COLERGs-constrained safe reinforcement learning for realising MASS's risk-informed collision avoidance decision making

Chengbo Wang [a, b, e], Xinyu Zhang [b, *], Hongbo Gao [a, c, d, *], Musa Bashir [e], Huanhuan Li [e], Zaili Yang [e]

[a] Department of Automation, School of Information Science and Technology, University of Science and Technology of China, Hefei, China

[b] Navigation College, Dalian Maritime University, Dalian, China

[c] Institute of Advanced Technology, University of Science and Technology of China, Hefei, China

[d] Nanyang Technological University, 639798, Singapore

[e] Liverpool Logistics, Offshore and Marine Research Institute (LOOM), Liverpool John Moores University, Liverpool, UK

Abstract: Maritime autonomous surface ship (MASS) represents a significant advancement in maritime technology, offering the potential for increased efficiency, reduced operational costs, and enhanced maritime traffic safety. However, MASS navigation in complex maritime traffic and congested water areas presents challenges, especially in Collision Avoidance Decision Making (CADM) during multi-ship encounter scenarios. Through a robust risk assessment design for time-sequential and joint-target ships (TSs) encounter scenarios, a novel risk and reliability critic-enhanced safe hierarchical reinforcement learning (RA-SHRL), constrained by the International Regulations for Preventing Collisions at Sea (COLREGs), is proposed to realize the autonomous navigation and CADM of MASS. Finally, experimental simulations are conducted against a time-sequenced obstacle avoidance scenario and a swarm obstacle avoidance scenario. The experimental results demonstrate that RA-SHRL generates safe, efficient, and reliable collision avoidance strategies in both time-sequential dynamic obstacles and mixed joint-TSs environments. Additionally, the RA-SHRL is capable of assessing risk and avoiding multiple joint-TSs. Compared with Deep Q-network (DQN) and Constrained Policy Optimization (CPO), the search efficiency of the algorithm proposed in this paper is improved by 40% and 12%, respectively. Moreover, it achieved a 91.3 % success rate of collision avoidance during training. The methodology could also benefit other autonomous systems in dynamic environments.

Keywords: Collision avoidance decision-making, Collision risk, Safe reinforcement learning (SRL), COLREGs, Maritime autonomous surface ship (MASS).

* Corresponding author.
E-mail addresses: zhangxy@dlmu.edu.cn (X. Zhang), ghb48@ustc.edu.cn (H. Gao).

Nomenclature	
AI	Artificial Intelligence
AIS	Automatic Identification System
APF	Artificial Potential Field
CADM	Collision Avoidance Decision Making
COLREGs	Convention on the International Regulations for Preventing Collisions at Sea
CPA	Closest Point of Approach
CPO	Constrained Policy Optimization
CRF	Conditional Random Field
DCPA	Distance at the Closest Point of Approach
DL	Deep Learning
DQN	Deep Q-network
DT	Dispersion across Time
DWA	Dynamic Window Approach
FL	Fuzzy logic
IMO	International Maritime Organization
MASS	Maritime Autonomous Surface Ships
MDP	Markov Decision Process
ML	Machine learning
RA-SHRL	Safe Reinforcement Learning with Hierarchical Framework
RL	Reinforcement Learning
SRL	Safe Reinforcement Learning
TCPA	Time to the Closest Point of Approach
TSs	Target Ships
UG	Undirected Graphs
VO	Velocity Obstacle

1. Introduction

A common strategy used to investigate maritime transportation safety involves the study of autonomous navigation technology coupled with intelligent collision avoidance [1]. To facilitate the supervision and regulations of autonomous ships and smart vessels, these autonomous ships are collectively referred to as maritime autonomous surface ship (MASS) and characterized by their ability to operate independently to varying degrees without human intervention, as defined by the International Maritime Organization (IMO) [2-4]. In the initial phase of the regulatory scoping, MASS is categorized into four degrees of autonomy (i.e., ship with automated processes and decision support, remotely controlled ship with seafarers on board, remotely controlled ship without seafarers on board, and fully autonomous ship), which is non-hierarchical [5]. This classification system has been widely adopted in the fields of maritime transportation and ocean engineering.

In recent years, artificial intelligence and autonomous technology have had a significant impact on various domains, especially in the fields of unmanned ground vehicle (UGV) [6,7], unmanned aerial vehicle (UAV) [8,9] and robotics [10]. As opposed to these unmanned vehicles, MASS autonomous navigation technology is an integration of multiple technologies, including intelligent perception, global route optimization, intelligent collision avoidance decision-making, and intelligent control [11]. The navigation environment for MASS is inherently unstructured, encompassing not only surface navigation but also complex scenarios such as port operations and maritime traffic management. However, MASS needs to plan a safe trajectory from their starting point to the target point. Meanwhile, they are required to make real-time decisions to avoid obstacles when encountering other ships, referred to as target ship (TS), and when there is a risk of collision or imminent danger.

To realize the collision avoidance decision-making (CADM), various methods have been presented. These methods include the geometric collision avoidance method, which draws inspiration from migrating radar plotting principles [12], and methods adapted from field theory in the robotics domain, such as velocity obstacle (VO) and artificial potential field (APF)) [13,14]. Additionally, there are data-driven approaches involving deep learning for collision avoidance behavior feature extraction [15,16], as well as human-like methods represented by deep reinforcement learning (DRL) [17-20]. However, these methods have revealed some limitations in their applications as they heavily rely on the skills of the OOW, making it challenging to ensure collision avoidance in all situations. This challenge becomes even more pronounced in emergency scenarios, particularly in congested waters, disclosing a significant research problem to be addressed with urgency.

DRL models demonstrate remarkable adaptability by learning from their interactions with the environment to make real-time decisions in dynamic maritime settings. Nevertheless, existing DRL-based CADM research often overlooks the coupling effects among ships in crowded maritime areas. Several studies have focused on sequential obstacle avoidance. However, in cases where the distance between two ships is extremely close, it is advisable to treat them as a single entity, leading to coordinated decisions that ensure safe avoidance. Thus, the primary theoretical research question guiding this investigation is whether the joint-TSs obstacle avoidance is applicable in MASS encounter scenarios.

Following the introduction presented in Section 1, this paper is organized as follows: Section 2 provides a literature review of collision risk assessment and CADM approaches. Section 3 describes the methodology and model developed for risk assessment-based SRL and CADM with reference to the configured research problem. Section 4 presents the experimental results and analysis for testing and verification of the new methodology as presented in Section 3. Finally, Section 5 concludes the paper with recommendations for future directions.

2. Literature review

2.1. Collision risk assessment

Research on risk assessment has always been closely linked to CADM. To ensure the safety of intelligent decision-making, collision risk assessment is widely adopted in the decision-making system for MASS research [21,22]. The current focus of collision risk is mainly categorized into two areas: (1) macro-maritime traffic risk assessment or grounding risk based on Automatic Identification System (AIS) data [23-25] and (2) microcosmic collision risk assessment targeting individual ships [26]. While macro-maritime traffic risk assessment is more suitable for the safety of a specific area to justify maritime traffic regulation and/or long-duration mission planning, microcosmic collision risk assessment is better suited to the requirements of ship CADM research. The hierarchical solution for collision risk assessment in CADM problems consists of two key steps: (1) evaluating the collision risks of TSs surrounding the investigated MASS, and (2) developing a sequential anti-collision strategy for the MASS based on the results of the collision risk evaluation. The classic collision risk research methods are mainly Closest Point of Approach (CPA)-based and domain-based.

In the initial phase of research, the CPA-based method was proposed to calculate the collision risk. This approach was inspired by the availability of the radar plotting capability, which evaluates collision risk by calculating the distance to the closest point (for own ship) and time to the closest point (for TSs) between the MASS and TSs, as shown in Fig. 1 [27,28]. The CPA method is the most established in the ship collision risk analysis field. However, with the increasing demand for water traffic, a single Distance at the Closest Point of Approach (DCPA) and Time to the Closest Point of Approach (TCPA) may no longer meet the requirements of risk assessment accuracy, as evidenced by Kang et al. [29], Zhang et al. [21], and Wang et al. [13], in which more factors beyond DCPA and TCPA, including distance, bearing, and speed ratio, need to be considered to improve ship collision risk analysis results and determine the minimum safe encounter distance.

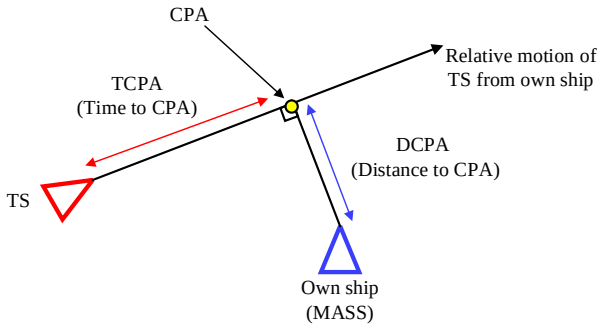


Fig. 1. Concepts of CPA.

In 1975, Goodwin, E.M. proposed the concept of a ship domain [30], and Davis et al. first introduced the concept of anti-collision arenas by improving the ship domain [31]. With the widespread applications of the ship domain model, domain-based ship collision risk has gained the attention of mainstream researchers in the field [32-34]. The aforementioned research provides a useful framework for calculating collision risk and offers a theoretical foundation for solving the CADM problem.

2.2. Collision avoidance decision-making

In the past several decades, CADM research has been a prominent topic in maritime transportation. There are two major theoretical and conceptual frameworks for CADM: the rule-based method and learning-based method. Most rule-based methods are however focused on explaining COLREGs or knowledge obtained from AIS data and integrating them into traditional CADM method frameworks [35-377]. Nevertheless, maritime environmental conditions and scenarios involving multi-ship encounters are complex and changeable. This complexity leads to numerous challenges related to scenario adaptability and avoidance strategy applicability. This problem may seem more straightforward when compared to the learning-based method. Consequently, an increasing number of researchers focus on the application of learning-based methods in CADM problem [38-40]. The most representative ones are reinforcement learning (RL) and deep learning (DL). With the aid of RL and DL methods, there have been several noteworthy studies addressing the problem of collision avoidance decision-making for MASS. Xu et al. [41] presented an intelligent hybrid collision avoidance algorithm using the DRL method combined with a risk assessment model based on the DCPA and TCPA. Jiang et al. [18] proposed a novel decision-making model with an attention-mechanism based on DRL, including ship collision risk assessment based on driver's attention distribution. Chun et al. [19] put forward a combination method of collision risk assessment based on CPA and collision avoidance path generation based on DRL. However, these methods are not without their challenges, particularly when it comes to fully addressing collision resolution in scenarios with multiple densely packed vessels. This is a complex task, requiring sequential temporal risk assessments and evasive maneuvers, especially in busy waterways.

The summary of the aforementioned relevant methods, COLREGs constraint, obstacles (i.e., static and dynamic), scenarios, collision avoidance strategy, risk assessment, objective function, computational time, and reliability analysis is presented in Table 1. The relevance of each factor is categorized as no, low, medium, or high. This overview highlights that current DRL-based methods primarily focus on ensuring safety through objective functions and rule constraints. However, the need to incorporate and address reliability analysis underscores a research gap in this study.

Table 1. Summary of related literature review.

Methods	FL	VO	APF	DWA	Learning-based method			
Author(s)	Perera rt al. [36]	Wang et al. [13]	Lyu et al. [14]	Guan et al. [37]	Chun et al. [19]	Jiang et al. [18]	Wang et al. [17]	This paper
COLREGs constraint	high①	med	high	med	med	med	high	high
Static obstacles (SOs)	no	no	circles	polygons	polygons	multiple circles	coastlines, polygons	multiple islands with circles②
Dynamic obstacles (TSs)	1	1~5	multiple	1	multiple	1~4	3	multiple⑥
Scenarios	low	high	med	low	med	med	med	high
CA strategy	sequence	sequence	sequence	sequence	sequence	sequence	sequence	sequence and joint
Risk assessment	ship domain	CPA③	ship domain	no	CPA	attention-mechanism	CPA	CPA, UG
Objective function	free collision	free collision	free collision	free collision	free collision	free collision	free collision	multiple objectives④
Computational time	-	-	very low	-	-	low	med⑤	high
Reliability analysis	no	no	no	no	no	no	no	yes

Note: The symbol "-" indicates that it is not mentioned or noted in the corresponding literature. "med" is the abbreviation of "medium". "CA" is the abbreviation of "Collision avoidance".
① It considered in detail the rules for "Give way" and "Stand on" vessels, and considered situations in which absence or negligence of actions from a "Give way" target vessel.
② The circle is the risk potential field area.
③ They used different factors, including DCPA, TCPA, distance, bearing, and speed radio.
④ Multiple objectives function consists of reaching the target point, collision-free, and efficiency.
⑤ The running time basically depends on scene complexity.
⑥ The complex scenario of intensive joint TSs of two ships is considered.

2.3. Review analysis and innovation

As mentioned before, current research in the literature on collision avoidance problems is mostly divided into risk assessment and CADM methods, with safety being a primary objective function. However, the scenarios concerned in the risk assessment are all based on time-series, and joint obstacle scenarios have yet been sufficiently taken into account. Meanwhile, even though DRL has great potential in solving CADM problems, there is limited research on the reliability of CADM methods based on DRL. It is therefore necessary to develop an adaptive and reliable intelligent DRL algorithm. Although safety parameters and key element values required for CADM can be calculated and analyzed through risk assessment, reliability analysis needs to be obtained through result verification analysis. Therefore, it is feasible to design a novel human-like method that combines safe reinforcement learning, risk assessment, and reliability analysis to address CADM problems comprehensively.

Motivated by the aforementioned research gaps, a novel integrated risk assessment and collision avoidance decision-making for MASS, which incorporates a novel safe reinforcement learning and reliability analysis method, is developed in this study. The main contributions of the study are outlined as follows:

(i) The development of a novel integrated risk assessment and CADM method for MASS, which can take into account both time-sequential and joint obstacle avoidance.

(ii) The creation of a probabilistic collision risk assessment approach involving DCPA, TCPA, risk area, and ship domain factors for joint-TSs encounter scenario, which is verified to be effective in mixed obstacle scenarios.

(iii) A new safety hierarchical reinforcement learning model, enhanced by a hybrid risk assessment and reliability analysis for CADM in uncertain environments. Several extended experiment studies illustrate how hybrid risk assessment and reliability analysis enable and improve the safety, accuracy, and reliability of developing collision avoidance strategies.

3. Problem formulation

It is assumed that the existence of an attention area with a radius of R_a, and a risk of collision is likely to exist in this area. In general, the safety of a MASS is typically determined by whether it can sail through this area without colliding with another nearby TSs ($TS_n, n = 1 \cdots N$) in a time window $[t_0, t_M]$. The formal method of this look-out can be formulated as follows within this time window:

$$\min_{t_0 \le t \le t_M} \left\| p_{MASS}(t) - p_{TS_n}(t) \right\| > \varepsilon_d, \forall n = 1 \cdots N \tag{1}$$

where P_{MASS} and P_{TS_n} are the MASS' and TS_n' positions. ε_d indicates a threshold distance between the MASS and TS_n. N expresses the total number of *TSs* in the attention area of the MASS. t denotes the current time.

After taking an avoidance action to ensure anti-collision, there is an equivalent constraint between threshold distance and remaining distance. The formal method can be updated as:

$$D_n(s_t, a_t) > \varepsilon_d, \quad \forall n = 1 \cdots N \tag{2}$$

where s_t represents the current environment state s at the time t. a_t indicates the action taken by the MASS at the time t. $D_n(s_t, a_t)$ expresses the remaining distance between the MASS and TS_n when the MASS acts a_t and keeps it until final past clear TS_n. Unless a new risk of collision occurs, a new avoidance action needs to be taken again.

Therefore, the entire collision avoidance strategy is obtained using Equation 3:

$$\pi' = \arg\max_{\pi} \sum_{n=1}^{N} D_n(s_t, \pi(s_t)) \tag{3}$$

However, in complex encounter scenarios, multiple TSs may be close to each other. Assessing the risk and determining collision avoidance strategies for each individual ship in real-time can be time-consuming. Moreover, when avoiding one ship TS_1, it may create a new risk with another dynamic ship TS_2. To address this issue, a model is needed to evaluate the overall collision risk of complex obstacle scenarios. And for these close-to-each-other scenarios, collision risks can be combined and avoided simultaneously. Therefore, it is necessary to make a more reliable and safe navigation strategy (Equation 4) to handle the close-to-each-other scenarios.

$$\begin{aligned} \pi^* &= \arg\min_{\pi} J(\pi(s_t)) \\ s.t. \quad & D_{n \cap n+1}(s_t, \pi(s_t)) > R_a, \ \forall n = 1 \cdots N \end{aligned} \tag{4}$$

where J is a judge function of strategy π in the state of the joint probability distribution of risks. $D_{n \cap n+1}$ is the remaining safe threshold distance between the MASS and joint parts between the TS_n with TS_{n+1}.

This judge function J is represented as:

$$\begin{aligned} J &= \left| \pi(s_t) - a_{\exp} \right| \\ s.t. \quad & D_{n \cap n+1}(s_t, \pi(s_t)) > R_a \\ & \forall n = 1 \cdots N \end{aligned} \tag{5}$$

where $a_{\exp}$ is the desired anti-collision action from good seamanship.

4. Methodology

RL is a widely used technique in ocean engineering research, particularly in developing optimal navigation strategies for MASS. Through trial-and-error, RL explores various actions in maritime encounters, optimizing anti-collision behavior according to environmental feedback and eventually training an optimal strategy with the highest cumulative reward expectation. However, during simulation training, unsafe behaviors such as small-angle avoidance and collisions may be explored. This can be dangerous in high-security areas like autonomous ships. Traditional methods such as punishments, negative rewards, action space limitations, and direct termination of unsafe actions or states can prevent some unsafe actions, but they do not solve the problem entirely. To address this problem, SRL techniques can maximize rewards while ensuring safety constraints are met [3,42,43]. In this section, we present a comprehensive design of the risk assessment based CADM approach, which includes risk assessment, and CADM algorithm design based on SRL.

4.1. Risk assessment

4.1.1. Sequential risk assessment of multiple obstacles

In the multiple TSs encounter scenario, we drew inspiration from radar plotting and DCPA/TCPA techniques and constructed safe and risk areas for assessing collision risk. Fig. 2 depicts the risk and safe areas for MASS based on the DCPA and TCPA information. If a TS enters the risk area, it poses a certain collision risk to the MASS. A collision will occur if TSs enter the safe area of the MASS.

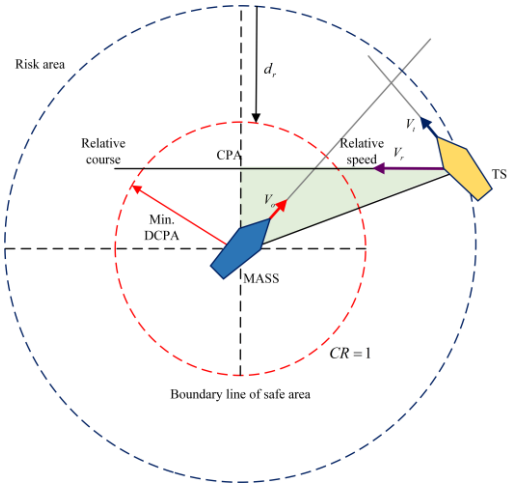


Fig. 2. Illustration of sequential risk assessment.

In the sequential risk assessment model, DCPA, TCPA, and the kinetic energy trend (i.e. the true speed of MASS V_0, true speed of TS V_t, and relative speed V_r) are further considered for the state-of-the-art

assessment. In the previous research [44,45], the coefficient of the relationship between encounter risk assessment and DCPA, TCPA, and the encounter angle is established through mathematical derivations:

$$CR = CR_0 \cdot F_{tcpa} \cdot F_{dcpa} \cdot F_{angle} \tag{6}$$

where CR_0 indicates a criterion risk which symbolizes the MASS's earliest anti-collision event. F_{tcpa}, F_{dcpa}, and F_{angle} are the relative multipliers of TCPA, DCPA, and angle between MASS and TSs.

However, it should be noted that Eq. (6) is not normalized in each encountered scenario. Therefore, this suggests that Eq. (6) can be improved by adopting Eq. (7) [19]:

$$CR = \exp(-DCPA / a) \cdot \exp(-TCPA / b) \tag{7}$$

where a, b indicates the normalization coefficient of DCPA and TCPA.

Consequently, this study introduces the domain theory to improve the method in which standard risks and coefficients of determination by constructing risk areas are calculated. Resulting from Fig. 2, for the case where $TCPA = 0, DCPA = d_r$, a is found to be $-d_r / \ln(CR_0)$; and for the case where $DCPA = 0, TCPA = d_r / V_t$, b is determined to be $-d_r / V_t \cdot \ln(CR_0)$. Thus, Eq. (7) can be converted into Eq. (8):

$$CR_{Seq} = \exp((DCPA + V_t \cdot TCPA) \cdot \ln(CR_0) / d_r) \tag{8}$$

where d_r represents the radius of the risk area between the MASS and TS. V_t expresses the true speed of TS.

This part of the method will be verified in the time-sequenced obstacle avoidance scenario in Section 5.2.1.

4.1.2. Risk assessment in joint-TSs encounter scenario

In complex multi-TSs encounter situations, the method of assessing collision risk in sequence increases computational cost, especially when the distance between two TSs is too close. Furthermore, avoiding one of TSs may increase the collision risk posed by other TSs. The collision avoidance process of MASS adheres to the principles of Markov Decision Process (MDP). While there are probabilistic dependencies among multiple TSs, these relationships do not necessarily indicate causality. To address this challenge, a novel risk joint probability distribution model is developed based on probabilistic undirected graphical models. As illustrated in Fig. 3, two dynamic obstacles, TS1 and TS2, have overlapping risk areas. The probabilistic undirected graphs of TS1 and TS2 are decomposed, including a clique of two nodes and a maximum clique of three nodes. The joint risk probability distribution of the probabilistic undirected graph model can be calculated by multiplying the function of the random variables on the largest clique.

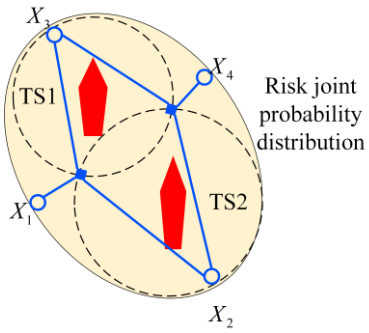


Fig. 3. Principles of joint probability distributions for probabilistic undirected graphical models.

$$Maximum\ clique: C = \{X_1, X_2, X_3\}, \{X_2, X_3, X_4\} \tag{9}$$

The potential function, $\Psi_C(Y_C) \to \mathbb{R}$, for the two largest cliques of the joint probability distribution is [46]:

$$\Psi_C(Y_C) = \exp\{-E(Y_C)\} \tag{10}$$

where Y_C indicates the random variable corresponding to C.

Then, the risk joint probability distribution based on the conditional random field (CRF) is calculated as [46]:

$$P(Y) = \frac{1}{Z}\prod_C \Psi_C(Y_C) \tag{11}$$

where z is a normalization factor, described as follows:

$$Z = \sum_Y \prod_C \Psi_C(Y_C) \tag{12}$$

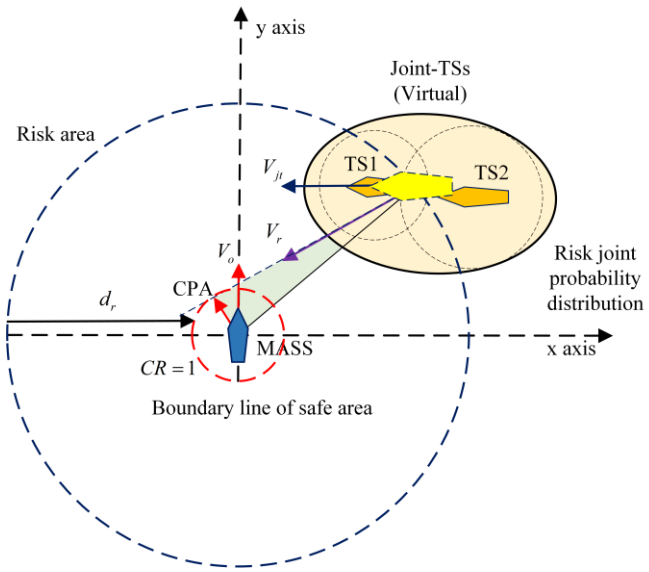


Fig. 4. Illustration of risk assessment in joint-TSs encounter scenario.

As shown in Fig. 4, the proposed risk assessment in joint-TSs encounter scenario is illustrated. By constructing the joint probability distribution of TS1 and TS2 using a CRF, the dotted yellow pentagon represents the corresponding virtual joint-TSs. TS1 and TS2 can be considered as a single overall dynamic obstacle, joint-TSs, for risk assessment and avoidance. The collision risk can be calculated when adjacent TSs are considered as a virtual joint-TSs obstacle by substituting the velocity of the joint-TSs into Eq. (8), Eq. (13) is obtained:

$$CR_{jt} = \exp((DCPA + V_{jt} \cdot TCPA) \cdot \ln(CR_0) / d_r) \tag{13}$$

where V_{jt} represents the speed of joint-TSs.

According to good seamanship and COLREGs, vessels should adopt the most collision risk calculation in the current situation. Therefore, the final collision risk is taking the maximum risk value calculated by Eq. (14):

$$CR = \max\left\{CR_{jt}, P(Y)\right\} \tag{14}$$

Based on the collision risk assessment conducted in various TSs encounter situations, the probability of ship collision in the two examined situations can be estimated. The pseudocode for various TSs encounter situation' collision risk assessment for MASS is summarized in Algorithm 1. This part of the method will be verified in the joint-TSs obstacle avoidance scenario in Section 5.2.2.

Algorithm 1. Pseudocode for collision risk assessment for MASS.

Input: $p_{MASS} \in \mathbb{R}^2$, $p_{TS} \in \mathbb{R}^2$, V_0, V_t, d_r, CR_0

Output: risk CR, the likelihood of collision $P(Y_C)$

1: **function** assess _risk (p_{MASS}, p_{TS}, V_t, d_r, CR_0)

2: j = judge _scenario ($p_{TS_n}, \forall n \in 1, \cdots, N$)

3: **if** j = Sequential scenario

4: $CR = \exp((DCPA + V_t \cdot TCPA) \cdot \ln(CR_0) / d_r)$

5: **elif** j_{TSs} = Scenario joint-TSs

6: $CR = \max\left\{CR_{jt}, P(Y)\right\} \leftarrow$ Eq. (9)

7: end if

8: **return** CR, $P(Y_C)$

9: **end function**

4.2. Collision avoidance decision-making based on SRL

In this section, a novel CADM method is developed based on SRL to achieve reliable and safe autonomous sailing by MASS. Unlike existing methods, the new approach introduces risk constraints into a reward function and an objective function and uses a neural network with reliability analysis to optimize collision avoidance actions.

4.2.1. State space

In the environmental state space described in this paper, the MASS perceives the relative orientation and position between itself and obstacles in real time through the sensors model. As shown in Fig. 5, this is the obstacle detection model for the MASS, where P_{MASS} represents the MASS's position; P_{target} indicates the target point; v_0 expresses the MASS's speed; φ_0 is the heading; P_{TS} denotes the obstacle's position; δ_0 indicates the relative orientation of the obstacle to the MASS; represents the distance from the MASS D_{A-P} to the target point; and D_{A-T} is the distance between the MASS and the obstacle.

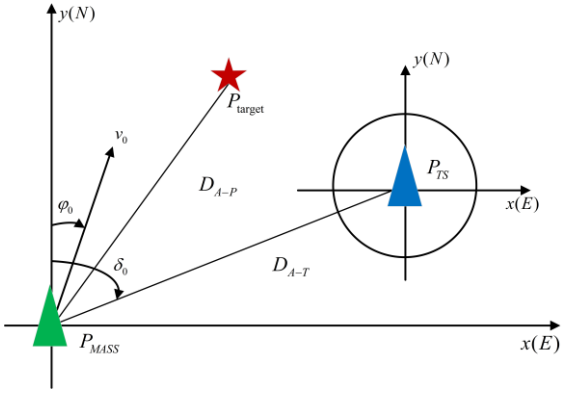


Fig. 5. Obstacle detection model.

Based on the obstacle detection model, the current environmental state space information perceived by the MASS can be represented as follows.

$$s_t = [v_0, \varphi_0, \delta_0, D_{A-P}, D_{A-T}]^T \tag{15}$$

4.2.2. Action space

According to the requirements of navigation practice and collision avoidance rules, the avoidance method based on turning is selected for MASS obstacle avoidance. Therefore, this paper adopts the compass point method to construct the collision avoidance action space.

$$A=\{E,\ \cdots,\ NE,\ \cdots,\ NNE,\ \cdots,\ N/E,\ \cdots,\ NW/N,\ N/W\} \tag{16}$$

where E,S,W,N represents the east, south, west, and north, separately.

4.2.3. Reward function design

To improve the consistency of the optimal learning strategy, this section proposes a reward function designed from two perspectives: the external revenue signal and the internal incentive signal. The external revenue signal is usually generated through interactions with the navigation environment, which can reflect the impact of the MASSs' behavior on the maritime navigational environment and are the main indicator of collision avoidance performance. The internal incentive signal can motivate the MASSs to try exploring unknown navigation states and anti-collision actions during the learning process, thereby promoting learning effectiveness.

In this study, external revenue signal rewards are categorized primarily into position reward, speed reward, target reward, and shaping reward. Position reward r_{pos} incentivizes the MASS based on its location within the encounter scenario. This serves to encourage MASS to steer clear of potential collisions with other TSs, thereby guiding it away from risk zones [17].

$$r_{pos}=\begin{cases}0 & if\ \ CR\le CR_0\\ \dfrac{1}{d_r}CR & if\ \ CR_0\le CR\le 1\\ CR & if\ \ CR\ge 1\end{cases} \tag{17}$$

where d_r denotes the radius of the risk area between the MASS and TS.

The speed reward r_v primarily assesses the stability of MASS during the avoidance process, taking into account factors such as the relative speed against the TS, variations in the heading angle of MASS, and changes in the rudder angle [47].

$$r_v=-\omega_{rv}|\Delta v_{rv}|+\omega_{heading}\Delta v_{heading}{}^2-\delta \tag{18}$$

where Δv_{rv} denotes the relative speed variation between the MASS and TS. $\Delta v_{heading}$ indicates the heading variation of the MASS. ω_{rv} is an influence coefficient, while $\omega_{heading}$ represents the importance of relative speed and heading change to stability, reflecting rewards in terms of steering rate. δ expresses a constant.

The target reward r_{target} primarily aims to expedite MASS in swiftly searching and navigating toward the target point, preventing the algorithm from inducing a circular behavior around the target point and ensuring more effective navigation [47].

$$r_{\text{target}} = 2^{\frac{1}{d_{\text{target}}}} - 2^{\frac{1}{d_{TS}}} \tag{19}$$

where d_{target} is the relative distance between the MASS and the target point, d_{TS} is the relative distance between the MASS and TS.

"Shaping rewards" is a concept in reinforcement learning, which refers to modifying the original reward structure or adding additional reward signals in order to guide the agent to learn effective strategies faster [48]. To tackle the issue of suboptimal initial performance and sluggish convergence speed of RL algorithms in continuous multi-target scenarios, this paper incorporates an Artificial Potential Field (APF) function to formulate the shaping reward r_{shaping}. This serves to guide MASS toward a particular behavior or strategy.

$$r_{\text{shaping}} = \gamma\Phi(s_{t+1}) - \Phi(s_t) \tag{20}$$

where $\Phi(s_t)$ represents the strength of the gravitational field from the target point to MASS at state s_t. γ is discount factor.

Therefore, the external revenue signal reward r_{t+1}^{e} can be calculated as follows:

$$r_{t+1}^{e} = r_{pos} + r_{v} + r_{\text{target}} + r_{\text{shaping}} \tag{21}$$

Internal incentive signals r_{t+1}^{i} are employed to enhance the exploration of self-reinforcement learning [49]. When selecting a random action, environment state signal feedback is utilized to bolster its target search capability and incentivize MASS to explore environments with a lower probability of state transitions.

$$r_{t+1}^{i} = \tau\left[1 - f^{a_t}(s_{t+1}|s_t)\right] \tag{22}$$

where f denotes the environment state transition probability.

Therefore, the total reward can be solved as Eq. (25):

$$R = r_{t+1}^{e} + r_{t+1}^{i} \tag{23}$$

4.2.4 Collision avoidance strategy objective function

To ensure reasonable reliability, safety, or meet specific constraints, a hierarchical SRL algorithm is formulated to divide each collision avoidance policy into sub-objective policies. The goal of SRL is to find an avoidance strategy under risk and reliability constraints that can achieve the maximum expected reward, defined as function, throughout the entire sailing process [50]:

$$Q^{\pi}(s,a) = \mathbb{E}^{\pi}\left[\sum_{t=0}^{\infty}\gamma^{t} r_t \,|\, s_0 = s, a_0 = a\right] \tag{24}$$

where s denotes the state, a indicates the action at time t, γ is discount factor.

The optimal $Q*$ is:

$$Q*(s,a) = \max_{\pi} Q^{\pi}(s,a) \tag{25}$$

A multi-objective SRL with a hierarchical structure is put forward. Fig. 6 illustrates the SRL algorithm with a hierarchical structure. The system collects various types of data, including the current environmental state S_t, rewards R_t, and risks of collision CR_t, and feedback data from a performance evaluation model. This data is then used to train the algorithm and enable it to learn. At the high-level, the objective with the largest cumulative return is divided into multiple sub-objective functions (e.g., safety $Q_t^{(1)}$, efficiency $Q_t^{(2)}$, reliability $Q_t^{(3)}$). One innovation we need to clarify here is that the evaluation advantages of DQN are combined in the Actor-Critic network. Specifically, the Q-value and E-value estimation methods similar to DQN are introduced in the Critic part of Actor-Critic to improve the performance of critic. Three sub-objective functions are introduced into the Actor-Critic network to discretize and update the critic's estimated value to improve the efficiency and overall performance of the algorithm.

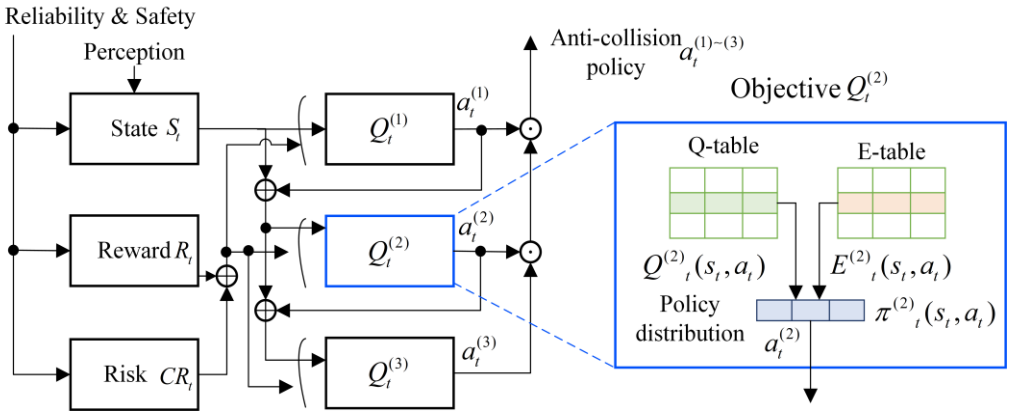


Fig. 6. Illustration of the SRL algorithm with hierarchical structure.

Specifically, the safety and reliability performance evaluation are introduced into the SRL network through a feedback channel. The evaluation values of safety and reliability are updated as follows:

$$\mathbf{E}_t(s_t, a_t^{(1)\sim(3)}) = \{\lambda_{CR}E(CR), \lambda_{DT}E(DT), \lambda_{LRT}E(LRT)\} \tag{26}$$

where the discount rate $\lambda_{CR}, \lambda_{DT}, \lambda_{LRT} \in [0,1]$ represents the impact ratio of risk, dispersion across time and long-term risk across time, separately.

Three sub-objective functions were designed, including reaching the target point, collision-free success rate, and efficiency. Then, the outputs of each objective function are integrated with their corresponding reward functions to create a reliable objective fusion and decision-making mechanism. The interests of multiple

objectives are linearly integrated to determine the collision avoidance actions corresponding to the optimal balance point.

$$F(Q) = \omega^{(1)}Q^{(1)} + \omega^{(2)}Q^{(2)} + \omega^{(3)}Q^{(3)} \tag{27}$$

where F is the objectives fusion function. $\omega^{(1)}$, $\omega^{(2)}$, $\omega^{(3)}$ indicate the fusion parameters of these corresponding three objectives, respectively.

4.3. Performance assessment metrics

In this study, the following performance indicators are mainly used to evaluate the performance of deep reinforcement learning models in terms of safety, convergence, and reliability. Table 2 is a summary of a performance assessment metrics.

(1) Safety performance: In this paper, the concept of risk generation analysis is mainly borrowed to evaluate safety performance by combining collision risk and collision rate.

(2) Convergence performance: In this paper, the convergence performance of the algorithm is reflected by the changing trend of the step size and reward with the epoch.

(3) Reliability performance: This paper uses the expected value error of reward and the number of collision risk interactions to evaluate the reliability of the proposed algorithm. The number of collision risk interactions mainly reflects the stability of the algorithm's decision-making. The error in the expected value of reward mainly reflects the consistency of the strategy.

Table 2. Summary of performance assessment metrics.

Performance	Indicators
Safety	Risk of collision *CR*
Convergence	Step-epoch
	Reward
Reliability	Reward expectation error
	Collision risk interactions

4.4. Integrated framework and algorithm implementation

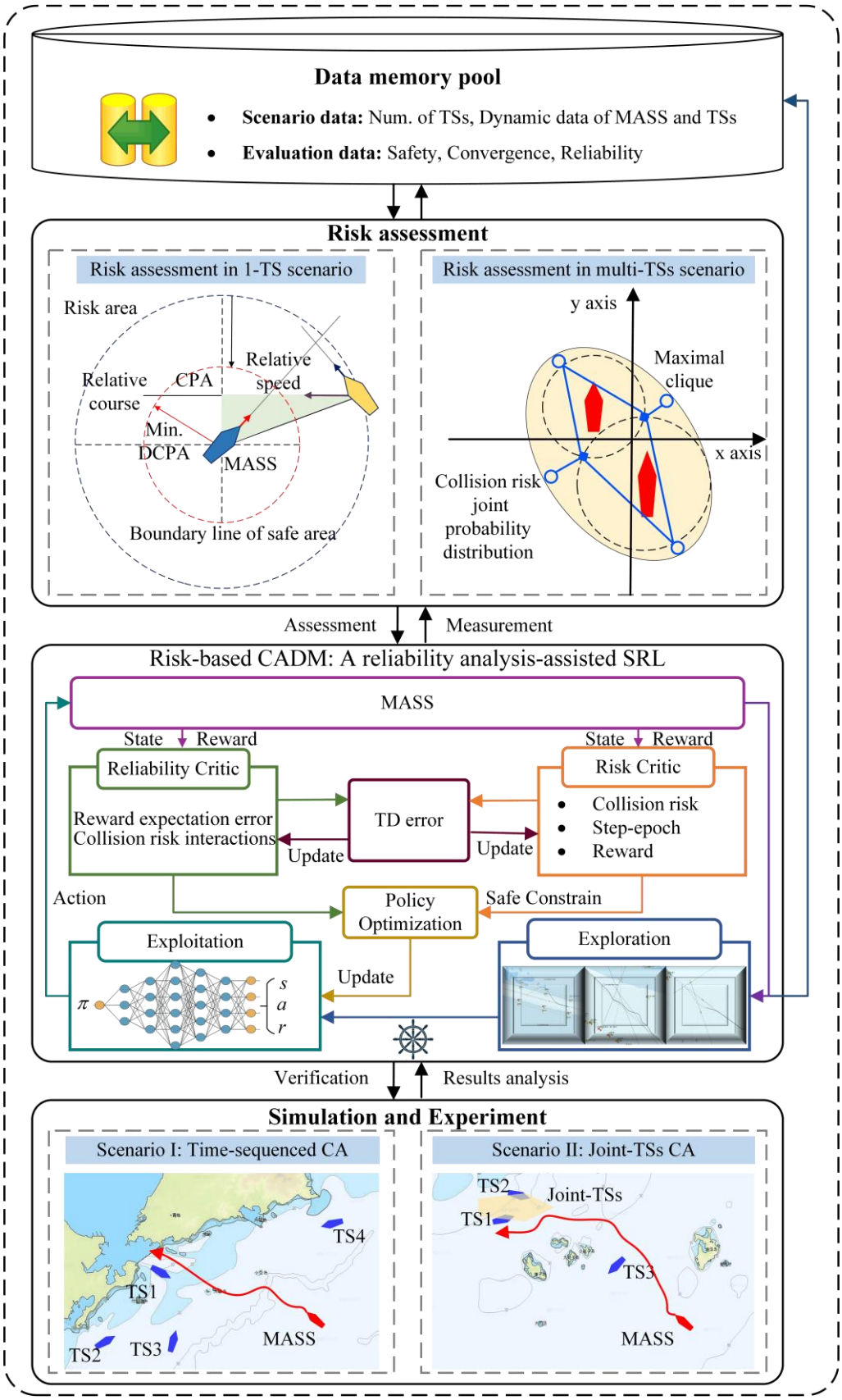


Fig. 7. The framework of risk-based CADM of MASS using a reliability analysis-assisted SRL method.

Algorithm 2. Pseudo-code of the integrated algorithm.

1: Initialize λ_{CR}, λ_{DT}, λ_{LRT}, w_{tp}, w_{ca}, w_{ef}, $\left[A^{i}_{t}\right]_{1\le i\le 3}$, S, $\left[\hat{Q}_0^{(i)}\right]_{1\le i\le 3}$, $\left[\hat{E}_0^{(i)}\right]_{1\le i\le 3}$

2: **for** $i=1,2,3$ **do**

3: $Q_t^{(i)}=\frac{1}{t}\sum_{1}^{t}\hat{Q}_t^{(i)}$

4: $E_t^{(i)}=\frac{1}{t}\sum_{1}^{t}\hat{E}_t^{(i)}$

5: **end for**

6: **for** $t=1,2,3,\cdots$ **do**

7: Observe s_t

8: **for** $i=1,2,3$ **do**

9: **if** $t=1$ **then**

10: Randomly choose $a_1\in[A]$

11: **else**

12: Randomly sample experience from memory pool and choose $a_i^{(1)\sim(3)}$

13: **end if**

14: Obtain the reward R_t based on the previous safety and reliability performance

15: Evaluate CR via (14)

16: Update $Q_t(s_t,a_t^{(1)\sim(3)})$

17: Update $E_t(s_t,a_t^{(1)\sim(3)})$

18: Evaluate J via (5), $Q_t'\leftarrow\sum Q_t^{(1)\sim(3)}$, $\pi'\leftarrow\pi_t^{*(1)\sim(3)}$

19: **end for**

20: **end for**

The overall algorithm framework of risk assessment-based collision avoidance decision-making of MASS using a COLREGs-assisted SRL method is shown in Fig. 7. The framework consists of four interconnected parts: the data memory pool, risk assessment, risk-based CADM, and simulation and experiment. The data memory pool stores scenario data (such as the number of TSs and dynamic information of MASS and each TS) and evaluation data (such as safety, convergence, and reliability performance) during training. In the risk assessment part, scenarios are categorized as either single-TS or joint-TSs. For joint-TSs scenarios, it

incorporates the concepts of undirected graphs (UG) to assess the risk. When the attention domains of two TSs overlap, a new risk is generated using the joint probability distribution method via the maximal clique of the synthetic risk area. The third part is the risk-based CADM, which introduces a risk of collision and reliability critical network into the actor-critic architecture. This step enhances the safety and reliability of CADM. The final part is the simulation and experiment, which indicates time-sequence CA scenario and joint-TSs CA scenario used to verify the performance of the proposed algorithm.

In summary, the solutions to anti-collision decision-making of MASS to avoid joint-TSs collision risk can be outlined as follows. First, the objective and evaluation functions are determined through parameter initialization. Then, MASS needs to observe the state of the situation with regard to different encounter scenarios. According to the state of the current time window, random trial and error avoidance behavior are undertaken, along with assessing risk and reliability. Until the evaluation function is finally satisfied, the optimal strategy is obtained. Further details about the CADM method via RA-SHRL are available in Algorithm 2.

5. Experimental results and comparative analysis

This section presents a simulation of a complex scenario where multiple ships encounter to validate the convergence efficiency, safety, and reliability of the RA-SHRL. In Section 5.1, the CADM results of the algorithm are validated by analyzing the avoidance trajectories. Then, the algorithm's performance is analyzed from the perspectives of convergence efficiency, safety, and reliability. The convergence efficiency is verified by the step-epochs curve, safety is verified by the success rate of anti-collision and interactions with collision risk, and reliability is verified.

5.1. Parameter setup and ablation Study

Referring to Mnih et al. [51] and Wang et al. [52]. The relevant parameters of the algorithm are set, as shown in Table 3. The parameter selection of the learning rate and discount factor is crucial to the algorithm training results. For the algorithm proposed in this paper, hyperparameters such as learning rate (α) and discount factor (γ) can influence the performance of the model. Referring to [51], we assign the learning rate to 0.1 and 0.01, and the discount factor to 0.9 and 0.99, and conduct ablation experiments. This section undertakes a sensitivity analysis by varying both the learning rate and discount factor and examining the training step sizes of the model under these distinct parameters. Fig. 8 illustrates the model's learning step size under varied learning rates and discount factors. Key observations include: 1) With a constant discount factor and a

decreasing learning rate (as visualized in the blue area of the diagram), the step size exhibits significant instability during the training process, and the step size is notably larger in the early phases; 2) Holding the learning rate steady and solely diminishing the discount factor (as represented in the green area of the diagram) leads MASS into local iterations early in training, with the step size prone to more pronounced fluctuations; 3) Modifying both the learning rate and discount factor exerts minimal impact on the final iteration convergence step size; 4) By the conclusion of the training process, the optimal training outcomes are achieved as $\alpha_0 = 0.01,\ \gamma = 0.99$. This aligns with conventional default values designated for RL algorithm parameters, further proving the rationale of the model.

Table 3. The parameters setting.

Parameters	Values	Description
Learning rate	0.01	The learning rate is set to o.o1 by ablation study.
Discount factor	0.99	The learning rate is set to o.99 by reference [49] and ablation study.
Training epoch	2500	Train fully to balance efficiency and effectiveness.
Exploration	1-0.1	Initial value to final value of $\varepsilon-$greedy exploration.
Update frequency	5	It means that updated every 5-time steps, which can balance the consumption of computing resources and the efficiency of learning.

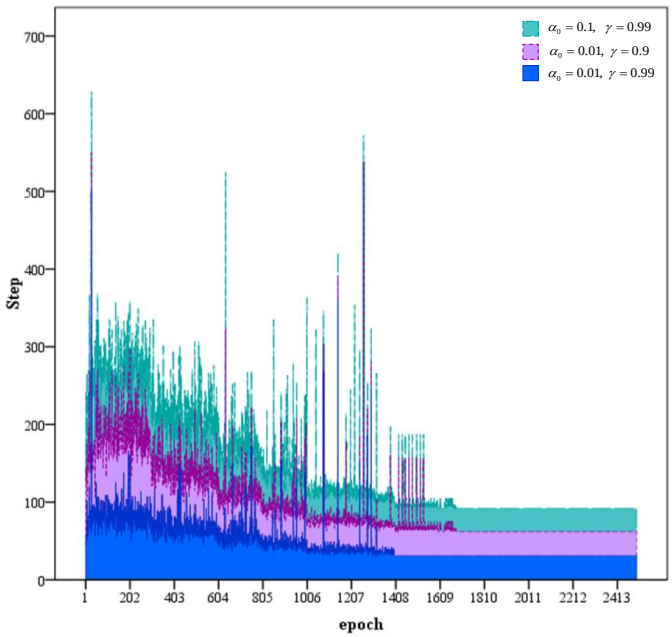


Fig. 8. Model learning step size under different learning rates and discount factors.

5.2. Validation of collision avoidance decision-making results

In this section, two scenarios were designed: one is a five-TSs encounter situation, which includes typical head-on, crossing, and overtaking operations. The second is a three-TSs encounter situation, which includes crossing and two ships' joint-TSs operations. The collision avoidance decision-making model presented in this paper is verified for safety and reliability in the simulation environments of both cases. PyGame, a two-dimensional (2D) programming library in Python, was used to construct the multi-TSs encounter scenario for testing the proposed CADM algorithm.

5.2.1. Scenario I: a time-sequenced obstacle avoidance scenario

In this section, the efficacy of the proposed model is evaluated in environments with multiple dynamic obstacles and uncertainties. Drawing from COLREGs, a comprehensive navigation environment has been designed, encompassing typical encounter scenarios such as head-on, crossing, and overtaking situations. Table 4 presents the initial states of the MASS and all TSs for this scenario.

Table 4. The initial environment state settings under Scenario I.

Obstacles	Initial point	Goal point	Speed	Risk area radius
TS1	(238, 256)	(601, 446)	(12, 7)	20
TS2	(151, 444)	(383, 386)	(8, -2)	10
TS3	(259, 473)	(402, 172)	(5, -10)	15
TS4	(733, 141)	(332, 227)	(-14, 3)	12
MASS	(590, 415)	(190, 195)	variable speed	15

After iterative training of the algorithm, MASS successfully avoids each TS, with the decision-making avoidance trajectory illustrated in Fig. 9. The trajectory reveals that MASS identified a head-on situation with TS1, subsequently opting for a right-turn avoidance action at the outset, which is consistent with COLREGs. Further into the journey, a potential collision risk arose due to a crossing situation with TS3 and TS4. In this context, all TSs were set to display no avoidance behavior. As a result, during this period, MASS relied on the adaptability of the RA-SHRL algorithm to execute two separate avoidance maneuvers, ensuring safe passage to the destination.

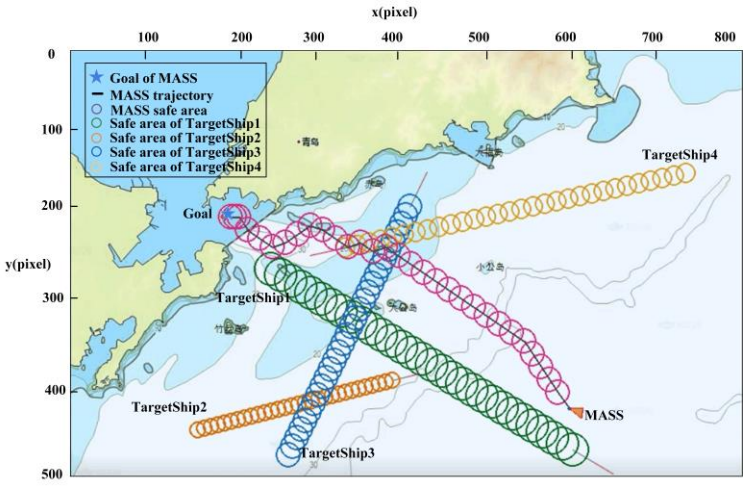


Fig. 9. Collision avoidance results.

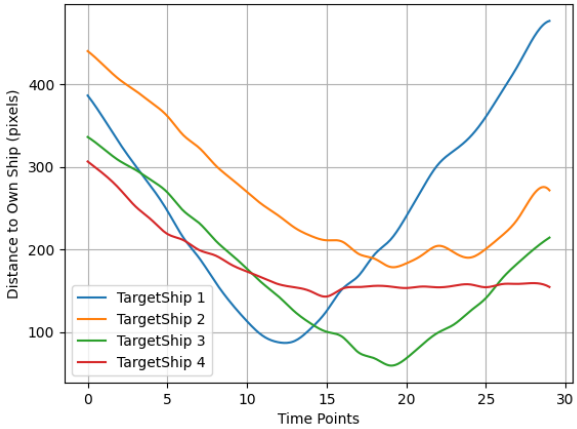


Fig. 10. Distance variation of TSs relative to MASS.

Additionally, the relative distances between MASS and each TS during the entire collision avoidance process are depicted in Fig. 10. This figure elucidates the dynamic variations throughout the avoidance sequence. Notably, towards the end of the voyage, an overtaking situation emerged between TS4 and the MASS. The curve for change in relative distance suggests that this was the most challenging avoidance scenario in the entire scene. This underscores that in intricate environments, sequential avoidance poses sustained risks. Hence, a holistic consideration of all situational factors is imperative, particularly when navigating clusters of closely situated dynamic obstacles.

The model proposed in this paper has strong adaptability when solving the decision-making problem of collision avoidance in complex situations, although showing uncertainties brought by the need of a balance between exploration and exploitation in the process of interacting with the environment and the difference between the real environment and the simulated environment. The impact of these uncertainties on the model will be reduced under the collaborative drive of knowledge and data with the provision of this advance model as the foundation.

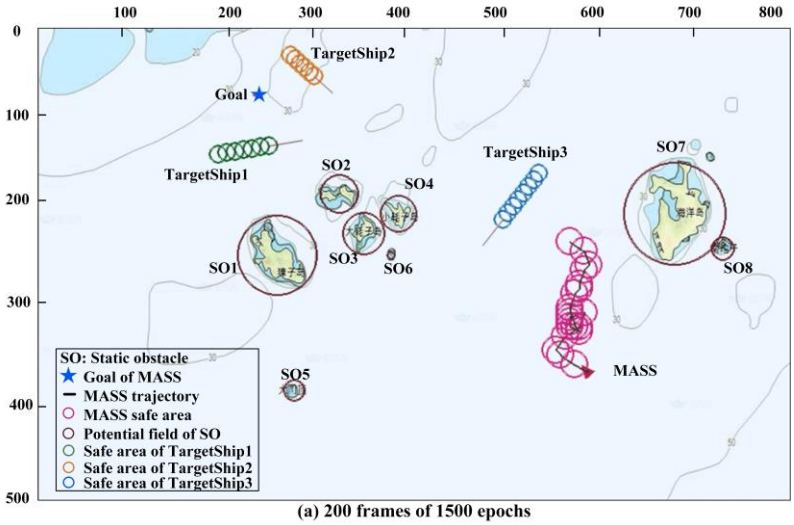
5.2.2. Scenario II: a joint-TSs obstacle avoidance scenario

In this case study, the model proposed in this paper is validated as effective in a joint-TSs environment. The Zhangzi Island water areas are taken as an example to simulate a multiple static-dynamic mixed environment with dimensions of 800 by 500 pixels. The environment state settings are outlined in Table 5, encompassing three TSs and eight static obstacles (SOs).

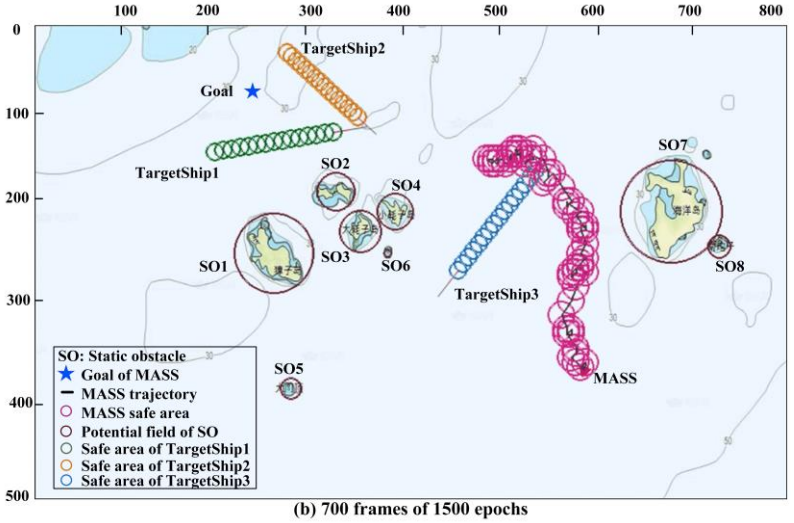
Table 5. The initial environment state settings under Scenario II.

Obstacles	Initial point	Goal point	Speed	Risk area radius
TS1	(191, 133)	(446, 89)	(18, -3)	20
TS2	(268, 28)	(393, 141)	(10, 9)	20
TS3	(499, 195)	(344, 397)	(-11, 14)	20
SO1	(257, 245)	—	(0, 0)	28
SO2	(320, 175)	—	(0, 0)	10
SO3	(346, 217)	—	(0, 0)	15
SO4	(383, 196)	—	(0, 0)	13
SO5	(272, 383)	—	(0, 0)	8
SO6	(375, 240)	—	(0, 0)	4
SO7	(677, 193)	—	(0, 0)	35
SO8	(728, 233)	—	(0, 0)	8

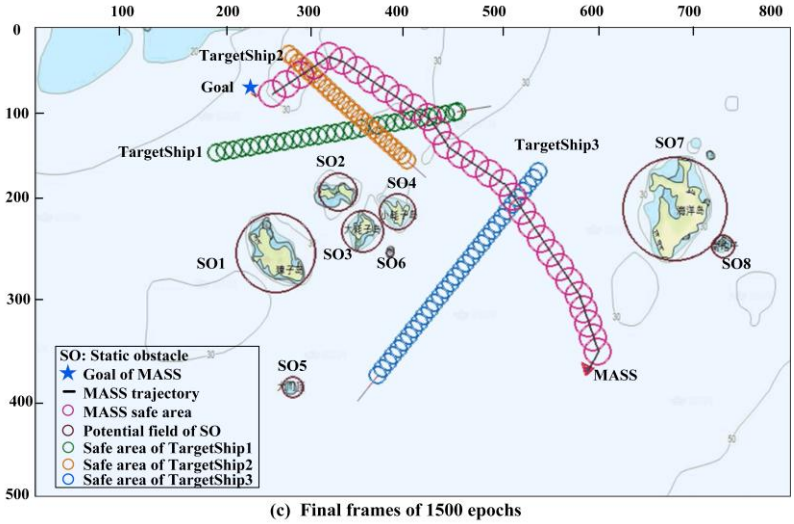
The potential field of SOs defines the collision zone, signifying that if MASS enters this zone, a collision is imminent. Conversely, for the risk and attention areas, this study has devised a specialized reward-driven interaction mechanisms to shape MASS's avoidance strategy. In the context of the ship trial scenario, as MASS navigates from the initial position to the destination, it encounters two distinct scenarios featuring three TSs. The first scenario entails a crossing situation between MASS and TS3, while the second scenario encompasses both a head-on encounter and a crossing situation between MASS and TS1 and TS2, respectively.



449 (a) 200 frames of 1500 epochs



450 (b) 700 frames of 1500 epochs

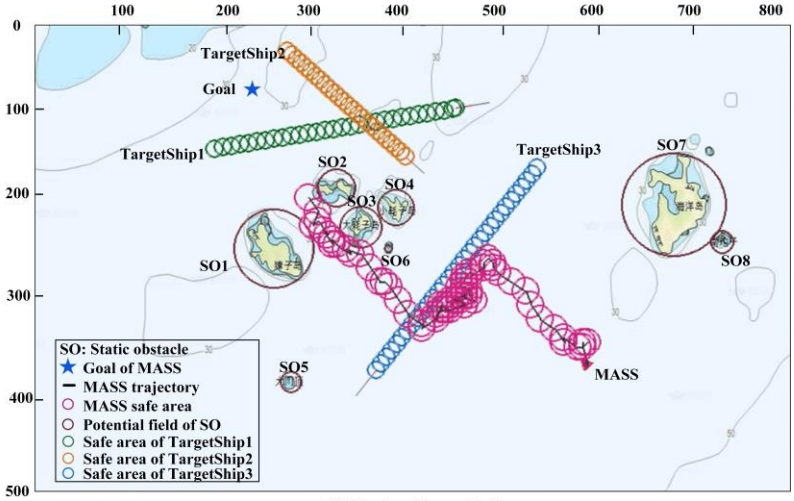


451 (c) Final frames of 1500 epochs

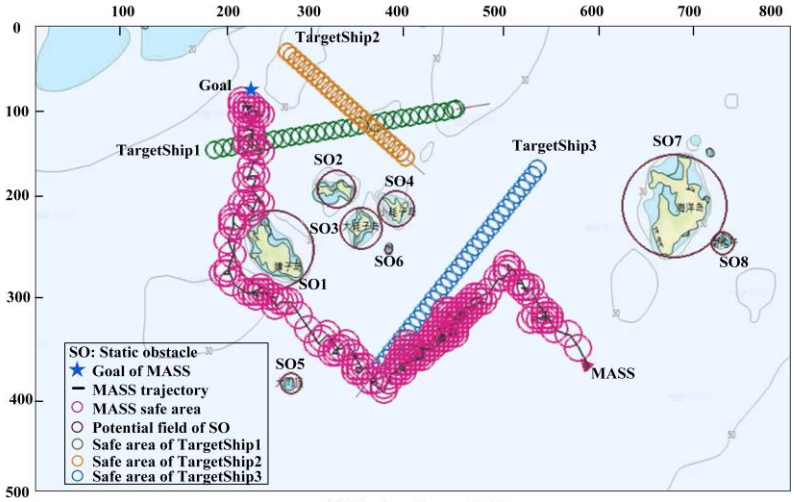
452 **Fig. 11.** Simulation results of collision avoidance decision-making.

453 In this instance, MASS's starting point and target point are designated as (581, 362) and (233, 72),
454 respectively. Following training for 1500 epochs, the algorithm yields a secure, steady, and reliable anti-

collision trajectory. Fig. 11 depicts the simulation outcomes of CADM based on RA-SHRL with three TSs and eight static obstacles. As evident in Fig. 11(a), when MASS encounters TS3, a collision risk emerges. Nevertheless, by the 200th epochs, MASS has not yet perceived the collision risk posed by TS1 and TS2. After many trials and errors of environmental interaction, MASS initially chose to turn slightly right to avoid TS3. However, this strategy, over time, heightens the risk of colliding with SO4 and complicates the avoidance of TS1 and TS2 in the future. By the 700th epochs, illustrated in Fig. 11(b), MASS executes a significant right turn at the outset to evade TS3 and SOs, minimizing the need for frequent turns and avoidance maneuvers while adhering to COLREGs. However, due to the perceived risk of complex collisions with TS1 and TS2, multiple trial-and-error waypoints searches were conducted. In the scenario where TS1 and TS2 encounter MASS, sequential avoidance exposes MASS to a new collision risk with TS1 after successfully avoiding TS2. Consequently, TS1 and TS2 are treated as joint-TSs in this study. As depicted in Fig. 11(c), a one-time avoidance is performed based on the joint-TSs' overall risk joint probability distribution.



(a) Exploration path 1



(b) Exploration path 2

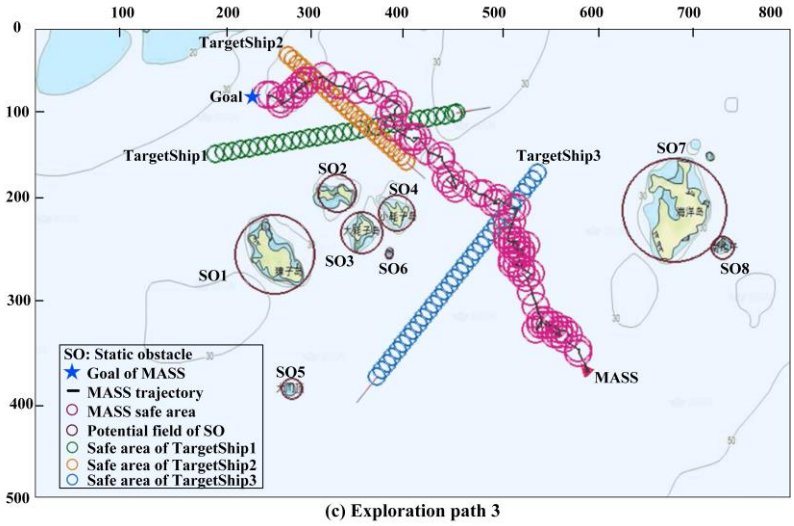


Fig. 12. Exploration path comparison.

In this experiment, the static obstacles are modeled as circles of varying sizes. This approach expedites decision-making and calculations, even in the presence of uncertainties, by providing a simplified representation. It addresses challenges such as overly conservative obstacle avoidance decisions and wasted space. Moreover, this paper primarily addresses the issue of joint risk assessment and collision avoidance decision-making in complex, short-distance dynamic obstacle environments. The impact of uncertainties on the primary decision-making objectives and performance is considered to be minimal.

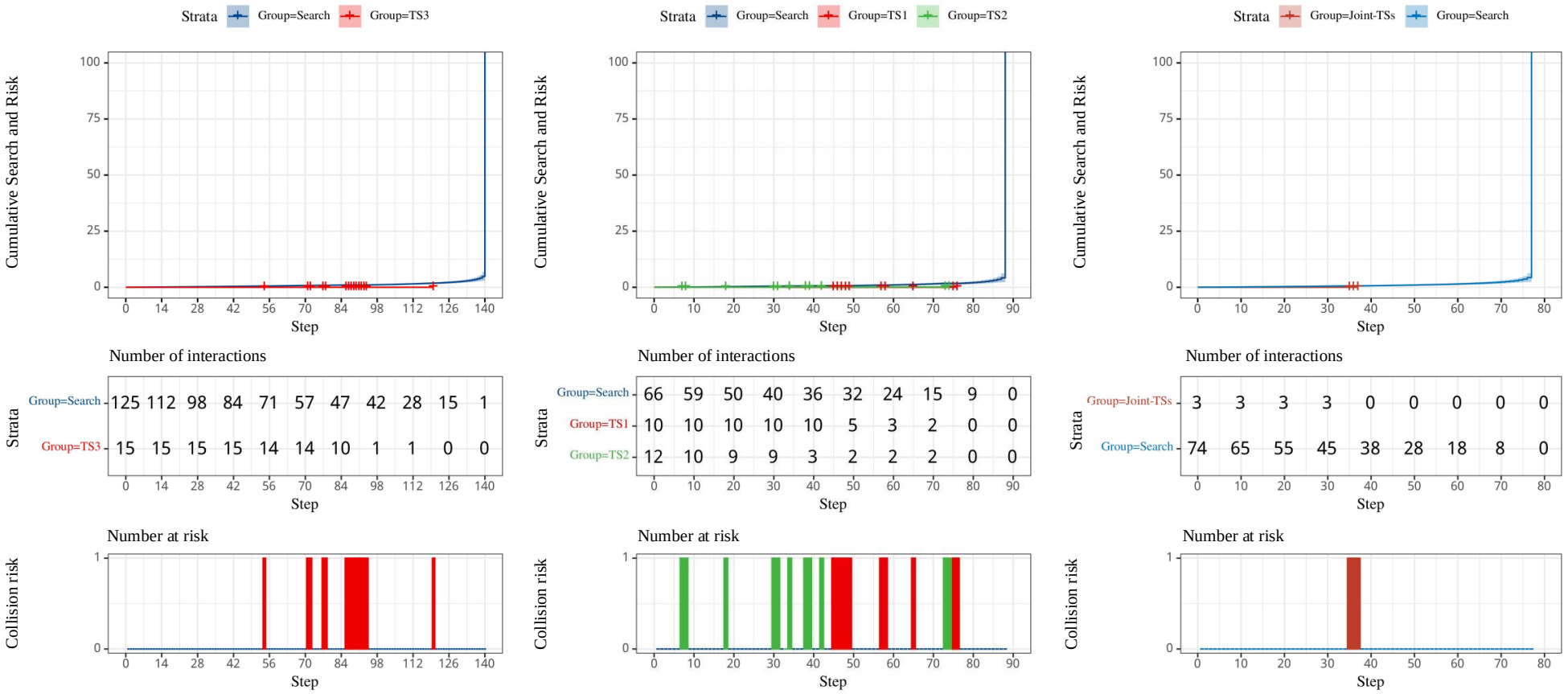
In addition, the exploration path which from the right and left sides of the static obstacle SO1 are intercepted, as well as from the right side of the static obstacle SO4 for the first time, in order to compare the rationality of the output collision avoidance path. As shown in Fig. 12(a) and Fig. 12(b), whether the algorithm searches between multiple obstacles or turns left from the left side of SO1, it will encounter multiple avoidance maneuvers with the dynamic TS3, resulting in particularly unstable trajectories during the first half of the path. Fig. 12(c) shows the path of the algorithm's first right-turn avoidance of TS3 and subsequent search for the target from the right side of the static obstacle SO4. It is evident that the algorithm's decision to make a right-turn avoidance of TS3 not only complies with the COLREGs, but also results in better performance in terms of path stability, step length, and collision frequency compared to the two paths resulting from left-turn searches.

5.3. Safety performance of the proposed method

To improve the clarity of safety performance analysis for the proposed algorithm, separate survival analyses on search and collision risks during obstacle avoidance were conducted. Survival analysis is a statistical method that explores the time of occurrence of events [53]. The percentage of outcome events at the

anti-collision search for analysis, as shown in Fig. 13, was determined. Specifically, Fig. 13(a) presents the search and risk statistics analysis for avoiding TS3, while Fig. 13(b) illustrates the search and risk statistics analysis for the anti-collision process when the algorithm uses TS1 and TS2 as independent dynamic obstacles. Finally, Fig. 13(c) provides an analysis of the anti-collision process search and risk statistics for the joint-TSs (TS1 and TS2 combined).

The search curve reaching 100% indicates that obstacle avoidance has been successfully achieved in the final step. For the avoidance of TS3, MASS completed the avoidance process in 140 search steps. When avoiding TS1 and TS2 separately, the avoidance was completed at Step 88. However, when considering TS1 and TS2 as a joint-TSs, the avoidance was completed at step 76. Additionally, the number of interactions between MASS and search, obstacles, as well as the number of risks during the entire anti-collision process was counted. The number of interactions during the exploration stage can be interpreted as the number of trials and errors, which objectively reflects the robustness of the algorithm. The convergence performance of the algorithm can be judged from the number of searches. At the same time, the number of interactions with TS during the search process represents the number of trial and error, the less the number of interactions, the lower the collision risk, and the more stable the avoidance strategy. Comparing the number of interactions in Fig. 13 (c) and Fig. 13 (b), after TS1 and TS2 are jointly processed, the number of trial-and-error interactions is witnessed with great reduction. The number at risk indicates the number of collision risks with corresponding obstacles on the timeline of steps. The vertical stick represents a specific collision risk at a particular step, and the thickness indicates the severity of the risk. In the risk statistics section of Fig. 13, the green color represents the risk of collision between the MASS and TS2, while the red color indicates the risk of collision between the MASS and TS1. During the avoidance of TS1, a new collision risk with TS2 arises, even though it has already been avoided before. However, it is evident that both the time and degree of collision risks are improved after combining TS1 and TS2 as joint-TSs.



(a) TS3 (b) TS1&TS2 (c) Joint-TSs

Fig. 13. Survival analysis on search and collision risk during the anti-collision process.

5.4. Convergence performance of the newly proposed method

In this section, numerous experiments are carried out to evaluate its performance across various epochs. The algorithm's states and actions are systematically recorded at each iteration, and the rewards obtained from executing these actions are thoroughly analyzed. In this research, the assessment of the algorithm's convergence performance primarily relies on the analysis of average step data.

To underscore the superior convergence of the proposed algorithm, comparative experiments with three benchmark algorithms: Deep Q-network (DQN), Constrained Policy Optimization (CPO), and RA-SHRL were conducted. Fig. 14 illustrates a step trend line chart that displays the evolution of steps for DQN, CPO, and RA-SHRL. The blue dashed line represents the step size variations in DQN, while the brown dashed line reflects the trend in CPO's step size. The green dashed line illustrates the progression of the step size for the RA-SHRL algorithm introduced in this paper. The observation reveals that, in terms of the number of random searches and iteration time, the convergence performance of RA-SHRL surpasses that of DQN and CPO. Furthermore, the average step lengths of the three benchmark algorithms during the iteration process are calculated and plotted in Fig. 14. From these values, it is evident that the proposed RA-SHRL algorithm exhibits the smallest average search step length, indicating better performance in aspect convergence.

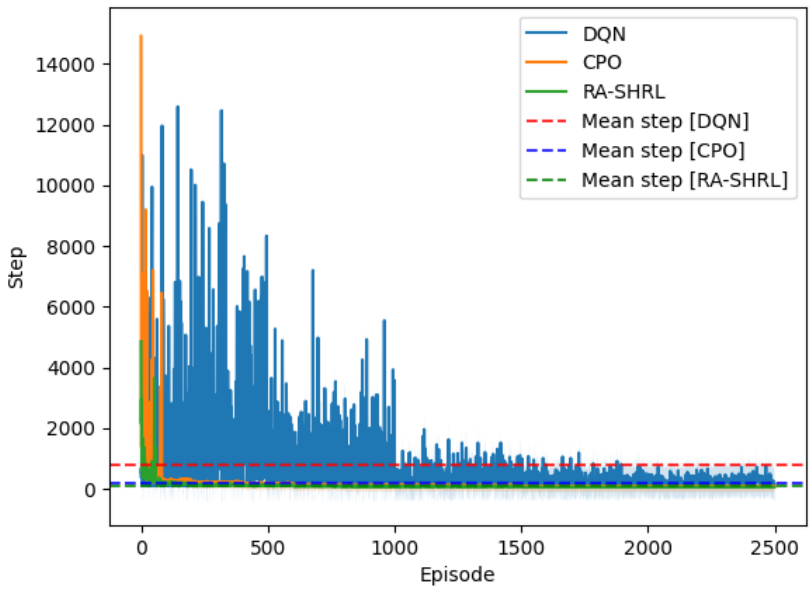


Fig. 14. The step trend line chart of DQN, CPO, and RA-SHRL.

5.5. Reliability performance of the proposed RA-SHRL method

To reflect the reliability of the three groups of algorithms more clearly, Fig. 15 shows the error of the three groups of algorithms' reward values. The results distinctly indicate that the reward values generated by this proposed algorithm exhibit a larger magnitude, accompanied by smaller errors. This observation underscores the superior stability of our algorithm in terms of cumulative returns, even amidst the variations over time. The pronounced difference in reward values and reduced error associated with our algorithm suggests its robustness and consistent performance across different scenarios. The larger reward values imply more effective learning and decision-making, while the smaller errors indicate a reduced degree of variability in the algorithm's performance. This heightened stability is particularly crucial in dynamic environments where the system needs to adapt and make decisions over time. In summary, the comprehensive assessment of reward value errors emphasizes the superior stability and reliability of our proposed algorithm, highlighting its efficacy in delivering consistent and favorable cumulative returns over extended durations.

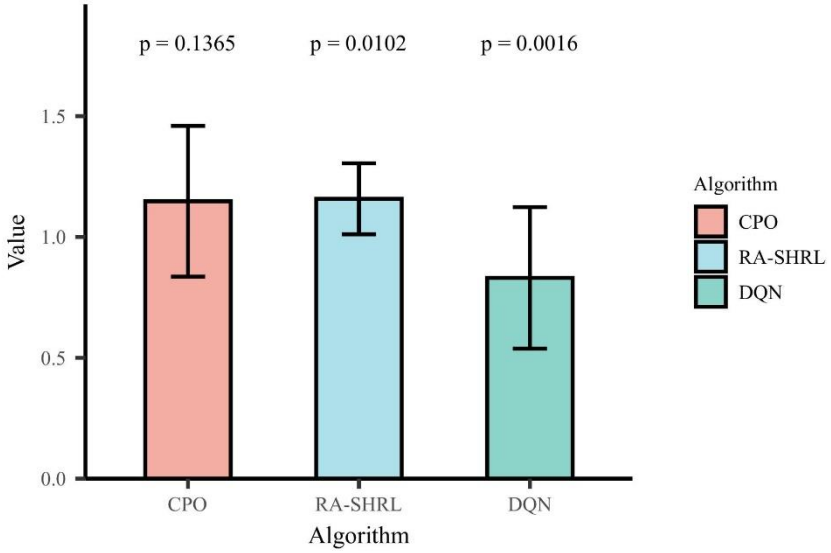


Fig. 15. Error bar histogram.

On the other hand, the number of collision risk interactions in each epoch was counted, as shown in Fig. 16. In the process of CADM for the joint TSs, the number of collision risks becomes less and less with the training of the algorithm. It is clear that during the initial search process, MASS had more than 50 instances of collision risk interaction with obstacles in the environment. After accumulating a significant amount of learning experience during the early exploration process, MASS fully leveraged its experience in the later stages, allowing it to rapidly resolve collision conflicts. Even after the algorithm's learning had converged, MASS only had minimal necessary risk interactions with the environment. As shown in Fig. 17, the training reward curve

increases and gradually stabilizes with the increase of epochs. In the early stages of training, although the cumulative reward is negative, the curve rises rapidly. This indicates that the algorithm has learned effective strategies in the initial phase and has initially mastered the collision avoidance task. In the mid-stage, the cumulative reward curve continues to rise, indicating that the algorithm is constantly improving its strategies and learning better solutions. In the later stages, the cumulative reward curve tends to stabilize, showing that the algorithm has essentially converged and found a stable strategy. According to the above-mentioned, this indicates that the algorithm has learned an effective collision avoidance strategy, and the results have a certain degree of reliability.

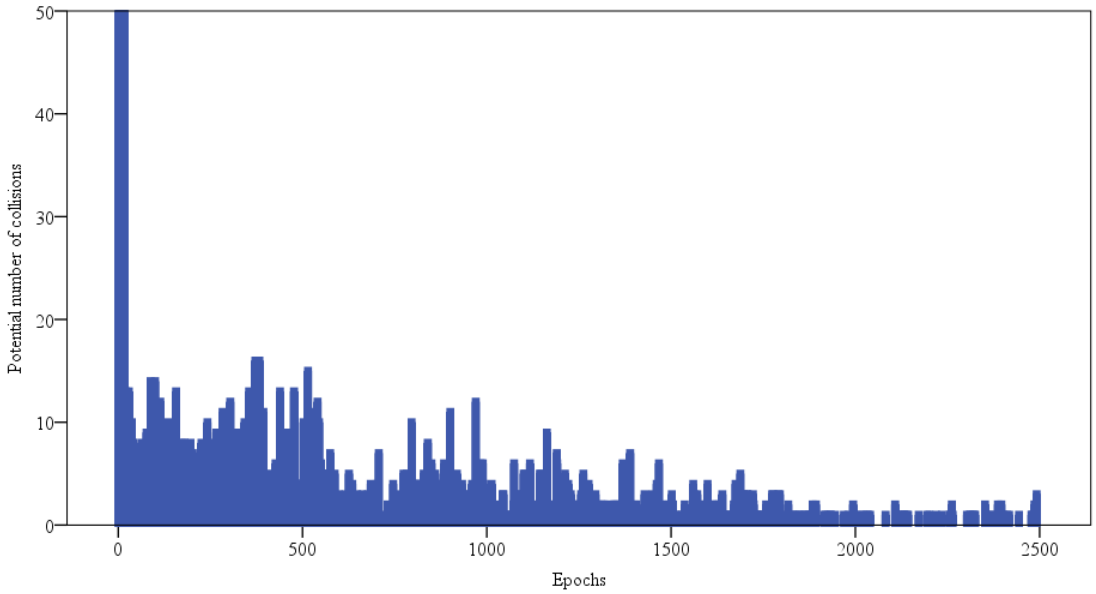


Fig. 16. Number of the collision risk count in each epoch.

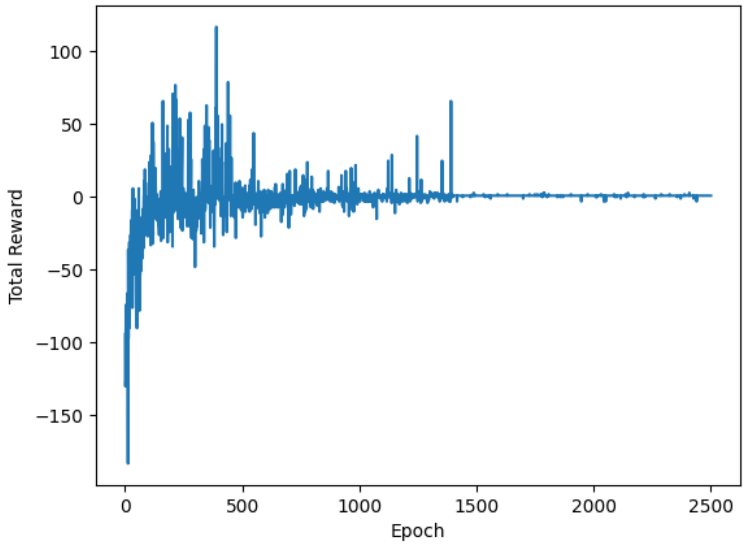


Fig. 17. Training rewards.

6. Conclusion

In this paper, a risk informed CADM model is proposed to realize the safe navigation of MASS in waters of complicated traffic scenarios. First, a novel collision risk assessment model is designed by combining CPA-based sequential collision risk assessment improved by safety area and joint-TSs collision risk assessment based on the conditional random field. Based on risk assessment, a CADM algorithm with multi-objectives is formulated by using the new RA-SHRL with reliability measurement. Two complex dynamic obstacle scenarios are used to examine the safety, convergence, and reliability of the proposed method. Multiple sets of algorithmic comparative experiments have been conducted. Compared with DQN, the search efficiency of the algorithm proposed in this paper is improved by 40%. It achieved a 91.3 % success rate of collision avoidance during training. The results illustrate that the innovative approach is capable of achieving CADM in joint-TSs environments. After jointly assessing the risks and avoiding multiple TSs, the MASS can pass through and avoid them clearly with just one interaction. It serves as a systematic method for evaluating the performance of learning-based algorithms across safety, convergence, and reliability metrics. This research lays the groundwork for the development of safer and more reliable autonomous maritime systems, with far-reaching implications for the maritime industry's transition towards autonomous and intelligent navigation. Furthermore, the proposed algorithm provides a valuable contribution to the field by offering a tangible solution for implementing safety-aware decision-making for other autonomous systems in dynamic environments. One limitation of the proposed method is related to its restriction by the quantification of the reward mechanism in good seamanship.

Future endeavors will focus on optimizing and constructing data-driven models for RA-SHRL using both expert demonstration data and historical real-world data. Additionally, there is a need to investigate the development of a comprehensive reliability evaluation system for learning-based algorithms, particularly DRL algorithms. This system aims to assess the stability, robustness, and reliability of CADM algorithms in diverse navigation scenarios, ultimately mitigating uncertainties and risks associated with learning-based algorithms. This holistic approach strives to enhance the practicality and security of CADM algorithms in real-world applications.

CRediT authorship contribution statement:

Chengbo Wang: Conceptualization, Methodology, Formal analysis, Data curation, Writing – original draft, Writing – review & editing, Software, Visualization, Validation. **Xinyu Zhang:** Conceptualization, Methodology, Supervision, Writing – original draft, Writing – review & editing, Funding acquisition. **Hongbo Gao:** Conceptualization, Investigation, Supervision, Writing – review & editing, Funding acquisition. **Musa Bashir:** Formal analysis, Writing – review & editing.

Huanhuan Li: Formal analysis, Writing – review & editing. **Zaili Yang:** Methodology, Writing – original draft, Writing – review & editing, Funding acquisition.

Acknowledgements: The work is supported by the National Natural Science Foundation of China (52371359), in part by the National Natural Science Foundation of China (Grant Nos. U20A20225 and U2013601), Anhui Province Natural Science Funds for Distinguished Young Scholar (Grant No. 2308085J02), the Science and Technology Innovation 2030 - " New Generation Artificial Intelligence" Major Project (Grant No. 2022ZD0116305), Innovation Leading Talent of Anhui Province TeZhi plan, the Natural Science Foundation of Hefei, China (Grant No. 202321), the CAAI-Huawei Mind Spore Open Fund (Grant No. CAAIXSJLJJ-2022-011A); This work is also financially supported by the European Research Council project (TRUST CoG 2019 864724).

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.

References

[1] Xin, Xuri, et al., Multi-scale collision risk estimation for maritime traffic in complex port waters, Reliab. Eng. Syst. Safe 240 (2023) 109554, https://doi.org/10.1016/j.ress.2023.109554.

[2] Hannaford, E., Maes, P. & Van Hassel, E. Autonomous ships and the collision avoidance regulations: a licensed deck officer survey. WMU J Marit Affairs. 21 (2022) 233–266. https://doi.org/10.1007/s13437-022-00269-z.

[3] Zhang, Xinyu, et al., Collision-avoidance navigation systems for Maritime Autonomous Surface Ships: A state of the art survey, Ocean Engineering. 235 (2021) 109380. https://doi.org/10.1016/j.oceaneng.2021.109380.

[4] Li, Huanhuan, et al., Deep bi-directional information-empowered ship trajectory prediction for maritime autonomous surface ships, Transport. Res. Part E: Logist Transport. Rev. 181 (2024) 103367, https://doi.org/10.1016/j.tre.2023.103367.

[5] IMO 'IMO takes first steps to address autonomous ships' 25/05/2018, http://www.imo.org/en/mediacentre/pressbriefings/pages/08-msc-99-mass-scoping.aspx [accessed December 2021].

[6] Wu, Xiangdong, et al., Ameliorated equilibrium optimizer with application in smooth path planning oriented unmanned ground vehicle, Knowledge-Based Systems. 260 (2023) 110148. https://doi.org/10.1016/j.knosys.2022.110148.

[7] Jiang, Yue, et al., Model free predictive path tracking control of variable-configuration unmanned ground vehicle, ISA transactions. 129 (2022) 485-494. https://doi.org/10.1016/j.isatra.2022.01.026.

[8] Zhang, Xiangyin, et al., Multi-objective particle swarm optimization with multi-mode collaboration based on reinforcement learning for path planning of unmanned air vehicles, Knowledge-Based Systems. 250 (2022) 109075. https://doi.org/10.1016/j.knosys.2022.109075.

[9] Rao, Jinjun, et al., Path planning for dual UAVs cooperative suspension transport based on artificial potential field-A* algorithm, Knowledge-Based Systems. 277 (2023) 110797. https://doi.org/10.1016/j.knosys.2023.110797.

[10] Chen, Zuoyi, et al., Residual shrinkage transformer relation network for intelligent fault detection of industrial robot with zero-fault samples, Knowledge-Based Systems. 268 (2023) 110452. https://doi.org/10.1016/j.knosys.2023.110452.

[11] Wang, C., Wang, N., Xie, G., & Su, S. F., Survey on collision-avoidance navigation of Maritime autonomous surface ships. In Offshore Robotics: Volume I Issue 1, 2021 (pp. 1-33). Springer Singapore. https://doi.org/10.1007/978-981-16-2078-2_1.

[12] Dong, Wenhai, Jie Li, and Xinghua Tu., A study on the development and trend of COLREGs–A broader perspective, The Journal of Navigation. (2023) 1-13. https://doi.org/10.1017/S0373463323000279.

[13] Shaobo, Wang, Zhang Yingjun, and Li Lianbo., A collision avoidance decision-making system for autonomous ship based on modified velocity obstacle method, Ocean Engineering. 215 (2020) 107910. https://doi.org/10.1016/j.oceaneng.2020.107910.

[14] Lyu, Hongguang, and Yong Yin., COLREGS-constrained real-time path planning for autonomous ships using modified artificial potential fields, The Journal of navigation. 72.3 (2019) 588-608. https://doi.org/10.1017/S0373463318000796.

[15] Li, Huanhuan, Jiao, Hang, and Yang, Zaili, AIS data-driven ship trajectory prediction modelling and analysis based on machine learning and deep learning methods, Transportation Research Part E: Logistics and Transportation Review. 175 (2023) 103152. https://doi.org/10.1016/j.tre.2023.103152.
[16] Zhang, Mingyang, et al., A machine learning method for the prediction of ship motion trajectories in real operational conditions, Ocean Engineering. 283 (2023) 114905. https://doi.org/10.1016/j.oceaneng.2023.114905.
[17] Wang C, Zhang X, Yang Z, et al., Collision avoidance for autonomous ship using deep reinforcement learning and prior-knowledge-based approximate representation., Front Mar Sci. 9(2023) 1084763. http://dx.doi.org/10.3389/fmars.2022.1084763.
[18] Jiang, L., An, L., Zhang, X., Wang, C., & Wang, X., A human-like collision avoidance method for autonomous ship with attention-based deep reinforcement learning, Ocean Eng. 264(2022) 112378. https://doi.org/10.1016/j.oceaneng.2022.112378.
[19] Chun, D. H., Roh, M. I., Lee, H. W., Ha, J., & Yu, D., Deep reinforcement learning-based collision avoidance for an autonomous ship, Ocean Eng. 234(2021) 109216. https://doi.org/10.1016/j.oceaneng.2021.109216.
[20] Wu, Xing, et al., The autonomous navigation and obstacle avoidance for USVs with ANOA deep reinforcement learning method, Knowledge-Based Systems. 196 (2020) 105201. https://doi.org/10.1016/j.knosys.2019.105201.
[21] Zhang M, Montewka J, Manderbacka T, et al., A big data analytics method for the evaluation of ship-ship collision risk reflecting hydrometeorological conditions, Reliab Eng Syst Safe. 213(2021) 107674. https://doi.org/10.1016/j.ress.2021.107674.
[22] Bakdi A, Vanem E., Fullest COLREGs evaluation using fuzzy logic for collaborative decision-making analysis of autonomous ships in complex situations, IEEE T Intell Transp. 23(10) (2022) 18433-18445. https://doi.org/10.1109/TITS.2022.3151826.
[23] Xin, X., Liu, K., Li, H., & Yang, Z. Maritime traffic partitioning: An adaptive semi-supervised spectral regularization approach for leveraging multi-graph evolutionary traffic interactions. Transport Res C-Emer, (2024) 164, 104670. https://doi.org/10.1016/j.trc.2024.104670.
[24] Zhang M, Kujala P, Hirdaris S., A machine learning method for the evaluation of ship grounding risk in real operational conditions, Reliab Eng Syst Safe. 226(2022) 108697. https://doi.org/10.1016/j.ress.2022.108697.
[25] Zhang M, Zhang D, Fu S, et al., A predictive analytics method for maritime traffic flow complexity estimation in inland waterways, Reliab Eng Syst Safe. 220(2022) 108317. https://doi.org/10.1016/j.ress.2021.108317.
[26] Yuan, X., Zhang, D., Zhang, J., Zhang, M., & Soares, C. G., A novel real-time collision risk awareness method based on velocity obstacle considering uncertainties in ship dynamics, Ocean Eng. 220(2021) 108436. https://doi.org/10.1016/j.oceaneng.2020.108436.
[27] KEARON, J., 1977 Computer programs for collision avoidance and traffic keeping. In Conference on Mathematical Aspects on Marine Traffic. London: Academic Press.
[28] International Maritime Organization. Convention on the International Regulations for Preventing Collisions at Sea, 1972 (COLREGs), 1972.
[29] Kang, L., Lu, Z., Meng, Q., Gao, S., & Wang, F., Maritime simulator based determination of minimum DCPA and TCPA in head-on ship-to-ship collision avoidance in confined waters, Transportmetrica A. 15(2) (2019) 1124-1144. https://doi.org/10.1080/23249935.2019.1567617.
[30] Goodwin, E. M., A statistical study of ship domains, J Navigation. 28(3) (1975) 328-344. https://doi.org/10.1017/S0373463300041230.

[31] Davis, P. V., Dove, M. J., & Stockel, C. T., A computer simulation of marine traffic using domains and arenas, J Navigation.;33(2) (1980) 215-222. https://doi.org/10.1017/S0373463300035220.
[32] Szlapczynski, R., & Szlapczynska, J., An analysis of domain-based ship collision risk parameters, Ocean Eng.126(2016) 47-56. https://doi.org/10.1016/j.oceaneng.2016.08.030.
[33] Qu, X., Meng, Q., & Suyi, L., Ship collision risk assessment for the Singapore Strait, Accident Anal Prev. 43(6) (2011) 2030-2036. https://doi.org/10.1016/j.aap.2011.05.022.
[34] Zhang, L., & Meng, Q., Probabilistic ship domain with applications to ship collision risk assessment, Ocean Eng. 186(2019) 106130. https://doi.org/10.1016/j.oceaneng.2019.106130.
[35] Zhao-Lin W. Quantification of action to avoid collision. J Navigation 1984; 37(3): 420-430. https://doi.org/10.1017/S0373463300019949.
[36] Perera L P, Carvalho J P, Guedes Soares C., Fuzzy logic based decision making system for collision avoidance of ocean navigation under critical collision conditions, J Mar Sci Technol. 16(2011) 84-99. https://doi.org/10.1007/s00773-010-0106-x.
[37] Zhang, J., Liu, J., Hirdaris, S., Zhang, M., & Tian, W. An interpretable knowledge-based decision support method for ship collision avoidance using AIS data. Reliab Eng Syst Safe, (2023) 230, 108919. https://doi.org/10.1016/j.ress.2022.108919.
[38] Zhang X, Wang C, Liu Y, et al., Decision-making for the autonomous navigation of Maritime autonomous surface ships based on scene division and deep reinforcement learning, Sensors-Basel. 19(18) (2019) 4055. https://doi.org/10.3390/s19184055.
[39] Sawada R, Sato K, Majima T., Automatic ship collision avoidance using deep reinforcement learning with LSTM in continuous action spaces, J Mar Sci Tech-Japan. 26(2021) 509-524. https://doi.org/10.1007/s00773-020-00755-0.
[40] X. Zhang, K. Zheng, C. Wang, et al., A novel deep reinforcement learning for POMDP-based autonomous ship collision decision-making, Neural Comput & Applic (2023) 1–15.
[41] Xu, Xinli, et al., Path planning and dynamic collision avoidance algorithm under COLREGs via deep reinforcement learning, Neurocomputing. 468 (2022): 181-197. https://doi.org/10.1016/j.neucom.2021.09.071.
[42] Y. Wang, H. Xu, H. Feng, J. He, H. Yang, F. Li, Z. Yang, Deep reinforcement learning based collision avoidance system for autonomous ships, Ocean Eng. 292 (2024) 116527, https://doi.org/10.1016/j.oceaneng.2023.116527.
[43] L. Wen, J. Duan, S.E. Li, S. Xu, H. Peng, Safe reinforcement learning for autonomous vehicles through parallel constrained policy optimization, in: 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC), IEEE, 2020, pp. 1–7, https://doi.org/10.1109/ITSC45102.2020.9294262.
[44] Mou, Jun Min, Cees Van der Tak, and Han Ligteringen, Study on collision avoidance in busy waterways by using AIS data, Ocean Eng. 37.5-6 (2010) 483-490. https://doi.org/10.1016/j.oceaneng.2010.01.012.
[45] Ha, Jisang, Myung-Il Roh, and Hye-Won Lee, Quantitative calculation method of the collision risk for collision avoidance in ship navigation using the CPA and ship domain, Journal of Computational Design and Engineering. 8.3 (2021) 894-909. https://doi.org/10.1093/jcde/qwab021.
[46] Sutton, Charles, and Andrew McCallum., An introduction to conditional random fields, Foundations and Trends® in Machine Learning. 4.4 (2012) 267-373. http://dx.doi.org/10.1561/2200000013.

[47] Wang, C., Zhang, X., Gao, H., Su, H., Zheng, K., & Wang, W. (2022, October). Efficient Reinforcement Learning for Autonomous Ship Collision Avoidance under Learning Experience Reuse. In 2022 IEEE International Conference on Unmanned Systems (ICUS) (pp. 1563-1568). IEEE. https://doi.org/10.1109/ICUS55513.2022.9986793.

[48] Grześ, Marek, and Daniel Kudenko, Online learning of shaping rewards in reinforcement learning, Neural networks. 23.4 (2010) 541-550. https://doi.org/10.1016/j.neunet.2010.01.001.

[49] Schultz, Wolfram., Neuronal reward and decision signals: from theories to data, Physiological reviews. 95.3 (2015) 853-951. https://doi.org/10.1152/physrev.00023.2014.

[50] J. Garcıa, F. Fernandez, A comprehensive survey on safe reinforcement learning, J. Machine Learn. Res. 16.1 (2015) 1437–1480.

[51] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... & Hassabis, D. Human-level control through deep reinforcement learning. Nature, (2015) 518(7540), 529-533. https://doi.org/10.1038/nature14236.

[52] Wang, C., Zhang, X., Gao, H., Bashir, M., Li, H., & Yang, Z. Optimizing anti-collision strategy for MASS: A safe reinforcement learning approach to improve maritime traffic safety. Ocean Coast Manage, 253, (2024) 107161. https://doi.org/10.1016/j.ocecoaman.2024.107161.

[53] A.C. Cameron, P.K Trivedi, Microeconometrics: Methods and Applications, Cambridge university press, 2005. ISBN: 978-0-521-84805-3.