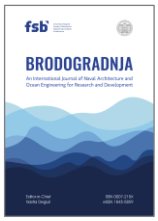
Brodogradnja Volume 77 Number 4 (2026) 77406

University of Zagreb
Faculty of Mechanical
Engineering and Naval
Architecture

journal homepage: www.brodogradnja.fsb.hr

Brodogradnja

An International Journal of Naval Architecture and
Ocean Engineering for Research and Development



A robust stereo-vision system for ship detection and localization on unmanned surface vehicles

Bingqi Ding, Yingjun Zhang, Lai Wei, Hongrui Lu, Zhuolin Wang, Haoze Zhang*

Navigation College, Dalian Maritime University, Dalian, 116026, China

ARTICLE INFO

Keywords:
Stereo vision
Ship detection
Cascaded filtering localization
Unmanned surface vehicles

ABSTRACT

Enhancing the environmental perception of Unmanned Surface Vehicles (USVs) in complex waters is among the primary approaches to ensuring the safety of autonomous navigation. This paper proposes a robust stereo-vision system that integrates ship detection and localization for maritime scenarios. In terms of detection, GS-YOLO is proposed to address the problems of large model parameters and low detection accuracy for small target ships. Based on YOLO11n, the Global-Local Spatial Attention (GLSA) module, Bidirectional Feature Pyramid Network (BiFPN), and SIoU loss function are introduced to improve detection performance while maintaining model lightness. For localization, a cascaded filtering localization algorithm is proposed to address the instability of distance measurement caused by dynamic interference. The algorithm takes depth data generated by RAFT-Stereo as input and applies temporal smoothing by sequentially combining median filtering and Kalman filtering. This significantly enhances robustness against dynamic interference. Experimental results show that GS-YOLO achieves a mean average precision of 93.5 % on the Mcships dataset while reducing parameters by 17.8 % compared with YOLO11n. It achieves an optimal balance between detection accuracy and model lightweighting. Additionally, compared with traditional methods, the cascaded filtering localization algorithm significantly reduces positioning error. Within the range of 50 m, the standard deviation of the error is reduced by 89.3 %; within an 80 m range, the positioning error is maintained below 2.3 %. These results demonstrate that the proposed stereo-vision system provides accurate and stable perception data for the autonomous navigation of USVs.

1. Introduction

Unmanned surface vehicles (USVs), owing to their small size, low cost, and high maneuverability, are widely deployed for ocean monitoring, environmental surveys, maritime search and rescue, and security patrols [1-3]. In these tasks, the environmental perception capability of USVs directly determines the stability and safety of their autonomous navigation. Accurate target detection and localization are fundamental to ensuring the safe autonomous operation of USVs [4]. However, in the complex maritime environment, static obstacles such as reefs and buoys, as well as dynamic targets such as fishing ships and merchant ships

* Corresponding author.
E-mail address: zhangyj@dlmu.edu.cn

https://doi.org/10.21278/brod77406
Received 18 November 2025; Received in revised form 14 April 2026; Accepted 23 April 2026
Available online 20 May 2026
ISSN 0007-215X; eISSN 1845-5859

underway, can pose threats to the navigational safety of USVs. Therefore, under variable meteorological and sea state conditions, rapidly and accurately perceiving targets and estimating their positions is crucial to improving the autonomous navigation capability of USVs [5].

At present, maritime target detection primarily relies on active sensing devices such as radar, sonar, and Automatic Identification System (AIS) [6-8]. Although these devices perform well in target detection, they have certain limitations in practical applications as they lack detailed texture information and cannot operate in specific scenarios such as radio silence [9]. In contrast, vision perception technology based on optical cameras can acquire high-resolution images without transmitting signals [10]. This technology is not only suitable for tasks with strict requirements on the electromagnetic environment but also provides rich texture and shape information, enabling precise target recognition and localization [11, 12].

Among visual perception approaches, binocular stereo vision, due to its simple architecture, low cost, and ability to provide high precision depth information, has become an important means of acquiring three-dimensional (3D) information. This technique reconstructs the 3D structure of a scene by computing the disparity between the left and right cameras It has been widely applied in autonomous driving, unmanned aerial vehicle inspection, and robotic navigation [13]. For USVs, vision-based target detection not only offers higher spatial resolution than radar but also effectively compensates for the blind spots of radar and sonar in target detection. In addition, binocular stereo vision can provide precise 3D information of targets, providing more accurate perception data for downstream decision-making tasks [14, 15]. Although some studies have applied binocular stereo vision to unmanned systems, the following problems still exist in maritime scenarios applications: (1) Existing research integrating deep learning-based detection and 3D reconstruction remains inadequate to simultaneously meet the requirements for accurate, efficient, and real-time target localization. (2) Most existing detection models rely on large-scale network architectures, resulting in high computational costs that make them unsuitable for real-time applications on resource-constrained USV platforms. (3) During navigation, ocean waves and water surface disturbances affect USVs. These factors cause variations in roll, pitch, and lateral forces. These attitude changes can cause localized fluctuations or abrupt variations in localization results, thereby degrading the stability of visual depth estimation and reducing the temporal consistency of the perception system [16].

These problems reduce the reliability of the perception system in dynamic scenes and threaten the navigational safety of USVs. To address these issues, this paper develops an integrated stereo-vision system that combines lightweight ship detection with robust localization. The lightweight ship detection approach optimizes target detection efficiency, while the cascaded filtering localization algorithm enhances spatial localization accuracy. These two components work synergistically, providing highly reliable environmental perception data support for autonomous navigation systems.

The main contributions are as follows:

(1) An integrated stereo-vision perception system is developed to achieve both ship detection and localization for USVs. The system is validated through real-ship experiments, demonstrating its effectiveness and applicability in complex maritime environments.

(2) The GS-YOLO model is built on YOLO11, with improvements to the neck section through the introduction of a Global-Local Spatial Attention (GLSA) mechanism. The original feature fusion module is replaced by a Bidirectional Feature Pyramid Network (BiFPN), and bounding box regression is refined using the SCYLLA-Intersection over Union (SIoU) loss function. These improvements enhance small object detection performance in complex scenes while maintaining a lightweight model design.

(3) To address the problem of unstable distance measurements caused by ship motion during navigation, a cascaded filtering localization algorithm is proposed. The algorithm applies median filtering and Kalman filtering to temporally smooth the depth sequences output by RAFT-Stereo, ensuring continuous and stable distance estimation of the target.

Subsequent chapters are organized as follows. Section 2 reviews related work. Section 3 presents the proposed system, including the lightweight GS-YOLO detection model and the cascaded filtering localization

algorithm. Section 4 evaluates detection performance and verifies localization accuracy using real-ship data. Section 5 presents the discussion and conclusions of this work.

2. Related work

2.1 Ship detection based on deep learning

With the development of artificial intelligence (AI), traditional detection methods based on manually crafted features such as Histogram of Oriented Gradients (HOG) and Scale Invariant Feature Transform (SIFT) are gradually being replaced by deep learning approaches using Convolutional Neural Networks (CNNs). Deep learning methods exhibit superior capabilities in feature extraction and generalization, which make them the mainstream approach in ship detection. Deep learning-based detection methods are categorized into two-stage and one-stage methods. Two-stage methods, such as Faster R-CNN, first generate candidate regions and then perform object classification and bounding box regression within those regions [17]. The intermediate step of candidate region reduces the overall detection speed of two-stage methods, resulting in insufficient performance for real-time ship detection. In contrast, one-stage methods such as You Only Look Once (YOLO) [18] integrate classification and regression into a single network and make predictions directly from the feature map. Compared with two-stage methods, they provide faster detection and are better suited for ship detection tasks.

The uniqueness of ship target detection lies in the large variation in object scales, frequent occlusion, and the concentration of targets near the sea-sky line, which significantly increases its difficulty [19]. To address these challenges, existing studies have proposed a variety of improvement strategies. Liu et al. [20] integrated semantic features from multiple levels and introduced Cross-Layer Modulated Deformable Convolution (CLMD-Conv) to enhance target localization performance. They further employed the Spatially Informed Multi-Scale Feature Refinement (SIMSFR) module to improve detection accuracy for ship targets near the sea-sky line by enhancing the discrimination of multi-scale features. Si et al. [21] proposed a ship detection method based on multi-scale feature fusion. They constructed a spatial pyramid for remote sensing images and applied dual-branch and context-aware networks. This enhanced the detection of small ships and improved the separation capability for densely distributed targets. Wang et al. [22] proposed a lightweight real-time ship detection model called ALF-YOLO. Based on YOLOv8n, the model incorporates the Asymptotic Feature Pyramid Network (AFPN), Large Selective Kernel Attention Mechanism (LSK), and an additional detection head, which extracts appropriate discriminative features and improve recognition performance in occluded scenes.

In summary, existing methods have made significant progress in target detection. However, there are still shortcomings in improving the detection accuracy of small targets and suppressing interference from sea-sky background while maintaining model lightweight. Therefore, this paper proposes an enhanced lightweight algorithm based on the YOLO11n framework, aiming to improve the detection performance of small targets in maritime scenes and promote more efficient deployment.

2.2 Ship localization based stereo vision

For the autonomous navigation of unmanned systems, precise 3D position information is crucial for path planning and obstacle avoidance [23]. In recent years, binocular stereo vision has been widely applied to target localization tasks because it can obtain both images and depth information within a single vision system [24]. Jung et al. [25] developed a stereo vision-based system that enables mobile platforms to acquire real-time 3D information, enhancing object recognition and spatial localization in dynamic environments. Benacer et al. [26] proposed an obstacle detection algorithm based on U-V disparity maps, which analyses disparity structure to achieve accurate foreground segmentation, improving stability and detection accuracy in complex scenes. These studies have laid a solid foundation for applying stereo vision to localization in dynamic environments. Meanwhile, with the rapid development of deep learning in visual perception, several studies have explored the use of deep learning-based image matching to enhance vision-based localization. Khurshid et al. [27] proposed a vision-based 3D localization method using deep image matching, where deep neural

networks are employed to extract robust visual features and establish reliable correspondences between images, enabling accurate spatial localization of UAVs. To address the challenges encountered in maritime environments, Zheng et al. [28] proposed a method that combines FSRCNN super-resolution enhancement with sub-pixel ORB matching. Firstly, FSRCNN is utilized to enhance image resolution and feature details. Then, sub-pixel localization improves matching accuracy and optimizes disparity estimation. The method significantly improves the accuracy of depth estimation for maritime targets, effectively addressing the issues of low resolution and sparse matching in long-range stereo ranging. Shang et al. [29] introduced an IMU attitude compensation mechanism in the ranging system. Real-time pitch and roll angles data of the vessel are used to correct image distortion to enhance the robustness of stereo matching in dynamic conditions. This effectively reduces depth errors caused by attitude fluctuations.

These studies demonstrate the potential of binocular stereo vision in dynamic maritime environments. However, the accuracy and stability of binocular localization depend not only on the performance of the stereo matching algorithm but also on the application scenario. In real maritime scenes, sea surface reflections, low texture sky, and the dynamic change of waves often cause the depth output to contain noise, holes, and transient outliers, which undermines the stability and reliability of continuous localization. Accordingly, this paper proposes a ship localization method that builds on high precision stereo matching results and incorporates cascaded filtering to enhance the stability and continuity of localization.

3. Proposed system

3.1 Overall framework

This paper develops a robust stereo-vision system for ship detection and spatial localization on USVs. The system consists of two main modules: a ship detection module and a localization module. As illustrated in Figure 1, the integrated framework enables an end-to-end perception pipeline that covers image acquisition, target detection, and 3D distance estimation. By combining these modules, the system provides a unified and reliable perception capability for maritime environments.

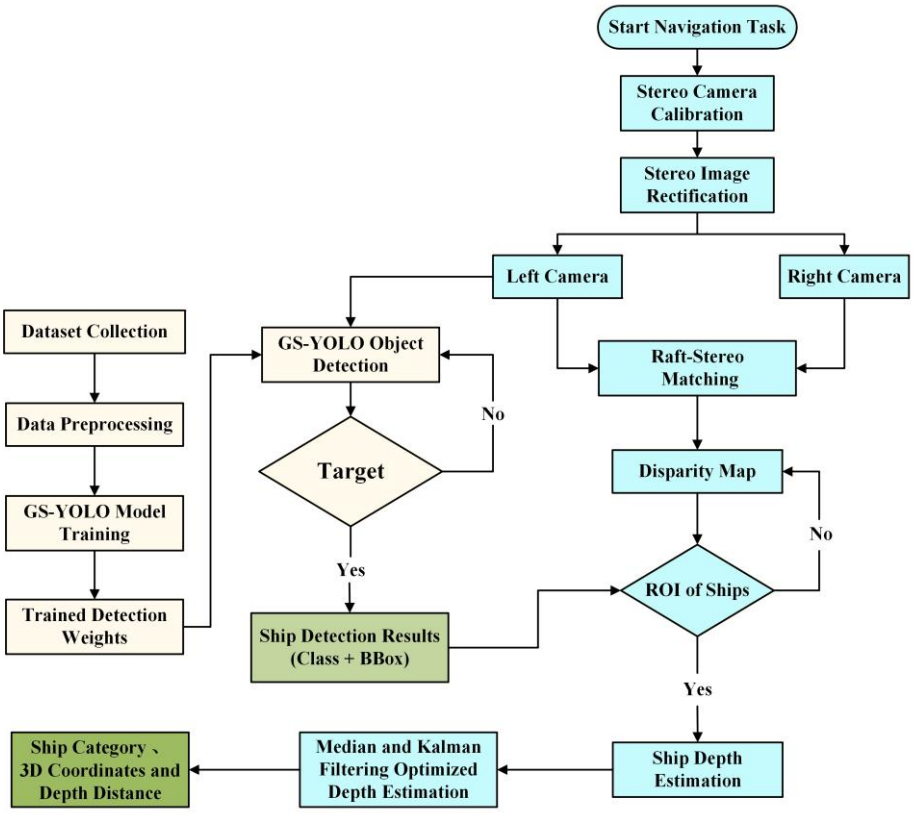


Fig. 1 Overall structure of the stereo-vision system

The ship detection module receives real time images from the left camera and uses the proposed GS-YOLO model to detect and classify ships. The detection output includes the target class, bounding box, and confidence score. The pixel coordinates of the bounding box are extracted and passed to the localization module as input. The localization module takes synchronized images from the left and right cameras. It uses the region of interest (ROI) provided by the detection module as a spatial constraint and applies RAFT-Stereo to estimate the disparity map of the target region. For each frame, the average disparity within the central region of the ROI is extracted to reduce background interference near the bounding box edges, and the 3D coordinates of the target are then computed based on stereo imaging geometry. Based on the estimated depth information, the spatial position of the target can then be determined. To improve the smoothness and stability of the depth sequence, the postprocessing first applies median filtering to remove transient outliers and then applies Kalman filtering for temporal smoothing, thereby yielding a continuous and stable 3D localization output.

The system generates high-confidence semantic information via the target detection module and achieves precise 3D localization through the localization module. This integration significantly enhances the autonomous perception capabilities of USVs in complex environments, providing reliable perception data for downstream tasks such as path planning and obstacle avoidance.

3.2 Ship detection module

The YOLO11 algorithm is an advanced model that enables target detection and classification. The model is divided into three functional stages: the backbone network for early feature processing, the neck network for multi-scale fusion, and the head for prediction tasks [30]. The backbone incorporates the C3k2 module to replace the C2f structure from YOLOv8, using smaller convolutional kernels to construct cross-stage residual connections. Together with the lightweight Conv-BN-SiLU (CBS) module and the Spatial Pyramid Pooling-Fast (SPPF) module, this design enhances the receptive field and feature extraction capability of the model. The neck adopts a bidirectional feature fusion structure based on the Path Aggregation Network and Feature Pyramid Network (PAN-FPN). The head part employs an anchor-free decoupled detection head, which separates classification from regression tasks, thereby improving detection accuracy, especially for small targets.

To address the numerous challenges of ship detection, particularly in terms of sea-sky background interference, poor detection performance on small targets and constrained deployment resources, this paper proposes a lightweight detection model called GS-YOLO, based on YOLO11. The model is designed to improve detection accuracy, reduce missed detections, and enhance computational efficiency and real-time performance, which facilitates effective deployment on USVs for maritime operations.

The architecture of the proposed GS-YOLO model is illustrated in Figure 2.

1) Attention mechanism enhancement: The GLSA attention module is introduced between backbone and neck to enhance ship target features and suppress the sea-sky background interference.

2) Neck network optimization: The original PAN-FPN is replaced with BiFPN, which adaptively fuses multi-scale features through a weighted connection mechanism to enhance the detection of small targets.

3) Loss function optimization: The Complete Intersection over Union (CIoU) loss is replaced with SIoU, which incorporates angle deviation, aspect ratio, and center direction information. This strengthens the geometric constraints of bounding box regression, improving regression accuracy and convergence efficiency.

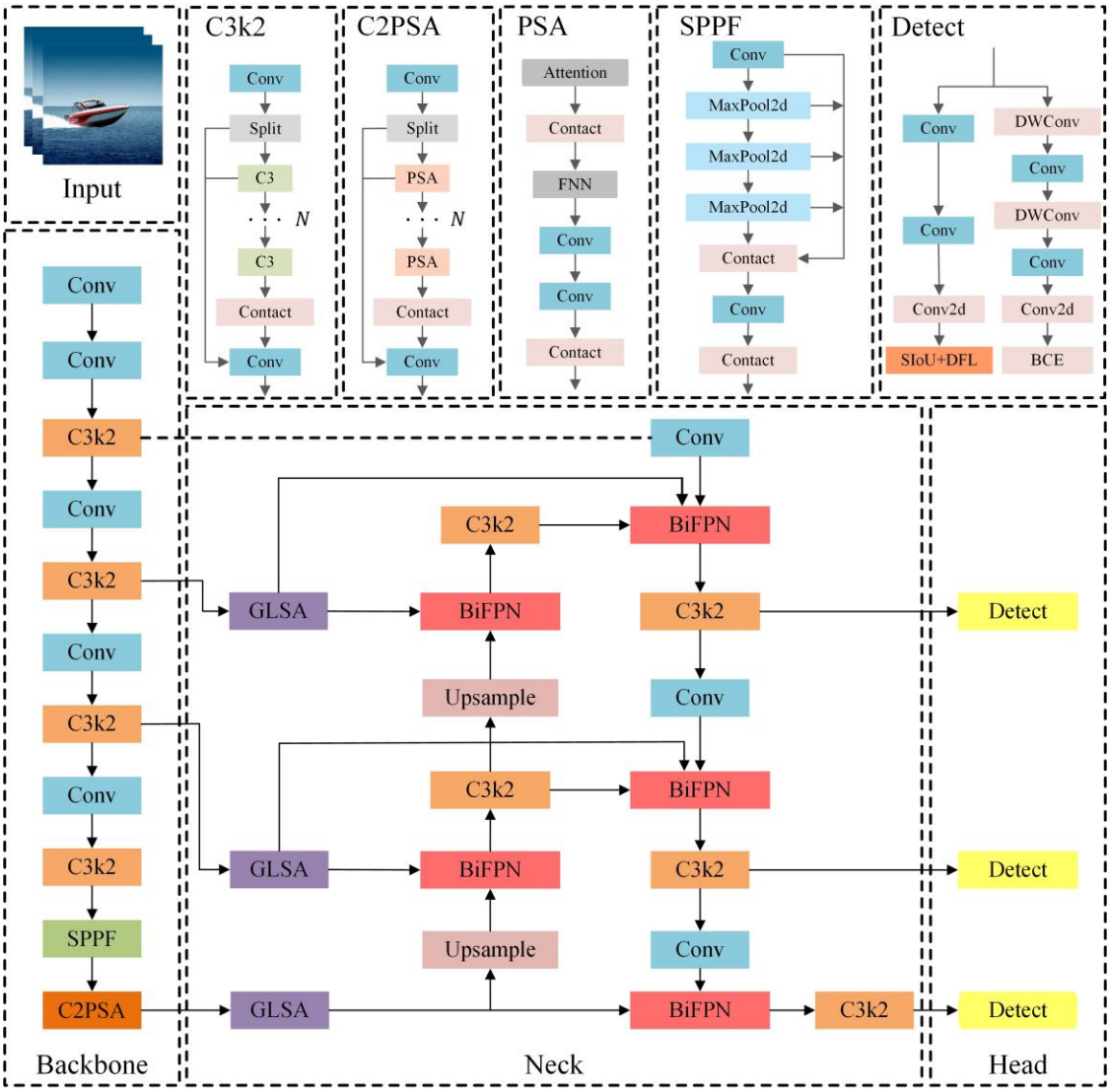


Fig. 2 Network architecture of GS-YOLO. GS-YOLO improves the YOLO11 architecture by integrating GLSA for the fusion of global and local information, BiFPN for multiscale feature fusion, and SIoU for enhanced localization accuracy

3.2.1 GLSA feature enhancement module

In the YOLO11 architecture, the connection between the backbone and neck serves as a critical feature transmission layer but merely forwards feature data, lacking semantic reconstruction and contextual enhancement of the feature maps. In ship detection tasks, feature maps from different down sampling stages of the backbone differ significantly in semantic richness and spatial accuracy. Shallow features preserve fine details but weak semantics, while deeper features contain stronger semantics but vague spatial resolution. Directly feeding these features into the neck for fusion may lead to semantic inconsistency and imbalanced feature responses, ultimately compromising overall detection accuracy. To address this problem, the GLSA attention module [31] is introduced between the backbone and neck as an intermediate feature enhancement layer in the YOLO11 network.

This module adopts a two-dimensional collaborative enhancement strategy to improve the expressiveness of multi-scale feature maps through spatial response and semantic information. In maritime scenes, the sea–sky background often forms large homogeneous regions with weak texture, which can easily cause background confusion during detection. By jointly modeling global contextual dependencies and local spatial responses, the GLSA module suppresses irrelevant background activations and highlights ship-related structural features. This process strengthens the representation of ship targets and improves robustness against sea–sky background interference.

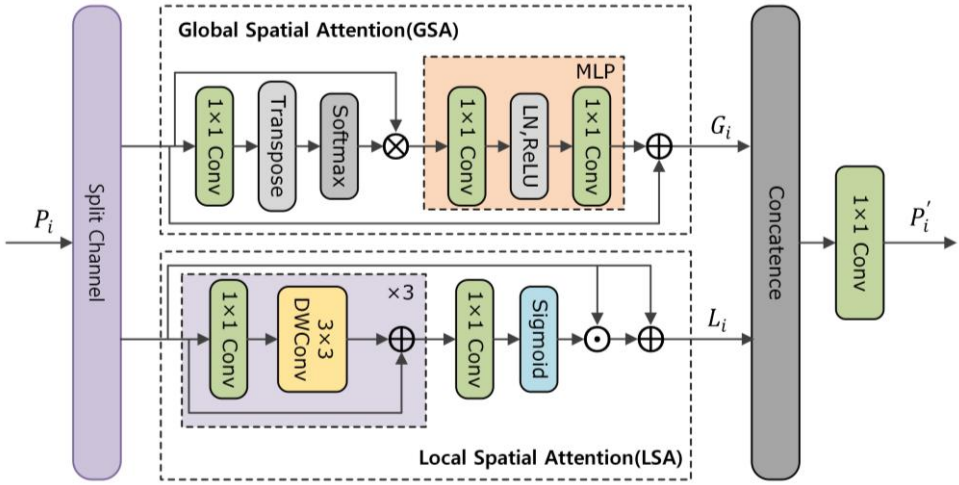


Fig. 3 The detailed structure of the GLSA attention module

The GLSA attention module comprises two parallel submodules: Global Spatial Attention (GSA) and Local Spatial Attention (LSA), which are responsible for enhancing contextual information and emphasizing local salient responses within the input features, respectively. Its structure is shown in Figure 3. To strengthen the representation of spatial information within each feature map, the GLSA attention module is applied individually to every feature map P_i, which is then divided along the channel dimension into two sub-feature maps, and the operation can be expressed as:

$$P_i^g, P_i^l = \text{split}(P_i), P^g, P^l \in \mathbb{R}^{\frac{C_i}{2}\times H_i\times W_i} \tag{1}$$

where, P_i^g and P_i^l denote the sub-feature maps used as inputs to the GSA and LSA branches, respectively. C_i, H_i and W_i denote the channel number, height, and width of the feature map P_i, respectively. $\text{split}(\cdot)$ represents the operation of dividing the feature map along the channel dimension into two equal parts.

The GSA branch enhances semantically dependent regions in the feature map using a spatial attention mechanism. A transpose operation aligns the feature map dimensions to compute the global spatial correlation matrix. This process improves contextual awareness and actively filters out the widespread homogeneous sea or sky noise based on long range dependencies. LSA branch uses depth wise separable convolutions to extract local structures and edge features, while combining spatial attention weights to emphasize salient regions, such as the distinct boundary contours between the ship and the horizon, preventing the target from blurring into local wave textures. Following concatenation along the channel dimension, a 1×1 convolution is used to fuse the output from both branches for feature integration and channel restoration, resulting in an enhanced output feature map that effectively isolates the ship target from the complex maritime background. The fusion process is defined as follows:

$$P_i' = C_{1\times1}\big(\text{concat}(G_i, L_i)\big), \qquad P_i' \in \mathbb{R}^{C_i\times H_i\times W_i} \tag{2}$$

where, P_i' denotes the enhanced output feature map corresponding to pyramid level i. G_i and L_i represent the output feature maps of the GSA and LSA branches, respectively. $\text{concat}(\cdot)$ denotes the channel-wise concatenation operation. $C_{1\times1}(\cdot)$ represents the 1×1 convolution used for feature fusion and channel restoration.

3.2.2 Enhanced BiFPN for multi-scale feature fusion

The neck network of YOLO11 adopts a combined structure of Feature Pyramid Network (FPN) and Path Aggregation Network (PANet) to achieve multi-scale fusion. As shown in Figure 4(a), FPN enables top-down semantic propagation by progressively up sampling deep features and integrating them with shallow ones. PANet introduces bottom-up pathways on the basis of FPN to supplement detailed information, forming a bidirectional feature flow, as shown in Figure 4(b). PANet achieves cross-level feature interaction through

bidirectional information structure, yet its fusion strategy exhibits notable limitations. This simplistic feature superposition approach fails to effectively distinguish between the importance of shallow-level details and deep-level semantics, while simultaneously increasing computational complexity.

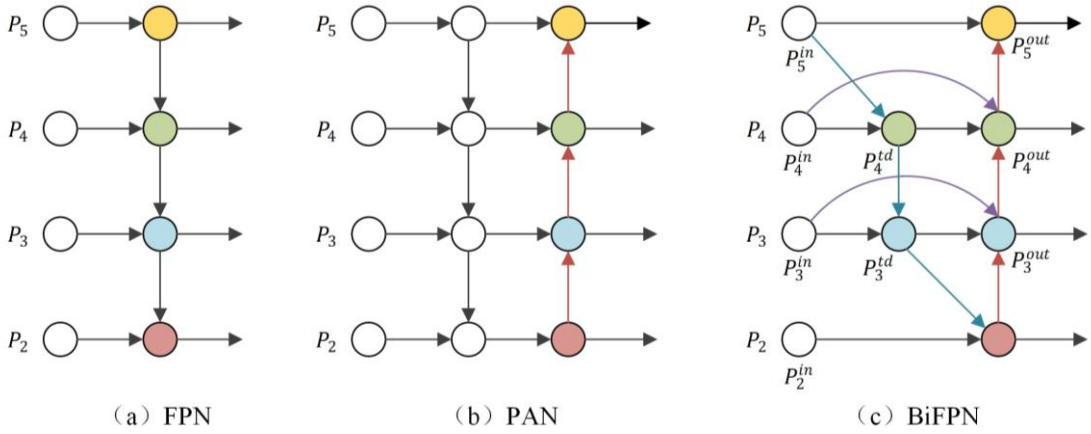


Fig. 4 Comparison of different neck network structures: (a) Feature Pyramid Network (FPN), (b) Path Aggregation Network (PANet), and (c) Bidirectional Feature Pyramid Network (BiFPN)

Therefore, the BiFPN [32] is introduced as the feature fusion module, as shown in Figure 4(c). In this structural, the feature levels are arranged from top to bottom. The feature map P_5 corresponds to deeper layers with lower spatial resolution but stronger semantic information, while P_2 corresponds to shallower layers with higher spatial resolution and richer spatial details. BiFPN employs a learnable weighted fusion mechanism that adaptively adjusts the contribution of each input feature map to the fused output. In addition, the original PANet structure is further refined by removing nodes with only single-path inputs, thereby simplifying the fusion process and improving efficiency. Additional skip connections are also introduced between the input and output nodes at the same pyramid level, as indicated by the curved arrows in Figure 4(c). This design enables the direct propagation of input features and preserves important spatial details during bidirectional multi-scale feature fusion.

In the feature fusion stage, the contributions of different input features to the output vary due to their differing resolutions. BiFPN employs a fast normalized fusion mechanism to balance the weights of different input features. This mechanism is specifically designed to extract deep structural information from ship targets and ultimately reduce missed or false detections in complex maritime environments. The fusion process is as follows:

$$O = \sum_i \frac{\omega_i}{\varepsilon + \sum_j \omega_i} \cdot I_i \tag{3}$$

where, O denotes the fused output feature map. ω_i represents the learnable weight associated with the input feature I_i, and ReLU activation is applied to ensure $\omega_i \geq 0$. ε is a small constant added to avoid numerical instability during normalization.

To enhance the detection capability for long-range small targets, P_2^{in} is integrated into the BiFPN fusion pathway. The added path enables low-level features to flow from shallow to deeper layers through cross-level connections, thereby strengthening the representation of distant ship targets with fine-grained details from earlier layers. This modification mitigates the limitations of the original structure in capturing small targets while introducing only modest computational cost.

Taking feature layer P_4 as an example, the intermediate feature P_4^{td} and output feature P_4^{out} are computed as follows:

$$P_4^{td} = \mathrm{conv}\left(\frac{\omega_1 \cdot P_4^{in} + \omega_2 \cdot \mathrm{resize}\left(P_5^{in}\right)}{\omega_1 + \omega_2 + \epsilon}\right) \tag{4}$$

$$P_4^{out} = \text{conv}\left(\frac{\omega_1' \cdot P_4^{in} + \omega_2' \cdot P_4^{td} + \omega_3' \cdot \text{resize}(P_3^{out})}{\omega_1' + \omega_2' + \omega_3' + \epsilon}\right) \tag{5}$$

where, $\text{conv}(\cdot)$ represents the convolution operation, $\text{resize}(\cdot)$ denotes the up sampling or down sampling operation. The learned feature weights along the paths leading to node P_4^{td} are ω_1, ω_2, while those leading to node P_4^{out} are $\omega_1', \omega_2', \omega_3'$.

3.2.3 Loss function optimization

While lightweight object detection models reduce computational costs, they often lead to a decrease in detection accuracy. To compensate for this drawback, researchers focused on improving localization performance through enhancements beyond architectural design. Advanced regression loss functions such as EIoU, WiSE-IoU, and SIoU have been introduced to strengthen bounding box regression, thereby improving localization accuracy and overall detection performance.

Current bounding box regression methods have gradually incorporated factors such as center distance and aspect ratio for optimization based on Intersection over Union (IoU). However, these improvements still focus on predicting the relative geometric relationship between the predicted and ground truth boxes. There is still insufficient consideration of directional and other vector information. YOLO11 adopts CIoU as its regression loss function. Its regression path may deviate when detecting rotating or laterally moving ships. This limits the model's localization accuracy.

Consequently, the SCYLLA-Intersection over Union SIoU [33] loss function is incorporated to increase sensitivity to directional information, further improving accuracy of bounding box regression. Specifically, the SIoU loss decomposes the regression process into three components: angle, distance, and shape matching. It effectively addresses the lack of directional awareness in CIoU. As shown in Figure 5, the multi-dimensional constraint offers clearer gradient guidance for optimizing the model. This leads to faster convergence and improved localization accuracy when detecting ships with different orientations.

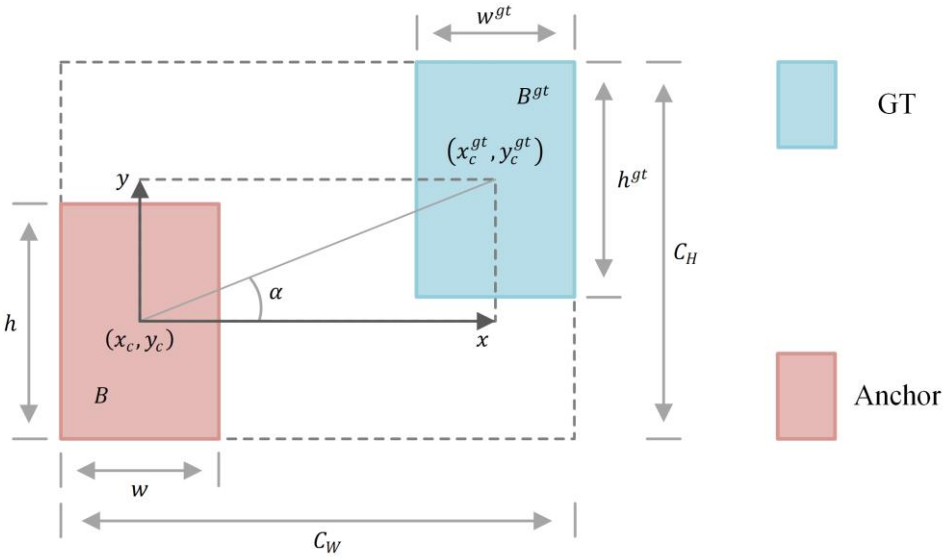


Fig. 5 The SIoU principal diagram

The SIoU loss function includes four components: angle cost, distance cost, shape cost, and IoU cost. The computation of these components is defined as follows:

$$\Lambda = 1 - 2 \cdot \sin^2\left(\arcsin(\alpha) - \frac{\pi}{4}\right) \tag{6}$$

$$\Delta = 2 - \mathrm{e}^{-(\Lambda-2)} \cdot \left(\frac{x_c - x_c^{gt}}{C_W}\right)^2 - \mathrm{e}^{-(\Lambda-2)} \cdot \left(\frac{y_c - y_c^{gt}}{C_H}\right)^2 \tag{7}$$

$$\Omega = \left(1 - \mathrm{e}^{-\left(\frac{|w-w^{gt}|}{max(w,w^{gt})}\right)^{\theta}}\right) + \left(1 - \mathrm{e}^{-\left(\frac{|h-h^{gt}|}{\max(h,h^{gt})}\right)^{\theta}}\right) \tag{8}$$

$$IoU = \frac{|B \cap B^{gt}|}{|B \cup B^{gt}|} \tag{9}$$

where, B and B^{gt} are the predicted box and the ground truth box, respectively. w^{gt} and h^{gt} are the width and height of the ground truth box, while w and h are the width and height of the predicted box. (x_c, y_c) and $\left(x_c^{gt}, y_c^{gt}\right)$ are the center coordinates of the predicted and ground truth boxes, respectively. α is the angle between the line connecting the centers and the x-axis. C_W and C_H are the width and height of the minimum enclosing rectangle covering both boxes, which are used for normalization in the distance loss term. θ is a shape-sensitivity factor controlling the nonlinearity of the shape cost.

Finally, the SIoU regression loss is expressed as:

$$L_{SIoU} = 1 - IoU + \frac{\Delta + \Omega}{2} \tag{10}$$

3.3 Ship localization module

3.3.1 Binocular stereo vision localization principle

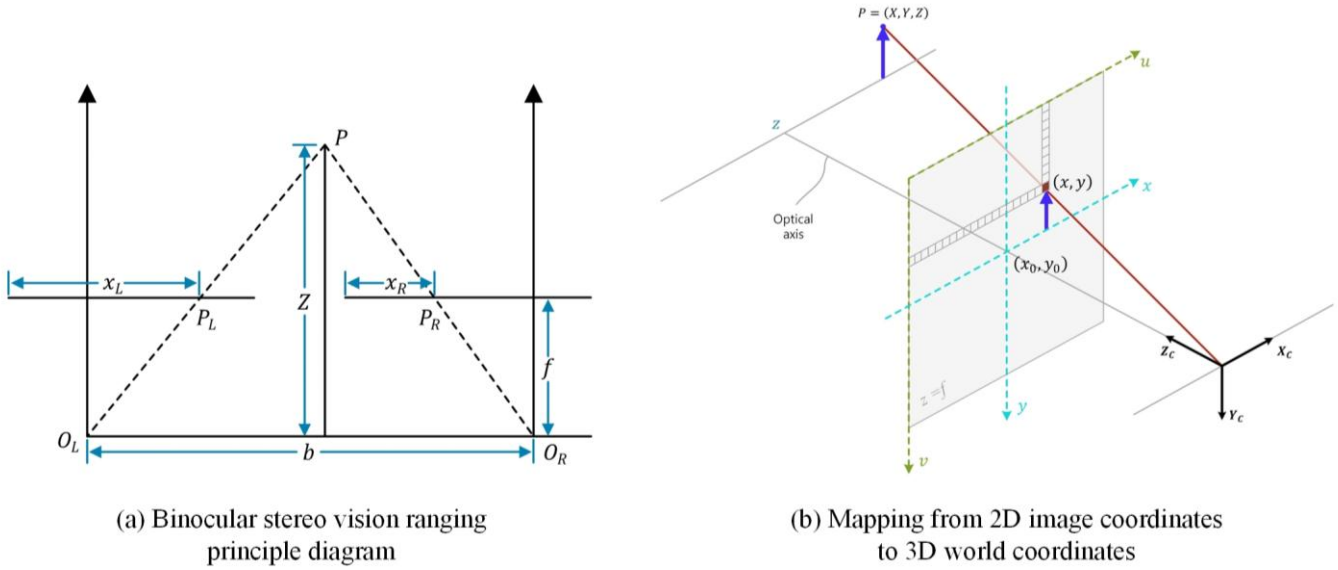
Binocular stereo vision captures synchronized images of the same scene using left and right cameras, and calculates depth based on the disparity between corresponding pixel points in the stereo images. Its principle of distance measurement principle is shown in Figure 6(a). For any point P in 3D space, O_L and O_R are the optical centers of the left and right cameras, respectively. b is the distance between the two camera centers, and f is the focal length of the camera. x_L and x_R are the pixel coordinates of point P on the left and right image planes. Z is the forward depth of point P relative to the left camera.

Based on the principle of similar triangles, the geometric relationship between depth Z and disparity d is derived as follows:

$$Z = \frac{fb}{|x_L - x_R|} = \frac{fb}{d} \tag{11}$$

Based on the pinhole camera model and the depth Z obtained from disparity, 2D image coordinates can be mapped to 3D spatial coordinates. The mapping process is shown in Figure 6(b). Taking the left camera as reference, the principal point is denoted as (x_0, y_0). According to the principle of similar triangles, the 3D coordinates (X, Y, Z) of point P can be derived as follows:

$$\begin{cases} X = \dfrac{x - x_0}{f} Z = \dfrac{b(x - x_0)}{d} \\ Y = \dfrac{y - y_0}{f} Z = \dfrac{b(y - y_0)}{d} \\ Z = \dfrac{fb}{d} \end{cases} \tag{12}$$



(a) Binocular stereo vision ranging principle diagram

(b) Mapping from 2D image coordinates to 3D world coordinates

Fig. 6 Binocular stereo vision ranging principle and 3D coordinate mapping

3.3.2 Stereo matching algorithm

Stereo vision systems rely heavily on the accuracy of the matching algorithm, which directly determines both disparity quality and spatial localization precision. Traditional stereo matching methods predominantly depend on pixel color and local gradients, particularly in challenging maritime environments. This dependence makes them vulnerable to water surface reflections and low-texture regions such as the sky, ultimately reducing matching accuracy. Compared to traditional methods, deep learning-based stereo matching leverages CNNs to automatically extract useful features from images. These models are generally less affected by noise and produce more accurate disparity estimations.

Currently, deep learning methods for stereo matching can be categorized into two categories. One is cost volume-based regression methods, such as the Geometry and Context Network (GC-Net) [34] and the Pyramid Stereo Matching Network (PSMNet) [35]. The core idea is to construct a matching cost volume that contains all potential disparities. Then, a 3D convolutional network is used to perform global information aggregation on this cost volume, ultimately, generating a dense disparity map. This global optimization structure outperforms traditional local or block matching methods in handling object edges and occlusions. Another category is based on iterative optimization, represented by RAFT-Stereo [36]. This method is inspired by ideas from the optical flow domain. It builds gated recurrent units (GRU) on multi-scale feature maps and continuously refines disparity estimation through iterative updates. Compared with cost volume methods, this architecture is more efficient and suitable for real-time processing of high-resolution images.

For performance evaluation in real marine scenarios, the Semi-Global Block Matching (SGBM) [37] algorithm is selected along with representative deep learning methods such as NMRF-Stereo [38], IGEV-Stereo [39], and RAFT-Stereo. The comparison is conducted using disparity estimation on real stereo image pairs with a resolution of 1920×1080 px.

Figure 7 presents the disparity maps generated by different stereo matching algorithms across three sets of port scenes. The traditional SGBM algorithm performs poorly in weak textures and strong reflections regions of the sea surface, resulting in significant noise and blocky artifacts in the disparity maps. Furthermore, as shown in Table 1, its processing speed on high resolution images fails to meet real-time requirements. Deep learning methods perform better overall but differ in their balance between accuracy and speed. NMRF-Stereo offers faster inference but struggles to recover fine details around object edges and boundaries. IGEV-Stereo achieves high accuracy and clearly reconstructs ship structures, but its long inference time limits real-time applicability. Compared with other methods, RAFT-Stereo provides the best balance of accuracy and inference speed. It significantly outperforms other methods with an inference time of just 0.326 s. Moreover, its disparity maps feature sharp edges and precise depth layering, providing high quality depth information for

downstream localization. Therefore, RAFT-Stereo is selected for 3D localization and distance estimation of ship targets in this study.

Table 1 Matching time of different stereo matching algorithms

Algorithm	Matching Time (s)
SGBM	1.232
NMRF-Stereo	0.528
IGEV-Stereo	2.362
RAFT-Stereo	0.326

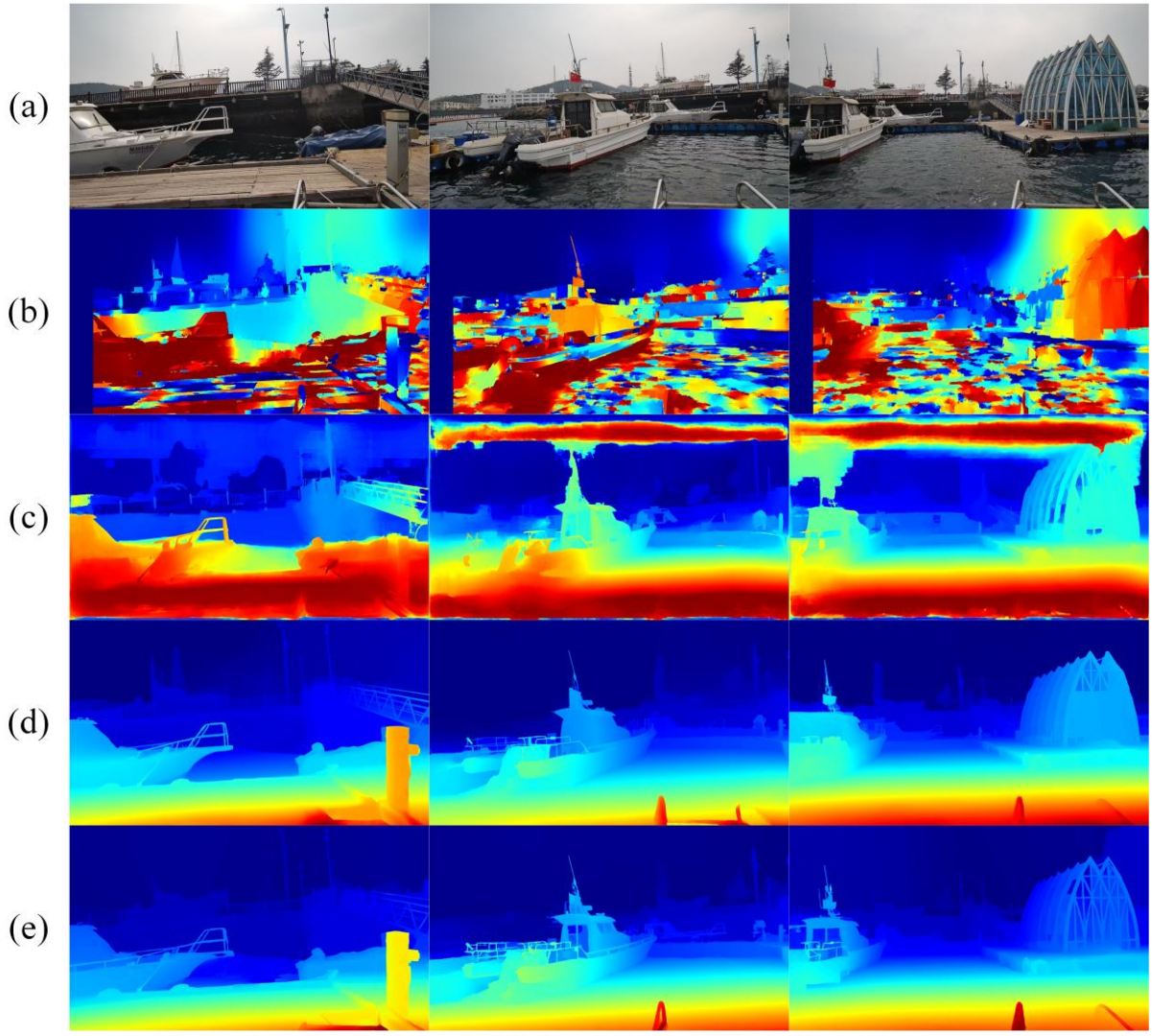


Fig. 7 Visual comparison of disparity maps generated by different stereo matching algorithms on maritime scenes: (a) Left input images; (b) SGBM; (c) NMRF-Stereo; (d) IGEV-Stereo; (e) RAFT-Stereo

3.3.3 Cascaded filtering for robust localization

During autonomous navigation, significant shaking of the USV will alter the pose of the binocular stereo camera. This causes transient deviations in disparity estimation and introduces measurement errors. If these noisy depth values with noise are used directly in decision-making, they may lead to incorrect responses from the USV control system.

To suppress depth fluctuations, this paper proposes a cascaded filtering localization algorithm based on binocular stereo vision. The algorithm combines a median filter and a Kalman filter to reduce the impact of dynamic disturbances. First, median filtering is used to remove extreme outliers in depth estimation, followed by Kalman filtering to suppress sudden changes using prior state estimates, resulting in temporal smoothing of the depth sequence. This approach effectively transforms single-frame, unstable spatial measurements into continuous and smooth state estimation, providing more reliable perception data for downstream decision-making systems.

First, the raw depth sequence Z_k obtained from RAFT-Stereo is processed using a temporal median filter with a sliding window of length w to remove isolated outliers caused by environmental noise. The filtering step produces a smoothed observation $\tilde{Z}_k$, which is then used as the measurement input for the subsequent recursive estimation. A Kalman filter is then employed to combine the prior state estimate with the filtered observation $\tilde{Z}_k$ to obtain a stable distance estimate.

The state transition and observation models are defined as follows:

$$D_k = D_{k-1} + w_k \tag{13}$$

$$\tilde{Z}_k = D_k + v_k \tag{14}$$

where, D_k represents the recursively estimated distance state at frame k. w_k and v_k represent the process and observation noise, respectively, reflecting the system dynamic perturbation and measurement error.

The Kalman filter operates recursively through two stages: prediction and update. In prediction stage, a priori estimate of the current frame is made based on the previous estimate. The corresponding state computations are defined as follows:

$$\widehat{D}_{k|k-1} = \widehat{D}_{k-1|k-1} \tag{15}$$

$$P_{k|k-1} = P_{k-1|k-1} + Q \tag{16}$$

where, $\widehat{D}_{k|k-1}$ denotes the predicted value at frame k. $P_{k|k-1}$ denotes the predicted error covariance. Q denotes the process noise variance.

In the update stage, the prior prediction is adjusted using the current observation to obtain a more accurate state estimate. This process involves computing the Kalman gain, updating the state estimate, and adjusting the error covariance. The corresponding update equations are given below:

$$K_k = \frac{P_{k|k-1}}{P_{k|k-1} + R} \tag{17}$$

$$\widehat{D}_{k|k} = \widehat{D}_{k|k-1} + K_k\left(\tilde{Z}_k - \widehat{D}_{k|k-1}\right) \tag{18}$$

$$P_{k|k} = (1 - K_k)P_{k|k-1} \tag{19}$$

where, $P_{k|k}$ represents the updated error covariance. R denotes the variance of the observation noise. K_k denotes the Kalman gain. When K_k is large, the filter relies more on the observed value; otherwise, it depends more on the prior prediction value. $\widehat{D}_{k|k}$ denotes the optimal estimate obtained by fusing the prior state and the current observation. This estimate serves as the final output after applying median filtering for noise reduction and Kalman filtering for recursive state optimization.

4. Experimental results and analysis

4.1 Experiments setup and equipment

The experimental platform is a small, motorized ship approximately 4 m long, with good maneuverability and navigation stability, as shown in Figure 8(a). A Hikvision binocular stereo camera is mounted at the front of the ship, with a baseline of 870 mm, the focal length set to 3.2 mm, and a resolution of 1920×1080 px. It is installed at a height of 1.5 m, with the pitch angle aligned with the heading direction of the ship. The data acquisition system uses a host computer running Ubuntu 20.04, equipped with an NVIDIA RTX 4060 GPU and an Intel i7-14650HX CPU. It is also equipped with a Trimble MPS 865 GNSS Heading Receiver (RTK-GNSS) to validate the localization accuracy, with a theoretical positioning error of better than ±0.01 m. The data acquisition equipment is shown in Figure 8(b).

The experiment was conducted at Xiaopingdao Wharf in Dalian, Liaoning Province. The site comprised typical port conditions, including varying lighting, complex backgrounds, and dynamic water surface disturbances. During the experiment, the image data and RTK positioning information were strictly time synchronized and spatially calibrated. The recorded video sequences captured scenes at various distances and under diverse conditions.

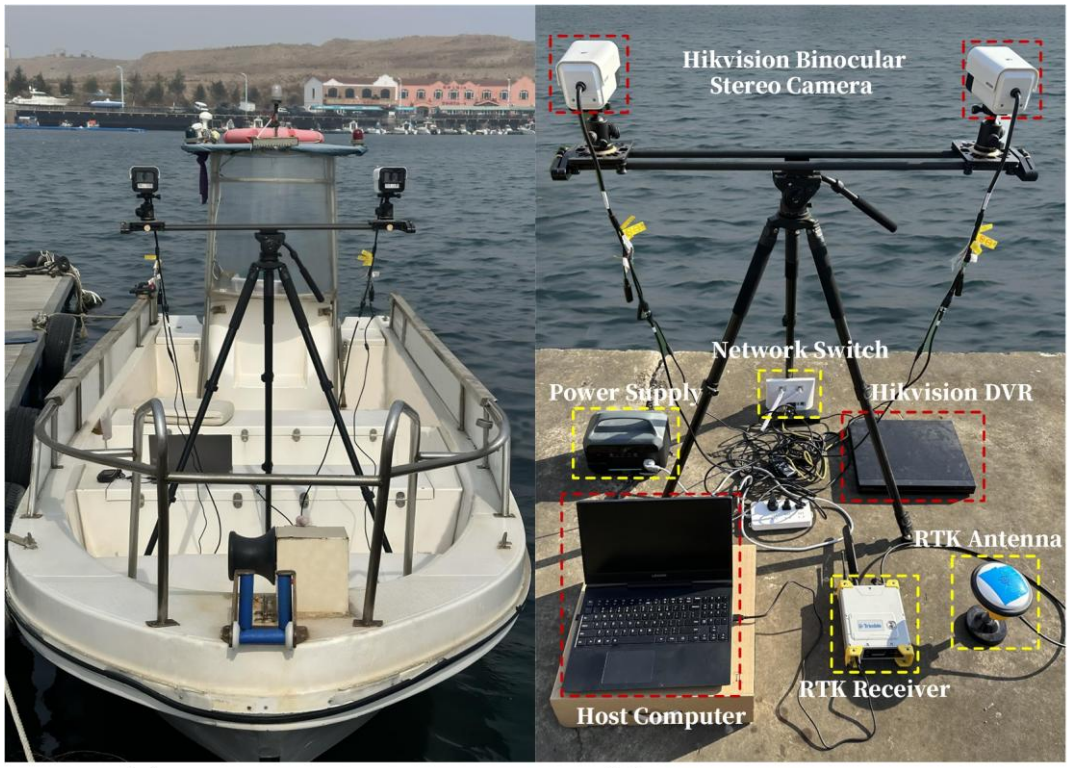


Fig. 8 Experimental platform for ship-based stereo vision

4.2 Camera calibration and image rectification

To ensure the accuracy of subsequent stereo matching and distance estimation, the stereo camera pair was calibrated, and the images were rectified before the experiments. The Zhengyou Zhang planar calibration method was adopted for stereo camera calibration [40]. A stereo camera captured 40 sets of checkerboard images from multiple viewpoints, and 30 higher quality sets were retained for calibration. The calibration target was a black and white checkerboard with an 8×6 grid, and each square measured 50 mm. The calibration of the stereo cameras was completed with the MATLAB camera calibration toolbox. The intrinsic and extrinsic parameters are listed in Tables 2 and 3. The calibration interface is shown in Figure 9.

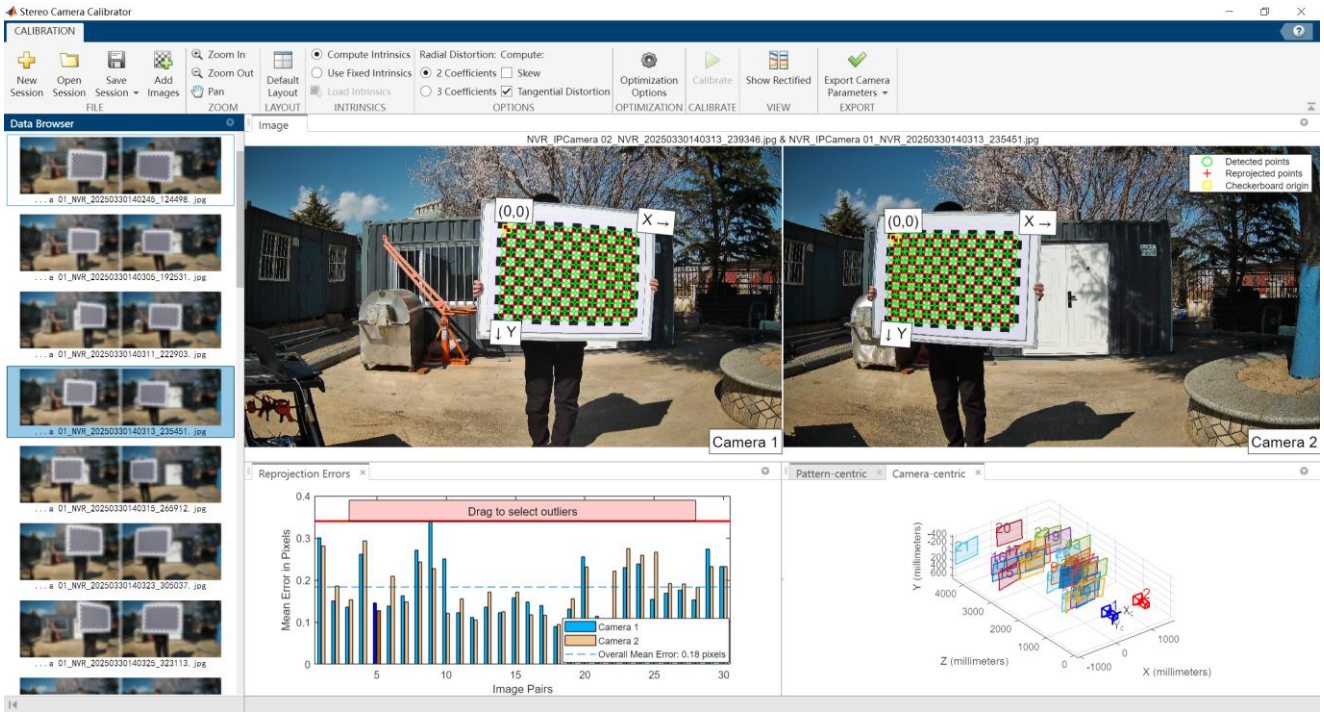


Fig. 9 Calibration interface

Table 2 Parameters of left and right cameras

Camera	Left Camera	Right Camera
Intrinsic matrix	$\begin{bmatrix} 1268.56 & 0 & 967.33 \\ 0 & 1268.1 & 509.99 \\ 0 & 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 1264.71 & 0 & 962.68 \\ 0 & 1264.23 & 513.32 \\ 0 & 0 & 1 \end{bmatrix}$
Radial distortion	$[-0.41148 \quad 0.16644]$	$[-0.41102 \quad 0.16431]$
Tangential distortion	$[0.00048 \quad -0.00127]$	$[0.00008 \quad -0.00005]$

Table 3 Relative relationship of binocular cameras

Stereo Camera	Parameter Value
Rotation matrix	$\begin{bmatrix} 0.99997 & -0.00035 & 0.00674 \\ 0.00045 & 0.99989 & -0.01479 \\ -0.00673 & 0.01479 & 0.99986 \end{bmatrix}$
Translation vector	$[873.12 \quad -3.5927 \quad -19.464]$

In practical applications, slight installation differences between the left and right cameras may cause the two image planes of the stereo system to be non-coplanar, which can significantly affect the accuracy of stereo measurement and localization. This paper applies the Bouguet algorithm [41] to rectify the left and right images, aligning the image planes and removing disparity distortion, thereby ensuring accurate disparity computation. Figures 10(a) and 10(b) show the images before and after rectification, respectively.

Fig. 10 Stereo image rectification results

4.3 Ship detection experiment

4.3.1 Dataset and parameter settings

The publicly available Mcships dataset [42] is utilized for the evaluation of the proposed GS-YOLO model. The Mcships dataset contains 6 classes of warships and 7 classes of civilian ships. The warship classes include aircraft carriers, auxiliary ships, destroyers, landing ships, missile boats, and submarines. And the civilian ship classes encompass container ships, fishing boats, passenger ships, sailboats, speedboats, tugboats, and support vessels. Compared with datasets such as Seaships, which contain only civilian ships and limited viewpoints, Mcships dataset incorporates diverse viewing angles, scale variations, complex weather conditions, lighting changes, occlusions, and cluttered backgrounds. These factors significantly increase the difficulty of both detection and fine-grained classification tasks. To ensure balanced training, 7,996 valid images were randomly divided into training, validation, and test sets with a ratio of 7:2:1. All evaluation results are reported on the test set.

To further broaden the experimental scope, the Singapore Maritime Dataset [43] is introduced as an additional dataset. This dataset consists of maritime surveillance videos captured using visible-light cameras in the Singapore maritime region. From the original videos, 6,350 images are selected for the experiments, and each image contains the corresponding bounding box annotations and category labels. The dataset includes nine maritime object categories, namely Boat, Buoy, Ferry, Flying bird-plane, Kayak, Other, Sail boat, Speed boat, and Vessel-ship. The dataset is divided into 4,445 images for training, 1,270 images for validation, and 635 images for testing.

Model training on both datasets was conducted using the PyTorch 2.0.1 deep learning framework, with CUDA 11.8 and Python 3.8. During training, all input images were resized to 640×640 px. The batch size was set to 16 to ensure stable optimization while satisfying GPU memory constraints. The model was optimized using the SGD optimizer with a momentum of 0.937, which improves convergence stability and reduces oscillations during training. The initial learning rate was set to 0.01, and a learning rate decay factor of 0.001 was applied to facilitate gradual convergence in the later stages of training.

4.3.2 Ablation experiment

To validate the superiority of the GS-YOLO model, a series of ablation experiments were conducted using YOLO11n as the baseline. The precision(P), recall(R), mean average precision (mAP), model size (M), and frames per second (FPS) were used as evaluation metrics.

Table 4 reports the results of the ablation experiments on the Mcships dataset, where "✓" and "×" indicate the specific modules enabled versus disabled in the comparison methods. As shown in Table 4, integrating GLSA, BiFPN, and SIoU modules into YOLO11 leads to varying levels of improvement in ship detection accuracy and inference speed. GS-YOLO incorporates the GLSA attention module between backbone and neck networks, enhancing feature extraction capability in dense ship regions, with improvements of 2.1 and 0.7 percentage points in precision and recall, respectively. BiFPN feature fusion network deployed in the neck section facilitates effective integration of deep and shallow features, enabling more effective multi-scale representation. Consequently, compared with the baseline, the model achieves a 1.6 percentage point increase in precision and a 1.4 percentage point increase in mAP@0.5, while reducing parameters by 25.6 %. Additionally, integrating SIoU into YOLO11 introduces almost no change in the number of model parameters, while improving precision and mAP@0.5:0.95 by 1.3 and 0.8 percentage points, respectively. The inclusion of the angle component in SIoU helps improve the optimization of bounding box regression. Finally, compared to the baseline, GS-YOLO achieves improvements of 1.6, 2.0, 1.4, and 1.6 percentage points in precision, recall, mAP@0.5, and mAP@0.5:0.95, respectively, while reducing the model parameters by 17.8 %. These results demonstrate that the proposed network achieves better detection performance while preserving favorable computational efficiency.

Table 4 Comparison of experimental results between GS-YOLO and different ablation models on the Mcships dataset

Model	GLSA	BiFPN	SIoU	Precision	Recall	mAP@0.5	mAP@0.5:0.95	Parameters(M)	FPS
YOLO11n	×	×	×	0.905	0.853	0.921	0.664	2.58	122.8
1	✓	×	×	**0.926**	0.860	0.928	0.674	3.73	86.4
2	×	✓	×	0.921	0.858	0.925	0.670	**1.92**	120.1
3	×	×	✓	0.918	0.862	0.924	0.672	2.62	121.9
4	✓	✓	×	0.915	0.870	0.932	0.674	2.07	102.3
5	✓	×	✓	0.916	0.854	0.927	0.676	3.64	90.6
6	×	✓	✓	0.918	0.866	0.924	0.672	1.95	**132.2**
GS-YOLO (Ours)	✓	✓	✓	0.921	**0.873**	**0.935**	**0.680**	2.12	118.3

As shown in Figure 11, to better illustrate the performance of GS-YOLO in ship detection tasks, four evaluation metric curves during training are visualized. The compared models include the baseline, models with only GLSA, BiFPN, or SIoU, and improved versions with different module combinations. GS-YOLO is indicated by the red curve. The figure shows that the proposed model converges faster and achieves higher final accuracy than the other models throughout the training process. In particular, the red curve shows a clear advantage during the early epochs for both mAP@0.5 and mAP@0.5:0.95, indicating that the designed feature fusion and attention mechanisms enable the model to learn target features more quickly and accurately.

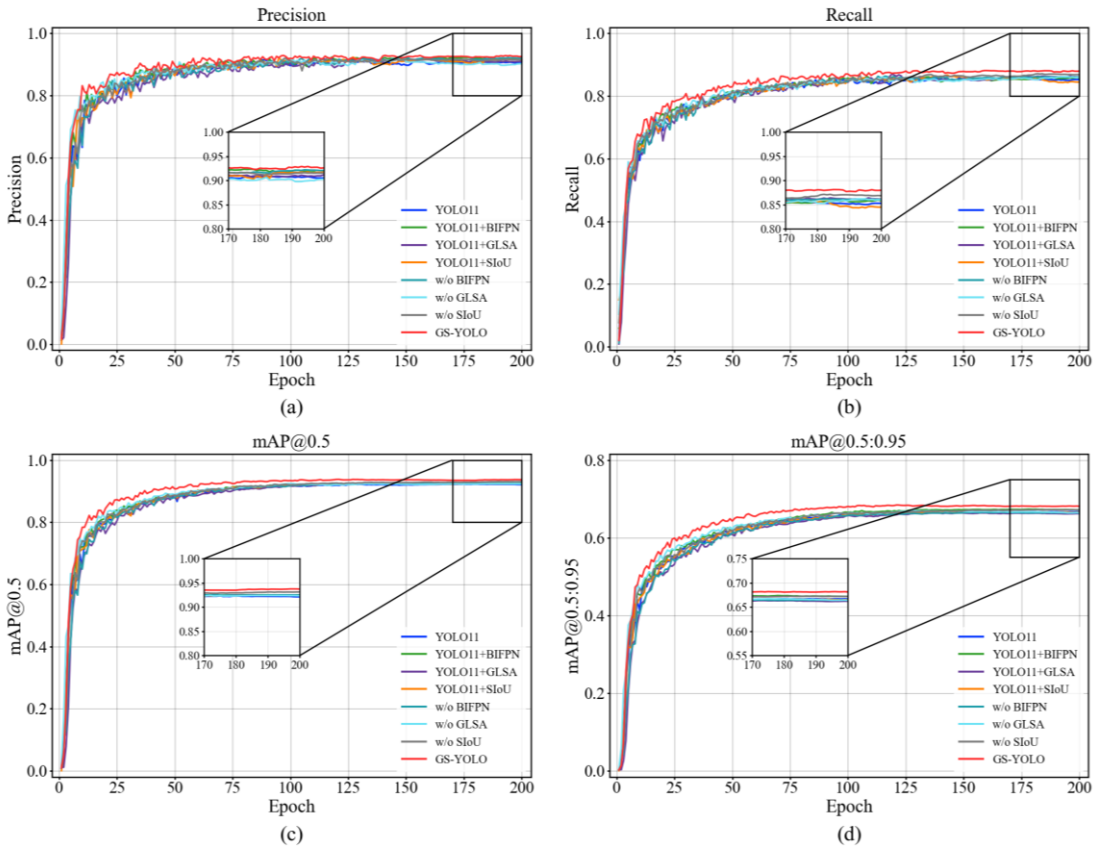


Fig. 11 Training process of different ablation models on Mcships dataset: (a) curve of precision, (b) curve of recall, (c) curve of mAP@0.5, (d) curve of mAP@0.5:0.95

As shown in Table 5, GS-YOLO also demonstrates superior performance across the evaluation metrics of the training results on the Singapore Maritime Dataset. After integrating the GLSA module, although mAP@0.5 shows a slight decrease, the precision increases by 2.2 percentage points compared with the baseline model. With the introduction of the SIoU loss function and the BiFPN module, the model ultimately outperforms the baseline across all accuracy-related metrics, confirming the effectiveness of each proposed improvement. On the Singapore Maritime Dataset, GS-YOLO improves precision, recall, mAP@0.5, and mAP@0.5:0.95 by 2.7, 3.7, 2.5, and 5.0 percentage points, respectively, compared with the baseline model. The experimental results on both datasets demonstrate that GS-YOLO reduces model parameters while maintaining high detection accuracy and competitive inference speed, making the proposed model suitable for deployment on edge computing devices in intelligent vessels.

Table 5 Comparison of experimental results between GS-YOLO and different ablation models on the Singapore Maritime Dataset

Model	GLSA	BiFPN	SIoU	Precision	Recall	mAP@0.5	mAP@0.5:0.95	Parameters(M)	FPS
YOLO11n	×	×	×	0.906	0.845	0.906	0.633	2.58	122.8
1	✓	×	×	0.928	0.847	0.900	0.641	3.73	86.4
2	×	✓	×	0.914	0.868	0.924	0.648	**1.92**	120.1
3	×	×	✓	0.912	0.852	0.908	0.662	2.62	121.9
4	✓	✓	×	0.930	0.880	0.930	0.647	2.07	102.3
5	✓	×	✓	0.919	0.851	0.925	0.676	3.64	90.6
6	×	✓	✓	0.905	0.874	0.920	0.662	1.95	**132.2**
GS-YOLO (Ours)	✓	✓	✓	**0.933**	**0.882**	**0.931**	**0.683**	2.12	118.3

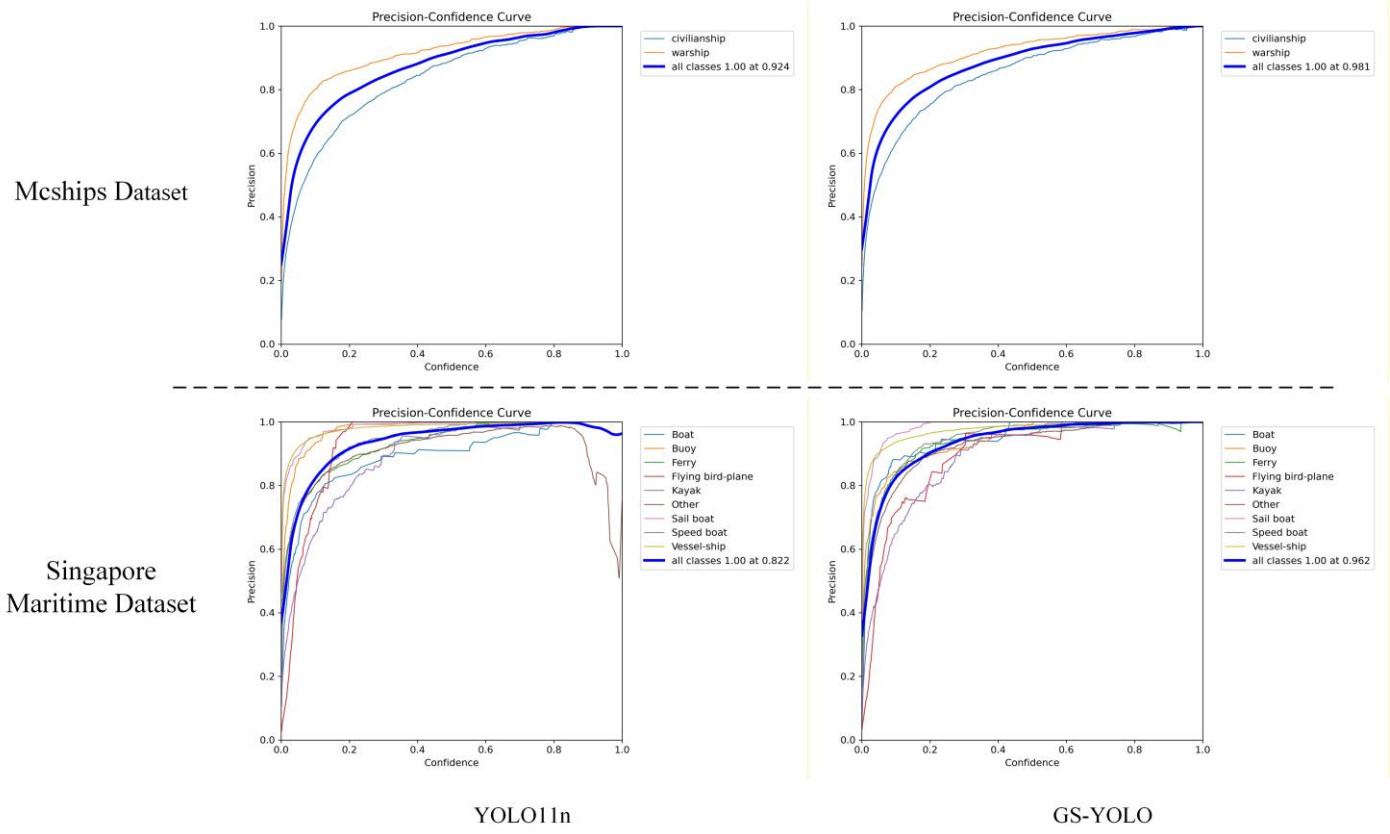


Fig. 12 Comparison of Precision-Confidence curves

Figure 12 presents the precision–confidence curves of the baseline model and the GS-YOLO model on both the Mcships dataset and the Singapore Maritime Dataset. Compared with the baseline model, GS-YOLO achieves higher precision for most categories on both datasets, and the overall curve of all classes is also consistently improved. The GS-YOLO curves are smoother and maintain more stable precision across a wide range of confidence thresholds, indicating stronger robustness and improved detection reliability. In particular, the performance gain on the Singapore Maritime Dataset is more pronounced, demonstrating that the proposed model provides more reliable detection performance across different ship categories and maritime scenes.

Figure 13 presents a visual comparison of the detection results for the baseline model, several ablation models, and the GS-YOLO model across a variety of representative ship scenarios. In this figure, red solid boxes denote false negatives (FN), while red dashed boxes denote false positives (FP). In Figures 13(a) and 13(b), YOLO11n exhibits noticeable missed detections under conditions of long distance and complex backgrounds, as indicated by the FN annotations. The introduction of the GLSA and BiFPN modules improves the model capability for multi-scale feature extraction, resulting in more accurate detection. However, when ships are partially occluded, duplicate bounding boxes may occur, as shown by the FP annotations in Figure 13(d). With the integration of SIoU, GS-YOLO can detect and accurately recognize target ships with high confidence across various complex scenarios, while reducing both false negatives and false positives.

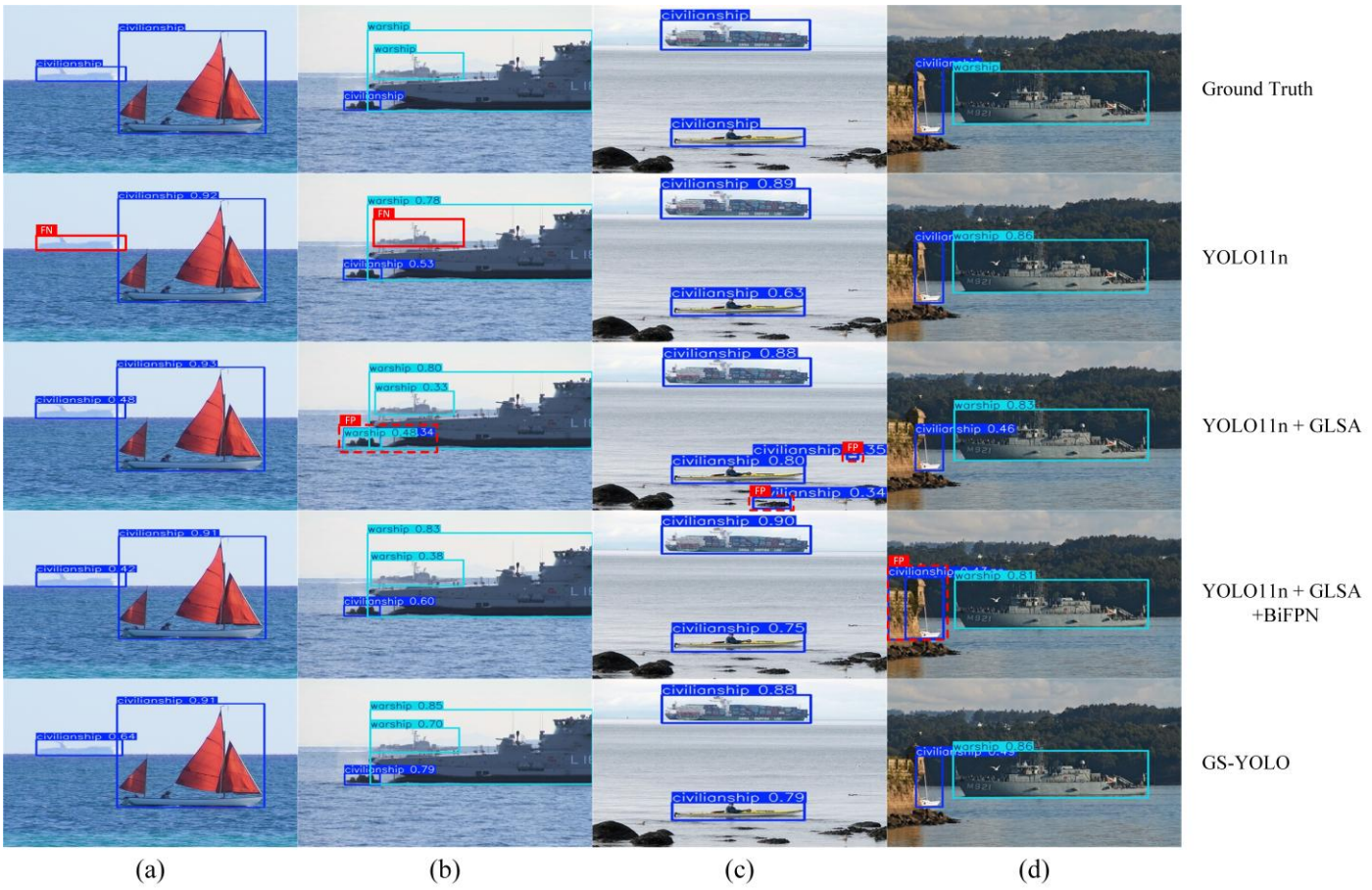


Fig. 13 Comparison of visual detection results from different ablation models on the Mcships test set

To further investigate the impact of the GLSA attention module at different feature scales, fine-grained ablation experiments were conducted on the Mcships dataset. Specifically, the GLSA module was introduced separately at different levels of the feature pyramid P_3, P_4 and P_5. The results were then compared with the configuration where GLSA was applied to all feature scales simultaneously.

As shown in Table 6, when the GLSA module is applied to a single feature scale, the model achieves improvements in precision, recall, and mAP, but the overall gain remains relatively limited. This result indicates that introducing an attention mechanism at a single scale enhances local feature representation, but its effectiveness is constrained by the limited feature hierarchy. Although applying GLSA at all scales introduces additional computational cost, the model achieves an mAP@0.5 of 0.928 and a precision of 0.926, indicating a significant overall performance improvement. These results demonstrate that incorporating the GLSA attention mechanism across multi-scale feature maps enhances feature representation at different scales and improves detection performance in complex maritime environments.

Table 6 Fine-grained ablation study of GLSA insertion positions on Mcships dataset

Model	Precision	Recall	mAP@0.5	mAP@0.5:0.95	Parameters(M)
YOLO11n	0.905	0.853	0.921	0.664	**2.58**
YOLO11n+GLSA+P_3	0.912	0.860	0.925	0.668	3.11
YOLO11n+GLSA+P_4	0.909	**0.866**	0.926	0.667	3.12
YOLO11n+GLSA+P_5	0.915	0.857	0.922	0.671	3.11
YOLO11n+GLSA+All	**0.926**	0.860	**0.928**	**0.674**	3.73

In large-scale maritime environments, small ships at long distances often occupy only a few pixels, causing critical information to be easily lost during repeated down sampling. To address this issue, a high-

resolution shallow feature branch P_2^{in} was introduced in the BiFPN architecture to enhance the model's sensitivity to small-scale targets. Table 7 presents the ablation experiments conducted on the Mcships dataset, comparing the performance of the standard BiFPN architecture with BiFPN incorporating P_2^{in}, in order to evaluate the effectiveness of this design. The experimental results show that without P_2^{in}, the model achieves a recall of 0.851 and an mAP@0.5 of 0.920. After introducing P_2^{in}, recall increases to 0.858 and mAP@0.5 to 0.925. These results demonstrate that the P_2^{in} structure effectively enhances the utilization of high-resolution features, thereby improving detection performance for small targets at long distance.

Table 7 Ablation study of the P_2^{in} component in the BiFPN module on Mcships dataset

Model	Precision	Recall	mAP@0.5	mAP@0.5:0.95	Parameters(M)
YOLO11n	0.905	0.853	0.921	0.664	2.58
YOLO11n+BiFPN(without P_2^{in})	0.912	0.851	0.920	**0.670**	1.86
YOLO11n+BiFPN(with P_2^{in})	**0.921**	**0.858**	**0.925**	**0.670**	**1.92**

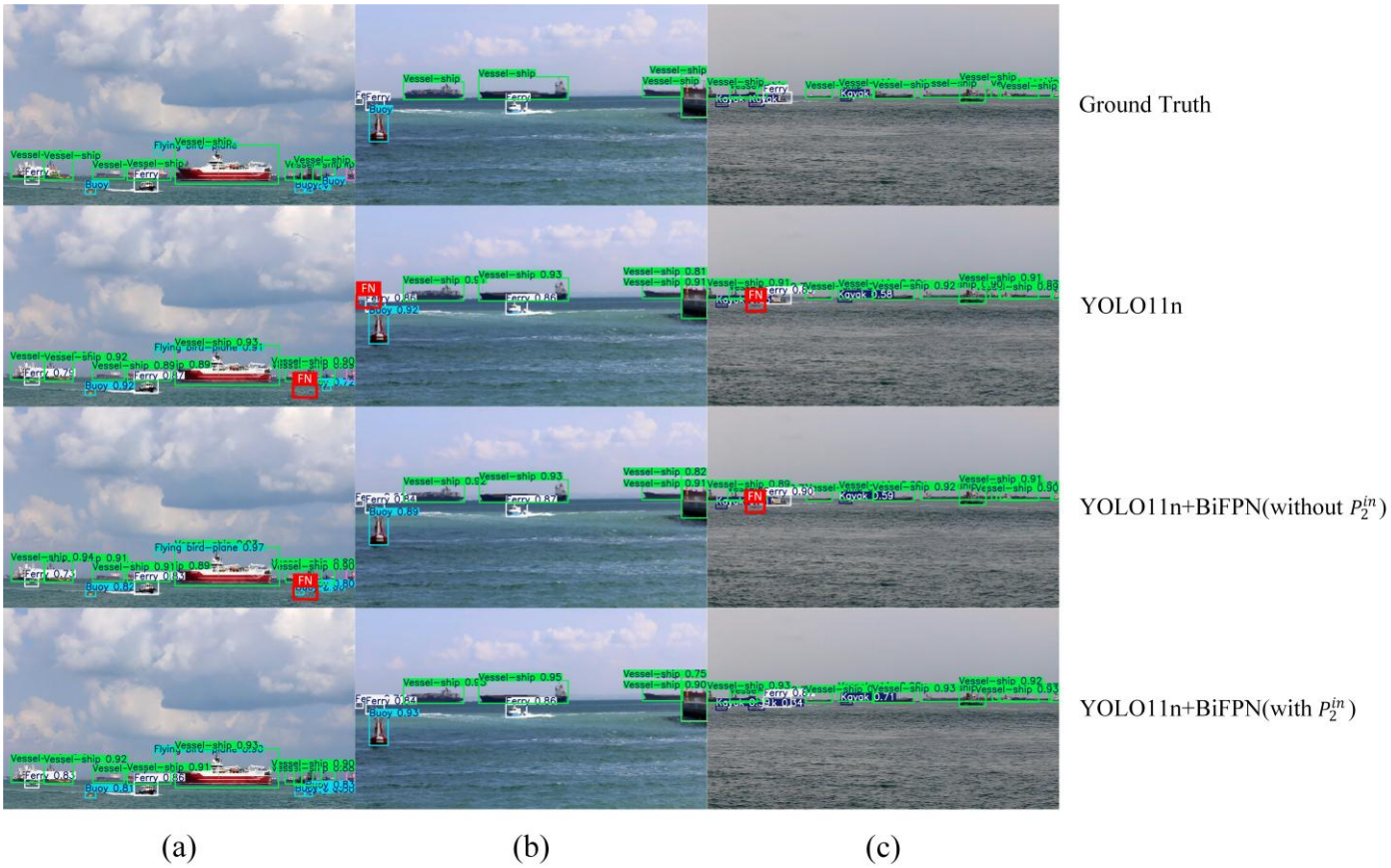


Fig. 14 Comparison of visual detection results of the P_2^{in} branch in BiFPN on the Singapore Maritime Dataset

As shown in Figure 14, the ablation experiment is visualized using the Singapore Maritime Dataset. This dataset contains a large number of small targets and is well suited for evaluating the small object detection capability of the model. Figure 14(a) illustrates a multi-category scene containing large ships, small vessels, and buoys. Compared with the baseline model, the introduction of the BiFPN structure enables the detection of some tiny buoys, but some missed detections remain, as marked by the FN annotations. After further incorporating the P_2^{in} branch, the detection of small targets is further improved. As shown in Figures 14(b) and 14(c), the introduction of P_2^{in} improves the detection performance for ships near image boundaries and small vessels such as kayaks, demonstrating the effectiveness of this branch for small target detection.

4.3.3 Comparative experiment

We compare GS-YOLO with several object detection methods on the Mcships dataset, including YOLO based single stage detection methods such as YOLOv5, YOLOv8, and YOLOv10, as well as Transformer based end to end detection methods RT-DETR [44] and DINO [45]. In addition, the DAMO-YOLO proposed by Xu et al. [46] was also included for comparison. To ensure a fair comparison, all YOLO based methods were implemented using the YOLO Nano versions, while the Transformer based methods adopted their lightweight configurations, and identical training hyperparameters were used across all models. To provide a more comprehensive evaluation, the metrics AP^s, AP^m, and AP^l were introduced to measure the detection performance for small, medium, and large objects, corresponding to target areas smaller than 32×32 px, between 32×32 and 96×96 px, and larger than 96×96 px, respectively. Table 8 presents the experimental results of different detection models on the Mcships dataset.

Based on the quantitative analysis in Table 8, GS-YOLO achieves an mAP@0.5 of 0.935 on the Mcships dataset. Compared with DAMO-YOLO, GS-YOLO achieves improvements of 1.7, 3.9, and 2.5 percentage points in precision, recall, and mAP@0.5. Compared to YOLO series methods (YOLOv5, YOLOv8, and YOLOv10), GS-YOLO demonstrates the highest precision and recall, with mAP@0.5 improvements of 4.2, 2.1, and 1.7 percentage points, respectively, while maintaining the lowest number of parameters among all models.

GS-YOLO achieves the best AP^s value of 0.301 for small object detection, improving by 3.0 and 6.3 percentage points compared with DAMO-YOLO and RT-DETR, respectively. This result indicates that the proposed feature fusion and attention mechanisms can more effectively preserve high resolution feature information, thereby significantly improving detection performance for small ship targets at long distance. Meanwhile, GS-YOLO also maintains strong performance on the AP^m and AP^l metrics, indicating good stability in detecting objects across different scales. Compared with DINO, although it shows a slight advantage in precision, it requires significantly more model parameters and exhibits a noticeable decrease in inference speed. In contrast, GS-YOLO maintains high detection accuracy while achieving a smaller model size and faster inference speed, demonstrating superior real time performance and deployment efficiency.

Table 8 Comparison of experimental results between GS-YOLO and other detection methods on the Mcships dataset

Model	Precision	Recall	mAP@0.5	mAP@0.5:0.95	AP^s	AP^m	AP^l	Parameters(M)	FPS
YOLOv5	0.871	0.831	0.893	0.581	0.209	0.432	0.716	2.31	129.4
YOLOv8	0.900	0.857	0.914	0.651	0.283	0.475	0.728	3.16	132.6
YOLOv10	0.911	0.839	0.918	0.665	0.239	0.466	0.698	2.71	110.4
DAMO-YOLO	0.904	0.834	0.910	0.658	0.271	0.421	0.703	3.35	**136.2**
RT-DETR	0.910	0.839	0.889	0.638	0.238	0.476	**0.735**	20.4	52.6
DINO	**0.924**	0.866	0.930	0.676	0.279	0.452	0.741	47.1	28.4
GS-YOLO (Ours)	0.921	**0.873**	**0.935**	**0.680**	**0.301**	**0.483**	0.732	**2.12**	118.3

To evaluate the generalization capability of the proposed method in real maritime environments, images collected from actual port waters were used to perform a visual comparison among different object detection methods. Figure 15 presents the detection results of different methods on representative real-world images. Figure 15(a) illustrates a nearshore scene with a complex background. YOLOv5 and YOLOv8 are more sensitive to background interference, resulting in both missed detections and false detections, as indicated by the FN and FP annotations, whereas GS-YOLO accurately identifies targets under the same complex conditions. Figure 15(b) illustrates a long-range scene with partial occlusion. YOLOv5 and RT-DETR exhibit missed detections for small distant objects, while DINO shows relatively limited stability in detecting occluded targets. In contrast, GS-YOLO maintains reliable detection of small and partially occluded targets, demonstrating strong robustness in both long-distance and occluded scenarios. Figure 15(c) illustrates a

medium-scale ship scene. DAMO-YOLO suffers from duplicate detections and overlapping boundaries, as indicated by the FP annotations, which reduce classification accuracy. GS-YOLO performs better in distinguishing targets and locating boundaries. In Figure 15(d), the target ship exhibits low contrast against the sea background and appears near the image border. Most methods either miss the detection or exhibit bounding box misalignment, whereas GS-YOLO still achieves accurate bounding box regression, highlighting its robustness in edge and low-contrast conditions.

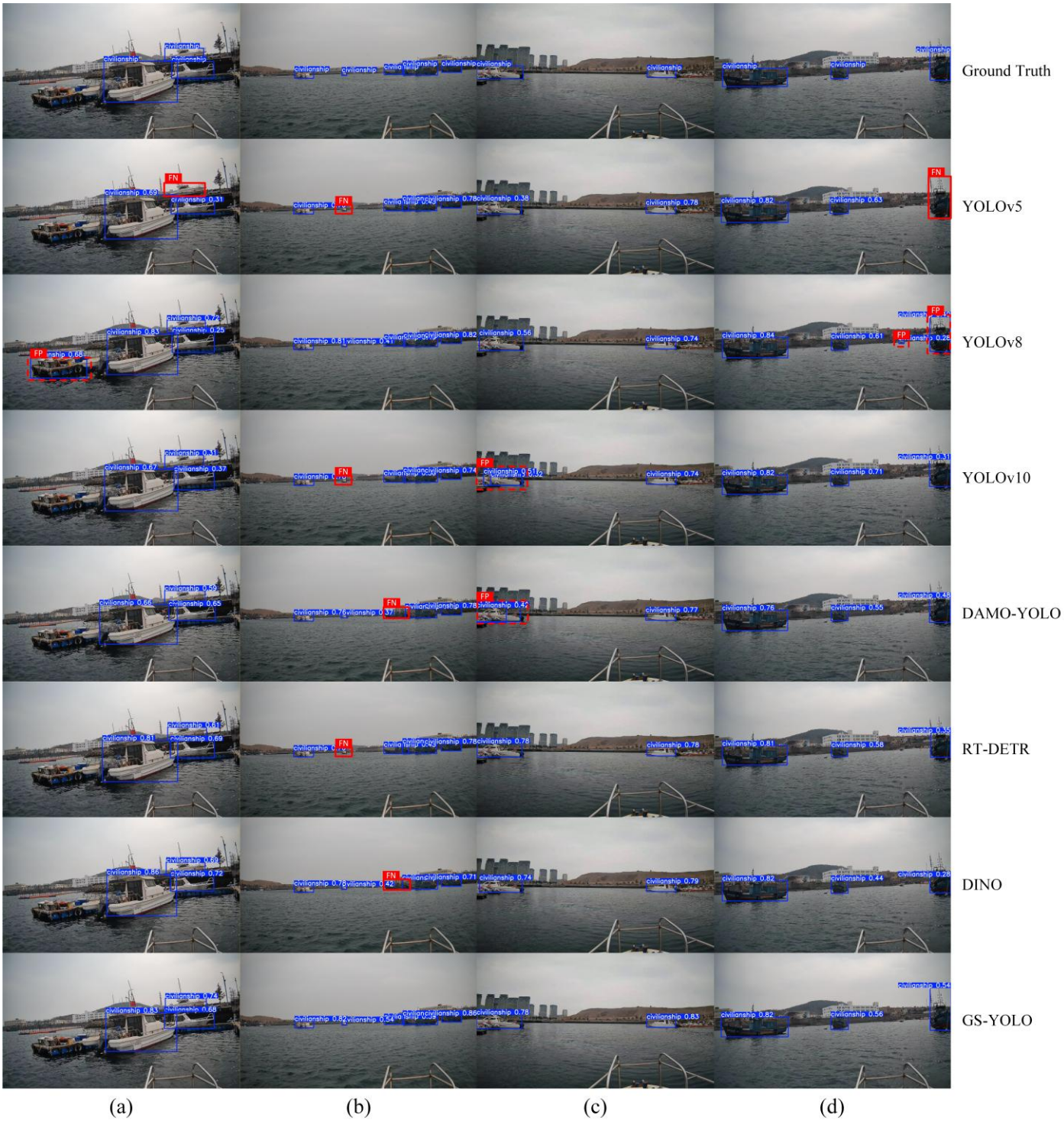


Fig. 15 Comparison of visual detection results from different methods on the real ship test scenes

4.4 Ship localization experiment

To evaluate the stability and robustness of the localization algorithm, ship localization experiments were conducted in a port environment. A video segment capturing the test ship approaching the target ship from approximately 40 to 80 m was selected. The target ship remained stationary throughout the video. During this sequence, the USV trajectory included moderate steering maneuvers rather than a strictly straight approach, which resulted in gradual changes in the relative viewing angle between the camera and the target vessel. Such conditions introduce additional variations in visual observations and better reflect practical navigation situations in port environments. This experiment validates the localization capability of the proposed approach in real-world scenarios. Since the ground truth 3D coordinates of the target point in the left camera coordinate system were unavailable, the Euclidean distance between the left camera and the target point was used as the localization estimate. In the experiment, the Euclidean distance between the test ship and the target ship measured by RTK served as the reference value. The localization accuracy was evaluated by comparing the stereo vision distance estimates with the RTK measurements.

Prior to the localization experiment, the parameter settings of the cascaded filtering localization algorithm are summarized in Table 9. These parameters were determined based on physical motion limits and empirical statistical analysis. The median filter window size was set to 7 frames, corresponding to a temporal window of approximately 0.28 s given the camera frame rate of 25 fps. This window length effectively suppresses transient outliers caused by water-surface reflections while maintaining sufficient responsiveness to actual changes in target distance. For the Kalman filter, the process noise variance Q is set to 0.05. This value reflects the expected variation of the target distance between consecutive frames due to ship motion, ensuring that the filter remains responsive to genuine motion changes while avoiding excessive smoothing. The observation noise variance R is set to 4.0, which is estimated from the variance of static stereo distance measurement errors at a distance of 80 m. Finally, the initial error covariance P_0 is set to 1.0 to ensure rapid convergence during initial tracking.

Table 9 Parameters of the cascaded filtering localization algorithm

Parameter	Value
w	7 frames
Q	0.05 m^2
R	4.0 m^2
P_0	1.0 m^2

To analyze the localization accuracy of the algorithm, this paper employs the Standard Deviation of the Error (*SDE*) and Mean Relative Error (*MRE*) as evaluation metrics. These metrics quantitatively assess the stability and accuracy of the distance estimation. The *SDE* is calculated as the standard deviation of the differences between the predicted and true values, as shown in Equation (20). It quantifies the variability of the estimation errors, where smaller values indicate more stable predictions. The *MRE* indicates the relative deviation between the predicted and true values, with smaller values corresponding to higher accuracy. Its calculation is shown in Equation (21).

$$SDE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\left(D_i^{pred} - D_i^{true}\right) - \frac{1}{N}\sum_{j=1}^{N}\left(D_j^{pred} - D_j^{true}\right)\right)^2} \tag{20}$$

$$MRE = \frac{1}{N}\sum_{j=1}^{N}\left|\frac{D_i^{pred} - D_i^{true}}{D_i^{true}}\right| \tag{21}$$

where, D^{true} denotes the ground truth distance measured by RTK, D^{pred} represents the distance estimated by the localization algorithm.

4.4.1 Comparison of stereo matching algorithms

As discussed in the stereo matching analysis, RAFT-Stereo was selected as the core spatial perception module because of its superior disparity quality and inference speed. To quantitatively validate this choice in terms of localization accuracy, comparative distance estimation experiments were conducted using several stereo matching methods. Specifically, the traditional SGBM algorithm and representative deep learning methods, including IGEV-Stereo and RAFT-Stereo, were evaluated.

To evaluate the distance estimation accuracy of the stereo matching algorithms, a discrete sampling approach was adopted. Representative keyframes capturing the target ship were uniformly extracted across the distance range from 40 to 80 m and divided into four distance intervals of 10 m. These keyframes were specifically selected to minimize the influence of severe motion blur and sudden camera shake present in the continuous video, thereby isolating the baseline spatial matching performance of each algorithm. The target distance was computed directly from the disparity maps generated by each method on these discrete frames. The MRE was used to quantitatively evaluate their accuracy.

Table 10 presents the quantitative comparison of distance estimation accuracy among the different stereo matching algorithms. The traditional SGBM algorithm exhibits the largest estimation errors, primarily because local matching costs struggle with strong reflections and texture less regions on the water surface. Although IGEV-Stereo improves its accuracy, RAFT-Stereo consistently achieves the lowest MRE across all distance intervals. This numerical evaluation confirms that RAFT-Stereo provides the most stable and accurate raw spatial perception baseline, laying a reliable foundation for the subsequent filtering module.

Table 10 MRE of distance estimates among different stereo matching algorithms

Method	40-50 m	50-60 m	60-70 m	70-80 m
SGBM	3.32 %	4.68 %	5.13 %	7.72 %
IGEV-Stereo	1.92 %	2.91 %	3.85 %	5.41 %
RAFT-Stereo	1.32 %	1.98 %	3.37 %	4.53 %

In addition, to quantitatively evaluate the efficiency gain brought by the proposed ROI constraint, the runtime performance of the complete localization system was analyzed. Table 11 presents the end-to-end throughput for different stereo matching algorithms under full image processing and ROI constrained processing. The throughput measurement includes target detection, disparity estimation, and temporal filtering. With the ROI constraint, disparity estimation is performed only within the target region defined by the detected bounding box. This reduces redundant computation in background regions and shortens the runtime of the stereo matching stage. As a result, the end-to-end throughput of the RAFT-Stereo based localization system increases from 3.1 to 13.1 fps. These results demonstrate that the proposed ROI-constrained strategy significantly improves runtime efficiency in the localization stage.

Table 11 Processing throughput of different stereo matching algorithms with and without ROI constraint

Method	Full Image Throughput (FPS)	ROI Constrained Throughput (FPS)
SGBM	0.8	3.3
IGEV-Stereo	0.5	2.6
RAFT-Stereo	3.1	13.1

4.4.2 Verification of localization accuracy

Throughout the experiment, the system performed target detection and localization on a frame-by-frame basis. The dense disparity map was obtained using the RAFT-Stereo algorithm, and the distance was estimated by averaging disparity values within the central region of the detected bounding box. To qualitatively demonstrate the continuous tracking capability of the complete system, Figure 16 shows the detection and localization results at different frames in the video, corresponding to varying distances between the test ship and the stationary target.

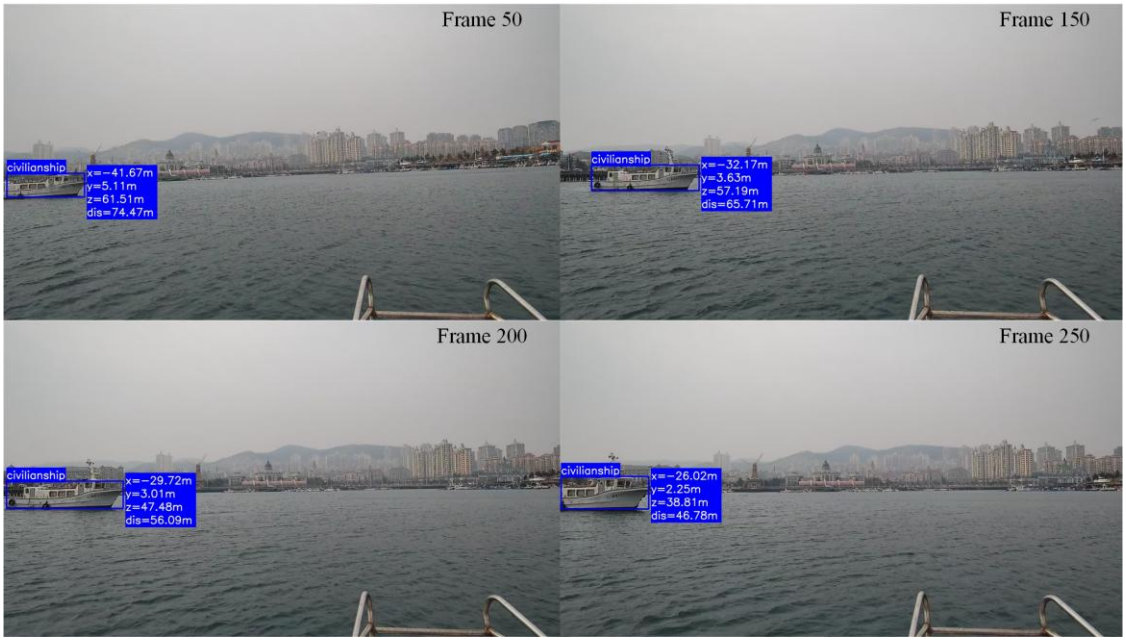


Fig. 16 Visualization of ship detection and localization at different distances

Figure 17 shows a comparison of the distance estimation results among the raw measurements, conventional filtering methods, and the proposed cascaded filtering algorithm, together with the RTK reference values. The green curve represents the ground truth distance measured by RTK. It can be observed that the raw measurements exhibit significant fluctuations throughout the entire process, especially between frames 0-75, where vessel motion and wave disturbances introduce noticeable fluctuations in the estimated distance. Although the moving average (MA) mitigates some high-frequency noise, it still suffers from local fluctuations and phase delays. The standard Kalman filter (KF), while producing a smooth trajectory, is strongly affected by extreme outliers without a preceding rejection mechanism, resulting in a prolonged deviation from the reference trajectory. In contrast, the results obtained using the proposed cascaded filtering algorithm are smoother and more stable, effectively reducing errors caused by platform motion while maintaining responsiveness to actual distance changes.

To reduce the influence of random measurement errors, this paper statistically analyzes the positioning errors across different distance ranges. As shown in Figure 18(a), the proposed cascaded filtering localization algorithm significantly reduces error fluctuations across all distance ranges compared with the other comparison methods. Table 12 presents the *SDE* values for different distance ranges. In particular, within the 40-50 m range, the *SDE* of the localization results was reduced to only 0.12 m, representing an 89.3 % reduction compared to the raw measurement. Although the standard KF also suppresses fluctuations to some extent, the cascaded filtering method consistently achieves lower *SDE* values across all distance ranges, indicating better robustness and stability.

Figure 18(b) illustrates the trend of *MRE* with increasing distance. Although the errors of all methods increase as the distance grows, the error growth of the cascaded filtering method is more gradual and consistently remains lower than that of the other comparison methods. This indicates that the cascaded filtering localization algorithm maintains superior accuracy across the entire distance range.

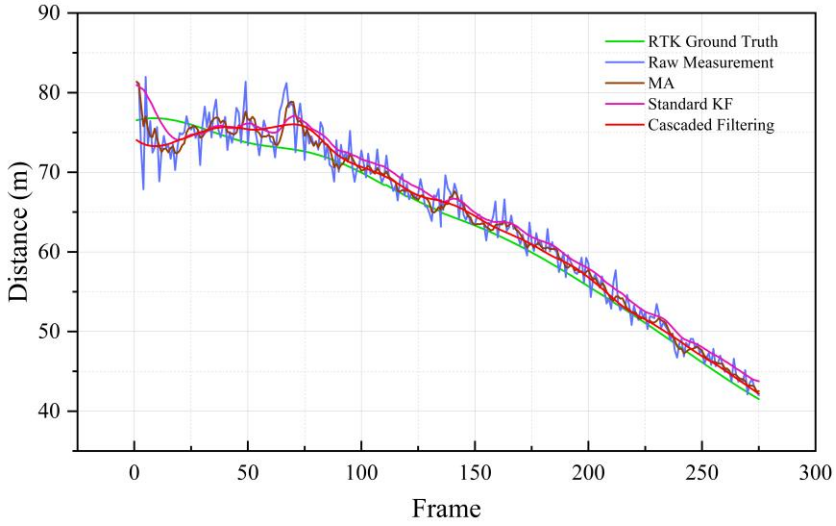


Fig. 17 Comparison of distance estimation with and without cascaded filtering

Table 13 presents a comparison of the *MRE* across different distance ranges. In the four ranges of 40-50, 50-60, 60-70, and 70-80 m, the proposed cascaded filtering localization algorithm reduces the MRE by 0.71, 0.82, 1.46, and 2.38 %, respectively, compared with the raw measurement. Compared with the conventional MA and standard KF filters, the proposed method also exhibits consistent accuracy improvements. In the closer range of 40-60 m, the *MRE* of the cascaded filtering localization algorithm remains below 1.3 %. In the more distant range of 70-80 m, although vessel motion and longer distance increase the difficulty of stereo matching, the algorithm still maintains an *MRE* of 2.30 %, demonstrating its high localization accuracy in complex environments.

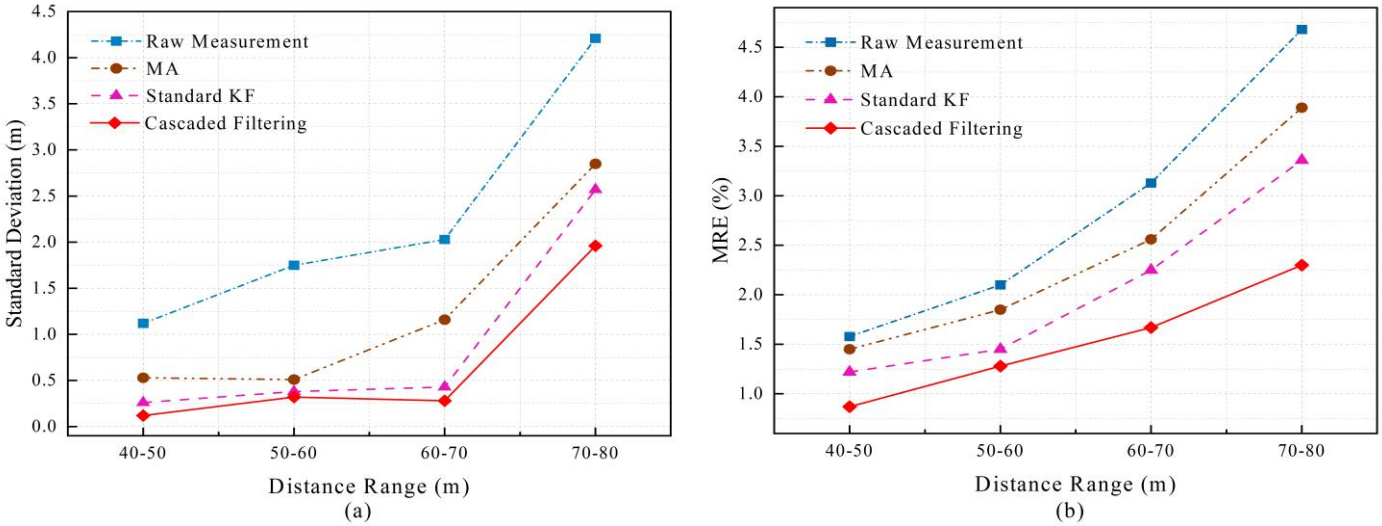


Fig. 18 Comparison of *SDE* and *MRE* in distance estimates with and without cascaded filtering

Table 12 *SDE* of distance estimates among different localization filtering systems

Distance Range (m)	Raw Measurement (m)	MA (m)	Standard KF (m)	Cascaded Filtering (m)
40-50	1.12	0.53	0.26	0.12
50-60	1.75	0.51	0.38	0.32
60-70	2.03	1.16	0.43	0.28
70-80	4.21	2.85	2.57	1.96

Table 13 *MRE* of distance estimates among different localization filtering systems

Distance Range (m)	Raw Measurement (%)	MA (%)	Standard KF (%)	Cascaded Filtering (%)
40-50	1.58	1.45	1.22	0.87
50-60	2.10	1.85	1.45	1.28
60-70	3.13	2.56	2.25	1.67
70-80	4.68	3.89	3.36	2.30

4.4.3 Verification of dynamic performance

To evaluate the performance of the cascaded filtering localization algorithm in dynamic scenarios, the relative velocity is used as a metric for assessing dynamic performance. In autonomous USV navigation, stable and accurate estimation of the target ship's velocity is important for downstream navigation and obstacle avoidance tasks. Unstable distance measurements can cause significant fluctuations and distortions in velocity estimation, thereby affecting downstream decision-making and the reliability of the control system. Therefore, the instantaneous radial relative velocity of the target with respect to the test ship is used to evaluate dynamic performance.

Figure 19 compares the velocity estimates derived from the raw measurements and the cascaded filtering method against the RTK ground truth. The velocity curve obtained from the raw distance measurements exhibits severe fluctuations and numerous outliers, with peak values exceeding 300 m/s. These values clearly fail to reflect the true motion state of the target ship. In contrast, the velocity curve obtained using cascaded filtering is considerably smoother, and measurement noise as well as abnormal fluctuations are effectively suppressed. As shown more clearly in the inset, the filtered velocity curve follows the overall trend of the RTK reference more closely, providing a more realistic representation of the relative motion between the two ships.

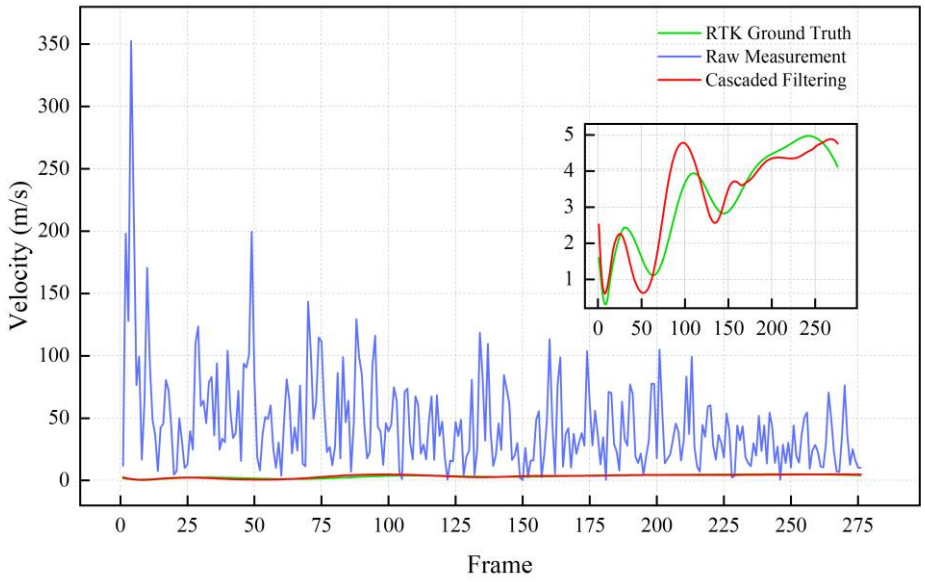


Fig. 19 Comparison of velocity estimates derived from raw measurements and cascaded filtering

The experimental results demonstrate that the proposed cascaded filtering localization algorithm provides clear improvements in dynamic scenarios. By applying median filtering and Kalman filtering to the raw measurement data, the algorithm transforms noisy and highly fluctuating measurements into smoother and more stable state estimates. This significantly improves the continuity and stability of the localization results, providing more reliable perception data for downstream USV functions such as trajectory prediction and dynamic obstacle avoidance.

5. Conclusion

To address low ship detection accuracy and poor localization stability for USVs in complex scenarios, a robust stereo-vision system is proposed. The system comprises a ship detection module and a localization module. In the detection module, the proposed GS-YOLO model integrates the GLSA attention module and a lightweight BiFPN feature fusion structure. This innovative design not only enhances detection capability but also reduces the number of model parameters. Furthermore, the SIoU loss function is introduced to improve the accuracy of bounding box regression. For the localization module, cascaded median and Kalman filters perform temporal smoothing on depth sequences obtained by the RAFT-Stereo algorithm, enhancing positioning stability in maritime environments.

Experimental results demonstrate that the proposed GS-YOLO achieves an inference speed of 118.3 FPS on the Mcships dataset, with 0.921 precision, 0.873 recall, and 0.935 mAP@0.5, while reducing parameters by 17.8 % compared to YOLO11n. This confirms its capability to provide a reliable foundation for localization. The cascaded filtering localization algorithm achieves a positioning error standard deviation of 0.12 m at a range of 50 m and maintains positioning errors within 2.3 % within 80 m. These results confirm that the proposed stereo-vision system provides accurate, stable, and efficient perception data for autonomous USV navigation.

Nevertheless, the system has certain limitations. Due to constraints in the experimental environment and data collection conditions, the impact of weather conditions on algorithm performance has not been evaluated. In addition, reliance solely on visible-light data without infrared integration limits its applicability in low-visibility or nighttime environments. These limitations will be the focus of future research.

ACKNOWLEDGMENTS

The authors would like to thank the editors and anonymous reviewers for their valuable comments and suggestions to improve the presentation of this paper. This work was supported by the National Natural Science Foundation of China [Grant No.52231014] and Team Project of Dalian Maritime University [Grant No.3132023511].

REFERENCES

[1] Ljulj, A., Slapničar, V., Brigić, J. 2022. Unmanned surface vehicle–tritor. *Brodogradnja*, 73(3), 135-150. https://doi.org/10.21278/brod73308

[2] Liu, Z., Zhang, Y., Yu, X., Yuan, C. 2016. Unmanned Surface Vehicles: An Overview of Developments and Challenges. *Annual Reviews in Control*, 41, 71-93. https://doi.org/10.1016/j.arcontrol.2016.04.018

[3] Kim, J. 2020. Target Following and Close Monitoring Using an Unmanned Surface Vehicle. *IEEE Transactions on Systems Man Cybernetics-Systems*, 50(11), 4233-4242. https://doi.org/10.1109/tsmc.2018.2846602

[4] Liu, C., Xiang, X., Huang, J., Yang, S., Zhang, S., Su, X., Zhang, Y. 2022. Development of USV autonomy: architecture, implementation and sea trials. *Brodogradnja*, 73(1), 89-107. https://doi.org/10.21278/brod73105

[5] Fan, X., Yang, S., Xiang, X., Sun, S., Hashali, S.D. 2026. A PointPillars-based 3D point cloud object detector of USVs for small target detection in dynamic aquatic environments. *Brodogradnja*, 77(2), 77202. https://doi.org/10.21278/brod77202

[6] Sun, S., Lyu, H., Dong, C. 2023. AIS Aided Marine Radar Target Tracking in a Detection Occluded Environment. *Ocean Engineering*, 288, 116133. https://doi.org/10.1016/j.oceaneng.2023.116133

[7] Emmens T., Amrit C., Abdi A., Ghosh M. 2021. The Promises and Perils of Automatic Identification System Data. *Expert Systems with Applications*, 178, 114975. https://doi.org/10.1016/j.eswa.2021.114975

[8] Zacchini, L., Topini, A., Franchi, M., Secciani, N., Manzari, V., Bazzarello, L., Stifani, M., Ridolfi, A. 2022. Autonomous Underwater Environment Perceiving and Modeling: An Experimental Campaign with Feelhippo AUV for Forward Looking Sonar-Based Automatic Target Recognition and Data Association. *IEEE Journal of Oceanic Engineering*, 48(2), 277-296. https://doi.org/10.1109/JOE.2022.3209719

[9] Liu, R.W., Guo, Y., Nie, J., Hu, Q., Xiong, Z., Yu, H., Guizani, M. 2022. Intelligent Edge-Enabled Efficient Multi-Source Data Fusion for Autonomous Surface Vehicles in Maritime Internet of Things. *IEEE Transactions on Green Communications and Networking*, 6(3), 1574-1587. https://doi.org/10.1109/tgcn.2022.3158004

[10] Alexandre, C., Devillers, R., Mouillot, D., Seguin, R., Catry, T. 2024. Ship Detection with Sar C-Band Satellite Images: A Systematic Review. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17, 14353-14367. https://doi.org/10.1109/jstars.2024.3437187

[11] Nomura, Y., Yamamoto, S., Hashimoto, T. 2021. Study of 3D Measurement of Ships Using Dense Stereo Vision: Towards Application in Automatic Berthing Systems. *Journal of Marine Science and Technology*, 26(2), 573-581. https://doi.org/10.1007/s00773-020-00761-2

[12] Sun, Z., Dai, M., Leng, X., Lei, Y., Xiong, B., Ji, K., Kuang, G. 2021. An Anchor-Free Detection Method for Ship Targets in High-Resolution Sar Images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14, 7799-7816. https://doi.org/10.1109/JSTARS.2021.3099483

[13] Fan, R., Jiao, J., Pan, J., Huang, H., Shen, S., Liu, M. 2019. Real-Time Dense Stereo Embedded in a UAV for Road Inspection. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16-17 June, Long Beach, CA, USA. https://doi.org/10.1109/cvprw.2019.00079

[14] Huntsberger, T., Aghazarian, H., Howard, A., Trotz, D.C. 2011. Stereo Vision-Based Navigation for Autonomous Surface Vessels. *Journal of Field Robotics*, 28(1), 3-18. https://doi.org/10.1002/rob.20380

[15] Guan, W., Xi, Z., Cui, Z., Zhang, X. 2025. Adaptive Trajectory Controller Design for Unmanned Surface Vehicles Based on SAC-PID. *Brodogradnja*, 76(2), 76206. https://doi.org/10.21278/brod76206

[16] Xiang, G., Rao, K., Wang, H., Xiang, X., Zuo, M., Wang, S., Soares, C.G. 2026. Numerical investigation of the hydrodynamic characteristics and interactions of an unmanned surface vehicle chasing a mother ship during recovery. *Brodogradnja*, 77(1), 77105. https://doi.org/10.21278/brod77105

[17] Ren, S., He, K., Girshick, R., Sun, J. 2016. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6), 1137-1149. https://doi.org/10.1109/TPAMI.2016.2577031

[18] Redmon, J., Divvala, S., Girshick, R., Farhadi, A. 2016. You Only Look Once: Unified, Real-Time Object Detection. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 27-30 June, Las Vegas, NV, USA. 779-788. https://doi.org/10.1109/cvpr.2016.91

[19] Lyu, H., Shao, Z., Cheng, T., Yin, Y., Gao, X. 2023. Sea-Surface Object Detection Based on Electro-Optical Sensors: A Review. *IEEE Intelligent Transportation Systems Magazine*, 15(2), 190-216. https://doi.org/10.1109/mits.2022.3198334

[20] Liu, L., Fu, L., Zhang, Y., Ni, W., Wu, B., Li, Y., Shang, C., Shen, Q. 2024. CLFR-Det: Cross-Level Feature Refinement Detector for Tiny-Ship Detection in SAR Images. *Knowledge-Based Systems*, 284, 111284. https://doi.org/10.1016/j.knosys.2023.111284

[21] Si, J., Song, B., Wu, J., Lin, W., Huang, W., Chen, S. 2023. Maritime Ship Detection Method for Satellite Images Based on Multiscale Feature Fusion. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 16, 6642-6655. https://doi.org/10.1109/jstars.2023.3296898

[22] Wang, S., Li, Y., Qiao, S. 2024. ALF-YOLO: Enhanced YOLOv8 Based on Multiscale Attention Feature Fusion for Ship Detection. *Ocean Engineering*, 308, 118233. https://doi.org/10.1016/j.oceaneng.2024.118233

[23] Thombre, S., Zhao, Z., Ramm-Schmidt, H., Vallet Garcia, J.M., Malkamaki, T., Nikolskiy, S., et al. 2022. Sensors and AI Techniques for Situational Awareness in Autonomous Ships: A Review. *IEEE Transactions on Intelligent Transportation Systems*, 23(1), 64-83. https://doi.org/10.1109/tits.2020.3023957

[24] Poggi, M., Tosi, F., Batsos, K., Mordohai, P., Mattoccia, S. 2021. On the Synergies between Machine Learning and Binocular Stereo for Depth Estimation from Images: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9), 5314-5334. https://doi.org/10.1109/TPAMI.2021.3070917

[25] Jung, B., Sukhatme, G.S. 2010. Real-Time Motion Tracking from a Mobile Robot. *International Journal of Social Robotics*, 2, 63-78. https://doi.org/10.1109/SSRR.2014.7017674

[26] Benacer, I., Hamissi, A., Khouas, A. 2015. A Novel Stereovision Algorithm for Obstacles Detection Based on U-V-Disparity Approach. *2015 IEEE International Symposium on Circuits and Systems (ISCAS)*, 24-27 May, Lisbon, Portugal. 369-372. https://doi.org/10.1109/ISCAS.2015.7168647

[27] Khurshid, M., Shahzad, M., Khattak, H.A., Malik, M.I., Fraz, M.M. 2024. Vision-Based 3-D Localization of UAV Using Deep Image Matching. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17, 12020-12030. https://doi.org/10.1109/JSTARS.2024.3422310

[28] Zheng, Y., Liu, P., Qian, L., Qin, S., Liu, X., Ma, Y., Cheng, G. 2022. Recognition and Depth Estimation of Ships Based on Binocular Stereo Vision. *Journal of Marine Science and Engineering*, 10(8), 1153. https://doi.org/10.3390/jmse10081153

[29] Shang, Y., Yu, W., Zeng, G., Li, H., Wu, Y. 2024. StereoYOLO: A Stereo Vision-Based Method for Maritime Object Recognition and Localization. *Journal of Marine Science and Engineering*, 12(1), 197. https://doi.org/10.3390/jmse12010197

[30] Khanam, R., Hussain, M. 2024. YOLOv11: An Overview of the Key Architectural Enhancements. *arXiv preprint arXiv:241017725*. https://doi.org/10.48550/arXiv.2410.17725

[31] Tang, F., Xu, Z., Huang, Q., Wang, J., Hou, X., Su, J., Liu, J. 2023. DuAT: Dual-Aggregation Transformer Network for Medical Image Segmentation. *2023 Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, 13-15 October, Xiamen, China. 343-356. https://doi.org/10.1007/978-981-99-8469-5_27

[32] Tan, M., Pang, R., Le, Q.V. 2020. EfficientDet: Scalable and Efficient Object Detection. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13-19 June, Seattle, WA, USA. 10778-10787. https://doi.org/10.1109/cvpr42600.2020.01079

[33] Gevorgyan, Z. 2022. SIoU Loss: More Powerful Learning for Bounding Box Regression. *arXiv preprint arXiv:220512740*. https://doi.org/10.48550/arXiv.2205.12740

[34] Kendall, A., Martirosyan, H., Dasgupta, S., Henry, P., Kennedy, R., Bachrach, A., Bry, A. 2017. End-to-End Learning of Geometry and Context for Deep Stereo Regression. *2017 IEEE International Conference on Computer Vision (ICCV)*, 22-29 October, Venice, Italy. 66-75. https://doi.org/10.1109/ICCV.2017.17

[35] Chang, J.-R., Chen, Y.-S. 2018. Pyramid Stereo Matching Network. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 18-23 June, Salt Lake City, UT, USA. https://doi.org/10.1109/cvpr.2018.00567

[36] Lipson, L., Teed, Z., Deng, J., Soc, I.C. 2021. RAFT-Stereo: Multilevel Recurrent Field Transforms for Stereo Matching. *2021 International Conference on 3D Vision (3DV)*, 1-3 December, London, UK. 218-227. https://doi.org/10.1109/3dv53792.2021.00032

[37] Žbontar, J., LeCun, Y. 2016. Stereo Matching by Training a Convolutional Neural Network to Compare Image Patches. *Journal of Machine Learning Research*, 17, 1-32.

[38] Guan, T., Wang, C., Liu, Y.-H., 2024. Neural Markov Random Field for Stereo Matching. *2024 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 16-22 June, Seattle, WA, USA. 5459-5469. https://doi.org/10.1109/cvpr52733.2024.00522

[39] Xu, G., Wang, X., Ding, X., Yang, X. 2023. Iterative Geometry Encoding Volume for Stereo Matching. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 17-24 June, Vancouver, BC, Canada. 21919-21928. https://doi.org/10.1109/CVPR52729.2023.02099

[40] Zhang, Z. 2002. A Flexible New Technique for Camera Calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11), 1330-1334. https://doi.org/10.1109/34.888718

[41] Bouguet, J.-Y. 2010. Camera Calibration Toolbox for Matlab. California Institute of Technology, Pasadena, USA. https://doi.org/10.22002/D1.20164

[42] Zheng, Y., Zhang, S. 2020. Mcships: A Large-Scale Ship Dataset for Detection and Fine-Grained Categorization in the Wild. *2020 IEEE International Conference on Multimedia and Expo (ICME)*, 6-10 July, London, UK. 1-6. https://doi.org/10.1109/ICME46284.2020.9102907

[43] Prasad, D.K., Rajan, D., Rachmawati, L., Rajabally, E., Quek, C. 2017. Video Processing from Electro-Optical Sensors for Object Detection and Tracking in a Maritime Environment: A Survey. *IEEE Transactions on Intelligent Transportation Systems*, 18(8), 1993-2016. https://doi.org/10.1109/TITS.2016.2634580

[44] Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., et al. 2024. DETRs Beat YOLOs on Real-Time Object Detection. *2024 IEEE/CVF conference on computer vision and pattern recognition*, 16-22 June, Seattle, WA, USA. https://doi.org/10.1109/CVPR52733.2024.01605

[45] Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., et al. 2022. DINO: DETR with Improved Denoising Anchor Boxes for End-to-End Object Detection. *arXiv preprint arXiv:220303605*. https://doi.org/10.48550/arXiv.2203.03605

[46] Xu, X., Jiang, Y., Chen, W., Huang, Y., Zhang, Y., Sun, X. 2022. DAMO-YOLO: A Report on Real-Time Object Detection Design. *arXiv preprint arXiv:221115444*. https://doi.org/10.48550/arXiv.2211.15444