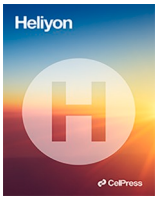
Heliyon 10 (2024) e25105

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

A machine vision method for the evaluation of ship-to-ship collision risk

Zhiqiang Jiang, Lingyu Zhang, Weijia Li *

Huazhong University of Science and Technology, School of Naval Architecture and Ocean Engineering, 1037 Luoyu Road, Hongshan District, Wuhan, Hubei, 430074, China

ARTICLE INFO

Keywords:
Machine vision
Object tracking
Trajectory estimation
Collision warning

ABSTRACT

The development of ship technology and information technology has been driving the continuous improvement of ship intelligence, with safety being an inevitable requirement in the shipping industry. A machine vision-based ship collision warning method is proposed for high monitoring system cost and limited information acquisition in safety design of autonomous ship navigation. The method combines machine learning with image algorithms. Firstly, the backbone of YOLOv7 detector is replaced by EfficientFormerV2 network to achieve model lightweight while ensuring detection accuracy. Public datasets *SeaShips*, *Flow* and self-made ship pictures are combined, and the network is trained on this dataset. *StrongSORT* is used for target tracking. Secondly, a data fusion algorithm is introduced to determine the target point at the bow-bottom of the ship based on the time-varying attitude of the camera and the time-series features of the bounding boxes. Ship navigation trajectory estimation is performed using image algorithms. Finally, a collision evaluation model is established to calculate the collision risk index. Experimental results demonstrate that the improved YOLOv7 network maintains similar mAP.5 and Recall compared to the original model, while reducing the parameters by 31.2 % and GFLOPs by 58.4 %. The accuracy of target ship trajectory estimation is high, with MAE values below 1.5 % and RMSE values below 2 % in experiments. In ship collision warning experiments, the proposed method accurately identifies navigating vessels, estimates the trajectories, and provides timely warnings for imminent collision accidents. Compared to traditional ship collision warning methods, this paper offers a more intelligent and lightweight solution.

1. Introduction

With the continuous growth of global maritime trade, ensuring the safety of ship navigation has become a critical concern [1]. Ship collisions are particularly serious accidents in maritime transportation, posing threats not only to the involved ships but also to the safety of crew members [2]. Additionally, fuel leakage resulting from damage to the ships fuel tanks can lead to severe environmental pollution [3]. Consequently, the development of ship collision warning systems is imperative in preventing collision accidents.

Existing research on collision warning for navigating ships and floating objects, the focus of information acquisition has mostly focused on radar monitoring and wireless communication. Lee et al. [4] utilized millimeter-wave communication terminals to provide collision warning for fishing boat operations and compared various maritime risk assessment models. Kazimierski et al. proposed a

* Corresponding author.
E-mail address: liweijia@hust.edu.cn (W. Li).

https://doi.org/10.1016/j.heliyon.2024.e25105
Received 18 September 2023; Received in revised form 12 January 2024; Accepted 20 January 2024
Available online 22 January 2024

tracking algorithm based on data fusion, utilizing Automatic Identification System (AIS) for radar information processing [5]. However, these methods rely on costly equipment and suffer from low radar image resolution or limited target classification capabilities in complex scenes [6]. Liu et al. combined static and dynamic information of AIS to evaluate ship collision risk by integrating the relative position vector and the relative velocity [7]. In recent years, with the advancement of computer vision and machine learning technologies, visual detection has emerged as an efficient, accurate, and automated detection approach. The intelligentization of ships, encompassing information perception, communication, and navigation, is a crucial direction for the future of the maritime industry.

Compared with expensive detection equipment such as lidar, cameras are more lightweight and low-cost. Water surface target detection is predominantly conducted in two approaches: traditional image processing and deep learning [8]. Traditional image algorithms rely on water surface background features and filtering theory for target recognition, providing fast detection and recognition speeds. However, they are susceptible to complex environmental factors such as target scale and lighting variations, leading to challenges in meeting robustness requirements. On the other hand, deep learning-based methods primarily extract target features through convolutional neural networks (CNNs). Minami et al. proposed a floating object detection algorithm based on image color space gradients, enabling recognition of river debris through the color characteristics [9]. Liu and Zhu developed an improved convolutional neural network-based ship detection algorithm by incorporating channel attention mechanism (CAM) and weighted bidirectional feature fusion network (BiFPN), facilitating efficient detection and recognition of naval vessels on the sea surface [10]. These technologies enable automated detection and recognition of floating objects by analyzing features such as shape and texture. Zhang et al. utilized Faster R–CNN to detect small clustered vessels in coastal and inland water areas using high-resolution remote sensing images [11]. Based on Retinex theory, Guo et al. proposed a LVENet model to guarantee reliable vessel detection under low-visibility conditions [12]. Visual detection techniques enable real-time detection and tracking of objects such as ships on the water surface. In the presence of potential collision risks, the system issues timely alerts for necessary measures, effectively preventing collision accidents from navigating ships. Kristan et al. performed unsupervised floating obstacle detection using video footage captured by unmanned surface vehicles (USVs), achieving rapid and continuous detection of water surface targets [13]. Machine vision has been widely applied in land-based scenarios. By utilizing cameras, sensors, and image processing algorithms, machine vision technology enables real-time monitoring and analysis of the environment and obstacles, providing timely collision warnings [14]. Target detection research significantly contributes to maritime safety, offering profound scientific implications and extensive application prospects.

Traditional ship collision warning system relies on positioning data provided by the AIS system [15]. AIS (Automatic Identification System) is a shipborne broadcast response system. Through the AIS system, vessels can continuously transmit their identity, position, heading, speed, and other data to nearby ships and shore authorities on the VHF public wireless channel. For certain non-AIS intelligent ship collision warning scenarios, such as small vessels without equipped lidar or instances where the AIS system malfunctions, a low-cost and lightweight ship collision warning solution needs to be proposed.

In order to facilitate deployment in resource-constrained scenarios and address the challenges of high monitoring system cost and limited information acquisition in ship collision warning, an intelligent and cost-effective machine vision-based ship collision warning method is proposed in this paper. The method utilizes an improved lightweight YOLOv7 detection model and *StrongSORT* to analyze images captured by cameras fixed on the ship, detecting and tracking navigating ships or floating objects in the scene. Moreover, based on the bounding box and camera attitude, spatial coordinate calculation is performed to identify potential collision targets. Finally, the collision risk index (CRI) is computed to recognize and alert imminent collision behavior, ensuring the safety of ship navigation.

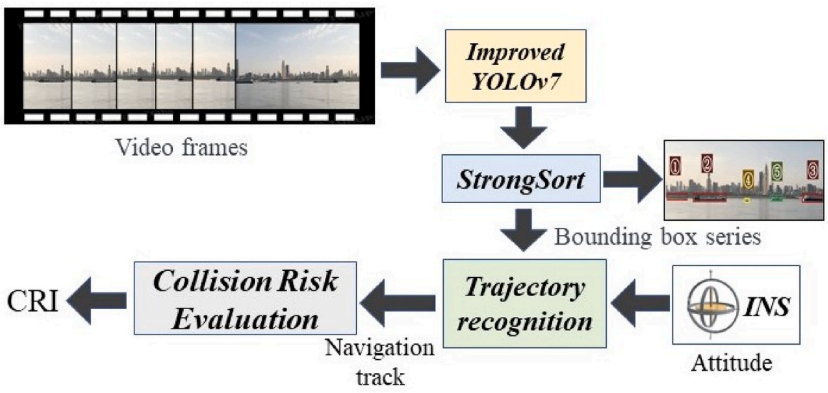


Fig. 1. Method framework.

2. Methodology

2.1. Framework

The proposed ship collision warning method based on monocular vision comprises three main parts: object detection and tracking, spatiotemporal feature extraction, and collision risk evaluation. In object detection and tracking, an improved lightweight YOLOv7 detection model is proposed to identify texture features of target ships. The convolutional feature extraction layers in the backbone were replaced with the EfficientFormerV2 network, reducing the model's parameters while maintaining high detection accuracy. The *StrongSORT* is used for tracking after object detection.

In spatiotemporal feature extraction, a data fusion-based target point selection algorithm is proposed. This algorithm performs spatiotemporal feature extraction of navigating ships by utilizing the bounding box information and camera attitudes, allowing for the calculation of navigation trajectory. Finally, a model is applied to calculate the collision risk index between two ships, evaluate the risk level, and issue corresponding warnings.

The method can be summarized as the following processes: 1) The improved lightweight YOLOv7 and *StrongSORT* algorithms are utilized to detect and track target ships, obtaining their image information. 2) Based on the acquired image information and camera attitude provided by Inertial Navigation System (INS), the coordinates of the target ships are estimated. Additionally, the ship's speed and direction are estimated. 3) Through mathematics model, the collision risk index is calculated to provide warnings for potential collisions. The overall algorithm flow is illustrated in Fig. 1.

2.2. Improved YOLOv7 and StrongSORT

YOLOv7 is an advanced object detection model that shows significant improvements in detection accuracy, speed, and improvements over previous versions like YOLOv5. YOLOv7 incorporates data augmentation and network structure optimizations to enhance model performance [16]. It also utilizes adaptive training strategies to dynamically adjust the learning rate and batch size during training, allowing better adaptation to different data distributions and scenarios.

In CNN-based neural network architectures, convolution operations are used to extract essential features from images. However, due to the local receptive field of CNN, multiple convolutional layers need to be stacked to increase the receptive field [17], leading to increased network complexity. In the recognition of navigating ships, adding attention mechanisms allows the neural network to focus more on specific texture features of ships or other large floating objects, while disregarding irrelevant information such as complex backgrounds on the sea surface. Compared to CNN convolutional models, Transformers can capture long-range contextual dependencies with the help of self-attention mechanisms [18]. However, the original Transformer architecture has drawbacks such as high memory consumption and computational cost. In this study, images are input to the network with a size of 640x640x3, and the feature extraction layers were replaced with the EfficientFormerV2 network, incorporating an attention mechanism to enhance feature extraction capability while reducing model parameters. This enables efficient deployment on resource-constrained hardware [19]. EfficientFormerV2 utilizes a fine-grained architecture search algorithm to jointly optimize model size and speed, balancing between model size and detection speed. The performance evaluation metric, Mobile Efficiency Score (MES), is calculated by Equation (1):

$$MES = Score \cdot \prod_i \left(\frac{M_i}{U_i}\right)^{-\alpha i} \tag{1}$$

where M_i is the corresponding metric, U_i is the corresponding unit, *Score* is the predefined base score, αi is the corresponding weights. In this paper, the YOLOv7 backbone feature extraction network is replaced with the EfficientNetV2 network, achieving model lightweight while balancing detection speed and accuracy.

StrongSORT is a multi-object tracking algorithm that enables efficient and accurate object tracking in high-density target scenarios [20]. It is designed to handle complex situations such as target occlusion, scale variation, and motion blur. The algorithm adopts an end-to-end multi-object tracking framework that formulates the tracking problem as an optimization task. *StrongSORT* utilizes algorithms for sample resampling and reassignment to achieve efficient target tracking. Additionally, A confidence-based object detection strategy is incorporated, which improves target identification and enhances tracking accuracy. The framework of *StrongSORT* is shown

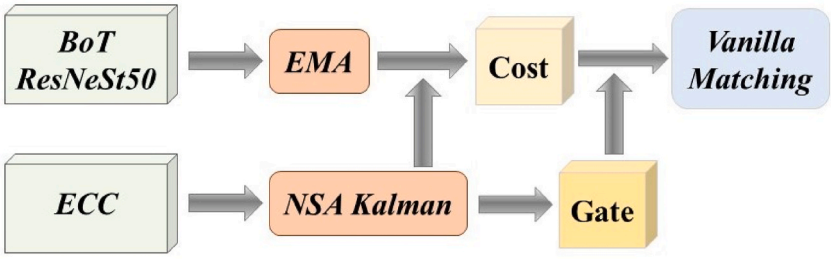


Fig. 2. *StrongSORT* framework.

in Fig. 2.

Compared to *DeepSORT* (the previous version of *StrongSORT*), the feature extractor BoT and ResNeSt50 are used to replace the CNN to enhance feature extraction capabilities [21,22]. Although *DeepSORT* can preserve long-term information, it is sensitive to detection noise. To suppress noise, the feature library mechanism is replaced by exponential moving average (EMA). The enhanced correlation coefficient maximization (ECC) model is used for camera motion compensation. ECC is a parametric image alignment technique that estimates global rotation and translation between adjacent frames. In *DeepSORT*, ordinary Kalman filters are susceptible to low-quality detections and ignore information on the scale of detection noise. To solve this problem, the NSA Kalman algorithm in GIAOTracker is used in *StrongSORT*. Matching cascades are also replaced by vanilla global linear assignments to improve matching accuracy [23].

2.3. Trajectory recognition algorithm

In the process of transforming a point from the pixel coordinate system captured by a camera to the earth coordinate system, four coordinate systems are involved: the earth coordinate system, camera coordinate system, image coordinate system, and pixel coordinate system. Assuming the imaging point of the camera fixed on the ship serves as the origin of the earth coordinate system. It is feasible to calculate the coordinates in the earth coordinate system for any point on the image. Fig. 3 provides a schematic diagram illustrating the transformations between these coordinate systems.

In the scenario where the camera is fixed on a ship or buoy, with the ship in a stationary state beneath calm water. The origin of the earth coordinate system coincides with the origin of the camera coordinate system. The camera's extrinsic rotation matrix is an identity matrix, and the translation vector is a zero vector. While the ship experiences movement due to wind and waves, the camera's position also changes, and is always in a non-stationary state. Any point in space has a specific set of coordinates in the earth coordinate system. In the camera coordinate system, the mapping of this point onto the camera photosensitive surface determines the relationship between the camera coordinate system and the image coordinate system. Finally, this is transformed into a digital signal recognizable by the computer, establishing the relationship between the image coordinate system and the pixel coordinate system [24,25]. Therefore, the transformation from the spatial coordinate system to the pixel coordinate system involves three transformations and four coordinate systems.

The transformation matrix between the earth coordinate system and the camera coordinate system is:

$$\begin{bmatrix} X_C \\ Y_C \\ Z_C \end{bmatrix} = R_t \begin{bmatrix} X_W \\ Y_W \\ Z_W \end{bmatrix} + T_t \tag{2}$$

where $[X_C, Y_C, Z_C]$ is the coordinates of a point in the camera coordinate system, $[X_W, Y_W, Z_W]$ is the coordinates of the same point in the earth coordinate. R_t, T_t denote the camera's time-varying rotation matrix and translation vector, respectively. These parameters can be measured using attitude sensors such as an Inertial Navigation System (INS).

Without considering lens distortion, the camera is treated as an ideal pinhole model. The projection of a point $[X_C, Y_C, Z_C]$ in space onto the entire two-dimensional image space in the camera coordinate system yields the projected point $[X_p, Y_p]$ in the image coordinate system based on the geometric relationships of imaging:

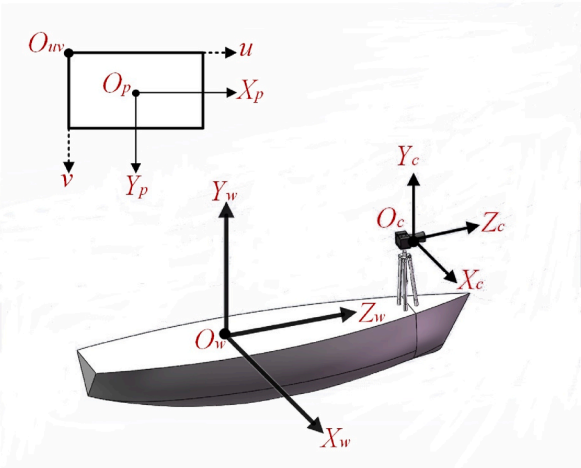


Fig. 3. Coordinate transformation. $O_W - X_W Y_W Z_W$: earth coordinate system; $O_C - X_C Y_C Z_C$: camera coordinate system; $X_p O_p Y_p$: image coordinate system; $uO_{uv}v$: pixel coordinate system.

$$\begin{cases} X_p = f\dfrac{X_C}{Z_C} \\ Y_p = f\dfrac{Y_C}{Z_C} \end{cases} \tag{3}$$

The conversion of simulated signals in the image coordinate system to computer digital signals involves the projection of points $[X_p, Y_p]$ in the image coordinate system onto points $[u, v]$ in the pixel coordinate system. This relationship can be expressed as:

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & u_0 \\ 0 & f_y & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X_p \\ Y_p \\ 1 \end{bmatrix} \tag{4}$$

From solving the system of Equation (2) to Equation (4), Equation (5) is utilized to transform a point from the pixel coordinate system $[u, v]$ to the earth coordinate system $[X_W, Y_W, Z_W]$ is:

$$Z_c \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & u_0 & 0 \\ 0 & f_y & v_0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} R_t^T & -T_t \\ \overrightarrow{0} & 1 \end{bmatrix} \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix} \tag{5}$$

where $\begin{bmatrix} f_x & 0 & u_0 \\ 0 & f_y & v_0 \\ 0 & 0 & 1 \end{bmatrix}$ is the camera intrinsic matrix, which can be determined using the Zhang's calibration method [26]. f_x, f_y represents the use of pixels to describe the length of the focal length in the XY direction, that is, the focal length of the image in both directions. $[u_0, v_0]$ is the principal point, which is a point on the image plane that the optical axis intersects. It is often considered the center of the image. u_0 represents the coordinate of the principal point in the horizontal direction of the image, and v_0 represents the coordinate of the principal point in the vertical direction of the image. $\begin{bmatrix} R_t^T & -T_t \\ \overrightarrow{0} & 1 \end{bmatrix}$ is the extrinsic matrix. In calm water, it is reasonable to assume that $Y_w = H$, where H is the distance between the camera and the water surface. In this paper, the selection of target points is performed based on the position and movement direction of target bounding boxes. However, since the camera may experience motion due to ship vibrations, it becomes necessary to project the coordinates of the target bounding box onto the camera coordinate system aligned with the calm water surface. In the *i*-th frame, the calm-water pixel coordinates of the top-left corner of the bounding box $[u_i, v_i]$ can be calculated as Equations (6) and (7):

$$Z_W \begin{bmatrix} u_i \\ v_i \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & u_0 & 0 \\ 0 & f_y & v_0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} X_{Wi} \\ -H \\ Z_{Wi} \\ 1 \end{bmatrix} \tag{6}$$

The image resolution is $m \times n$. and $[w_i, h_i]$ represents the size of bounding boxes. The target research point $[u_t, v_t]$ selected further analysis can be calculated as follows:

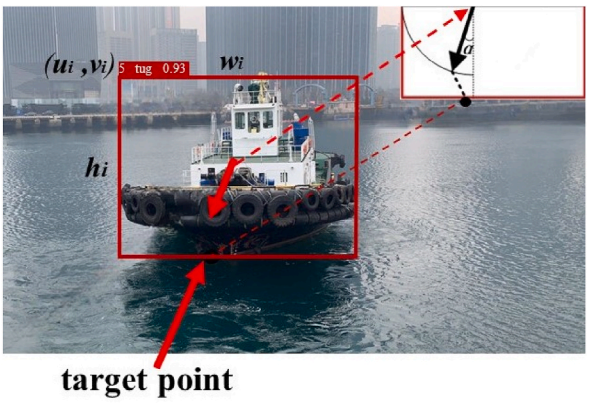


Fig. 4. target point selection.

$$\begin{cases} u_t = u_i + w_i C_u + D_u[(1 - C_u)H(\alpha) + C_u H(-\alpha)] \\ v_t = v_i + h_i \end{cases} \tag{7}$$

where C_u is the u-coordinate position coefficient; D_u is the motion direction coefficient, α is the angle between the motion direction of the target bounding box and the $v-$axis, and $H(\alpha)$ is the unit step function. The parameters above can be calculated by Equations (8)–(11), and Fig. 4 is the schematic diagram of the algorithm.

$$C_u = \frac{2u_i + w_i}{2m} \tag{8}$$

$$D_u = \frac{2w_i\alpha}{\pi} \tag{9}$$

$$\alpha = \tan^{-1}\frac{\Delta u_i}{\Delta v_i} \tag{10}$$

$$H(\alpha) = \begin{cases} 1, \alpha > 0 \\ 0, \alpha \le 0 \end{cases} \tag{11}$$

By projecting the target point $[u_t, v_t]$ into the earth coordinate system, the earth coordinate of the target point $[X_{wt}, -H, Z_{wt}]$ can be obtained by combining Equations (6) and (7). Further, the velocity of the target point can be obtained by Equation (12):

$$\begin{cases} v_{rx} = \dfrac{X_{Wt} - X_{Wt-1}}{\Delta t} \\ v_{rz} = \dfrac{Z_{Wt} - Z_{Wt-1}}{\Delta t} \end{cases} \tag{12}$$

However, due to the differences in the size of the bounding boxes in different frames, applying smoothing techniques to the trajectory is urged. In this paper, a second-order Butterworth filter is used to perform low-pass filtering on the target point sequence. The transfer function of the filter is defined as Equation (13):

$$|H(\omega)|^2 = \frac{1}{1 + \left(\frac{\omega}{\omega_c}\right)^{2n}} \tag{13}$$

where n is the order of the filter; ω_c is the cutoff frequency.

2.4. Collision risk assessment model

The visually detected ship's navigation trajectory is utilized for ship collision prevention through collision warning systems. The degree of collision likelihood between ships is commonly represented by the collision risk index, which ranges from 0 to 1. In this paper, the collision risk index of the target ship is determined by weighted summing the membership degrees of various risk factors, as expressed in Equation (14) [27]:

$$CRI = a_{DC}u_{DC} + a_{TC}u_{TC} + a_D u_D + a_\theta u_\theta + a_K u_K \tag{14}$$

where a_{DC}、a_{TC}、a_D、a_θ、a_K are weights of the parameter; u_{DC}、u_{TC}、u_D、u_θ、u_K are the membership degrees of distance closest point of approach (DCPA)、time closest point of approach (TCPA)、distance (D)、relative bearing angle (θ) and speed ratio (K). The formulas for calculating DCPA and TCPA are as Equation (15):

$$\begin{cases} DCPA = D \cdot \sin(Q_r) \\ TCPA = D \cdot \dfrac{\cos(Q_r)}{v_r} \end{cases} \tag{15}$$

where Q_r represents the relative bow angle between the two ships; v_r represents the relative velocity between the two ships.

The collision risk index can be calculated from the target ship's navigation trajectory coordinates over time, with the imaging point of the camera fixed on the ship being as the origin of the earth coordinate system. The membership degrees for each risk factor are obtained using the calculations described in Equation (16) to Equation (20) [28].

$$u_{DC} = \begin{cases} 1, DCPA \le d_1 \\ \frac{1}{2} - \frac{1}{2} sin\left[\frac{\pi}{d_2 - d_1}\left(DCPA - \frac{d_1 + d_2}{2}\right)\right], d_1 \le DCPA \le d_2 \\ 0, DCPA \ge d_2 \end{cases} \tag{16}$$

$$u_{TC} = \begin{cases} 1, |TCPA| \le t_1 \\ \left(\frac{t_2 - |TCPA|}{t_2 - t_1}\right)^2, t_1 \le |TCPA| \le t_2 \\ 0, |TCPA| \ge t_2 \end{cases} \tag{17}$$

$$u_D = \begin{cases} 1, D \le d_1 \\ \left(\frac{d_2 - D}{d_2 - d_1}\right)^2, d_1 \le D \le d_2 \\ 0, D \ge d_2 \end{cases} \tag{18}$$

$$u_\theta = 0.5 \times \left[cos(\theta - 19^\circ) + \sqrt{\frac{440}{289} + cos^2(\theta - 19^\circ)}\right] - \frac{5}{17} \tag{19}$$

$$u_K = \frac{1}{1 + \frac{2}{K\sqrt{K^2 + 2KsinC + 1}}} \tag{20}$$

where d_1 is the minimum safe meeting distance, and d_2 is the minimum distance between the striking ship and struck ship; t_1 is the collision time, and t_2 is the collision avoidance time; C is the collision angle and can be calculated as $C = \theta_T - \theta_0$, where θ_T and θ_0 are the course of the striking ship and struck ship. Each parameter can be calculated as Equation (21) to Equation (23):

$$d_1 = \begin{cases} 1.1 - \frac{0.2\theta}{180^\circ}, 0^\circ \le \theta \le 112.5^\circ \\ 1.0 - \frac{0.4\theta}{180^\circ}, 112.5^\circ \le \theta \le 180^\circ \\ 1.0 - \frac{0.4 \times (360^\circ - \theta)}{180^\circ}, 180^\circ \le \theta \le 247.5^\circ \\ 1.1 - \frac{0.2 \times (360^\circ - \theta)}{180^\circ}, 247.5^\circ \le \theta \le 360^\circ \end{cases} \tag{21}$$

$$t_1 = \frac{\sqrt{d_1^2 - DCPA^2}}{S_r} \tag{22}$$

$$t_2 = \frac{\sqrt{d_2^2 - DCPA^2}}{S_r} \tag{23}$$

Based on extensive statistical research, the weights of the factors influencing collision risk index are determined as follows: $a_{DC} = 0.40$、 $a_{TC} = 0.367$、 $a_D = 0.167$、 $a_\theta = 0.033$、 $a_K = 0.033$.

After performing the calculations in the above equations, the collision risk index of the target ship can be obtained, with values ranging [0, 1]. A collision risk index of 1 signifies the highest level of collision threat between the ships, indicating an imminent collision risk. Conversely, a collision risk index of 0 indicates no collision threat between the ships. Based on the magnitude of the collision risk index, the collision warning algorithm categorizes the collision risk into five levels. The 5th level collision warning indicates the lowest level of risk, with only a possibility of collision between the two ships. On the other hand, the 1st level collision warning signifies the highest level of danger, indicating a high probability of collision and necessitating immediate action.

3. Dataset and parameters

3.1. Dataset source and introduction

This paper focuses on the recognition of sailing ships and floating objects on the sea surface in complex environments. In order to increase the robustness and anti-interference ability of the model, the public datasets *SeaShips* and *FloW* [29,30], are randomly sampled and combined. The ship model pictures used in the subsequent experiments were also added to the dataset for model training.

The images were annotated using the annotation tool *LabelImg*. Subsequently, the dataset was randomly split into training, validation, and testing sets. The training set consisted of 80 % of the images, while the validation and testing sets accounted for 10 % each. To meet the requirements of recognition in complex environments and enhance the robustness of the model, data augmentation was used in this experiment as shown in Fig. 5. The training set images were randomly subjected to rotation, noise addition, occlusion, and other processing methods. This approach increased the quantity of the dataset and enriched the diversity of the samples. The sample distribution of the dataset is shown in Table 1 (see Table 2).

3.2. Platform and parameter settings

The experiments in this study were conducted in Windows 10 Professional operating system. The CPU was AMD Ryzen 7 5800H@ 3.2 GHz. The GPU was Nvidia RTX3060 Ultra W OC 12 GB, and the CUDA version was 11.3. The program was implemented using Python 3.7 and based on the PyTorch. The hyperparameters for network training are shown in the table below.

4. Results and discussion

4.1. Networks comparative experiment

To comprehensively evaluate the model's performance, Recall (R) and Mean Average Precision (mAP) are used as performance metrics. Recall measures the proportion of correctly detected positive samples among all positive samples, reflecting the model's ability to detect positive samples. It can be calculated using Equation (24):

$$R = \frac{TP}{(TP + FN)} \tag{24}$$

where TP (True Positive) is the number of positive samples correctly detected, and FN (False Positive) refers to false negatives, indicating the number of positive samples that were not correctly detected.

Mean average precision (mAP) is a comprehensive metric that combines Precision and Recall. It represents average of the average precision (AP) scores calculated for multiple different sample categories. The mAP can be calculated as Equation (25):

$$mAP = \frac{1}{N}\sum_{1}^{N}\left(\int_{0}^{1} p_i(r)dr\right) \tag{25}$$

To evaluate the superiority of the improved YOLOv7 on the dataset, several object detection model such as SSD, Faster R–CNN, YOLOv7-tiny, and YOLOv7 were used for comparative experiments. The results are shown in Table 3.

The improved YOLOv7 model the other four models on the self-built dataset. Compared to SSD, Faster R–CNN, and YOLOv7, the improved lightweight YOLOv7 network achieves higher Recall and mAP.5 while having fewer parameters. The Recall improves by 8.0 %, 7.9 %, and 1.1 %, respectively, and the mAP.5 improves by 7.7 %, 5.9 %, and 0.1 %, respectively. The improved lightweight YOLOv7 network also reduces parameters by 29.9 %, 37.6 %, 31.2 %, and GFLOPs by 75.3 %, 76.4 %, 58.4 %, respectively. Although YOLOv7-tiny has a much smaller parameter size and GFLOPs, the Recall and mAP.5 decrease by 5.6 % and 4.3 %, respectively. The

Fig. 5. Dataset augmentation.

Table 1
Dataset distribution.

Name	Images	Sample Types	Train	Validation	Test
Value	8200	7	6560	820	820

Table 2
Hyperparameters in training.

Hyperparameters	Value
Initial learning Rate	0.01
Coefficient	0.1
Momentum	0.937
Weight Decay	0.0005
Batch size	8
Input size	640*640
Optimizer	Adam

Table 3
Model comparison experiment results.

Model	Input Size	Params	GFLOPs	Recall	mAP.5
SSD	640	36.5 M	179.8	0.874	0.898
Faster R–CNN	640	41.0 M	188.1	0.875	0.916
YOLOv7-tiny	640	**6.1M**	**13.9**	0.898	0.932
YOLOv7	640	37.2 M	106.5	0.943	0.974
Ours	640	25.6 M	44.3	**0.954**	**0.975**

Table 4
Accuracy experiment results.

Exp	MAE	RMSE
①	0.0073	0.0105
②	0.015	0.0161
③	0.0147	0.017
④	0.0148	0.0176
⑤	0.0149	0.0196

Table 5
Risk rating.

Risk Degree	CRI	Risk Assessment
Level I	0.8–1.0	Very Dangerous
Level II	0.6–0.8	Dangerous
Level III	0.4–0.6	Normal
Level IV	0.2–0.4	Safe
Level V	0~.02	Very Safe

accuracy of the model drops significantly. The experiments demonstrate that the improved YOLOv7 network, incorporating the Transformer attention mechanism, achieves smaller model size while maintaining detection accuracy, making it advantageous for ship detection.

4.2. Trajectory accuracy experiment

A scaled ship model was used for experiments in order to verify the vision algorithm proposed in this paper. Since it is difficult to accurately measure the earth coordinates of the ship model outdoors, an indoor trajectory accuracy experiment and an outdoor collision warning experiment were set in this paper for cross-validating the performance of the algorithm. A trajectory accuracy experiment was conducted in laboratory using a high-precision Stewart platform as shown in Fig. 6. The model ship was securely mounted on the Stewart platform, and the platform motion was precisely controlled by program. An industrial camera was fixed at a specified observation point for recording. In the laboratory, the parameters of the Stewart platform and the coordinates of the observation point were known or measured in advance. The high-precision Stewart platform had a maximum pose error of 0.05mm/

Fig. 6. Experiment equipment.

0.001°, ensuring the accuracy of the actual trajectory of the model ship.

Based on the known parameters, the vertical distance between the camera lens and the upper surface of the Stewart platform (H) was measured as *0.513 m*. The platform executed circular motion with a radius of *0.4 m* in the X_WOZ_W plane. The camera captured the motion of the model ship, and trajectory calculations were performed. The entire time series trajectory of the model ship's motion is depicted in Fig. 7.

For the trajectory accuracy experiment, the model ship was fixed at five different positions on the upper platform, and the platform executed the same circular motion with a radius of R = 0.4 m for each experiment. Due to the presence of radial distortion in the camera lens, image correction was performed during the calculations. The trajectory accuracy experiment and resulting trajectories are presented in Fig. 8.

In each experiment, several circular trajectories with a radius of R = 0.4 m were selected for analysis. Fig. 9 compares the visually calculated trajectories from the five experiments with the actual trajectories of the model ship. The accuracy of the visually calculated trajectories was evaluated with the Mean Absolute Error (MAE) as Equation (26) and Root Mean Square Error (RMSE) as Equation (27).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\sqrt{\left(X_{Wi} - X_{Wi}^{'}\right)^2 + \left(Z_{Wi} - Z_{Wi}^{'}\right)^2} \tag{26}$$

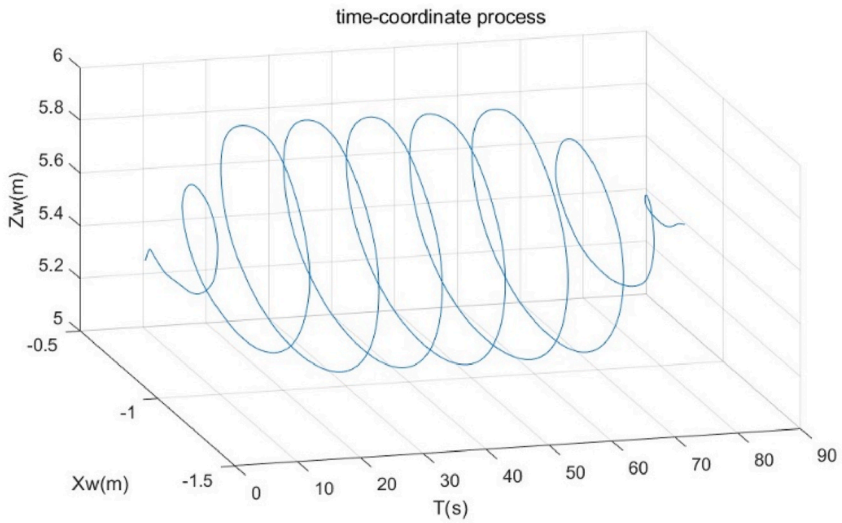


Fig. 7. Calculated ship model motion series. From the beginning to the end of the Stewart platform movement in a single experiment, the ship model space coordinate time series recognized by the image algorithm.

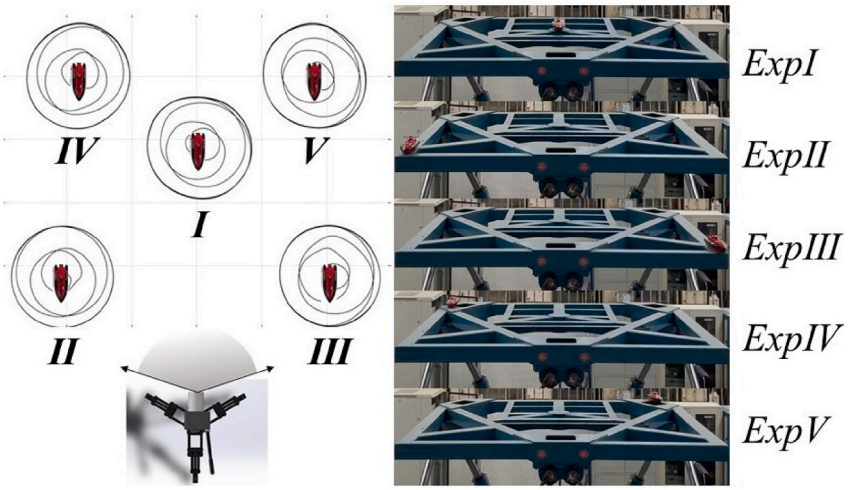


Fig. 8. Accuracy experiment and results.

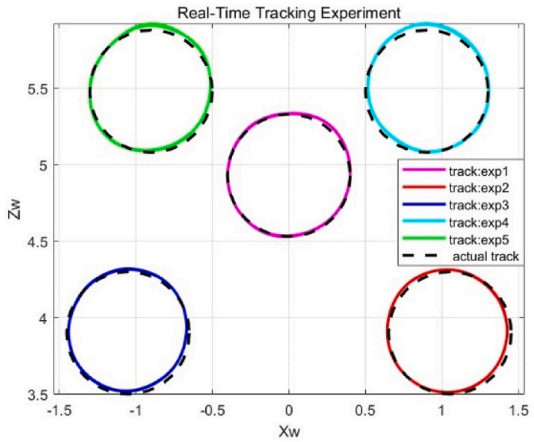


Fig. 9. Computed Trajectories and Real Trajectories. The dotted line is the actual displacement curve of the point on the Stewart platform, and the solid line is the coordinate average of the maximum circular motion of the ship model identified by the algorithm in the five experiments.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_{Wi} - X_{Wi}^{'}\right)^2 + \left(Z_{Wi} - Z_{Wi}^{'}\right)^2} \tag{27}$$

Fig. 10. Collision experiment and camera frames.

where $[X_{Wi}, Z_{Wi}]$ is the real trajectory coordinate, $[X'_{Wi}, Z'_{Wi}]$ is the identified trajectory coordinate by the algorithm. MAE is the average of the absolute errors, which can better reflect the actual situation of the predicted value error. RMSE is the square root of the sum of squares of the deviations between the observed value and the true value and the square root of the ratio of the number of observations m, which is used to measure the deviation between the observed value and the true value.

During the trajectory accuracy experiment, the Root Mean Square Error (RMSE) between the calculated trajectories and the actual trajectories remains within 2 %, and the Mean Absolute Error (MAE) remains within 1.5 % as shown in Table 4. This indicates that the visual coordinate estimation algorithm proposed in this paper achieves high accuracy. The consistency between the model-calculated trajectories and the actual measured trajectories validates the accuracy and reliability of the proposed method described in this paper.

4.3. Collision warning experiment

To validate the collision warning capability of the visual algorithm, two ship models were used for the outdoor experiment. As shown in Fig. 10, a pre-calibrated camera and an Inertial Navigation System (INS) were installed on a monitoring ship model to capture images and attitude data. The monitoring ship model remained stationary without any active manipulation. A target ship model was maneuvered to approach the monitoring ship. The visual tracking information was utilized to calculate the trajectory of the target ship for collision warning purposes.

The video frames captured by the camera were processed for object detection and tracking. Concurrently, the data collected by the INS was processed to convert the real-time 6-Dof attitude of the camera into a rotation matrix and translation vector. Both the camera and attitude sensor operated at a sampling frequency of 30Hz. In object detection, the size of the target in the image and the receptive field size of the network model determines the resolution of the input image. The receptive field refers to the region of the input image that influences each neuron in the neural network. In Convolutional Neural Networks (CNNs), the receptive field of a convolutional layer is determined by the size of the convolutional kernel. The resolution of the input image directly affects the number of receptive fields. Higher resolutions may require more convolutional layers to capture additional details and features in the image. Larger input image resolutions help the network comprehensively understand and capture the content of the image [31]. In collision warning experiments, the size of the target ship model gradually changes from far to near in the image. To reduce network computation while ensuring recognition accuracy, the resolution of the input video is resized to 800*800.

Based on the trajectory of the target ship model, collision warning was conducted by calculating the collision risk index. The monitoring ship model had a length of 0.32 m, and the camera lens center was positioned 0.18 m above the water surface. Figs. 11 and 12 present the results of the 6-DoF attitude of the monitoring ship model and the recognition outcomes of the target ship model trajectories in five experiments, respectively.

For the collision warning experiments, the influence of environmental factors such as weather, lighting, and human factors was minimized. The experiments were conducted on a calm lake under sunny conditions. It was assumed that visibility was good, the operator possessed sufficient skills, the water conditions were favorable, and both vessels had good maneuverability. The results of the five collision warning experiments are illustrated in Figs. 13 and 14.

From Fig. 14, it can be observed that in the five experiments, the target ship model gradually approached the monitoring ship model, resulting in an increasing collision risk (see Table 5). In Experiment 1, the azimuth of the target ship model exhibited a meandering path towards the monitoring ship model, resulting in increased fluctuations in the risk of collision. In Experiment 2, the azimuth of the target ship model was most directly pointed towards the monitoring ship model, leading to an early attainment of level 1 collision risk. The consistency of the collision risk indexes and ship model's trajectories in all five experiments demonstrates the feasibility and effectiveness of ship collision warning through the visual algorithm.

5. Conclusions and discussion

In the absence of AIS system assistance, maritime navigation is exposed to the potential threat of collision accidents. When ships operate in non-AIS states, lacking timely location and navigation data remains a challenge even in modern maritime environments. This absence can lead to an increased risk of collisions, as ships may not have a comprehensive understanding of the surrounding navigational environment. To address this potential safety concern, our research focuses on utilizing machine vision technology, particularly in scenarios without AISsupport, to develop an efficient ship collision warning system. Through this approach, we aim to enhance navigational safety and ensure secure maritime operations in situations where AIS assistance is unavailable. The research also aims to address the challenges of high monitoring system cost and limited information acquisition in safety design of autonomous ship navigation. The key research contributions are as follows.

1) A lightweight YOLOv7 network is proposed by replacing the backbone with EfficientFormerV2, achieving model lightweight without compromising detection accuracy. The improved YOLOv7 network outperforms SSD, Faster R–CNN, YOLOv7-tiny, and the original YOLOv7 models on the dataset. The Recall and mAP.5 demonstrate improvements of 8.0 %, 7.9 %, 5.6 %, 1.1 % and 7.7 %, 5.9 %, 4.3 %, 0.1 % respectively. The parameters are reduced by 31.2 % and GFLOPs are reduced by 58.4 %, making it more easily deployable in resource-constrained environments.
2) A data fusion-based spatiotemporal feature extraction algorithm is proposed for ship collision warning.
   a. The target point selection is based on the time-varying bounding boxes of the target ship, combined with the navigation direction characteristics.

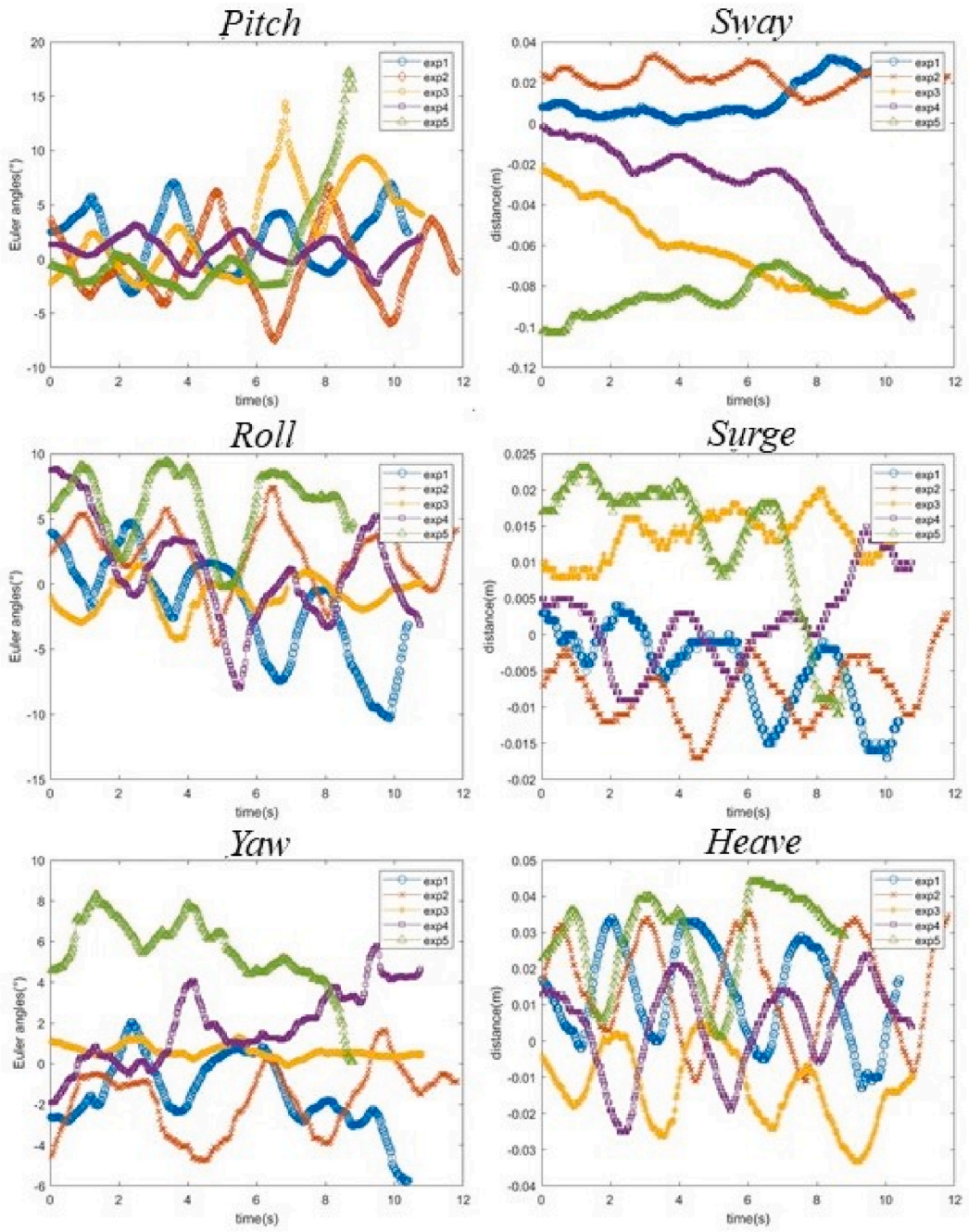


Fig. 11. Observed 6-dof attitude time series of the monitoring ship in five experiments. The five different-colored curves in each diagram represent the 6-dof motion of the monitoring ship in each experiment.

b. The ship's trajectory is computed in the earth coordinate system using the pixel coordinates of the target ship. A collision risk assessment model is established to calculate the ship's collision risk index (CRI), enabling timely warning of potential collision accidents.

However, the method proposed in this paper is simultaneously influenced by both hydro-meteorological factors and ship dynamic factors. Adverse weather conditions such as heavy fog, rain, and snow can reduce visibility, thereby decreasing the accuracy of image recognition. Additionally, variations in sea conditions directly affect the posture and motion of vessels, potentially causing noticeable swaying or pitching in waves, making them more challenging to identify visually. In subsequent research, in order to reduce the interference of these environmental factors on the recognition algorithm, the research direction can be carried out from aspects such as

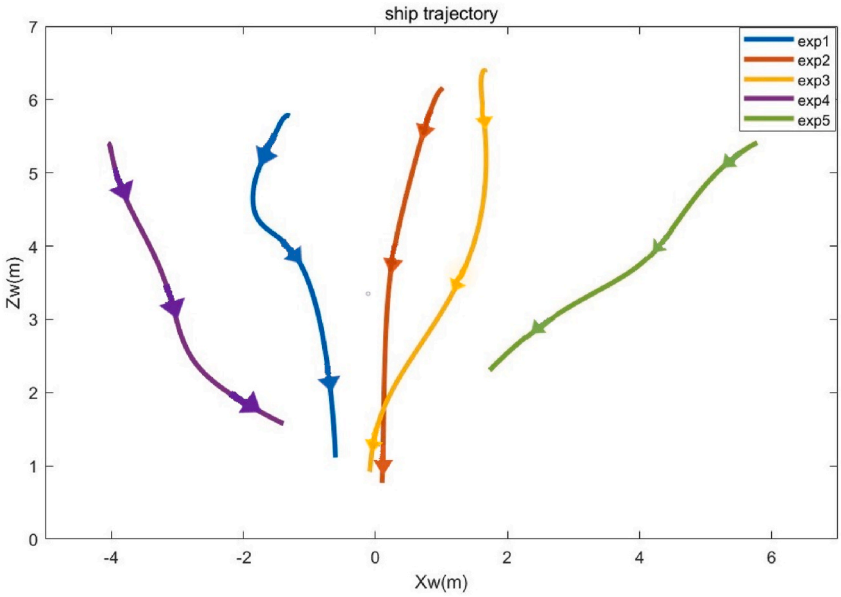


Fig. 12. Ship model trajectory in five experiments. In five experiments, the target ship sailed towards the monitoring ship from different starting points. The position of monitoring ship is at $(X_w = 0, Z_w = 0)$. The position of monitoring ship is at $(X_w = 0, Z_w = 0)$.

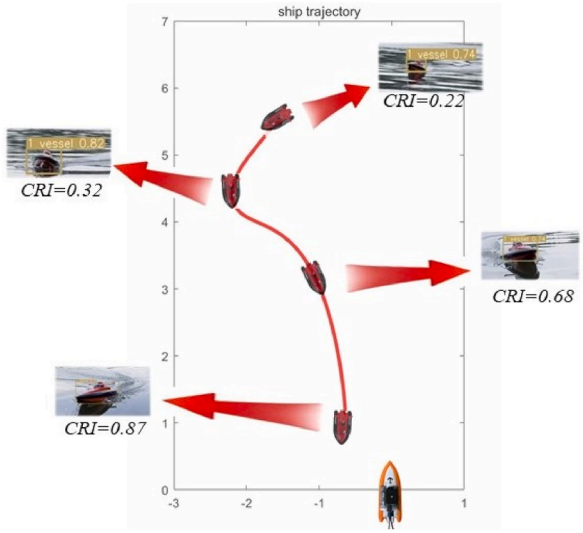


Fig. 13. Trajectory and collision risk index in Exp 1. As the target ship approaches, the risk of collision increases.

image defogging technology and attention mechanism.

Overall, the proposed method demonstrates the potential to enhance safety in ship navigation by effectively detecting and tracking maritime objects. The fusion of visual and spatial-temporal information provides a new low-cost method for ship collision warning, contributing to the prevention of collision accidents and improving maritime safety standards.

Data Availability statement

Data will be made available on request.

Ethics declarations

All participants provided informed consent to participate in the study.

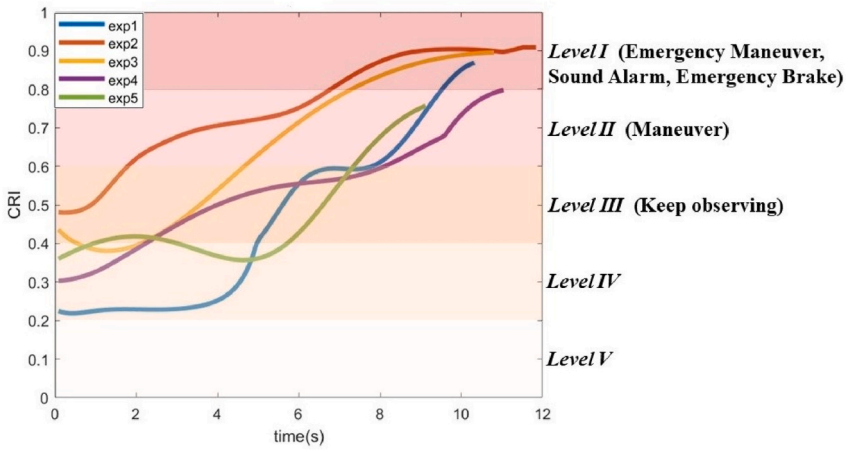


Fig. 14. collision risk index with risk levels in five experiments. Five different collision risk index curves represent the collision risk of the target ship in the five experiments.

CRediT authorship contribution statement

Zhiqiang Jiang: Writing – review & editing, Writing – original draft, Conceptualization. **Lingyu Zhang:** Investigation. **Weijia Li:** Supervision.

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] C. Chauvin, S. Lardjane, G. Morel, J.P. Clostermann, B. Langard, Human and organisational factors in maritime accidents: analysis of collisions at sea using the HFACS, ACCIDENT ANAL PREV 59 (2013) 26–37.
[2] S.S. Arici, E. Akyuz, O. Arslan, Application of fuzzy bow-tie risk analysis to maritime transportation: the case of ship collision during the STS operation, Ocean Eng. (2020) 217.
[3] D.D. Miller, K. Tooley, U.R. Sumaila, Large-scale oil spills and flag-use within the global tanker fleet, Environ. Conserv. 42 (2) (2015) 119–126.
[4] M.K. Lee, Y.S. Park, S. Park, E. Lee, M. Park, N.E. Kim, Application of collision warning algorithm Alarm in fishing vessel's waterway, APPL SCI-BASEL 11 (10) (2021).
[5] W. Kazimierski, A. Stateczny, IEEE: fusion of data from AIS and tracking radar for the needs of ECDIS, in: 2013 SIGNAL PROCESSING SYMPOSIUM (SPS), Signal Processing Symposium (SPS), 2013.
[6] A.C. Bukhari, I. Tusseyeva, B.G. Lee, Y.G. Kim, An intelligent real-time multi-vessel collision risk assessment system from VTS view point based on fuzzy inference system, Expert Syst. Appl. 40 (4) (2013) 1220–1230.
[7] Z. Liu, B. Zhang, M. Zhang, H. Wang, X. Fu, A quantitative method for the analysis of ship collision risk using AIS data, Ocean Eng. 272 (2023) 113906.
[8] G. Cheng, J.W. Han, A survey on object detection in optical remote sensing images, ISPRS J PHOTOGRAMM 117 (2016) 11–28.
[9] M. Minami, T. Onoi, Y. Nihei, T. Kataoka, H. Hinata, AN automatic and continuous monitoring system for floating-litter transport in river and its application to field survey in mogami river, in: A. Mynett (Ed.), PROCEEDINGS OF THE 36TH IAHR WORLD CONGRESS: DELTAS OF THE FUTURE AND WHAT HAPPENS UPSTREAM, 36th IAHR World Congress, 2015, pp. 1013–1020.
[10] C. Liu, W.G. Zhu, An improved algorithm for ship detection in SAR images based on CNN, in: Z. Pan, X. Hei (Eds.), TWELFTH INTERNATIONAL CONFERENCE ON GRAPHICS AND IMAGE PROCESSING (ICGIP 2020), vol. 11720, 12th International Conference on Graphics and Image Processing (ICGIP), 2021.
[11] S.M. Zhang, R.Z. Wu, K.Y. Xu, J.M. Wang, W.W. Sun, R-CNN-Based ship detection from high resolution remote sensing imagery, REMOTE SENS-BASEL 11 (6) (2019).
[12] Y. Guo, Y.X. Lu, R.W. Liu, Lightweight deep network-enabled real-time low-visibility enhancement for promoting vessel detection in maritime video surveillance, J NAVIGATION 75 (1) (2022) 230–250.
[13] M. Kristan, V.S. Kenk, S. Kovacic, J. Pers, Fast image-based obstacle detection from unmanned surface vehicles, IEEE T CYBERNETICS 46 (3) (2016) 641–654.
[14] Z.H. Sun, G. Bebis, R. Miller, On-road vehicle detection: a review, IEEE T PATTERN ANAL 28 (5) (2006) 694–711.
[15] L. Zhao, G. Shi, J. Yang, Ship trajectories pre-processing based on AIS data, J. Navig. 71 (5) (2018) 1210–1230.
[16] C. Wang, A. Bochkovskiy, H.M. Liao, YOLOv7: Trainable Bag-Of-Freebies Sets New State-Of-The-Art for Real-Time Object Detectors, 2022, pp. 2207–2696.
[17] M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: D. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars (Eds.), COMPUTER VISION - ECCV 2014, PT I, vol. 8689, 13th European Conference on Computer Vision (ECCV), 2014, pp. 818–833.
[18] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale, 2020, pp. 2010–11929.
[19] Y. Li, J. Hu, Y. Wen, G. Evangelidis, K. Salahi, Y. Wang, S. Tulyakov, J. Ren, Rethinking Vision Transformers for MobileNet Size and Speed, 2022, pp. 2212–8059.
[20] Y. Du, Z. Zhao, Y. Song, Y. Zhao, F. Su, T. Gong, H. Meng, StrongSORT: Make DeepSORT Great Again, 2022, pp. 2202–13514.
[21] N. Aharon, R. Orfaig, B. Bobrovsky, BoT-Sort, Robust Associations Multi-Pedestrian Tracking, 2022, pp. 2206–14651.
[22] K.M. He, X.Y. Zhang, S.Q. Ren, J. Sun, IEEE: deep residual learning for image recognition, in: 2016 *IEEE CONFERENCE on COMPUTER VISION and PATTERN RECOGNITION (CVPR)*. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.

[23] Y. Du, J. Wan, Y. Zhao, B. Zhang, Z. Tong, J. Dong, Giaotracker, A Comprehensive Framework for MCMOT with Global Information and Optimizing Strategies in VisDrone 2021, 2022, pp. 2202–11983.
[24] F. Itami, T. Yamazaki, A simple calibration procedure for a 2D LiDAR with respect to a camera, IEEE SENS J 19 (17) (2019) 7553–7564.
[25] R. Juarez-Salazar, J. Zheng, V.H. Diaz-Ramirez, Distorted pinhole camera modeling and calibration, APPL OPTICS 59 (36) (2020) 11310–11318.
[26] Z.Y. Zhang, A flexible new technique for camera calibration, IEEE T PATTERN ANAL 22 (11) (2000) 1330–1334.
[27] M. Abebe, Y. Noh, C. Seo, D. Kim, I. Lee, Developing a ship collision risk index estimation model based on dempster-shafer theory, Appl. Ocean Res. (2021) 113.
[28] M.Y. Zhang, J. Montewka, T. Manderbacka, P. Kujala, S. Hirdaris, A big data analytics method for the evaluation of ship - ship collision risk reflecting hydrometeorological conditions, RELIAB ENG SYST SAFE (2021) 213.
[29] Y.W. Cheng, J.N. Zhu, M.X. Jiang, J. Fu, C.S. Pang, P.D. Wang, K. Sankaran, O. Onabola, Y.M. Liu, D.B. Liu, et al., FloW: a dataset and benchmark for floating waste detection in inland waters, in: 2021 IEEE/CVF INTERNATIONAL CONFERENCE on COMPUTER VISION (ICCV 2021). 18th IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10933–10942.
[30] Z.F. Shao, W.J. Wu, Z.Y. Wang, W. Du, C.Y. Li, SeaShips: a large-scale precisely annotated dataset for ship detection, IEEE T MULTIMEDIA 20 (10) (2018) 2593–2604.
[31] L. Qi, J. Kuen, J.X. Gu, Z. Lin, Y. Wang, Y.K. Chen, Y.W. Li, J.Y. Jia, C.S. Ieee, Multi-scale aligned distillation for low-resolution detection, in: 2021 IEEE/CVF CONFERENCE on COMPUTER VISION and PATTERN RECOGNITION, CVPR 2021. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14438–14448.