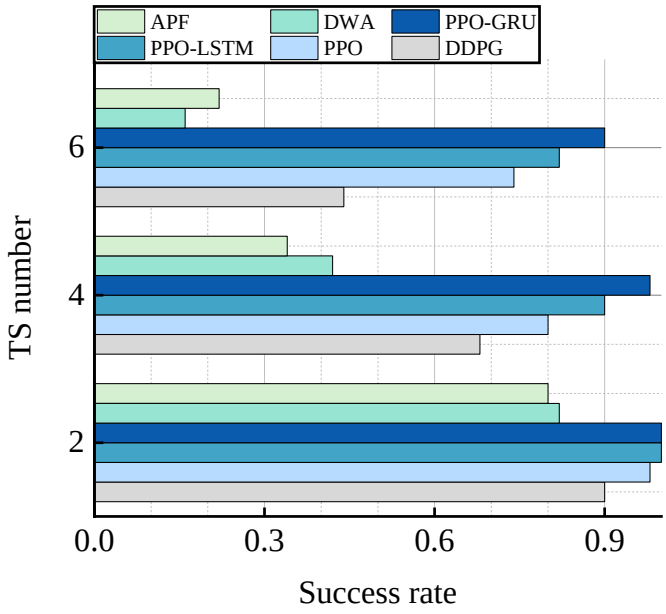


Figure 13. Success rate of different algorithms.

The experimental result demonstrates the superiority of the DRL algorithms over traditional algorithms in terms of overall performance. When the number of TS is two, both the PPO-LSTM algorithm and the PPO-GRU algorithm exhibit a remarkable success rate of 100%. With an increase in the number of TS, although there is a decline in the success rate of the PPO-GRU algorithm, it still outperforms other DRL algorithms and traditional algorithms by a significant margin. These findings highlight that the proposed PPO-GRU algorithm exhibits robust adaptability and decision-making capabilities within restricted waters.

5. Conclusions

This study proposes an autonomous collision avoidance decision-making method based on the improved PPO algorithm, which effectively enhances the learning rate and convergence efficiency of the algorithm by incorporating the GRU network. Also, to enhance the rationality of collision avoidance behavior, this study integrates the concepts of ship domain and action area into the decision-making mechanism and establishes a novel set of reward functions. These functions not only strictly comply with Article 13–17 of Chapter II of the COLREGs, but also consider path optimization for collision avoidance and immediate danger situations. Finally, the simulation results validate the compliance of the trained model's collision avoidance behavior with the COLREGs. Furthermore, a variety of restricted sea areas were selected to comprehensively assess the adaptability and practical application potential of the proposed autonomous collision avoidance decision-making method across varying levels of environmental complexity.

Also, this study has certain limitations that need to be addressed for future research. Firstly, the consideration of autonomy in TSs should be incorporated to enhance their intelligence. Secondly, the impact of different training environments and the quantities of TSs during the training process on the training results should be taken into account in subsequent studies. Lastly, it is recommended to include more environment disturbance in future experiments to simulate a more realistic marine environment.

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

Funding: The paper is partially supported by National Natural Science Foundation of China (No. 51409033, 52171342), Dalian Innovation Team Support Plan in the Key Research Field (2020RT08), and 2023 DMU navigation college first-class interdisciplinary research project, 2023JXA03.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data available on request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Wang, Y.; Xu, H.X.; Feng, H. Deep reinforcement learning based collision avoidance system for autonomous ships. *Ocean Eng.* **2024**, *292*, 19. [CrossRef]
2. EMSA. *Annual Overview of Marine Casualties and Incidents 2021*; EMSA: Lisbon, Portugal, 2021.
3. He, Z.B.; Chu, X.M.; Liu, C.G. A novel model predictive artificial potential field based ship motion planning method considering COLREGs for complex encounter scenarios. *ISA Trans.* **2023**, *134*, 58–73. [CrossRef] [PubMed]
4. Cui, Z.W.; Guan, W.; Zhang, X. Collision avoidance decision-making strategy for multiple USVs based on Deep Reinforcement Learning algorithm. *Ocean Eng.* **2024**, *308*, 15. [CrossRef]

5. Inan, T.; Baba, A.F. Building a hybrid algorithm based decision support system to prevent ship collisions. *J. Fac. Eng. Archit. Gazi Univ.* **2020**, *35*, 1213–1230.
6. Cheng, X.; Liu, Z.Y.; Soc, I.C. Trajectory optimization for ship navigation safety using genetic annealing algorithm. In Proceedings of the 3rd International Conference on Natural Computation, Haikou, China, 24–27 August 2007.
7. Li, Y.M.; Ma, Y.H.; Cao, J. An Obstacle Avoidance Strategy for AUV Based on State-Tracking Collision Detection and Improved Artificial Potential Field. *J. Mar. Sci. Eng.* **2024**, *12*, 695. [CrossRef]
8. Charalambopoulos, N.; Xidias, E.; Nearchou, A. Efficient ship weather routing using probabilistic roadmaps. *Ocean Eng.* **2023**, *273*, 15. [CrossRef]
9. Zaccone, R.; Martelli, M. A collision avoidance algorithm for ship guidance applications. *J. Mar. Eng. Technol.* **2020**, *19*, 62–75. [CrossRef]
10. He, Z.B.; Liu, C.G.; Chu, X.M. Dynamic anti-collision A-star algorithm for multi-ship encounter situations. *Appl. Ocean Res.* **2022**, *118*, 16. [CrossRef]
11. Yuan, X.Y.; Tong, C.C.; He, G.X. Unmanned Vessel Collision Avoidance Algorithm by Dynamic Window Approach Based on COLREGs Considering the Effects of the Wind and Wave. *J. Mar. Sci. Eng.* **2023**, *11*, 1831. [CrossRef]
12. Arul, S.H.; Manocha, D. V-RVO: Decentralized Multi-Agent Collision Avoidance using Voronoi Diagrams and Re-ciprocal Velocity Obstacles. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Prague, Czech Republic, 27 September–1 October 2021.
13. Sun, Z.R.; Lei, B.S.; Xie, P.J.; Liu, F.G. Multi-Risk-RRT: An Efficient Motion Planning Algorithm for Robotic Autonomous Luggage Trolley Collection at Airports. *IEEE Trans. Intell. Veh.* **2024**, *9*, 3450–3463. [CrossRef]
14. Votion, J.; Cao, Y.C. Diversity-Based Cooperative Multivehicle Path Planning for Risk Management in Costmap Environments. *IEEE Trans. Ind. Electron.* **2019**, *66*, 6117–6127. [CrossRef]
15. Chen, C.; Chen, X.Q.; Ma, F. A knowledge-free path planning approach for smart ships based on reinforcement learning. *Ocean Eng.* **2019**, *189*, 106299. [CrossRef]
16. Zhao, Y.J.; Qi, X.; Ma, Y. Path Following Optimization for an Underactuated USV Using Smoothly-Convergent Deep Reinforcement Learning. *IEEE Trans. Intell. Transp. Syst.* **2021**, *22*, 6208–6220. [CrossRef]
17. Du, Y.Q.; Zhang, X.G.; Cao, Z.Y. An Optimized Path Planning Method for Coastal Ships Based on Improved DDPG and DP. *J. Adv. Transp.* **2021**, *2021*, 23. [CrossRef]
18. Chun, D.; Roh, M.I.; Lee, H.W. Deep reinforcement learning-based collision avoidance for an autonomous ship. *Ocean Eng.* **2021**, *234*, 20. [CrossRef]
19. Zheng, K.J.; Zhang, X.Y.; Wang, C.B. Adaptive collision avoidance decisions in autonomous ship encounter scenarios through rule-guided vision supervised learning. *Ocean Eng.* **2024**, *297*, 15. [CrossRef]
20. Chun, D.H.; Roh, M.I.; Lee, H.W. Method for collision avoidance based on deep reinforcement learning with path-speed control for an autonomous ship. *Int. J. Nav. Archit. Ocean Eng.* **2024**, *16*, 19. [CrossRef]
21. Guan, W.; Han, H.S.; Cui, Z.W. Autonomous navigation of marine surface vessel in extreme encounter situation. *J. Mar. Sci. Technol.* **2024**, *29*, 167–180. [CrossRef]
22. Meyer, E.; Heiberg, A.; Rasheed, A. COLREG-Compliant Collision Avoidance for Unmanned Surface Vehicle Using Deep Reinforcement Learning. *IEEE Access* **2020**, *8*, 165344–165364. [CrossRef]
23. Wang, Y.Y.; Chin, H.C. An Empirically-Calibrated Ship Domain as a Safety Criterion for Navigation in Confined Waters. *J. Navig.* **2016**, *69*, 257–276. [CrossRef]
24. Dinh, G.H.; Im, N.-K. The combination of analytical and statistical method to define polygonal ship domain and reflect human experiences in estimating dangerous area. *Int. J. e-Navig. Marit. Econ.* **2016**, *4*, 97–108. [CrossRef]
25. Liu, Z.X.; Zhang, Y.M.; Yu, X. Unmanned surface vehicles: An overview of developments and challenges. *Annu. Rev. Control* **2016**, *41*, 71–93. [CrossRef]
26. Do, K.D.; Pan, J. *Control of Ships and Underwater Vehicles: Design for Underactuated and Nonlinear Marine Systems*; Spinger: London, UK, 2009; pp. 42–44.
27. Fossen, T.I. *Handbook of Marine Craft Hydrodynamics and Motion Control*; John Wiley & Sons: Chichester, UK, 2011; pp. 15–44.
28. Eriksen, B.-O.H.; Bitar, G.; Breivik, M. Hybrid Collision Avoidance for ASVs Compliant with COLREGs Rules 8 and 13–17. *Front. Robot. AI* **2020**, *7*, 11. [CrossRef]
29. Rongcai, Z.; Hongwei, X.; Kexin, Y. Autonomous collision avoidance system in a multi-ship environment based on proximal policy optimization method. *Ocean Eng.* **2023**, *272*, 113779. [CrossRef]
30. Cui, Z.W.; Guan, W.; Zhang, X.K. USV formation navigation decision-making through hybrid deep reinforcement learning using self-attention mechanism. *Expert Syst. Appl.* **2024**, *256*, 124906. [CrossRef]
31. Bingham, B.; Aguero, C.; Mccarrin, M.; Klamo, J. Toward Maritime Robotic Simulation in Gazebo. In Proceedings of the Oceans 2019 MTS/IEEE Seattle, Seattle, WA, USA, 27–31 October 2019.
32. Lillicrap, T.P.; Hunt, J.J.; Pritzel, A. Continuous control with deep reinforcement learning. *Comput. Ence* **2015**, *1509*, 02971.
33. Cui, Z.; Guan, W.; Zhang, X. Autonomous Navigation Decision-Making Method for a Smart Marine Surface Vessel Based on an Improved Soft Actor–Critic Algorithm. *J. Mar. Sci. Eng.* **2023**, *11*, 1554. [CrossRef]

34. Guan, W.; Wang, K. Autonomous Collision Avoidance of Unmanned Surface Vehicles Based on Improved A-Star and Dynamic Window Approach Algorithms. *IEEE Intell. Transp. Syst. Mag.* **2023**, *15*, 36–50. [CrossRef]
35. Lyu, H.G.; Yin, Y. COLREGs-constrained real-time path planning for autonomous ships using modified artificial potential fields. *J. Navig.* **2019**, *72*, 588–608. [CrossRef]