Journal of
Marine Science and Engineering

MDPI

Article

Target Ship Recognition and Tracking with Data Fusion Based on Bi-YOLO and OC-SORT Algorithms for Enhancing Ship Navigation Assistance

Shuai Chen , Miao Gao *, Peiru Shi, Xi Zeng and Anmin Zhang

School of Marine Science and Technology, Tianjin University, Tianjin 300072, China; 2022227022@tju.edu.cn (S.C.)
* Correspondence: gaomiao@tju.edu.cn

Abstract: With the ever-increasing volume of maritime traffic, the risks of ship navigation are becoming more significant, making the use of advanced multi-source perception strategies and AI technologies indispensable for obtaining information about ship navigation status. In this paper, first, the ship tracking system was optimized using the Bi-YOLO network based on the C2f_BiFormer module and the OC-SORT algorithms. Second, to extract the visual trajectory of the target ship without a reference object, an absolute position estimation method based on binocular stereo vision attitude information was proposed. Then, a perception data fusion framework based on ship spatio-temporal trajectory features (ST-TF) was proposed to match GPS-based ship information with corresponding visual target information. Finally, AR technology was integrated to fuse multi-source perceptual information into the real-world navigation view. Experimental results demonstrate that the proposed method achieves a mAP0.5:0.95 of 79.6% under challenging scenarios such as low resolution, noise interference, and low-light conditions. Moreover, in the presence of the nonlinear motion of the own ship, the average relative position error of target ship visual measurements is maintained below 8%, achieving accurate absolute position estimation without reference objects. Compared to existing navigation assistance, the AR-based navigation assistance system, which utilizes ship ST-TF-based perception data fusion mechanism, enhances ship traffic situational awareness and provides reliable decision-making support to further ensure the safety of ship navigation.

Keywords: data fusion; marine autonomous surface ships; ship traffic safety; ship recognition and tracking; machine vision; AR navigation assistance

check for updates

Academic Editor: Weicheng Cui

Received: 26 January 2025
Revised: 9 February 2025
Accepted: 12 February 2025
Published: 16 February 2025

Citation: Chen, S.; Gao, M.; Shi, P.; Zeng, X.; Zhang, A. Target Ship Recognition and Tracking with Data Fusion Based on Bi-YOLO and OC-SORT Algorithms for Enhancing Ship Navigation Assistance. *J. Mar. Sci. Eng.* **2025**, *13*, 366. https://doi.org/10.3390/jmse13020366

1. Introduction

With the rapid development of the shipping industry, the increasing number of ships and the complexity of maritime traffic situations have highlighted the issue of navigation safety. In 2021, the European Maritime Safety Agency published the "Annual Overview of Marine Casualties and Incidents [1]", which reported 15,481 maritime incidents between 2014 and 2020. The data indicates that approximately 53.5% of the investigated incidents were caused by improper human manipulation. In this context, Maritime Autonomous Surface Ships (MASSs) have emerged as the times require, utilizing advanced technologies such as the Internet of Things, big data, artificial intelligence, etc., to swiftly respond to maritime industry needs with little to no human intervention. The Maritime Safety Committee's (MSC) 100th session [2] formally defines MASSs from a legislative perspective as ships that can operate autonomously to varying degrees without human interaction, offering advantages such as low costs, low risks, and easy maintenance, thus garnering widespread attention from scholars around the world [3]. To promote the development of MASS technology, Rolls-Royce has proposed the vision of "smart ships", emphasizing the application of AR and data fusion technologies, which are believed to bring revolutionary progress to ship navigation systems. In line with this trend, AR technology has gradually become a key means to enhance the safety and operational efficiency of ships. Several companies and research institutions have made significant progress in integrating AR technology with the maritime sector. Specifically, the ClearCruise AR yacht navigation system, developed by the UK-based Raymarine company, captures the forward view in real time using high-definition cameras. It combines onboard sensor data and utilizes AR technology to overlay virtual information, providing decision support to the operator [4]. The HiNAS 2.0 navigation system, developed by the South Korean company Avikus, integrates AR technology. It has been successfully applied to the transoceanic navigation of the LNG ship "Prism Courage", significantly improving the safety and operational efficiency of autonomous ship navigation [5]. Although research on AR-assisted navigation systems in the maritime field is still in its early stages, its potential has been widely recognized. Researchers Laera et al. [6], in their study of 11 AR maritime solutions, pointed out that AR technology holds significant research value in maritime applications and is expected to further optimize ship navigation, situational awareness, and decision support functions, thereby having a profound impact on the safety and efficiency of the shipping industry.

Currently, mainstream ship navigation assistance systems mainly focus on visual scenarios. These systems fuse real-time perceived environmental information with heterogeneous data, such as the Automatic Identification System (AIS) [7,8], and present key navigation information in a visual format. However, these systems have shortcomings in terms of insufficient perception means, insufficient accuracy, real-time performance, single mutual communication information, and high demands on the operator's ability to interpret and understand the information, which make it difficult to meet the requirements of efficient navigation assistance in complex waterways. Therefore, many scholars have actively pursued strategies for integrating multiple perception methods, such as radar, sonar, Synthetic Aperture Radar (SAR), remote sensing, cameras, etc. [9–12], in order to comprehensively and accurately acquire information about the navigation environment. However, an effective and intelligent data fusion mechanism has not yet been proposed in the existing research, and at the same time, the method of expressing and understanding the sensing information is more demanding for the crew. Specifically, for the perception of ship target positions, existing research predominantly focuses on methods utilizing shore-based fixed-view cameras to project AIS position coordinates onto image coordinates for matching and fusion. However, this approach fails to address the issue of matching and fusion for ships with a deactivated AIS, as their specific position coordinates cannot be obtained [13]. Furthermore, current multi-source perception data fusion methods primarily concentrate on the relationship between target image positions and AIS positions at a single moment [14], neglecting the long-term trajectory characteristics of ships. Consequently, in busy waters with high ship density or when AIS signals are subject to interference, ensuring high-quality association and fusion of perception data becomes challenging. Additionally, the expression and interpretation methods of perceived information in existing navigation assistance systems impose high cognitive demands on crew members [15], potentially leading to decision-making errors. To address these challenges, in this study, a perception data fusion framework based on ship spatio-temporal trajectory features (ST-TF) was proposed. This framework was designed for application in ship AR navigation assistance systems, offering a more reliable and efficient solution for navigation and surveillance in complex waterways.

The main contributions of this paper are summarized as follows:

- Fusion Location Ship (FLShip) dataset: The FLShip dataset has been constructed based on data collected from six self-developed MASSs, which includes stereo and monocular image sequences, as well as high-frequency attitude and GPS data from the MASSs. A variety of navigational scenarios and environmental conditions are encompassed by the dataset, providing essential support for ship target detection, multi-source information fusion, and autonomous navigation research.
- Optimization of ship tracking system: To improve the detection performance of small targets while reducing the model's parameter count, a Bi-YOLO network based on BiFomer has been proposed, achieving more efficient and accurate ship identification. In response to complex navigational environments, the more powerful OC-SORT algorithm has been incorporated to ensure stable tracking of targets. The superiority of this method has been validated through comparative experiments with several advanced detection and tracking algorithms.
- Ship target absolute position estimation: In order to realize the position estimation of the target ship in the state of no reference, a visual localization model based on binocular imaging was constructed. By fusing the attitude information from a binocular camera, absolute position estimation of maritime targets at the moving end of a ship is successfully realized.
- Perception data fusion based on ship ST-TF: To effectively integrate information from multiple perception sources, a novel perception data fusion method based on ship ST-TF was proposed. This approach employs a dynamic time warping (DTW) algorithm to align multi-source perception data, achieving spatio-temporal synchronization between different sensory inputs, while aiming to ensure the accuracy and reliability of the fused perception information.

The rest of the paper is organized as follows. Section 2 begins with a review of state-of-the-art research on ship perception methods, visual target detection and tracking, and multi-sensor data fusion. In Section 3, the proposed perception data fusion framework based on ship ST-TF is described in detail. Then, in Section 4, field experiments are conducted to evaluate and validate the system performance to demonstrate the effectiveness of our framework. Finally, Section 5 summarizes the main contributions of this paper and discusses future perspectives. The organization of the sections is shown in Figure 1.

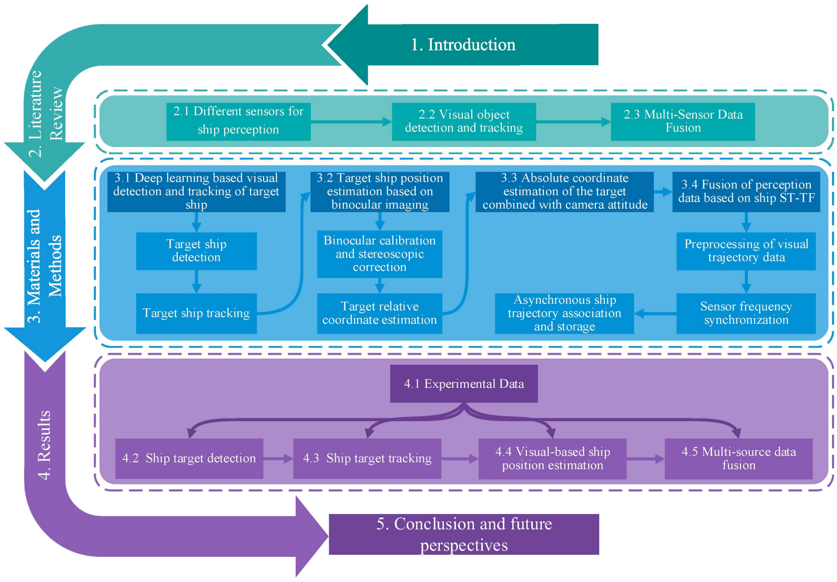

1. Introduction
2. Literature Review
2.1 Different sensors for ship perception
2.2 Visual object detection and tracking
2.3 Multi-Sensor Data Fusion
3. Materials and Methods
3.1 Deep learning based visual detection and tracking of target ship
Target ship detection
Target ship tracking
3.2 Target ship position estimation based on binocular imaging
Binocular calibration and stereoscopic correction
Target relative coordinate estimation
3.3 Absolute coordinate estimation of the target combined with camera attitude
3.4 Fusion of perception data based on ship ST-TF
Preprocessing of visual trajectory data
Sensor frequency synchronization
Asynchronous ship trajectory association and storage
4. Results
4.1 Experimental Data
4.2 Ship target detection
4.3 Ship target tracking
4.4 Visual-based ship position estimation
4.5 Multi-source data fusion
5. Conclusion and future perspectives

Figure 1. Organization diagram of the sections of this paper.

2. Literature Review

This section focuses on recent research related to our work, namely ship perception methods, visual target detection and tracking, and multi-sensor data fusion.

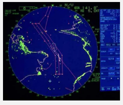
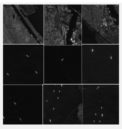
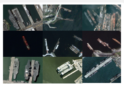
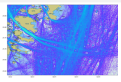
2.1. Different Sensors for Ship Perception

Currently, ship perception methods are mainly divided into two categories: active perception and passive perception. Active perception methods include mariner radar, sonar, SAR, etc., while passive perception methods include optical remote sensing, cameras, AISs, etc.

Various maritime perception methods have their own advantages and limitations. Mariner radar is virtually unaffected by visibility, lighting, and noise but it is expensive, has blind spots, and may experience reduced detection accuracy in the presence of sea clutter. In contrast, sonar performs well in detecting, positioning, and tracking underwater targets [16]. However, sonar images often suffer from poor quality, low contrast, and low resolution, requiring additional image enhancement processing for target identification and localization. Simultaneously, SAR systems have unique advantages in ship detection [17]. However, when dealing with both stationary and moving targets, SAR systems may exhibit image distortion and blurring [18], which makes the extraction of target behavior information more complicated. Optical remote sensing technology has garnered significant attention in recent years, especially in target identification and situational awareness [19]. However, remote sensing images are influenced by factors such as noise, seasonal climate, and light intensity, making them unable to work reliably in all weather conditions. In comparison to other sensors, cameras, as passive sensors similar to the human eye, operate by capturing information about the color, contours, and textures of objects. Cameras are widely used in areas such as driverless and target recognition [20]. Additionally, cameras can be used to measure the distance to the target and determine the direction and position of the target. Due to their advantages of high-precision measurement and not revealing the position of the detection system, cameras have demonstrated high application value in many fields [21]. AIS data contains a wealth of information to ensure the safety of maritime navigation. It is widely used in ship behavior analysis, route planning, and collision avoidance decision-making research [22,23].

Table 1 presents the characteristics of the primary ship perception methods. From this, we can conclude that relying solely on a single sensor has significant limitations. It fails to provide sufficient accuracy and completeness in capturing ship navigation dynamic data. In contrast, a multi-source data fusion framework leverages the strengths of multiple sensors, enabling a more comprehensive and precise perception of the navigation environment. This approach is particularly well suited for diverse and complex application scenarios.

Table 1. Comparison of different sensor perception methods.

Perception Methods			Target Recognitior Rate	Spatial Resolution	Perception Range	Robustness	Real-Time Performance	Data Fusion Compatibility	Power Consumption	Cost
ACTIVE	Radar		High	Medium	Large	High	High	Medium	Medium	High
	Sonar		Medium	Medium	Medium	Medium	Medium	Medium	Medium	Low
	SAR		Medium	Low	Medium	High	High	High	Low	Medium
PASSIVE	Remote Sensing		Medium	Medium	Medium	Medium	High	High	Low	Medium
	Camera		High	High	Small	Low	Low	Low	High	High
	AIS		High	Low	Large	High	Low	Low	Medium	High

2.2. Visual Object Detection and Tracking

With the continuous advancement of computer vision technology, visual detection and tracking methods based on deep learning have been widely applied in the field of ship detection and tracking, owing to their outstanding performance. The use of neural networks as the basis of algorithms has become a mainstream trend for classification and detection tasks, and such algorithms are divided into two main routes. The first comprises two-stage algorithms based on region recommendation, including R-CNN [24], Fast R-CNN [25], and Faster R-CNN [26]. This route essentially inherits the idea of traditional detection algorithms, where regions that may contain targets are first screened, followed by feature extraction and classification. The second route comprises end-to-end one-stage algorithms, such as the YOLO series [27–30] and SSD [31]. These algorithms directly transform the bounding box location problem into a regression problem, thus reducing computational complexity and enabling the network to detect targets with a lightweight architecture. Specifically, D. Zhang et al. used the Mask R-CNN algorithm and Faster R-CNN algorithm, respectively, to construct ship target feature extraction and a recognition model, and compared the accuracy and performance of these two algorithms in target detection and classification by classifying warships and civilian ships [32]. Aiming at the problem of traditional shipboard radar having difficulty performing the task of detecting near-small ships, Chen, Z et al. proposed a small-ship detection method based on an improved GAN and CNN, and the experimental results show that their method significantly improved the detection effect of small ships [33]. To address the challenges of missed detection and misidentification in ship target detection within SAR images under complex scenes, Chen, Z. et al. proposed a CSD-YOLO model based on YOLOv7. This model was validated on the SSDD and HRSID datasets, demonstrating significant improvements over Faster R-CNN, FCOS, YOLOv3, YOLOv5s, and YOLOv7 algorithms, with marked enhancements in multi-scale ship target feature extraction [34]. To address the challenges of low-visibility weather scenarios, Liu et al. developed an enhanced Convolutional Neural Network (CNN) to improve ship detection under different weather conditions, and experimental results under different monitoring conditions showed that their method outperforms SSD, Faster R-CNN, YOLOv2, and YOLOv3 in terms of detection accuracy [35].

In recent years, scholars have explored the application of visual detection and tracking technology for information mining in maritime videos. Zheng et al. applied binocular stereo vision technology to the recognition and estimation of the depth of inland waterway ships and used the MobileNetV1 network as the feature extraction module of the YOLOv4 model for ship detection, greatly reducing the amount of computation while ensuring accuracy. In addition, a sub-pixel-level feature point detection and matching algorithm was used for depth estimation, which enriched the ship's perception of the navigation environment [36]. Zhou et al. proposed a deep learning-based framework for extracting ship speeds in haze environments, used the YOLOv5 detector and Deep SORT tracker to detect and track the ship, and estimated the ship speed by calculating the mapping relationship between image space and real space, which provided important value for guaranteeing the safety of ship navigation [37].

2.3. Multi-Sensor Data Fusion

To obtain more accurate and consistent ship motion perception data, multi-source data fusion methods have become a focus of attention in the academic community [38,39]. Maritime radars, AISs, and cameras are the most widely used sensors in maritime transportation systems, each possessing distinct advantages and limitations. Therefore, effectively integrating data from these sensors to enhance ship detection and tracking accuracy has become a key research focus.

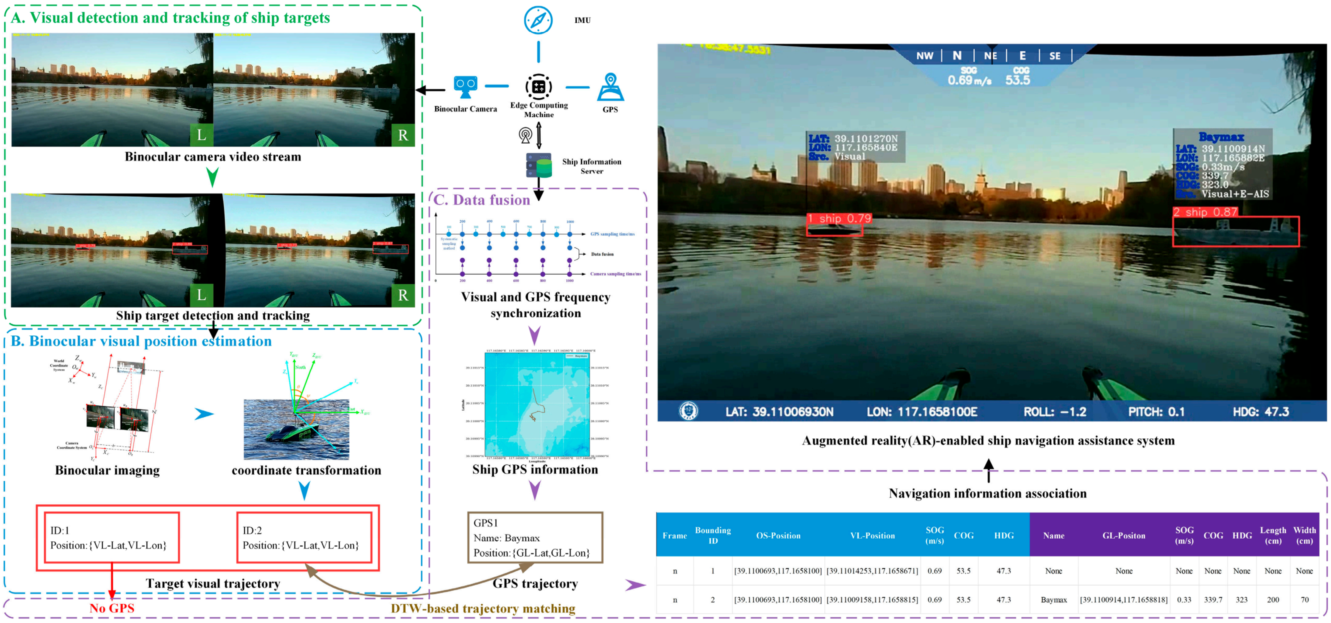
In previous studies, scholars have made a lot of effort in the fusion of maritime radars, AISs, and video surveillance cameras. For example, to improve the accuracy of AIS visual target association, Ding et al. proposed a three-step calibration method that projects image coordinates onto geographic coordinates [14]. By employing the Hungarian algorithm, they identified the optimal matches, effectively integrating AIS data with ship trajectories. Gülsoylu et al. introduced an image–AIS data fusion approach for fixed and periodically panning cameras [40]. This method utilizes the YOLOv5 model for ship detection, applies homography transformation for coordinate conversion, and assigns AIS messages to the nearest bounding box in the image using k-dimensional tree-based nearest neighbor search. The accuracy of the fusion framework was further validated. To improve ship tracking efficiency, Chen et al. integrated the AIS with maritime video surveillance. They adjusted the camera's attitude and focus based on AIS-reported positions to achieve stable tracking of specific ships [41]. Although this method performs well in single-target scenarios, it faces significant limitations in multi-ship scenarios and cannot achieve efficient multi-target tracking. Additionally, maritime radar images often contain false echoes, and AIS data transmission may experience loss, leading to uncertainties in estimating ship speed and heading. To address this issue, Xu et al. proposed a method that integrates radar sequence images with AIS data for estimating ship speed and heading [42]. This method was applied in the Yangtze River Basin, effectively reducing the uncertainty in ship speed and heading estimation. To improve waterway traffic monitoring in inland waterways, Lu et al. proposed a framework for ship identification through the fusion of visual detection and AIS data [43]. This method estimates the distance from the camera and the azimuth relative to the camera of the identified ship target based on the bounding box size in the shore-based monocular camera image and the ship length in the AIS data. However, this approach is only applicable for the distance estimation of ships with a side-facing orientation to the camera. Moreover, due to the effects of marine background, occlusion, and weather conditions, the bounding box size often varies, further impacting the accuracy of the estimation. For the fusion of ship trajectories, Qu et al.

proposed a vision-based framework for fusing ship and AIS trajectories [13] that uses the YOLOX detector and optimized Deep SORT tracker to extract ship image trajectories, preprocesses AIS data, and projects AIS coordinates to the image coordinate system. The Hungarian algorithm then associates AIS and visual data for intelligent maritime traffic monitoring in inland waterways. However, this method does not fully consider the limited computing resources of onboard devices, and the use of original YOLOX may result in slower inference speed. Zhang et al. proposed a method for matching AIS and video data by combining region division and an improved Kalman filter [44]. The method utilizes an enhanced LSTM network for accurate trajectory prediction of AIS data and applies a single-stage object detection model to analyze video data collected from buoys. This approach enables fast and efficient data fusion, significantly improving the accuracy and flexibility of channel detection.

The aforementioned studies indicate that, despite significant research focused on fusing multiple data sources for ship target perception, two unresolved issues remain. First, existing studies often fail to fully consider the limitations of onboard computational resources, leading to bottlenecks in inference speed, which affect system real-time performance and efficiency. Second, most research focuses on fusion perception methods for shore-based or fixed-angle cameras, while less attention is given to the challenges posed by onboard cameras in real-world navigation environments, such as the effects of fog interference and AIS data instability. Therefore, from the perspective of practical ship applications, optimizing the inference speed of perception models and the data fusion mechanism to improve the accuracy and usability of multi-source perception methods is an urgent problem that needs to be addressed.

3. Materials and Methods

This section provides a detailed description of the perceptual data fusion framework based on ship ST-TF, as shown in Figure 2. The framework includes ship detection and tracking, ship trajectory extraction based on binocular vision, and multi-source spatio-temporal data fusion. For the video data captured by the shipboard camera, the Bi-YOLO network is first used for ship target detection, and to address the ship occlusion problem in complex navigation environments, the more powerful OC-SORT algorithm is introduced to achieve stable target tracking. In the visual trajectory extraction stage, a visual localization model is constructed based on the principle of binocular imaging. This model is then integrated with the attitude information of the camera, enabling the shipborne terminal to perform absolute position estimation of ship targets. Subsequently, in the multi-source spatio-temporal data fusion, the visual trajectory is pre-processed to suppress the measurement noise, and then the DTW algorithm is used to match the visual trajectory with the GPS trajectory and align the spatio-temporal data points. After the trajectory matching, the data storage area is redesigned to jointly store the navigation information of the host ship and the target ship, thus improving the efficiency of the trajectory association. Finally, based on the results of multi-source perception data fusion, a ship AR navigation assistance system is developed, which improves the overall navigation experience of ship driving and enhances safety in the maritime field.



Frame	Bounding ID	OS-Position	VL-Position	SOG (m/s)	COG	HDG	Name	GL-Positon	SOG (m/s)	COG	HDG	Length (cm)	Width (cm)
n	1	[39.1100693,117.1658100]	[39.11014253,117.1658671]	0.69	53.5	47.3	None	None	None	None	None	None	None
n	2	[39.1100693,117.1658100]	[39.11009158,117.1658815]	0.69	53.5	47.3	Baymax	[39.1100914,117.1658818]	0.33	339.7	323	200	70

Figure 2. A perception data fusion framework based on ship ST-TF for ship AR navigation assistance.

3.1. Deep Learning-Based Visual Detection and Tracking of Target Ship

3.1.1. Target Ship Detection

YOLO (You Only Look Once) is a single-stage deep learning object detection algorithm. It has revolutionized real-time object detection by combining region proposals and classification into a unified neural network, significantly reducing computation time. Its unique design has made it highly popular in the deep learning community, offering superior accuracy and faster detection speeds compared to other object detection algorithms. The strength of YOLO is that it is not only limited to the original task, that is, the detection of 80 categories defined by the MS-COCO dataset, but also widely used in various detection fields, becoming the backbone model or benchmark for many tasks. YOLO11, the latest iteration of this series [45], represents the state-of-the-art in object detection technology, providing a solid and cutting-edge technical foundation for our research. In this study, the accuracy of ship target identification plays a crucial role in the final data fusion results. However, limited computational resources on the shipborne edge devices constrain this process. Furthermore, existing ship target detection models primarily focus on medium-to large-sized targets, and their performance for small targets is suboptimal. This makes them unsuitable for scenarios in open waters or nighttime navigation, where small, inconspicuous targets may be present. To enhance the detection of small targets while reducing the model's parameter count, a Bi-YOLO network has been proposed. This network incorporates the C2f_BiFormer module, which combines convolutional operations with the BiFormer block [46] into the backbone layer of YOLOv11, achieving more efficient and accurate ship identification. The structure of the Bi-YOLO network is shown in Figure 3.

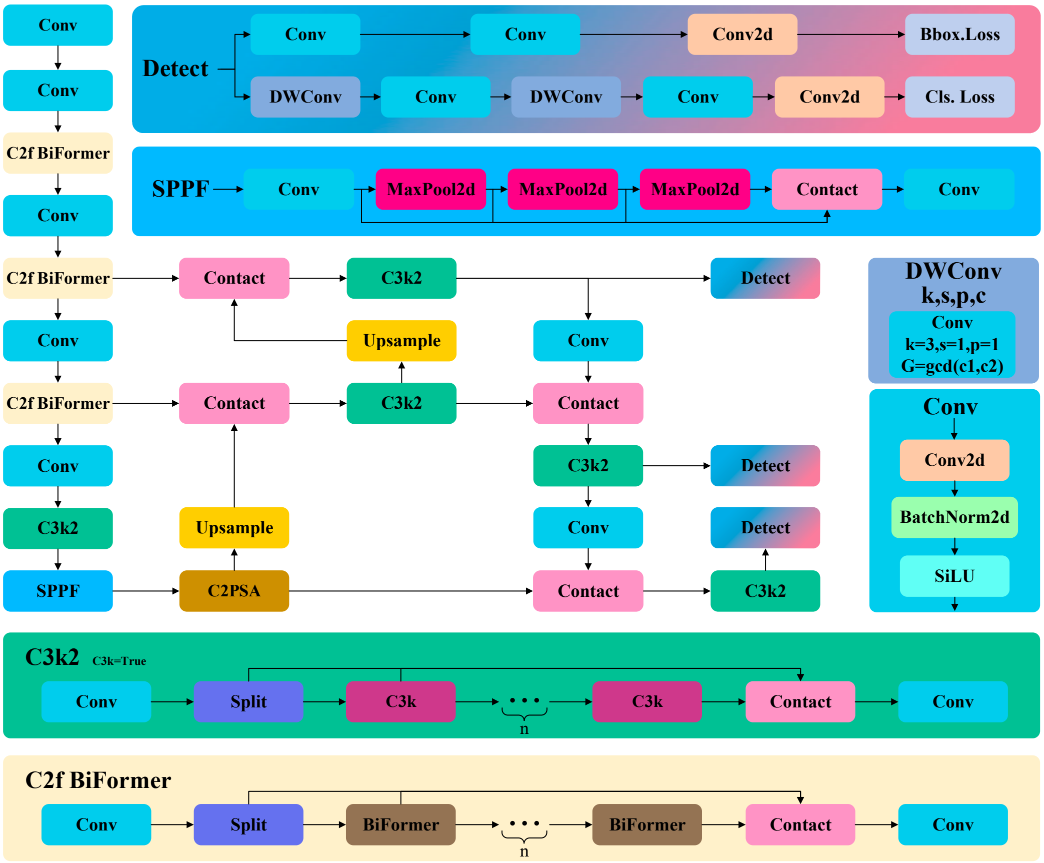


Figure 3. The structure of the Bi-YOLO network.

In YOLO11 network, the backbone utilizes Conv and C3k2 modules for downsampling, this can lead to the loss of contextual features to some extent, thereby reducing the resolution of the feature maps and losing important contextual information. The introduction of the C2f_BiFormer module helps mitigate the loss of small target information during downsampling, enabling the model to focus more on task-relevant information when processing inputs. This module employs a dual-level routing attention strategy to enhance feature representation through two key mechanisms: routing between regions via a directed graph, which helps the model focus on specific areas of the image; and token-to-token attention, which further refines the model's attention to these regions.

BiFormer block: The overall architecture of the BiFormer block is shown in Figure 4a. The BiFormer block is a transformer-based vision network architecture composed of several submodules, including deep convolution, layer normalization, Bi-Level Routing Attention (BRA), and MLP. The plus sign indicates the merging of two feature vectors. The submodules are executed sequentially to optimize the input features within each BiFormer block. The process begins with a 3 × 3 depthwise convolution, which slightly encodes positional information. Then, the BRA module is combined with a two-layer MLP with an expansion rate of e, further optimizing the features by capturing and modeling relationships across positions.

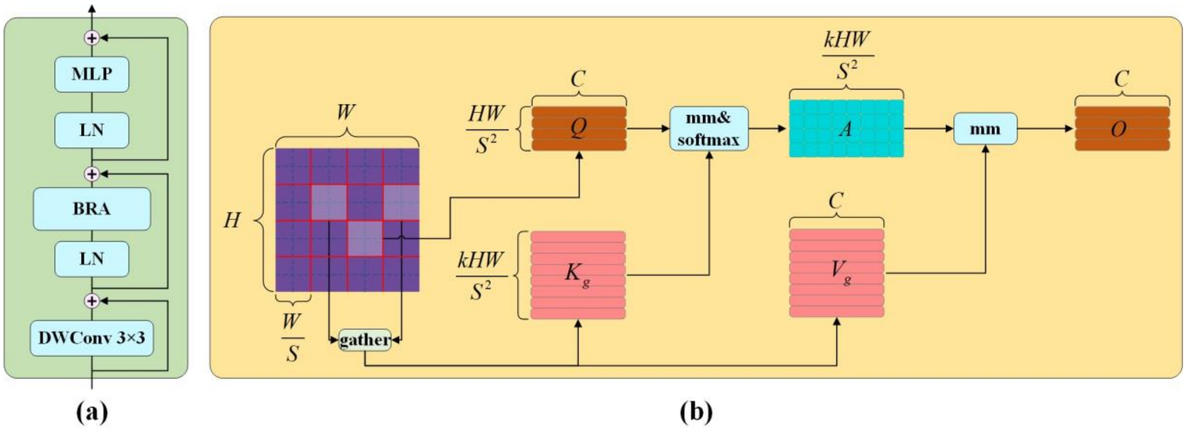


Figure 4. (**a**) Details of a BiFormer block; (**b**) Structure of the BRA.

BRA module: The operating principle of the BRA module is shown in Figure 4b. BRA is an advanced attention mechanism that employs a bi-level routing strategy to achieve dynamic, query-aware sparsity. The main steps of BRA include the following:

a. Region partition and input projection: Given a 2D input feature map $X \in \mathbb{R}^{H \times W \times C}$, where H is the height, W is the width, and C is the number of channels. We divide it into S×S nonoverlapping regions, such that each region contains $\frac{HW}{S^2}$ feature vectors. This is achieved by reshaping X into $X_r \in \mathbb{R}^{S^2 \times \frac{HW}{S^2} \times C}$. We then derive the query Q, key K, and value tensor $V \in \mathbb{R}^{S^2 \times \frac{HW}{S^2} \times C}$, using linear projections.

$$X_r \in reshape\left(X, S^2, \frac{HW}{S^2}, C\right) \tag{1}$$

$$Q = X_r W_q, K = X_r W_k, V = X_r W_v \tag{2}$$

where W_q, W_k, and W_v represent the learnable weight matrices.

b. Region-to-region routing with a directed graph: We then find the attending relationship by constructing a directed graph. Specifically, we first compute region-level queries and keys by applying an average to Q and K within each region, resulting in Q_r and $K_r \in \mathbb{R}^{S^2 \times C}$. Subsequently, we derive the adjacency matrix $A_r \in \mathbb{R}^{S^2 \times S^2}$ for the region-to-region affinity map by performing matrix multiplication between Q_r and the transposed K_r:

$$Q_r = mean(Q, 1), K_r = mean(K, 1) \tag{3}$$

$$A_r = Q_r K_r^T \tag{4}$$

The values in the adjacent matrix, A_r, quantify the semantic relationships between pairs of regions. The core step that we perform next is to prune the affinity graph by keeping only top-k connections for each region. To achieve this, we construct a routing index matrix, $I_r \in \mathbb{N}^{S^2 \times k}$, using a row-wise top-$k$ selection operation:

$$I_r = topIndex(A_r) \tag{5}$$

c. Token-to-token attention: Finally, token-to-token attention is performed within the specified regions. Each query token selectively attends to the query and context pairs corresponding to the defined routing areas. The attention function is then applied to these refined feature pairs to generate the final output. This process is expressed as follows:

$$K_g = \text{gather}(K, I_r), V_g = \text{gather}(V, I_r) \tag{6}$$

$$O = Atttention(Q, K_g, V_g) + LCE(V) \tag{7}$$

In this, K_g, V_g are the gathered key and value tensor, *Attention*$(\cdot)$ represents the attention function, and $LCE(\cdot)$ denotes the term that enhances local context, parameterized through deep convolution. The BiFormer block significantly reduces computational complexity by adopting a BRA strategy. Under a proper region partition factor S, its complexity is $O(S^2 \cdot k \cdot d)$. In contrast, the traditional global attention mechanism has a complexity of $O((HW)^2)$, while the axial attention mechanism has a complexity of $O((HW)^{\frac{3}{2}})$. This sparse computation strategy not only improves computational efficiency but also effectively reduces interference from irrelevant regions, thereby enhancing the robustness and accuracy of the model in handling small target detection tasks on the water surface.

3.1.2. Target Ship Tracking

Traditional motion-based tracking algorithms often use Kalman filters to predict the motion of the target. However, when the target exhibits nonlinear motion or encounters occlusion, long-term linear estimates may become highly inaccurate, leading to tracking failures. To address this problem, an Observation-Centric SORT (OC-SORT) [47] is used as the main tracking algorithm in this paper. Ship targets are tracked in real-time independently within the left and right images.

OC-SORT introduces three modules, namely Observation-Centric Online Smoothing (OOS), Observation-Centric Momentum (OCM), and Observation-Centric Recovery (OCR) to cope with the errors caused by target occlusion and nonlinear motion. Among them, Observation-Centric Re-Update (ORU) is proposed to reduce the cumulative error when an object is untracked and re-associated after a certain period of time, e.g., when the last observation before untracking is denoted as z_{t_1}, the observation that triggered the re-association is denoted as z_{t_2}, and the virtual trajectory $T_{virtual}$ is denoted as

$$\hat{z} = T_{virtual}(z_{t_1}, z_{t_2}, t), t_1 < t < t_2 \tag{8}$$

When the lost trajectory is re-associated along the virtual trajectory, OOS alternates between the prediction and update stages, backchecking the parameters of the Kalman filter back and forth, to prevent the accumulation of errors due to occlusion or missed detections. When constructing the association cost matrix, most multi-object tracking methods typically rely on two key points: motion features and appearance features. However, OC-SORT introduces OCM into the association cost during the tracking process, by incorporating the directional consistency of the trajectories into the association cost matrix. The association cost matrix can be represented as follows:

$$C(\hat{X}, \mathrm{Z}) = C_{\mathrm{IoU}}(\hat{X}, Z) + \lambda C_v(\mathcal{Z}, Z) \tag{9}$$

where $(\hat{X}, \mathrm{Z})$ is the set of estimated states and observations for the tracked object, λ is the weight factor, $\mathcal{Z}$ includes the observed trajectories of all existing tracks, C_{IoU} calculates the Intersection over Union (IOU), and C_v calculates the consistency of the orientation of the trajectory and the orientation of the historical and new observations of the trajectory.

Finally, after completing the association stage based on the predictions from the Kalman filter, OC-SORT applies OCR techniques to attempt to correlate the observations of mismatched trajectories with mismatched observations. This strategy helps reduce the generation of new trajectories, thereby enhancing the overall accuracy of multi-object tracking.

3.2. Target Ship Position Estimation Based on Binocular Imaging

3.2.1. Binocular Calibration and Stereoscopic Correction

Camera calibration is a prerequisite for stereo vision, where the imaging model typically assumes that the camera has no distortion under ideal conditions and describes the imaging process using a linear model. However, in practical applications, lens errors can cause distortion, affecting the imaging position of the images. Perspective, radial, and tangential distortion are the main types of distortion, with radial and tangential distortion having a significant impact on imaging. It has been shown that camera calibration [48] is an effective way to solve the camera lens distortion problem. Through the calibration process, we can obtain the parameters of the camera such as the distortion coefficients, rotation matrix, and translation vectors, which can be used as important input data for the next stereo correction. As the camera will be affected by the optical lens and its own process, the imaging process is prone to distortion, resulting in the left and right images not achieving coplanar and line alignment and requiring stereoscopic correction through left and right camera image plane reprojection, so that it is coplanar and aligned, so as to improve the accuracy of stereoscopic vision. Figure 5 shows a comparison of the effects before and after calibration and stereo rectification of our onboard stereo camera.

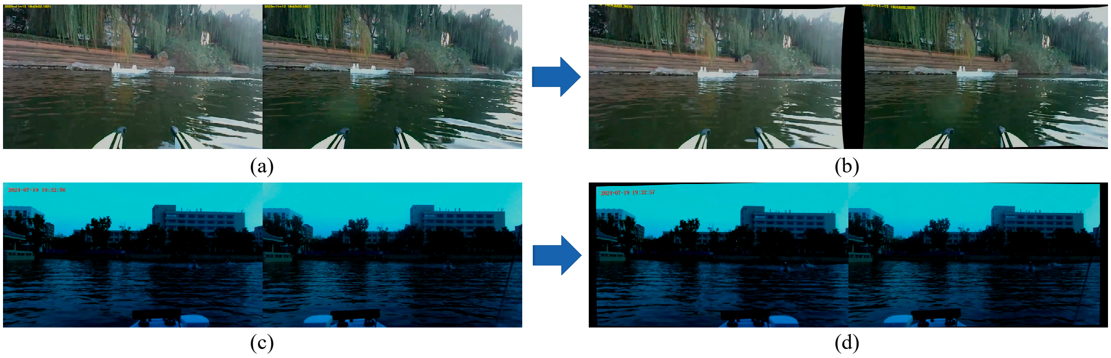


Figure 5. (**a**) to (**b**) illustrate the Driving-Leaves binocular camera before and after calibration and stereo rectification, and (**c**) to (**d**) illustrate the Baymax binocular camera before and after calibration and stereo rectification.

3.2.2. Target Relative Coordinate Estimation Based on Binocular Imaging

Under ideal conditions, the camera imaging model is assumed to be a central projection linear model. Based on the pinhole imaging model, visual target relative positioning can be achieved. In this process, we need to consider the world coordinate system $O_W - X_w Y_w Z_w$ (which represents the reference coordinate system for the camera and object position), the camera coordinate system $O_C - X_C Y_C Z_C$ (with the camera optical center as the origin and the optical axis direction as the Z-axis), the image coordinate system $o_i - x_i y_i$ (corresponding to the camera coordinate system with the origin at the intersection of the image plane and the optical axis of the camera), and the pixel coordinate system $o_P - u_i v_i$ (a two-dimensional plane with the origin at the top-left corner of the first pixel). These four coordinate systems are interrelated, forming the transformation process from the world coordinate system to the pixel coordinate system, providing the foundation for achieving visual target localization. The relationships among the four coordinate systems and the binocular imaging process are illustrated in Figure 6.

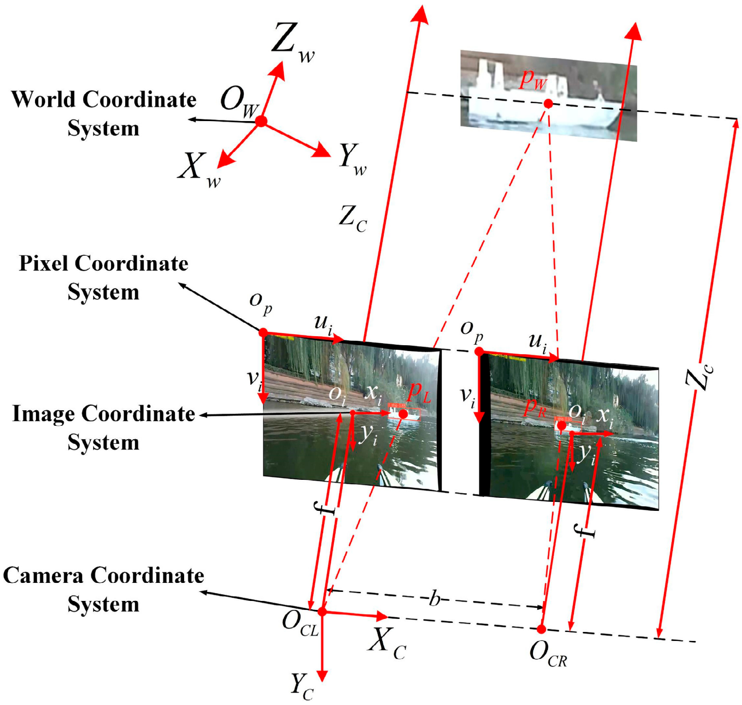


Figure 6. Conceptual diagram of the binocular imaging process.

According to the pinhole imaging principle, the transformation relationship between the pixel coordinate system and the world coordinate system is as follows:

$$Z_c \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} R & T \end{bmatrix} \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix} \tag{10}$$

where $\begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}$ is the intrinsic parameter matrix of the camera, which is obtained from the camera calibration above. R and T are the rotation matrix and translation vector, respectively.

Based on the binocular imaging principle (Figure 6) and the similarity of triangle proportions, we can derive the following:

$$Z_c = \frac{fb}{X_L - X_R} = \frac{fb}{(u_L - u_R)d_x} = \frac{bf_x}{u_L - u_R} \tag{11}$$

Here, X_L and X_R represent the physical dimensions corresponding to the target's u_L and u_R in the pixel coordinate system. d_x denotes the physical size of each pixel on the horizontal axis. b is the baseline of the binocular camera system.

Assume that the origin of the world coordinate system is the optical center of the left-eye camera. To transform from the camera coordinate system to the world coordinate system, a counterclockwise rotation around the X_c axis of the camera coordinate system by an angle α = 90° (viewed from the positive direction of the X_c axis towards the negative direction, rotating counterclockwise) is required. The transformation of the coordinate system can be expressed as follows:

$$Z_c \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X_w \\ -Z_w \\ Y_w \end{bmatrix} \tag{12}$$

(X_w, Y_w, Z_w) are the relative coordinates of the target ship relative to the world coordinate system with the optical center of the left-eye camera as the origin.

3.3. Absolute Coordinate Estimation of the Target Combined with Camera Attitude

The following study will provide a detailed description of the coordinate transformation process based on camera attitude. Through a multi-step transformation, the relative coordinates obtained from stereo vision (i.e., coordinates in the carrier coordinate system) will be converted to determine the precise position of the target on Earth. First, the relative coordinates of the target ship are converted to the east—north—up (ENU) coordinate by using the camera attitude information to accurately extract the target's position information relative to the carrier. Subsequently, through further coordinate mapping, it is converted to the Earth-centered, Earth-fixed (ECEF) coordinate, and finally, the conversion to the World Geodetic System 1984 (WGS-84) coordinate system is realized, i.e., the latitude and longitude coordinates of the target are obtained, and each coordinate system is shown in Figure 7. This conversion process can extract the accurate latitude and longitude coordinates of the target on the Earth's surface, which is the basis for realizing the global positioning of the target location. This part will introduce this conversion process in detail.

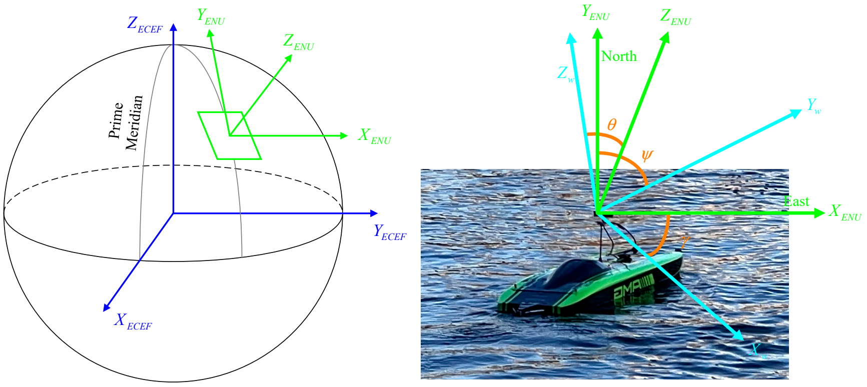


Figure 7. Illustration of coordinate system conversion.

Specifically, the transformation from the carrier coordinate system (i.e., the world coordinate system assumed in the previous text) to the ENU coordinate system is accomplished through the camera's attitude angles. The ENU coordinate system is a local coordinate system centered at the position of the camera, with the X_{ENU}-axis pointing east, the Y_{ENU}-axis pointing north, and the Z_{ENU}-axis pointing up. When there is an attitude angle of the shipboard camera, it is necessary to rotate the point $P_w(X_w, Y_w, Z_w)$ under the carrier coordinate system around the Y_w axis by the roll angle γ, around the X_w axis by the pitch angle θ, and around the Z_w axis by the yaw angle $-\psi$, in order to obtain the point $P_{ENU}(X_{ENU}, Y_{ENU}, Z_{ENU})$ under the ENU coordinate system. The transformation from the carrier coordinate system to the ENU coordinate system can be expressed as follows:

$$R_{CN} = \begin{bmatrix} \cos\psi\cos\gamma + \sin\psi\sin\theta\sin\gamma & \sin\psi\cos\theta & \cos\psi\sin\gamma - \sin\psi\sin\theta\cos\gamma \\ -\sin\psi\cos\gamma + \cos\psi\sin\theta\sin\gamma & \cos\psi\cos\theta & -\sin\psi\sin\gamma - \cos\psi\sin\theta\cos\gamma \\ -\cos\theta\sin\gamma & \sin\theta & \cos\theta\cos\gamma \end{bmatrix} \tag{13}$$

$$\begin{bmatrix} X_{ENU} \\ Y_{ENU} \\ Z_{ENU} \end{bmatrix} = R_{CN} \begin{bmatrix} X_w \\ Y_w \\ Z_w \end{bmatrix} \tag{14}$$

Here, R_{CN} represents the rotation matrix for transforming from the carrier coordinate system to the ENU coordinate system.

The ECEF coordinate system has its origin at the center of the Earth, represented by mutually orthogonal X, Y, and Z axes. The transformation relationship between point $P_{ENU}(X_{ENU}, Y_{ENU}, Z_{ENU})$ in the ENU coordinate system and point $P_{ECEF}(X_{ECEF}, Y_{ECEF}, Z_{ECEF})$ in the ECEF coordinate system involves the real-time latitude (lat) and longitude coordinates (lon) of the ship itself, as well as the coordinates of the ship's position in the ECEF coordinate system denoted as $(L_{ECEF}, L_{ECEF}, L_{ECEF})$. The transformation relation is as follows:

$$\begin{bmatrix} X_{ECEF} \\ Y_{ECEF} \\ Z_{ECEF} \end{bmatrix} = \begin{bmatrix} -\sin(lon) & -\cos(lon)\sin(lat) & \cos(lon)\cos(lat) \\ \cos(lon) & -\sin(lon)\sin(lat) & \sin(lon)\cos(lat) \\ 0 & \cos(lat) & \sin(lat) \end{bmatrix} \begin{bmatrix} X_{ENU} \\ Y_{ENU} \\ Z_{ENU} \end{bmatrix} + \begin{bmatrix} L_{ECEF} \\ L_{ECEF} \\ L_{ECEF} \end{bmatrix} \tag{15}$$

Finally, the WGS-84 coordinate system is used to express the position of a point on Earth in terms of geodetic latitude B, geodetic longitude L, and geodetic height H. The conversion process between point $P_{ECEF}(X_{ECEF}, Y_{ECEF}, Z_{ECEF})$ in the ECEF coordinate system and point $P_G(B, L, H)$ in the WGS-84 coordinate system can be obtained through iterative calculations. The conversion relationship here is the following:

$$\begin{cases} L = \arctan\frac{Y_{ECEF}}{X_{ECEF}} \\ B = \arctan\frac{Z+Ne^2\sin B}{\sqrt{X_{ECEF}^2+Y_{ECEF}^2}} \\ H = \frac{Z_{ECEF}}{\sin B} - N(1-e^2) \end{cases} \tag{16}$$

In the equation, N represents the radius of curvature in prime vertical, e is the first eccentricity of the reference ellipsoid, and B is the iterative quantity. Initially, the geodetic latitude B is computed based on the initial latitude B_1. Subsequently, B is updated iteratively until the convergence condition is met, resulting in the desired latitude and longitude coordinates. These steps will provide us with accurate positional information for the target on the Earth.

$$B_1 = \arctan\frac{Z}{\sqrt{X^2+Y^2}} \tag{17}$$

3.4. Fusion of Perception Data Based on Ship ST-TF

3.4.1. Preprocessing of Visual Trajectory Data

Due to factors such as ship navigation, camera shake, errors in bounding box size during manual annotation, and interference from the horizontal plane background, there may be small positional discrepancies between the detected bounding box size and the actual outline of the target ship. This can lead to incorrect estimation of the image coordinates for tracking the target ship, which subsequently affects the accuracy of visual localization estimates. To address these challenges effectively, we employ Kalman filtering to preprocess the position coordinates of the visual estimates, ensuring a smoother and more reliable trajectory. Kalman filtering is able to effectively suppress noise in measurements and processes, and better adapt to changes in the actual system, thereby improving the accuracy and stability of position estimation.

3.4.2. Visual Localization Frequency Synchronized with Target Ship GPS Frequency

When performing visual perceptual localization and GPS information fusion, it is critical to ensure that camera and GPS outputs are acquired at the same time to avoid significant errors in the information fusion results. Typically, the time for each sensor is independent, and the sampling frequencies vary. In this experiment, the target ship's GPS operates at frequencies of 5 Hz and 10 Hz, while the camera's operating frequency is set to 5 fps, meaning that it captures 5 frames per second with a 200 ms interval between each frame, which is the same as the GPS working frequency of the host ship. Here, we use the lower frequency camera sensor as a reference and apply systematic sampling to the 10 Hz GPS data to ensure synchronization between the world coordinates calculated by the camera and the GPS coordinates. This frequency synchronization strategy helps effectively integrate information from different sensors, enhancing the accuracy of visual perception localization and GPS information fusion. The frequency synchronization process is shown in Figure 8.

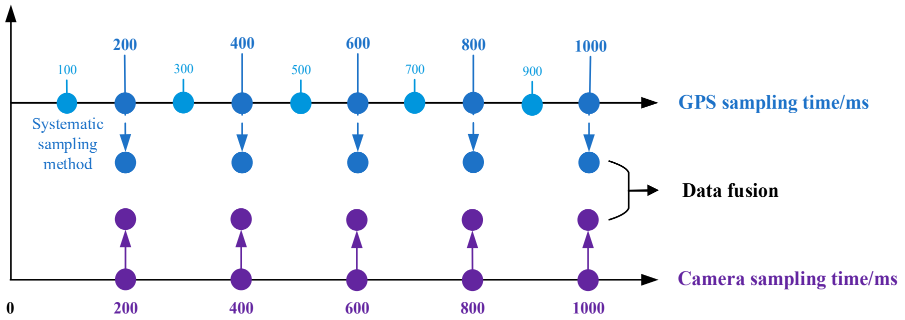


Figure 8. Synchronization process of different sensor frequencies.

3.4.3. Asynchronous Nonlinear Ship Trajectory Association and Storage Based on the DTW Algorithm

The core of data fusion is the matching association between visual and GPS trajectories. This section addresses the limitation of using cosine distance, which focuses solely on the directional information of vectors without accounting for positional differences between trajectory points. Therefore, Euclidean distance is employed as the similarity measure to globally assess vector similarity. The data sources used for matching in this study consist of two parts: First, high-frequency dynamic and static information uploaded to the server by the edge computing device of the experimental ship, referred to as the enhanced AIS (E-AIS). This information is collected through multi-sensors (e.g., GPS, IMU) with a high update frequency, making it suitable for the experimental environment. Second, visual positioning data of the target ship is computed in real time by the edge computing device onboard the ship. It is important to note that there are significant differences between E-AISs and traditional onboard AIS systems. While traditional AIS devices provide both static and dynamic information about the ship, their transmission frequency is irregular, and data loss may occur during communication, leading to delays or missing updates. Furthermore, visual positioning may be affected by ship occlusion in certain scenarios, resulting in trajectory loss and exhibiting nonlinear characteristics. To address these challenges, this study employs a Euclidean distance-based DTW algorithm for matching and aligning GPS data from the E-AIS or traditional AIS with the visual positioning data of the target ship, thereby achieving precise alignment and integration of trajectory points.

The DTW algorithm is a method for measuring trajectory similarity. It can be employed to compare the similarity between two nonlinear time series of varying lengths and to achieve optimal alignment between them [49]. Let the coordinate sequence of visual estimates be denoted as $n(n = 1, \ldots, N)$, while the edge computing device retrieves the historical GPS trajectory sequence $m(m = 1, \ldots, M)$ of surrounding ships by querying the ship information database. In practical applications, in order to reduce the computational burden, the edge computing device only retrieves the trajectory information of ships within the monitoring area of the host ship for the past 5 min. When Euclidean distance is used to measure similarity, a smaller distance indicates a higher similarity between two sequences. The formula for calculating similarity is as follows:

$$\begin{cases} DTW(n,m) = d(n,m) \ \ , n = 1\&m = 1 \\ DTW(n,m) = d(n,m) + \min\{DTW(n-1,m), DTW(n,m-1), DTW(n-1,m-1)\}, n = 1,2,\cdots,N\&m = 1,2,\cdots,M \end{cases} \tag{18}$$

$DTW(n,m)$ is calculated as the sum of the Euclidean distance $d(n,m)$ between current elements and the cumulative distance to the nearest element that can be reached. $DTW(N,M)$ serves as the result of the similarity measurement between two sequences. When the similarity measurement result between two trajectories falls below a preset similarity threshold, the visual-based ship trajectory and the GPS-based ship trajectory are directly associated, and data points are aligned. Figure 9 illustrates the asynchronous nonlinear ship trajectory sequence association using the DTW algorithm.

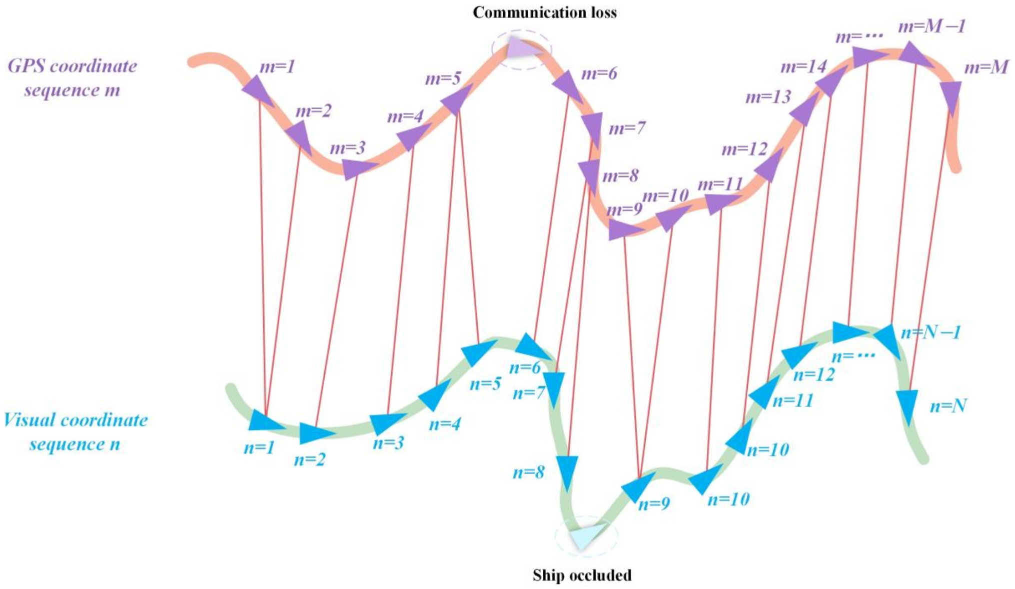
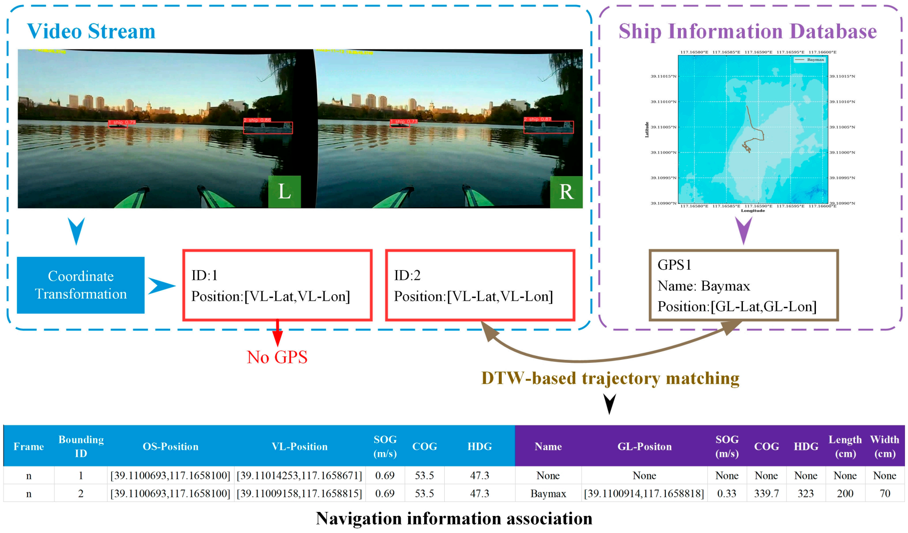


Figure 9. Asynchronous nonlinear ship trajectory sequence association based on the DTW algorithm.

Referring to the fusion data storage design in [13], we have redefined the method for storing correlated data based on our experimental approach. Specifically, after applying the DTW algorithm, the associated results are used to retrieve the target ship's static information (IMO, Name, Length, Breadth—note that the IMO number is a virtual number, as the experimental ship is a self-developed small MASS without an official IMO number, used solely for experimental purposes) and dynamic information (Time, SOG, COG, HDG) from the E-AIS database. These data are then integrated with the dynamic and static information of the experimental ship for joint storage. This combined storage approach aims to support AR-based ship navigation systems with multi-source data, as shown in Figure 10. The blue table represents the navigation information of the own ship and the estimated visual position of the target ship, while the purple table represents the correlated navigation information of the target ship. To enhance the efficiency of trajectory association, similarity measurements are conducted every 5 min for trajectories already associated in the data storage area. During this period, the system retains the association status of these trajectories by default and refrains from performing similarity measurements with other trajectories.



Frame	Bounding ID	OS-Position	VL-Position	SOG (m/s)	COG	HDG	Name	GL-Positon	SOG (m/s)	COG	HDG	Length (cm)	Width (cm)
n	1	[39.1100693,117.1658100]	[39.11014253,117.1658671]	0.69	53.5	47.3	None	None	None	None	None	None	None
n	2	[39.1100693,117.1658100]	[39.11009158,117.1658815]	0.69	53.5	47.3	Baymax	[39.1100914,117.1658818]	0.33	339.7	323	200	70

Navigation information association

Figure 10. Asynchronous ship trajectory association and joint data storage method.

4. Results

In this section, we conducted extensive experiments on ship target detection, tracking, and fusion localization to evaluate the proposed methods and validate their practicality. The experiments were conducted on the Ubuntu 20.04 operating system and Python version 3.8.10. The hardware configuration included an NVIDIA RTX 2080Ti GPU (manufactured by NVIDIA Corporation, Santa Clara, California, USA), Intel(R) Xeon(R) Platinum 8255C CPU (manufactured by Intel Corporation, Santa Clara, California, USA). The model framework was based on PyTorch 2.0.0.

4.1. Experimental Data

In this section, a benchmark dataset (FLShip) was constructed for ship target detection, tracking, and fused localization. Compared to existing open-source maritime datasets, FLShip introduces multidimensional innovations in its design. It not only includes continuous stereo and monocular image frames captured from a first-person perspective during navigation but also introduces high-frequency E-AIS information corresponding closely to the image data of both target and own MASSs for the first time. This feature significantly enhances the dataset's applicability in supporting multi-source information fusion research for MASSs, providing crucial data support for research in maritime target detection, multi-sensor data fusion, and autonomous navigation. The experiments utilized seven MASSs developed by Tianjin University as test subjects. The E-AIS data from the MASSs, including attitude and GPS information, are obtained by subscribing to the /mavros/imu/data topic for attitude data and the /mavros/global_position/global topic for positional data via the ROS system. These data are uploaded in real time to a server via 4G networks (which can be replaced by satellite communication as needed), ensuring efficient data transmission and storage. Figure 11 and Table 2 provide additional details about the MASSs used in this experiment, including their ID, length, breadth, equipped equipment, and levels of autonomy navigation. Among them, the edge computing devices equipped on the experimental ship include Raspberry Pi 4B (manufactured by Raspberry Pi Foundation, Cambridge, UK), Jetson Orin Nano Developer Kit, and Jetson Nano (manufactured by NVIDIA Corporation, Santa Clara, California, USA).

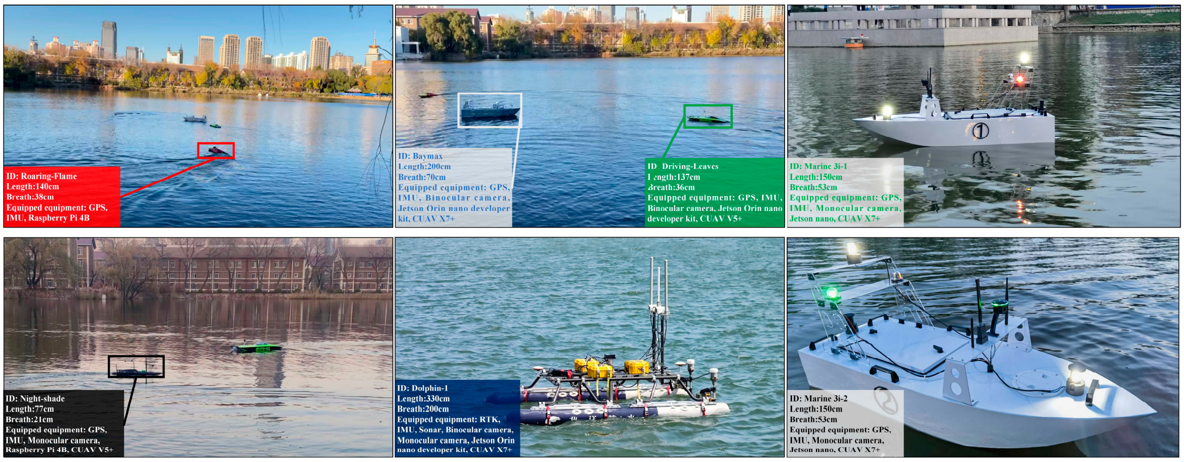

ID: Roaring-Flame
Length:140cm
Breath:38cm
Equipped equipment: GPS, IMU, Raspberry Pi 4B
ID: Baymax
Length:200cm
Breath:70cm
Equipped equipment: GPS, IMU, Binocular camera, Jetson Orin nano developer kit, CUAV X7+
ID: Driving-Leaves
Length:137cm
Breath:36cm
Equipped equipment: GPS, IMU, Binocular camera, Jetson Orin nano developer kit, CUAV V5+
ID: Marine 3i-1
Length:150cm
Breath:53cm
Equipped equipment: GPS, IMU, Monocular camera, Jetson nano, CUAV X7+
ID: Night-shade
Length:77cm
Breath:21cm
Equipped equipment: GPS, IMU, Monocular camera, Raspberry Pi 4B, CUAV V5+
ID: Dolphin-1
Length:330cm
Breath:200cm
Equipped equipment: RTK, IMU, Sonar, Binocular camera, Monocular camera, Jetson Orin nano developer kit, CUAV X7+
ID: Marine 3i-2
Length:150cm
Breath:53cm
Equipped equipment: GPS, IMU, Monocular camera, Jetson nano, CUAV X7+

Figure 11. The MASSs used in the experimental process.

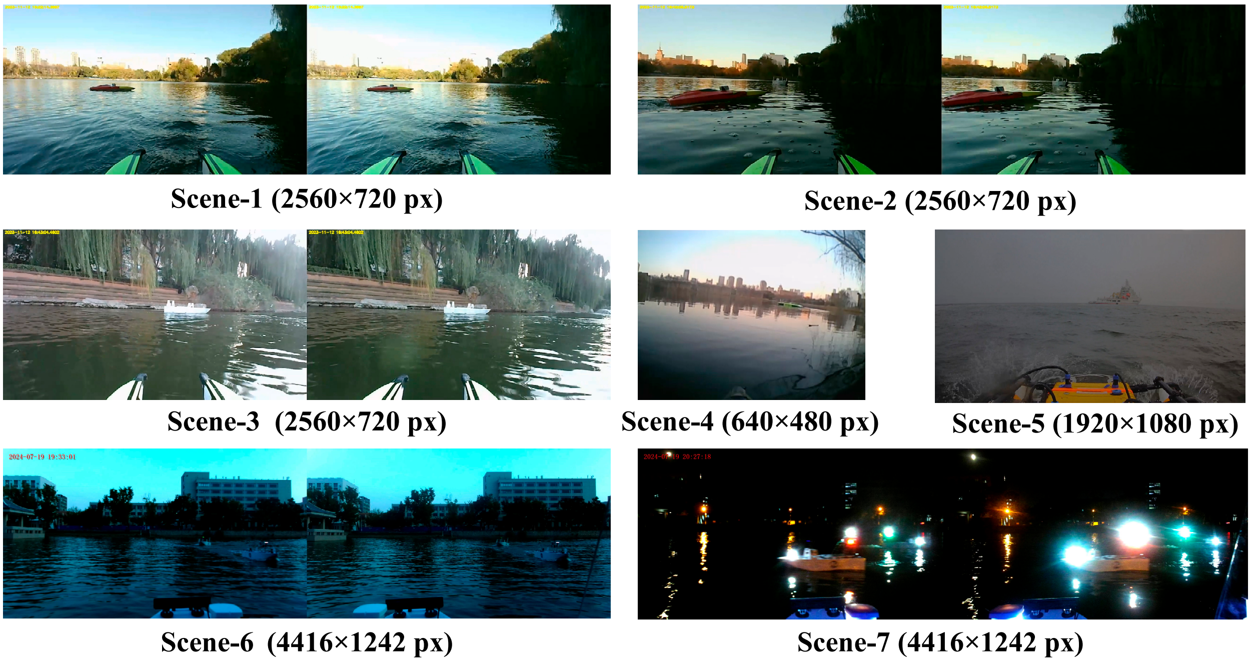
Table 2. Details of the MASS.

ID	**Equipped Equipment**					**Autonomy Level**
	GPS	**IMU**	**Camera**	**Edge Computing Box**	**Controller**	
Roaring-Flame	☑	☑	☒	Raspberry Pi 4B	Manual	Remote Control
Driving-Leaves	☑	☑	Binocular	Jetson Orin nano developer kit	CUAV V5+	Programmed Control
Baymax	☑	☑	Binocular	Jetson Orin nano developer kit	CUAV X7+	Programmed Control
Night-shade	☑	☑	Monocular	Raspberry Pi 4B	CUAV V5+	Programmed Control
Marine 3i-1	☑	☑	Monocular	Jetson nano	CUAV X7+	Programmed Control
Marine 3i-2	☑	☑	Monocular	Jetson nano	CUAV X7+	Programmed Control
Dolphin-1	☑	☑	Monocular and Binocular	Jetson Orin nano developer kit	CUAV X7+	Programmed Control

To validate the effectiveness of the proposed method, seven scenes with varying resolutions, noise levels, and visibility conditions were extracted from the dataset for ship target detection, tracking, and fused localization experiments, as illustrated in Figure 12. Table 3 presents additional details about the FLShip dataset, including the number of ships (NOS), the number of ships with E-AISs (NOE), the frequency of ship E-AISs (FOE), video quality, video frame rate, video resolution, and the average speed of the host ship.

Table 3. Detailed information of the FLShip dataset.

Video	**NOS**	**NOE**	**FOE**	**Video Quality**	**Video Frame Rate**	**Video Resolution (px)**	**Speed (m/s)**
Scene-1	2	2	5&10	Good	5	2560 × 720	0.65
Scene-2	3	2	5	Good	5	2560 × 720	0.42
Scene-3	2	2	5	Good	5	2560 × 720	0.11
Scene-4	2	2	1&5	Poor (Medium noise)	25	640 × 480	0.53
Scene-5	4	3	5	Poor (Fog and waves)	30	1920 × 1080	2.1
Scene-6	3	3	5	Fair (Cool color tone and low light)	5	4416 × 1242	0.41
Scene-7	3	3	5	Poor (Minimal light and high contrast)	5	4416 × 1242	0.15



Scene-1 (2560×720 px)

Scene-2 (2560×720 px)

Scene-3 (2560×720 px)

Scene-4 (640×480 px)

Scene-5 (1920×1080 px)

Scene-6 (4416×1242 px)

Scene-7 (4416×1242 px)

Figure 12. The data samples from the FLShip dataset.

4.2. Ship Target Detection Experiment

To evaluate the performance of the Bi-YOLO algorithm in complex scenarios, we conducted comparative experiments on the FLShip dataset using various benchmark algorithms. These included Deformable-DETR [50], RTMdet [51], YOLOv7 [52], YOLOv8s, and the original YOLO11s. Evaluation metrics included mAP@0.5, mAP@0.5:0.95, GFLOPs, Params, and model size.

The mAP is a comprehensive evaluation metric for object detection system performance, allowing for reliable comparison between different algorithms or models across various datasets. mAP@0.5 operates by calculating the average precision (AP) for all images in the first class when the IoU threshold is set to 0.5, and then repeating the process for the remaining classes. The average of these AP values gives the mAP score. mAP@0.5:0.95 is an extension of mAP, which calculates multiple AP values within the IoU range from 0.5 to 0.95 (with a step size of 0.05) and averages them. Specifically, mAP@0.5:0.95 provides a more comprehensive assessment of model performance across varying strictness conditions by considering multiple IoU thresholds. Compared to the single mAP@0.5, mAP@0.5:0.95 demands that the model maintain high detection accuracy at higher IoU thresholds, thus better reflecting the model's robustness and accuracy in complex scenarios. In addition to precision metrics, the computational efficiency of the model is an important factor in assessing its practical value. GFLOPs reflect the computational complexity of the model, while Params directly determine the model's storage and memory usage. In practical applications, especially on resource-constrained edge computing devices, achieving lower GFLOPs and Params is crucial for improving deployment efficiency.

This study conducted object detection model training based on the FLShip dataset. The dataset comprises 6286 images of MASSs captured in both inland and maritime environments. To enhance the model's generalization capability and reduce the risk of overfitting, data augmentation techniques, including random brightness adjustment (±30%) and Gaussian noise injection ($\sigma = 0.05$), were applied to improve the model's adaptability in complex environments. The dataset was split into training, validation, and test sets at a ratio of 76:16:8 to ensure sufficient training data. This experiment did not use external datasets for validation, focusing instead on the unique characteristics of the FLShip dataset. This dataset is highly representative of real-world application scenarios, containing a rich variety of MASS target information along with complementary E-AIS data, making the experimental results more practically valuable. As shown in Figure 13, we analyzed the performance of Bi-YOLO and mainstream detection models during training by examining the mAP@0.5 metric change curve. The experimental results demonstrate that, in the early stages of training, Bi-YOLO exhibited a performance improvement trend similar to YOLO11s and YOLOv8s, significantly outperforming YOLOv7, Deformable-DETR, and RTMdet-l. As training progressed, Bi-YOLO's detection performance gradually surpassed that of other algorithms, with a 0.4% mAP advantage over YOLO11s and a gap of more than 3.4% compared to traditional architectures like Deformable-DETR. Overall, Bi-YOLO not only showed a faster convergence speed during training but also demonstrated stronger competitiveness in final detection performance, confirming its advantages in object detection tasks.

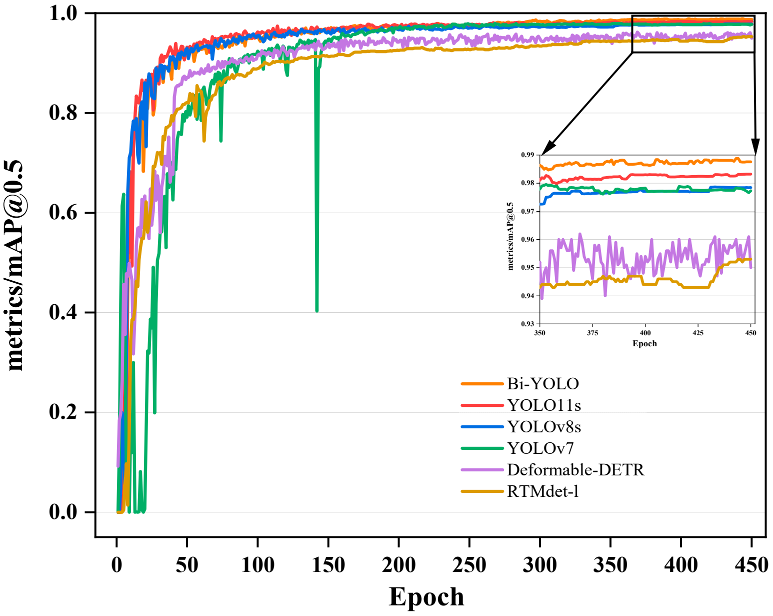


Figure 13. Training mAP@0.5 curves for Bi-YOLO and various object detection algorithms.

To further demonstrate the effectiveness of the proposed method, we compared the Bi-YOLO model with YOLO11s, YOLOv8s, YOLOv7, Deformable-DETR, and RTMdet on the FLShip validation set. The experimental results are shown in Table 4. As observed, Bi-YOLO outperforms the other models in both mAP@0.5 and mAP@0.5:0.95 metrics, achieving 98.7% in mAP@0.5 and 79.6% in mAP@0.5:0.95, with improvements of 0.4 and 0.5 percentage points over YOLO11s, respectively, demonstrating the model's comprehensive capability in target detection under various conditions. As shown in Figure 14, we present a comparison of the detection performance between YOLO11s and Bi-YOLO on the same set of images. In Figure 14c, YOLO11s missed detection because the target ship was far away and the light was dim, which caused the ship to be confused with the background; in Figure 14b,d, YOLO11s also failed to detect the target due to fog interference and the ship being far away. Bi-YOLO showed excellent robustness and adaptability when dealing with complex environments, especially in the detection of small targets at long distances in inland waters and in the sea environment due to interference caused by fog and waves. Bi-YOLO can still accurately identify all targets in these complex environments. This advantage is mainly attributed to the introduced C2f_BiFormer module, which enhances feature representation and fine-grained target feature extraction, enabling the model to perform more stably in detecting small, long-range targets, ensuring precise target capture even in complex, low-visibility scenarios.

Table 4. Evaluation results of ship detection algorithms on the FLShip validation set.

Models	Evaluation Metrics				
	mAP@0.5/% (↑)	**mAP@0.5:0.95/% (↑)**	**Model Size/MB (↓)**	**Params/106 (↓)**	**GFLOPs (↓)**
Bi-YOLO (ours)	98.7	79.6	17.4	8.9	19.4
YOLO11s	98.3	79.1	19.2	9.413187	21.3
YOLOv8s	97.8	78.3	21.3	11.135987	28.6
YOLOv7	97.7	75.5	71.3	36.479926	103.2
Deformable-DETR	95	61.3	646.5	40.099	123
RTMdet-l	95.3	74.3	825.1	52.255	79.951

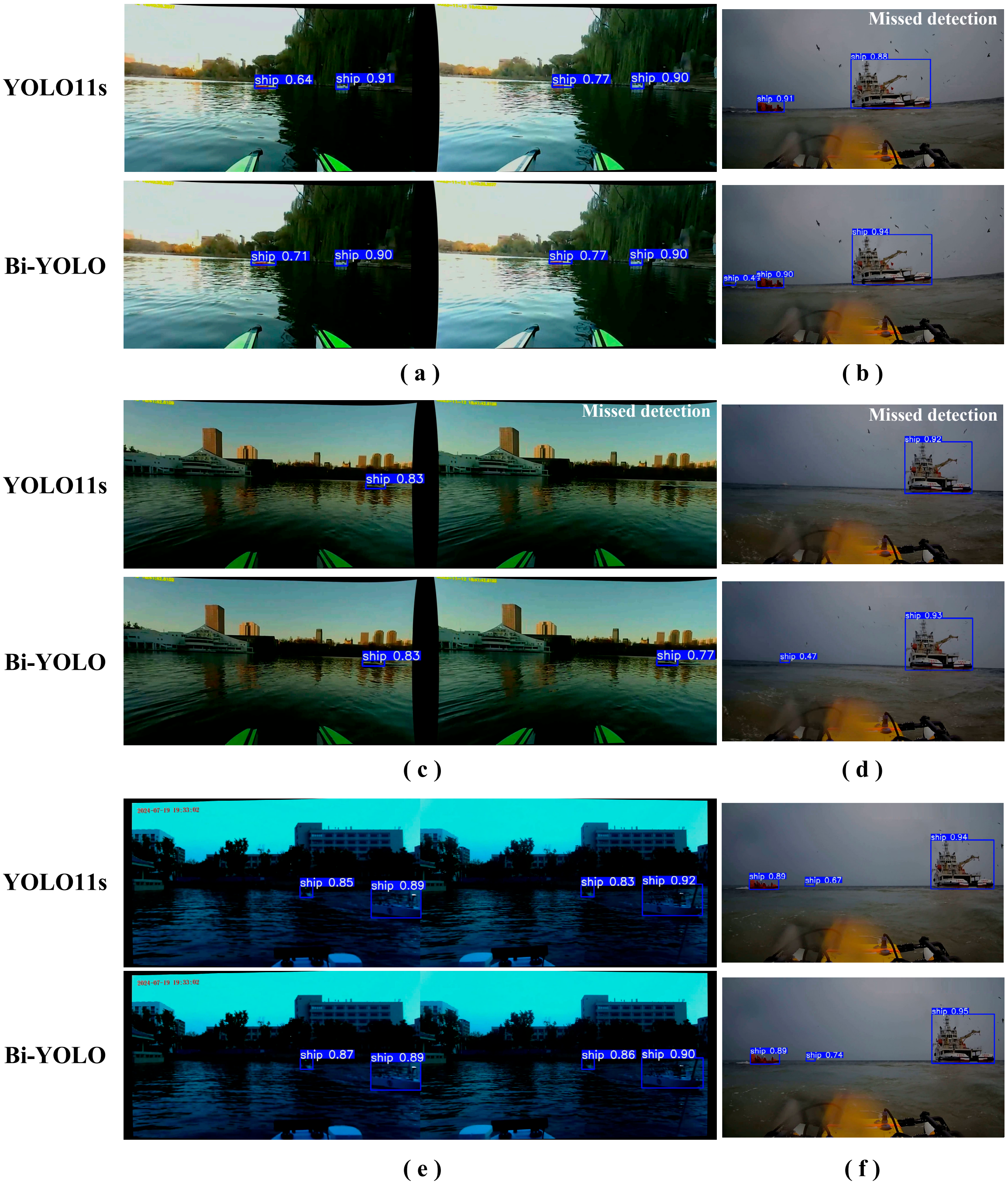


Figure 14. (**a**–**f**) respectively show the comparison of detection effects between YOLO11s and Bi-YOLO.

Compared to the classic Deformable-DETR and RTMdet-l, Bi-YOLO has demonstrated significant improvements in detection capabilities in complex scenarios. For instance, the mAP@0.5:0.95 metric improved by 18.3 percentage points over Deformable-DETR. This improvement is particularly important because mAP@0.5:0.95 better reflects the model's precision in locating small targets under high IoU conditions. Despite Bi-YOLO's excellent performance across multiple accuracy metrics, model complexity and real-time performance remain key considerations in practical applications, especially in embedded hardware environments. In the experiments, Bi-YOLO achieved 19.4 GFLOPs, with 8.9 M parameters and a model size of 17.4 MB, reaching an FPS of 120. Compared to the latest YOLO11s, Bi-YOLO achieves significant optimizations in model size, number of parameters, and computational complexity. Specifically, the model size of Bi-YOLO is reduced by 9.3%, the number of parameters is decreased by 5.4%, and the computational complexity is lowered by 8.9%. These optimizations enhance Bi-YOLO's application potential, making it particularly suitable for environments with limited computing resources.

4.3. Ship Target Tracking Experiment

In the ship target tracking experiments, to validate the tracking performance of the algorithm used in this study under complex scenarios, six scenes with varying resolutions, noise levels, brightness, and contrast were selected from the FLShip dataset for extensive tracking experiments.

Four state-of-the-art trackers, i.e., StrongSORT [53], OC-SORT, Byte Track [54], and BOT-SORT [55], are selected for the comparative tracking experiments in this paper. The performance of the trackers is evaluated using metrics such as Multi-Object Tracking Accuracy (MOTA), Identification F1 (IDF1), and Identity Switches (IDs). Additionally, we evaluated the speed of the tracker in terms of frames per second (FPS). Although the runtime may vary significantly with different hardware, the specific details are shown in Table 5. MOTA is calculated based on false positives (FPs), false negatives (FNs), and identity switches (IDs), emphasizing detection performance. In contrast, IDF1 better measures the consistency of ID matching, and IDs measure the performance of the multi-object tracking model in handling target identity switches. An upward arrow ↑ indicates that a higher value of the metric corresponds to better performance, while a downward arrow ↓ indicates that a lower value of the metric corresponds to better performance.

From Table 5, it can be observed that, under the same detector, OC-SORT demonstrates higher localization accuracy and tracking performance across different scenarios and complex environments. Compared to StrongSORT, ByteTrack, and BOT-SORT, OC-SORT maintains MOTA and IDF1 scores above 90%, even in conditions of severe noise interference or insufficient lighting, accompanied by faster inference speeds, allowing it to perform real-time perception updates at a higher frame rate. This implies that OC-SORT can handle input images more frequently and provide real-time and accurate target tracking information, thus providing reliable perception and decision support for subsequent applications.

Table 5. Multi-ship tracking evaluation results in six scenarios.

Video	Detector	Tracker	Evaluation Metrics			
			MOTA/% (↑)	IDF1/% (↑)	IDs (↓)	FPS/Hz (↑)
Scene-1	Bi-YOLO	StrongSORT	90.4	95.7	0	31
		OC-SORT (ours)	96.3	98	0	65
		ByteTrack	88.6	93.7	0	78
		BOT-SORT	94.1	97.3	0	18
Scene-2	Bi-YOLO	StrongSORT	89.7	94.1	1	26
		OC-SORT (ours)	93.3	96.4	0	57
		ByteTrack	91.8	95.9	0	75
		BOT-SORT	90.9	94.9	0	15
Scene-3	Bi-YOLO	StrongSORT	93.7	95.1	0	29
		OC-SORT (ours)	97.7	98.5	0	41
		ByteTrack	92.5	95.4	0	74
		BOT-SORT	96.2	96.3	0	14
Scene-4	Bi-YOLO	StrongSORT	95	95.5	1	36
		OC-SORT (ours)	93.8	97.2	0	78
		ByteTrack	93.6	96.7	0	86
		BOT-SORT	94.7	94.2	2	28
Scene-5	Bi-YOLO	StrongSORT	90.8	93.7	0	24
		OC-SORT (ours)	92.7	91.8	0	55
		ByteTrack	88.5	86.4	0	66
		BOT-SORT	88	89.9	4	10
Scene-6	Bi-YOLO	StrongSORT	87.3	93.2	1	27
		OC-SORT (ours)	90.6	96.2	0	52
		ByteTrack	85.1	91.5	0	67
		BOT-SORT	86.7	89.3	6	12

Figures 15–17 present the tracking results of four different trackers in Scene-2, Scene-4 with noise interference, and Scene-6 with low lighting conditions. It can be observed that the StrongSORT and BOT-SORT trackers experience multiple ID switches in Scene-2, Scene-4, and Scene-6. Additionally, ByteTrack exhibits lower IOU at several time points (e.g., T1 and T3 in Scene-4, and T4 in Scene-6), resulting in relatively lower bounding box localization accuracy. Comparative experiments show that OC-SORT achieves more stable tracking of the target ship while ensuring real-time performance, striking a good balance between real-time processing and stability. The combination of the Bi-YOLO detector and the OC-SORT tracker used in this study maintains high accuracy and stability in multi-ship tracking tasks, providing a solid foundation for precise and stable target localization in subsequent experiments.

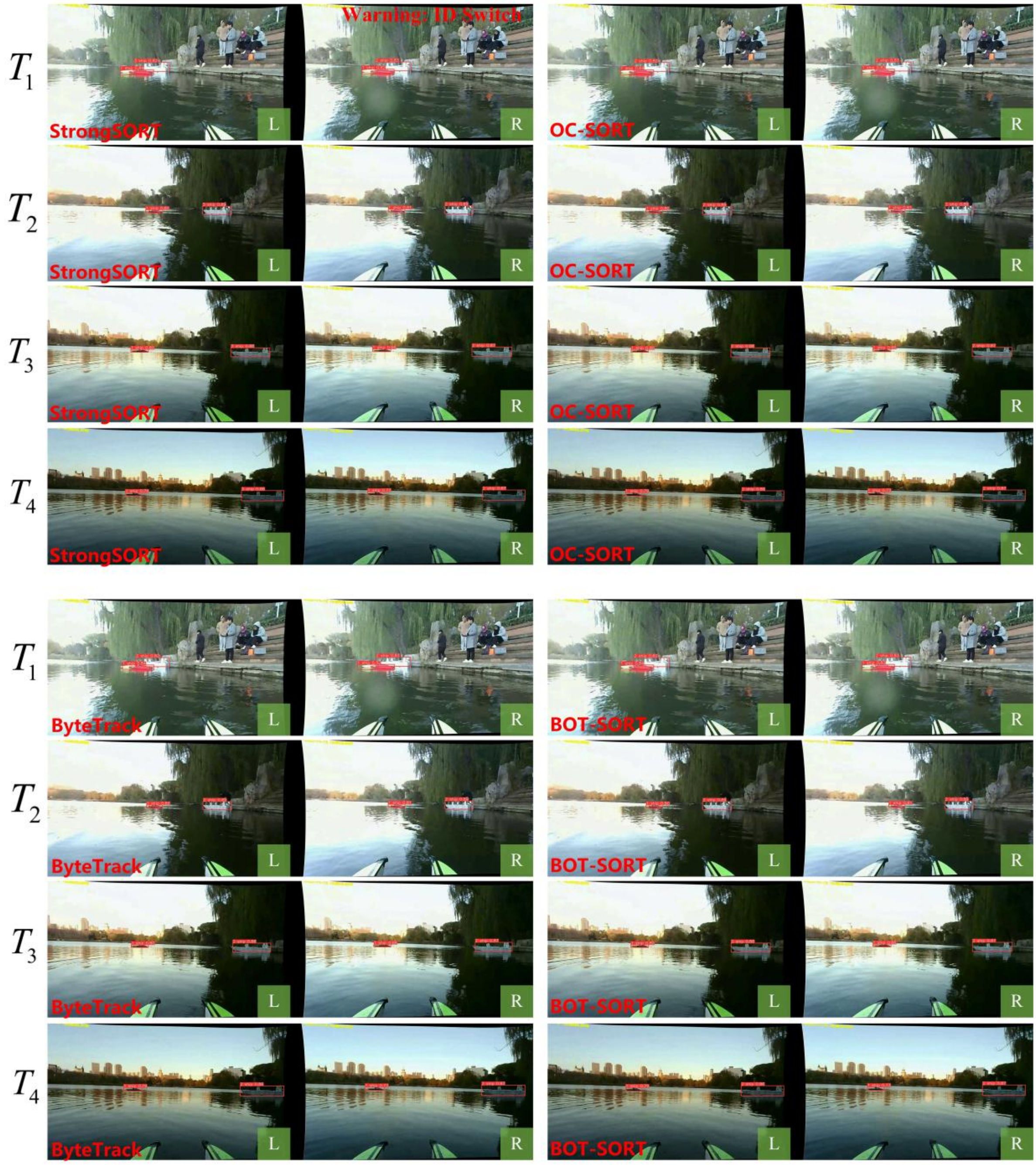


Figure 15. Tracking performance comparison of four state-of-the-art object trackers in Scene-2.

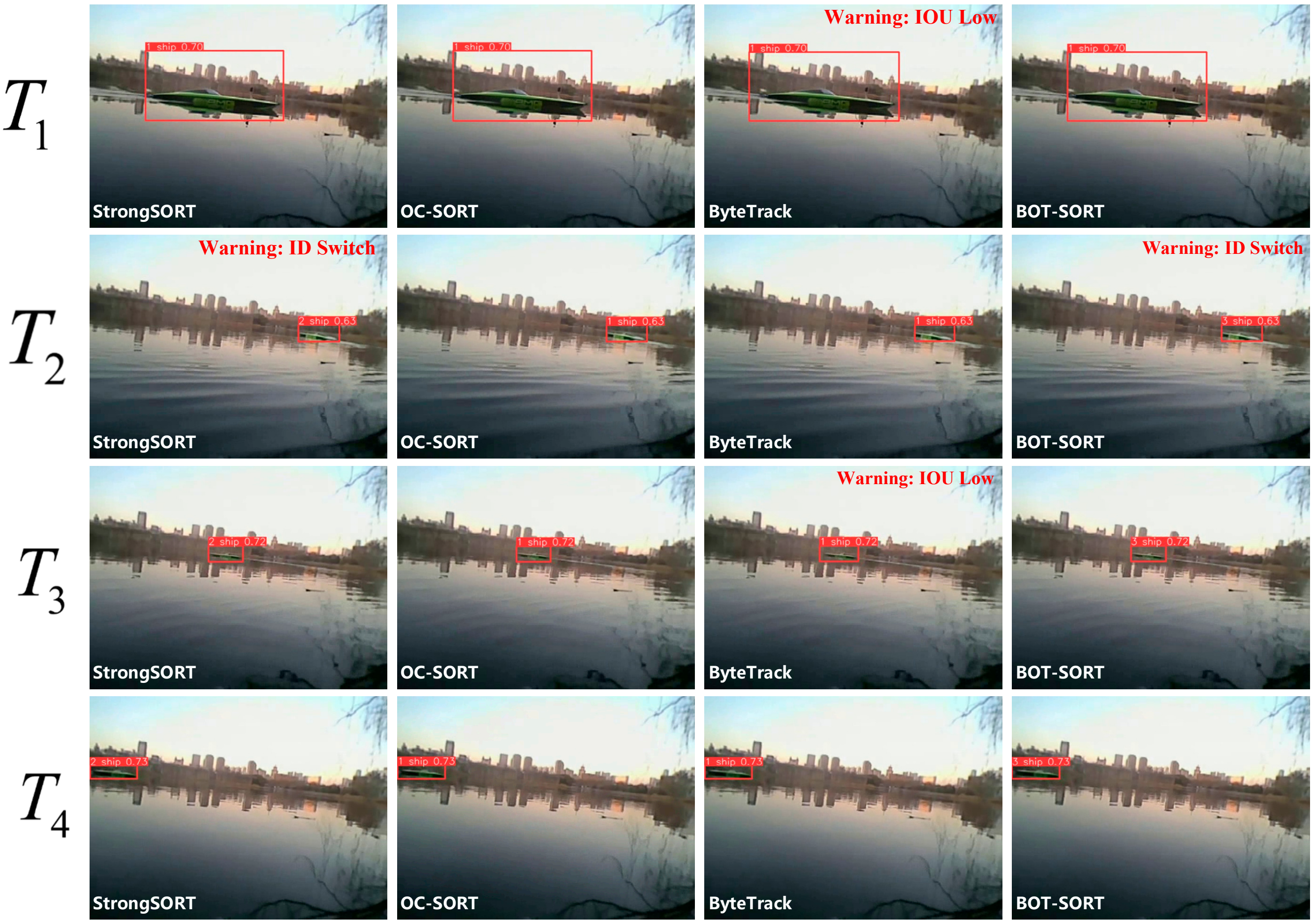


Figure 16. Tracking performance comparison of four state-of-the-art object trackers in Scene-4.

4.4. Visual-Based Ship Position Estimation Experiment

To validate the accuracy of the ship position estimation method in this study, we also recorded the real-time position and attitude data of the host ship and the target ship in real time during the video data collection and then uploaded them to the server for storage. The actual GPS data uploaded by the target ship serves as a ground truth for comparison. To facilitate the comparison between the real value of longitude and latitude and the measured value extracted by vision, systematic sampling is performed according to the GPS frequency of the target ship to ensure that the collected position information can be matched frame by frame. To verify the robustness and accuracy of the position estimation method, this paper selects videos recorded in different navigation states for position estimation experiments. In addition, the mean absolute error (MAE), mean square error (MSE), mean absolute percentage error (MAPE), and mean relative position error (MRPE) are also introduced in the experiment to comprehensively evaluate the position estimation method proposed in this paper.

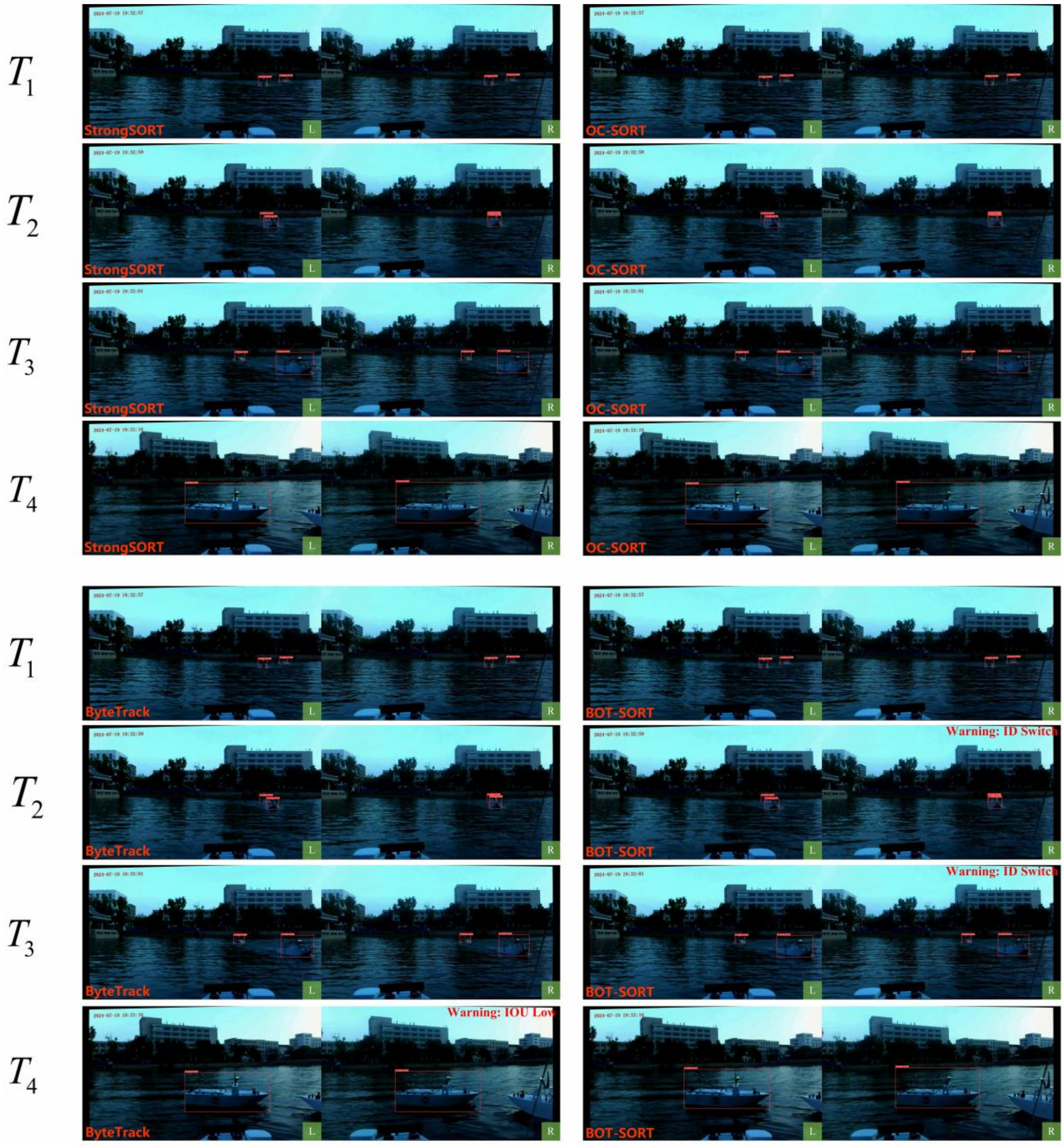


Figure 17. Tracking performance comparison of four state-of-the-art object trackers in Scene-6.

Figures 18–20 present the results of multi-target visual position extraction and comparison across various scenes. The purple trajectories represent the estimated positions of the target ship obtained using our framework through stereo images, while the blue trajectories indicate the actual GPS data of the target ship (i.e., its ground truth positions). The pink trajectories correspond to the GPS data of our own ship, which approximately represents the camera's position. Table 6 presents the MAE, MSE, MAPE, and MRPE values for visual latitude and longitude estimates across different scenes compared to the true values. In Scene-1, the own ship performed reverse motion and large-angle turning maneuvers. These movements caused significant vibrations and sensor noise, which notably impacted the accuracy of visual position estimation. Consequently, this scenario imposed higher precision requirements on the sensors. Under these conditions, the MRPE in Scene-1 was higher than that in Scene-2. Additionally, since stereo vision-based localization relies on disparity, the resolution of depth information decreases with increasing target distance as the disparity angle diminishes. This limitation resulted in lower position estimation accuracy in Scene-3 compared to other scenarios. Nevertheless, our framework demonstrated strong robustness in Scene-3, achieving an MAE of 5.6403 × $10^{-6\circ}$ and MSE of 4.0986 × $10^{-11\circ}$ and keeping the MRPE within 8.76%. These results indicate that the proposed framework can achieve high-precision target position estimation even in complex scenarios, providing stable and reliable trajectory features for subsequent multi-source data association.

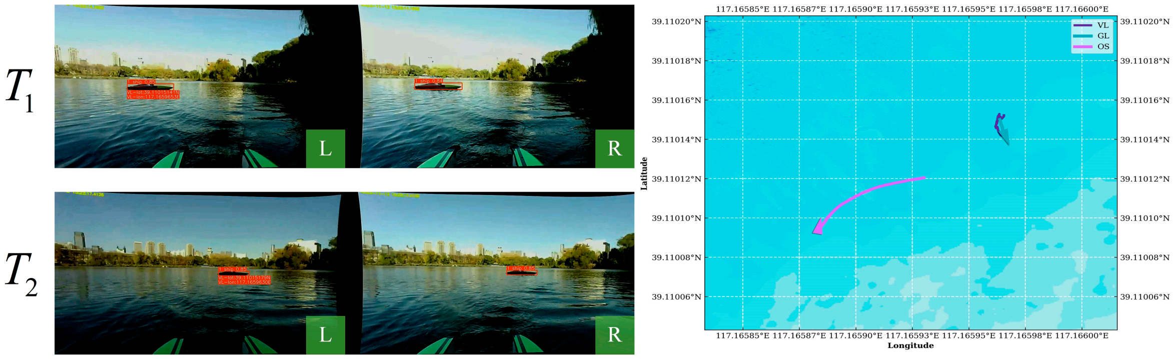


Figure 18. The visual position estimation results of the 'Roaring-Flame' MASS in Scene-1.

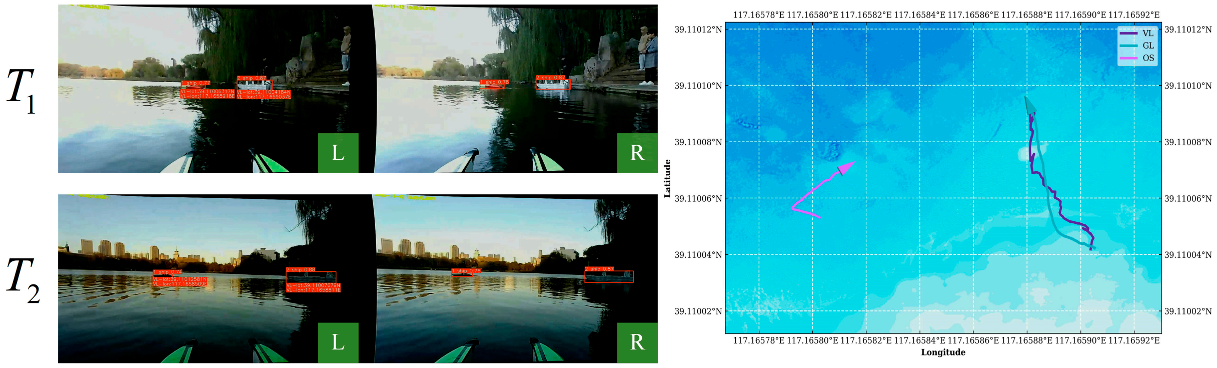


Figure 19. The visual position estimation result of the 'Baymax' MASS in scene-2.

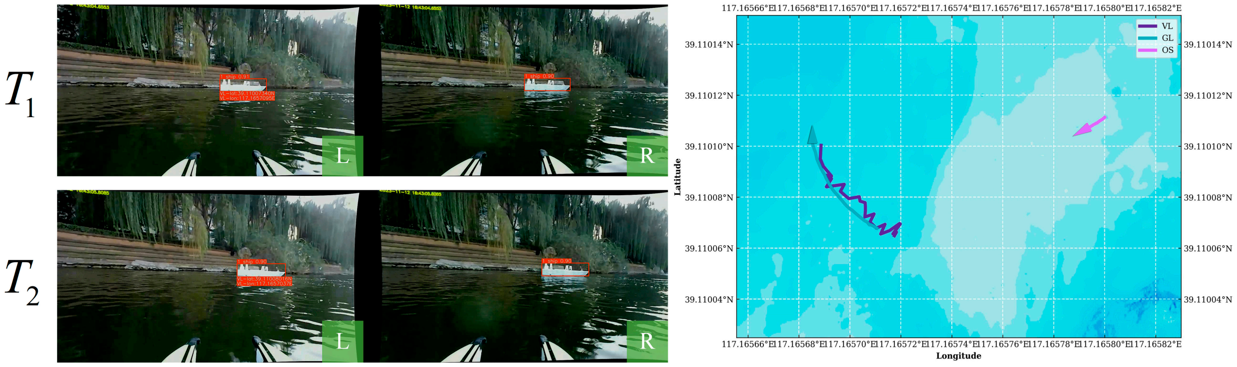


Figure 20. The visual position estimation results of the 'Baymax' MASS in Scene-3.

Table 6. The MAE, MSE, MAPE, and MRPE values for target ship position estimates in multiple scenes.

Video	Feature	MAE (°)	MSE (°)	MAPE (%)	MRPE (%)
Scene-1	Coordinates	1.68 × 10^{-6}	4.25 × 10^{-12}	3.37 × 10^{-8}	4.44
Scene-2	Coordinates	2.39 × 10^{-6}	8.16 × 10^{-12}	3.91 × 10^{-8}	4.4
Scene-3	Coordinates	5.64 × 10^{-6}	4.1 × 10^{-11}	8.2 × 10^{-8}	8.76

4.5. Multi-Source Data Fusion Experiment

In this section, we aim to verify the effectiveness of the proposed multi-source data fusion method. Specifically, we utilize a Euclidean distance-based DTW algorithm to associate the visual trajectory of the target ship (generated by fusing data from a camera, GPS, and IMU) with the GPS trajectory in the E-AIS database. The dynamic and static information of both the own ship and the target ship is then jointly stored. Considering the crew's sailing habits and information needs, we constructed a simple, intuitive, and easy-to-understand human–computer interface, and constructed the AR navigation assistance results in the left camera image, so that the crews can understand the ship's navigation status more comprehensively and intuitively and improve the efficiency of acquiring and understanding the navigation information. Figure 21 visually demonstrates the AR navigation assistance results for the ship based on the perceptual data fusion framework.

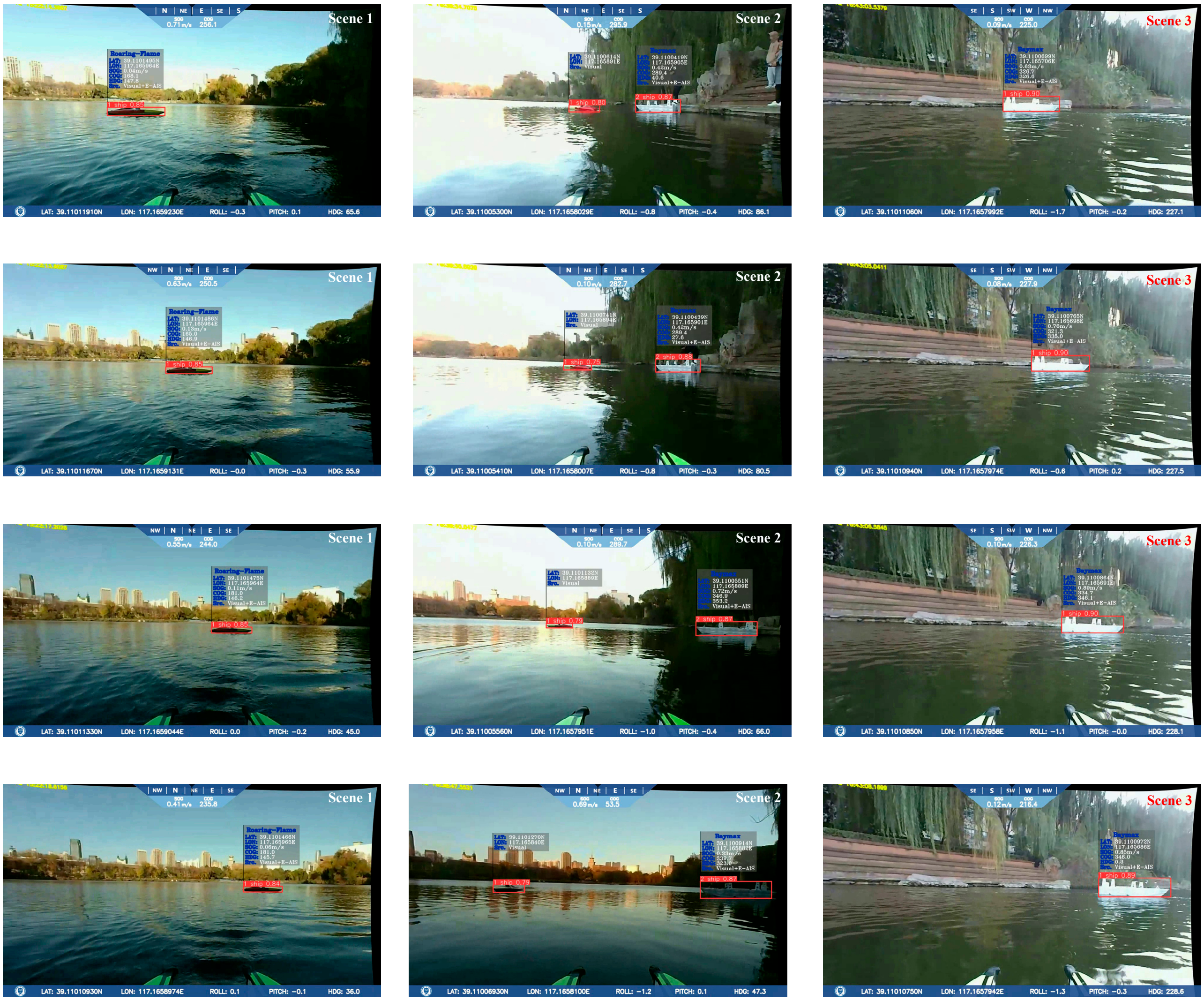


Figure 21. The AR navigation assistance effects of ships constructed at different timestamps in multiple scenes.

In Figure 21, the results of AR navigation assistance provide comprehensive information about the tracked ships, including static details (IMO, Name, Length, Breadth) and dynamic information (Lat, Lon, SOG, COG, HDG), as well as navigation information for the host ship (Lat, Lon, SOG, COG, HDG, etc.). The ship data fusion method proposed in this paper successfully and accurately matches the visual and GPS trajectories. The dynamic information is updated in real time with the movement of the ship, allowing the ship AR navigation system to more accurately reflect the current status. The navigation assistance system helps the crew to improve the efficiency of information acquisition and understanding during navigation and also provides the crew with richer navigation assistance information, so as to improve the overall navigation experience.

The unique feature of the perceptual data fusion method based on ship ST-TF is that it can respond to possible emergencies such as communication or other equipment failures by displaying the visual latitude and longitude coordinates (as shown by the left ship in Scene 2 of Figure 21) without being associated with the GPS trajectory. Even in these emergencies, the system is still able to provide basic location information to the crew, providing critical data support for emergency response, search, and rescue operations. In addition, for ships not associated with a GPS track, displaying their visual latitude and longitude coordinates can effectively monitor and identify illegal ship operations. In terms of maritime surveillance, this feature can make up for the defects of communication equipment not configured according to regulations or maliciously shut down, and enhance the maritime surveillance capability, thereby strengthening the security of the entire maritime field.

5. Conclusions and Future Perspectives

In this paper, a perceptual data fusion framework based on ship ST-TF was proposed for ship AR navigation assistance systems to ensure navigation safety and enhance maritime surveillance efficiency. The main contributions of this paper are as follows. First, the ship tracking system was optimized by introducing the Bi-YOLO network, which incorporates the C2f_BiFormer module to enhance the detection performance of small targets while reducing model parameters. Additionally, the more robust OC-SORT algorithm was integrated to achieve real-time and stable tracking of moving ship targets. Second, a visual localization model of maritime targets based on binocular imaging was constructed, and the absolute position estimation of maritime targets in the absence of a reference was achieved by fusing the attitude information of the camera. Subsequently, a perceptual data fusion method based on ship ST-TF was proposed. This method temporally and spatially fused the visual trajectory with the GPS trajectory from the E-AIS database. It also further integrated other dynamic and static information. Finally, a ship AR navigation assistance system was developed in conjunction with the proposed multi-source fusion perception framework. Comprehensive experiments on multi-scene MASS tracking, visual position estimation, and multi-source data fusion were conducted on the newly developed FLShip dataset to validate the robustness and effectiveness of the proposed perceptual data fusion framework based on ship ST-TF. The results of comparative experiments with various advanced object detection and tracking algorithms demonstrate that even in low-resolution, noisy, and low-brightness scenes, the framework achieves a mAP@0.5 of 98.7% in multi-ship detection while maintaining a MOTA between 90% and 98%, all while reducing the model's parameter count. The framework successfully extracted reliable visual trajectories of target ships and validated the accuracy of visual position estimation across multiple scenes, ensuring the accurate association between visual detection and E-AISs. The ship AR navigation assistance system based on the multi-source fusion perception framework proposed in this paper is able to enhance the ability of MASS traffic situational awareness, provide more comprehensive and intuitive navigation information for the crew, make it easier to cope with the challenges of complex waterways and ensure the safety of ship navigation. In addition, in the case of equipment failure or illegal operation of the target ship, the navigation assistance system is still able to provide critical dynamic data support for tasks such as emergency response, search and rescue, and maritime surveillance by presenting the visual latitude and longitude coordinates of the target, thus enhancing the safety level of the entire maritime field.

Although the proposed framework performs well in scenarios with noise, wave interference, and low visibility, vision-based perception strategies remain constrained by imaging conditions. Additionally, with the integration of multi-sensor perception systems, optimizing real-time performance and ensuring compatibility with existing ship navigation systems are critical prerequisites for practical deployment. To address these challenges, future research will focus on the following directions to enhance system robustness and adaptability, thereby ensuring effectiveness and reliability in real-world applications:

- Multi-sensor fusion technology: In-depth research on the fusion mechanisms of multi-source sensors, such as cameras, radars, and AISs, will be conducted. Dynamic sensor weight adjustment based on the navigation environment will provide more accurate ship motion information, further strengthening situational awareness in the era of MASS autonomous navigation and offering advanced and reliable solutions for maritime navigation.
- MASS self-organizing networks and data sharing: Research will focus on meeting the shared needs of MASS autonomous navigation and the construction of maritime spatial information. A highly efficient MASS self-organizing network architecture will be established to enable real-time sharing and collaborative processing of perception data, supporting comprehensive, multi-angle maritime environment awareness and laying the foundation for building smarter, more collaborative maritime traffic systems.
- Overcoming deployment and implementation challenges: Through model compression and hardware acceleration technologies, the computational demands will be reduced, and inference efficiency will be improved to ensure real-time performance on low-power embedded platforms. At the same time, efforts will be made to resolve compatibility issues between perception models and existing ship automation systems, navigation software, and international maritime regulations. The development of open interfaces compliant with IMO standards will facilitate seamless integration of technology with existing navigation systems, ensuring the alignment of performance, regulatory compliance, and safety, thus promoting stable and reliable deployment.

Author Contributions: S.C.: Conceptualization, Formal analysis, Investigation, Methodology, Validation, Visualization, Writing—original draft, Writing—review and editing. M.G.: Conceptualization, Data curation, Funding acquisition, Methodology, Writing—review and editing. P.S.: Visualization, Validation, Formal analysis. X.Z.: Data curation, Writing—review and editing, Visualization, Formal analysis. A.Z.: Supervision, Funding acquisition, Data curation. All authors have read and agreed to the published version of the manuscript.

Funding: This project was supported partially by the National Natural Science Foundation of China (52201414), Young Elite Scientists Sponsorship Program of the China Institute of Navigation (No. YESSCIN2023002), National R&D Program of China (2022YFC3006000, 2022YFB3207400), Natural Science Foundation of Tianjin (22JCQNJC00290) and Fundamental Research Funds for the Central Public Welfare Research Institutes (TKS20240305) and Tianjin Transportation Technology Development Plan Project (2023-47).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article.

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

References

1. European Maritime Safety Agency Annual Overview of Marine Casualties and Incidents. Available online: https://www.emsa.europa.eu/newsroom/latest-news/item/4266-annual-overview-of-marine-casualties-and-incidents-2020.html (accessed on 13 June 2024).
2. Maritime Safety Committee. *Report of the Maritime Safety Committee on Its One Hundredth Session*; IMO: London, UK, 2019.
3. Kretschmann, L.; Burmeister, H.-C.; Jahn, C. Analyzing the Economic Benefit of Unmanned Autonomous Ships: An Exploratory Cost-Comparison Between an Autonomous and a Conventional Bulk Carrier. *Res. Transp. Bus. Manag.* **2017**, *25*, 76–86. [CrossRef]
4. Teledyne FLIR. Available online: https://www.flir.cn/marine/navigation/ (accessed on 15 November 2024).
5. HD Hyundai's Avikus Successfully Conducts the World's First Transoceanic Voyage of a Large Merchant Ship Relying on Autonomous Navigation Technologies. Available online: https://avikus.ai/en-us/press/hd-hyundais-avikus-recognized-for-innovation-at-ces-for-second-consecutive-year-1-0-0-0-1 (accessed on 5 January 2025).
6. Laera, F.; Fiorentino, M.; Evangelista, A.; Boccaccio, A.; Manghisi, V.M.; Gabbard, J.; Gattullo, M.; Uva, A.E.; Foglia, M.M. Augmented Reality for Maritime Navigation Data Visualisation: A Systematic Review, Issues and Perspectives. *J. Navig.* **2021**, *74*, 1073–1090. [CrossRef]
7. Clunie, T.; DeFilippo, M.; Sacarny, M.; Robinette, P. Development of a Perception System for an Autonomous Surface Vehicle Using Monocular Camera, LIDAR, and Marine RADAR. In Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA), Xi'an, China, 30 May 2021; pp. 14112–14119.
8. Liu, R.W.; Guo, Y.; Nie, J.; Hu, Q.; Xiong, Z.; Yu, H.; Guizani, M. Intelligent Edge-Enabled Efficient Multi-Source Data Fusion for Autonomous Surface Vehicles in Maritime Internet of Things. *IEEE Trans. Green Commun. Netw.* **2022**, *6*, 1574–1587. [CrossRef]
9. Huo, G.; Wu, Z.; Li, J. Underwater Object Classification in Sidescan Sonar Images Using Deep Transfer Learning and Semisynthetic Training Data. *IEEE Access* **2020**, *8*, 47407–47418. [CrossRef]
10. Wei, S.; Zeng, X.; Qu, Q.; Wang, M.; Su, H.; Shi, J. HRSID: A High-Resolution SAR Images Dataset for Ship Detection and Instance Segmentation. *IEEE Access* **2020**, *8*, 120234–120254. [CrossRef]
11. Liu, Z.; Yuan, L.; Weng, L.; Yang, Y. A High Resolution Optical Satellite Image Dataset for Ship Recognition and Some New Baselines. In Proceedings of the 6th International Conference on Pattern Recognition Applications and Methods, Porto, Portugal, 24–26 February 2017; SCITEPRESS—Science and Technology Publications: Porto, Portugal, 2017; pp. 324–331.
12. Prasad, D.K.; Rajan, D.; Rachmawati, L.; Rajabally, E.; Quek, C. Video Processing From Electro-Optical Sensors for Object Detection and Tracking in a Maritime Environment: A Survey. *IEEE Trans. Intell. Transport. Syst.* **2017**, *18*, 1993–2016. [CrossRef]
13. Qu, J.; Liu, R.W.; Guo, Y.; Lu, Y.; Su, J.; Li, P. Improving Maritime Traffic Surveillance in Inland Waterways Using the Robust Fusion of AIS and Visual Data. *Ocean Eng.* **2023**, *275*, 114198. [CrossRef]
14. Ding, H.; Weng, J. A Robust Assessment of Inland Waterway Collision Risk Based on AIS and Visual Data Fusion. *Ocean Eng.* **2024**, *307*, 118242. [CrossRef]
15. Singh, A.; Zhou, J.; Lin, C.-T.; Lal, S.; Eidels, A.; Jiang, X.; Brown, S. Enhancing Marine Navigation Performance Using the Head-Up Interface. In Proceedings of the 2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Kuching, Malaysia,, 6–10 October 2024; pp. 496–502.
16. Wang, X.; Jiao, J.; Yin, J.; Zhao, W.; Han, X.; Sun, B. Underwater Sonar Image Classification Using Adaptive Weights Convolutional Neural Network. *Appl. Acoust.* **2019**, *146*, 145–154. [CrossRef]
17. Zhang, X.; Huo, C.; Xu, N.; Jiang, H.; Cao, Y.; Ni, L.; Pan, C. Multitask Learning for Ship Detection From Synthetic Aperture Radar Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 8048–8062. [CrossRef]
18. Skolnik, M.I. (Ed.) *Radar Handbook*, 3rd ed.; McGraw-Hill Education: New York, NY, USA, 2008; ISBN 978-0-07-148547-0.
19. Zhang, S.; Wu, R.; Xu, K.; Wang, J.; Sun, W. R-CNN-Based Ship Detection from High Resolution Remote Sensing Imagery. *Remote Sens.* **2019**, *11*, 631. [CrossRef]
20. Jin, H.-S.; Cho, H.; Jiafeng, H.; Lee, J.-H.; Kim, M.-J.; Jeong, S.-K.; Ji, D.-H.; Joo, K.; Jung, D.; Choi, H.-S. Hovering Control of UUV through Underwater Object Detection Based on Deep Learning. *Ocean Eng.* **2022**, *253*, 111321. [CrossRef]
21. Zhang, Y.; Li, Q.-Z.; Zang, F.-N. Ship Detection for Visual Maritime Surveillance from Non-Stationary Platforms. *Ocean Eng.* **2017**, *141*, 53–63. [CrossRef]
22. Gao, M.; Shi, G.-Y. Ship-Handling Behavior Pattern Recognition Using AIS Sub-Trajectory Clustering Analysis Based on the T-SNE and Spectral Clustering Algorithms. *Ocean Eng.* **2020**, *205*, 106919. [CrossRef]
23. Zhao, J.; Yan, Z.; Zhou, Z.; Chen, X.; Wu, B.; Wang, S. A Ship Trajectory Prediction Method Based on GAT and LSTM. *Ocean Eng.* **2023**, *289*, 116159. [CrossRef]
24. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; IEEE: New York, NY, USA, 2014; pp. 580–587.
25. Girshick, R. Fast R-CNN. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; IEEE: New York, NY, USA, 2015; pp. 1440–1448.

26. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
27. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; IEEE: New York, NY, USA, 2016; pp. 779–788.
28. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; IEEE: New York, NY, USA, 2017; pp. 6517–6525.
29. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. In *Computer Vision and Pattern Recognition*; Springer: Berlin/Heidelberg, Germany, 2018.
30. Bochkovskiy, A.; Wang, C.-Y.; Liao, H.-Y.M. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004.10934.
31. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. In *Computer Vision—ECCV 2016*; Lecture Notes in Computer Science; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Springer International Publishing: Cham, Switzerland, 2016; Volume 9905, pp. 21–37. ISBN 978-3-319-46447-3.
32. Zhang, D.; Zhan, J.; Tan, L.; Gao, Y.; Župan, R. Comparison of Two Deep Learning Methods for Ship Target Recognition with Optical Remotely Sensed Data. *Neural Comput. Applic.* **2021**, *33*, 4639–4649. [CrossRef]
33. Chen, Z.; Chen, D.; Zhang, Y.; Cheng, X.; Zhang, M.; Wu, C. Deep Learning for Autonomous Ship-Oriented Small Ship Detection. *Saf. Sci.* **2020**, *130*, 104812. [CrossRef]
34. Chen, Z.; Liu, C.; Filaretov, V.; Yukhimets, D. Multi-Scale Ship Detection Algorithm Based on YOLOv7 for Complex Scene SAR Images. *Remote Sens.* **2023**, *15*, 2071. [CrossRef]
35. Liu, R.W.; Yuan, W.; Chen, X.; Lu, Y. An Enhanced CNN-Enabled Learning Method for Promoting Ship Detection in Maritime Surveillance System. *Ocean Eng.* **2021**, *235*, 109435. [CrossRef]
36. Zheng, Y.; Liu, P.; Qian, L.; Qin, S.; Liu, X.; Ma, Y.; Cheng, G. Recognition and Depth Estimation of Ships Based on Binocular Stereo Vision. *J. Mar. Sci. Eng.* **2022**, *10*, 1153. [CrossRef]
37. Zhou, Z.; Zhao, J.; Chen, X.; Chen, Y. A Ship Tracking and Speed Extraction Framework in Hazy Weather Based on Deep Learning. *J. Mar. Sci. Eng.* **2023**, *11*, 1353. [CrossRef]
38. Chi Ming, W.; Yanan, L.; Lanxi, M.; Jiuhu, C.; Zhong, L.; Sunxin, S.; Yuanchao, Z.; Qianying, C.; Yugui, C.; Xiaoxue, D.; et al. Intelligent Marine Area Supervision Based on AIS and Radar Fusion. *Ocean Eng.* **2023**, *285*, 115373. [CrossRef]
39. Wu, Y.; Chu, X.; Deng, L.; Lei, J.; He, W.; Królczyk, G.; Li, Z. A New Multi-Sensor Fusion Approach for Integrated Ship Motion Perception in Inland Waterways. *Measurement* **2022**, *200*, 111630. [CrossRef]
40. Gülsoylu, E.; Koch, P.; Yildiz, M.; Constapel, M.; Kelm, A.P. Image and AIS Data Fusion Technique for Maritime Computer Vision Applications. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops, Waikoloa, HI, USA, 3–8 January 2024; pp. 859–868.
41. Chen, J.; Hu, Q.; Zhao, R.; Guojun, P.; Yang, C. Tracking a Vessel by Combining Video and AIS Reports. In Proceedings of the Second International Conference on Future Generation Communication and Networking, Hainan, China, 13–15 December 2008; IEEE: New York, NY, USA, 2008; Volume 2, pp. 374–378.
42. Xu, X.; Wu, B.; Teixeira, Â.P.; Yan, X.; Soares, C.G. Integration of Radar Sequential Images and AIS for Ship Speed and Heading Estimation Under Uncertainty. *IEEE Trans. Intell. Transport. Syst.* **2024**, *25*, 5688–5702. [CrossRef]
43. Lu, Y.; Ma, H.; Smart, E.; Vuksanovic, B.; Chiverton, J.; Prabhu, S.R.; Glaister, M.; Dunston, E.; Hancock, C. Fusion of Camera-Based Vessel Detection and Ais for Maritime Surveillance. In Proceedings of the 2021 26th International Conference on Automation and Computing (ICAC), Portsmouth, UK, 2–4 September 2021; IEEE: New York, NY, USA, 2021; pp. 1–6.
44. Zhang, R.; Zhao, C.; Liang, Y.; Hu, J.; Pan, M. Edge-Based Dynamic Spatiotemporal Data Fusion on Smart Buoys for Intelligent Surveillance of Inland Waterways. *J. Mar. Sci. Eng.* **2025**, *13*, 220. [CrossRef]
45. Jocher, G.; Chaurasia, A.; Qiu, J. Ultralytics YOLO 2024. Available online: https://github.com/ultralytics/ultralytics (accessed on 28 November 2024).
46. Zhu, L.; Wang, X.; Ke, Z.; Zhang, W.; Lau, R. BiFormer: Vision Transformer with Bi-Level Routing Attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023.
47. Cao, J.; Pang, J.; Weng, X.; Khirodkar, R.; Kitani, K. Observation-Centric Sort: Rethinking Sort for Robust Multi-Object Tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 9686–9696.
48. Song, L.; Wu, W.; Guo, J.; Li, X. Survey on Camera Calibration Technique. In Proceedings of the 2013 5th International Conference on Intelligent Human-Machine Systems and Cybernetics, Hangzhou, China, 26–27 August 2013; IEEE: New York, NY, USA, 2013; pp. 389–392.
49. Toohey, K.; Duckham, M. Trajectory Similarity Measures. *Sigspatial Spec.* **2015**, *7*, 43–50. [CrossRef]

50. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable DETR: Deformable Transformers for End-to-End Object Detection. *arXiv* **2020**, arXiv:2010.04159.
51. Lyu, C.; Zhang, W.; Huang, H.; Zhou, Y.; Wang, Y.; Liu, Y.; Zhang, S.; Chen, K. RTMDet: An Empirical Study of Designing Real-Time Object Detectors. *arXiv* **2022**, arXiv:2212.07784.
52. Wang, C.-Y.; Bochkovskiy, A.; Liao, H.-Y.M. YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022.
53. Du, Y.; Zhao, Z.; Song, Y.; Zhao, Y.; Su, F.; Gong, T.; Meng, H. Strongsort: Make Deepsort Great Again. *IEEE Trans. Multimed.* **2023**, *25*, 8725–8737. [CrossRef]
54. Zhang, Y.; Sun, P.; Jiang, Y.; Yu, D.; Weng, F.; Yuan, Z.; Luo, P.; Liu, W.; Wang, X. Bytetrack: Multi-Object Tracking by Associating Every Detection Box. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 1–21.
55. Aharon, N.; Orfaig, R.; Bobrovsky, B.-Z. BoT-SORT: Robust Associations Multi-Pedestrian Tracking. *arXiv* **2022**, arXiv:2206.14651. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.