

Article

Research on Ship Target Detection in Complex Sea Surface Scenarios Based on Improved YOLOv7

Zhuang Cai *  and Weina Zhou 

School of Information Engineering, Shanghai Maritime University, Shanghai 201306, China;
wnzhou@shmtu.edu.cn

* Correspondence: 202330310027@stu.shmtu.edu.cn

Abstract

Ships target detection plays a crucial role in safeguarding maritime transportation. However, affected by factors such as ocean waves, extreme weather, and target diversity (e.g., large size differences, arbitrary rotation, and occlusion), existing deep learning-based detection methods struggle to achieve a satisfactory balance among accuracy, speed, and model size in complex marine environments. To address this challenge, this paper proposes a real-time ship detection algorithm (C-YOLO) integrating global perception and multi-scale feature enhancement. First, a Transformer encoder is added before the detection head, which suppresses interference from sea clutter and cloud mist occlusion through long-range dependency modeling, improving the detection of small and occluded ships. Second, a Dual-Effect Focused Residual Fusion Module is designed to replace the backbone's multi-scale pooling structure, combining the advantages of CBAM (background noise suppression) and SK-Net (dynamic scale adaptation) to simultaneously capture features of ships of different sizes. Finally, a CZIoU loss function is proposed, which integrates constraints on angle, center point, vertex, and area to address rotation, deformation, and multi-scale issues in ship detection. Experimental results on the SeaShips 7000 dataset show that the proposed C-YOLO achieves a Recall of 0.842, mAP@50 of 0.797, and mAP@50:95 of 0.552, outperforming mainstream algorithms such as YOLOv7 (Recall = 0.785, mAP@50 = 0.781), YOLOv9s (Recall = 0.819, mAP@50 = 0.755), and SSD (Recall = 0.802, mAP@50 = 0.833). With 76.75 M parameters and an inference speed of 119 FPS, the model maintains efficient real-time performance while ensuring detection accuracy. This method effectively reduces false detection and missed detection rates in complex scenarios such as port monitoring and maritime traffic control, providing a reliable technical solution for intelligent maritime surveillance and safe navigation—with significant practical value for improving maritime transportation efficiency and reducing safety risks.

Keywords: global perception; dual-effect focused residual fusion module; CZIoU loss function; transformer encoder; ship target detection



Academic Editor: Atsushi Mase

Received: 30 December 2025

Revised: 24 January 2026

Accepted: 28 January 2026

Published: 11 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

The efficient detection of ship targets plays a crucial role in maritime transportation and the safe navigation of ships. However, it remains a challenging task in complex maritime environments [1]. To date, although numerous target detection methods have been proposed, few can be effectively applied to ship detection in maritime scenarios [2]. Ship target detection is confronted with interferences from complex marine environments, including sea clutter, weather variations such as haze, rain and snow, as well as lighting conditions

like reflection and low illumination. These factors significantly reduce the signal-to-noise ratio of images and increase the risks of false detection and missed detection. Meanwhile, the inherent diversity of targets—such as ships with large size differences, various attitudes and occlusion conditions, as well as low-contrast hulls (e.g., those with colors similar to seawater or stealth designs)—makes it difficult for conventional target detection algorithms to stably extract valid features. Current ship target detection methods are mainly divided into two categories: traditional method-based target detection and deep learning-based target detection. Traditional method-based ship target detection is an approach developed in the early stage of the computer vision field, which mainly relies on image processing and feature extraction technologies. In contrast, deep learning-based detection methods utilize convolutional neural networks for ship detection. Despite their relatively complex network architectures, these methods possess strong fitting capabilities and outperform traditional target detection methods by a wide margin in both speed and accuracy. In recent years, with the rise of deep learning, convolutional neural network-based methods have been widely adopted. Improved target detection algorithms (e.g., YOLOv8 [3], Faster R-CNN [4], DETR [5]) combined with Oriented Bounding Boxes (OBB [6]) are mainly used to address the issue of multi-angle and multi-scale ship detection. For small target detection, researchers have introduced Feature Pyramid Networks (FPN [7]), attention mechanisms (e.g., EMA [8]) and super-resolution reconstruction techniques (e.g., SRGAN [9]) to enhance the recognition rate of small ships in Synthetic Aperture Radar (SAR) or optical remote sensing images. Meanwhile, lightweight models (e.g., MobileNet-YOLO [10]) have been applied to edge devices (e.g., unmanned aerial vehicles and shipborne systems) to meet real-time requirements. DBW-YOLO [11] improves the performance of lightweight SAR ship detection models in complex backgrounds and small target scenarios by integrating deformable convolution and the BiFormer attention mechanism, yet challenges such as robustness in extreme weather and dense occlusion still need to be resolved. Ribeiro et al. [12] improved deep learning models to achieve real-time ship detection in maritime surveillance videos, but the detection robustness in complex weather and dense small target scenarios still requires enhancement. Sun et al. [13] proposed an anchor-free method based on key-point estimation, which improves the efficiency and generalization ability of ship detection in panchromatic satellite images, but it still faces problems such as occlusion of dense small targets in complex port scenarios and insufficient detection accuracy for multi-scale ships. Sun et al. [14] propose the KFIA-Net, which integrates domain knowledge, optimizes the attention mechanism and loss function, addresses the class imbalance issue in SAR ship detection, and enhances detection accuracy and efficiency in complex scenarios. Although the methods proposed in the above-mentioned studies have achieved improvements in target detection algorithms to varying degrees, existing algorithms for ship target detection still fail to achieve a satisfactory balance among performance metrics such as accuracy, speed and model size. Therefore, this paper proposes the C-YOLO network, whose contributions are as follows: 1. A Transformer encoder structure is added before the detection head, which significantly improves the performance of ship target detection. By correlating the occluded parts and visible regions of ships through long-range dependencies, the Transformer encoder enhances detection robustness in occlusion scenarios. 2. The designed Dual-Effect Focused Residual Fusion Module improves the performance of the proposed network in ship detection effectively by combining the dual advantages of Convolutional Block Attention Module (CBAM) and Selective Kernel Networks (SK-Net). This design addresses two major challenges in ship detection: CBAM can suppress irrelevant waves and light reflections, reducing background interference in complex sea conditions, while SK-Net can capture the features of both large and small ships simultaneously. 3. The proposed CZIoU loss function is specially optimized for the problems of rotation, deformation and

multi-scale in ship target detection. Compared with conventional detection methods, this function can accurately capture the rotation angle, main body size and spatial position of ships at the same time, significantly improving the detection accuracy of densely arranged and cross-berthed ships in scenarios such as port monitoring and maritime traffic control, and effectively overcoming typical problems such as “missed detection caused by rotation” and “adhesion of adjacent ships”.

2. Methodology

The ship detection network architecture of C-YOLO is based on YOLOv7. Choosing YOLOv7 as the base model is not due to a lack of awareness of the advancements in subsequent versions such as YOLOv8 to YOLOv11, but rather a deliberate consideration based on technical adaptability and practical application requirements. Multi-domain benchmark tests indicate that the performance improvement of the YOLO series is not monotonically increasing with version numbers; some newer versions even exhibit lower mAP performance than YOLOv7. The core ELAN architecture of YOLOv7 achieves an optimal balance between computational efficiency and inference stability through dynamic reparameterization and multi-scale feature fusion. Its parallel computing structure optimized for GPU devices can effectively reduce memory access costs, demonstrating more reliable compatibility in embedded hardware deployment. Compared with YOLOv8, which focuses on end-to-end multi-tasking, YOLOv9 specialized in small target detection, and YOLOv11 pursuing extreme generalization, YOLOv7 features higher architectural maturity, stable training convergence speed, and support for TensorRT optimization and quantization strategies. This makes it more suitable as a foundation for improvements to focus on performance enhancement in specific tasks. This choice not only avoids the adaptation risks of newer versions in specific scenarios but also provides a stable and competitive benchmark framework for verifying the innovations of the proposed method. Compared with other algorithms, YOLOv7 [15] features a modular structure and superior scalability, which allows customization for specific scenarios to meet diverse application requirements. The overall structure of C-YOLO is illustrated in Figure 1, and this architecture is suitable for ship target detection. In C-YOLO, multi-layer Dual-Effect Focused Residual Fusion Modules are inserted between the backbone network layers to replace the original multi-scale pooling structure and perform its corresponding functions. This module extracts target features from input data through fully connected layers. As the number of network layers increases, the network is capable of capturing more accurate target information. In addition, a Transformer Encoder is added before the network detection head, which captures global contextual relationships via a self-attention mechanism, thereby effectively improving the detection of occluded objects and small targets. Meanwhile, its dynamic feature weighting enhances the response of key regions, leading to higher confidence levels for blurry targets.

2.1. Transformer Encoder Structure

The encoder inside the Transformer [16] is illustrated in Figure 2.

The Transformer encoder is composed of a stack of identical encoder layers. Each layer is centered on a multi-head self-attention mechanism and a feed-forward network (FFN), with both sub-modules equipped with residual connections and layer normalization. At the input stage, positional encoding must be added to the embedding vectors to compensate for the lack of sequential information caused by the Transformer’s non-recurrent architecture. The multi-head self-attention maps the input into three sets of vectors, namely query, key and value, and calculates the correlation weights of different feature subspaces in parallel through multiple attention heads, thus achieving efficient fusion of global contextual

information. The subsequent FFN consists of two linear transformations and a ReLU activation function, performing independent non-linear transformations on the features at each position. The entire encoder structure abandons convolutional and recurrent operations and supports full parallel computation, which greatly improves the training efficiency of the model.

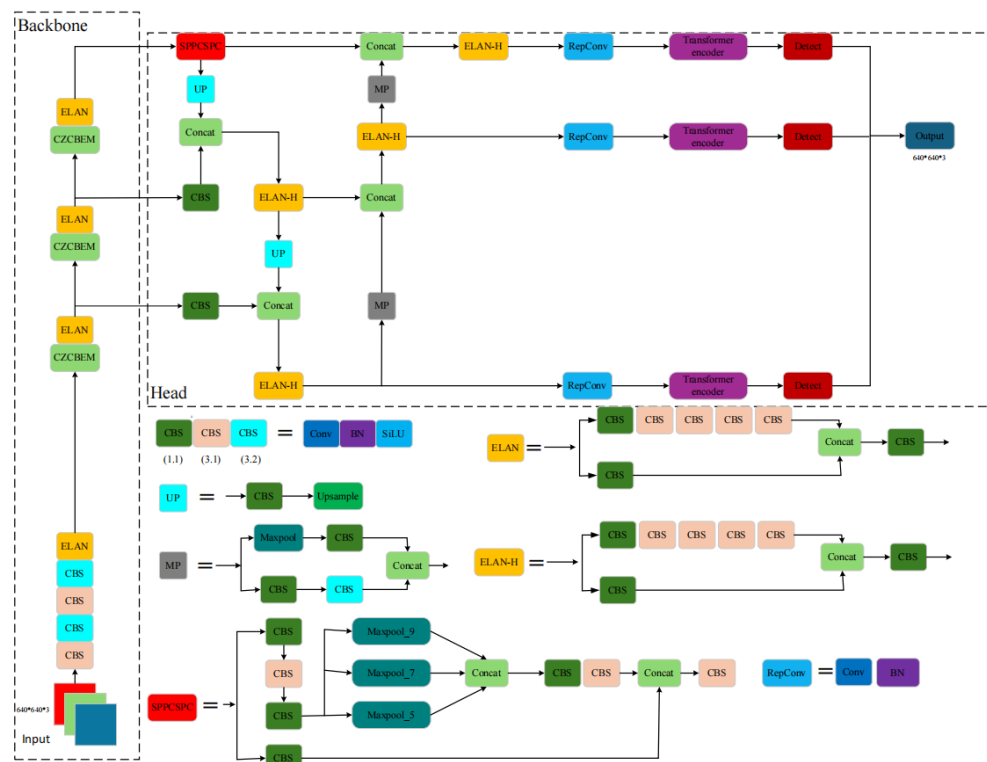


Figure 1. Overall structure of C-YOLO.

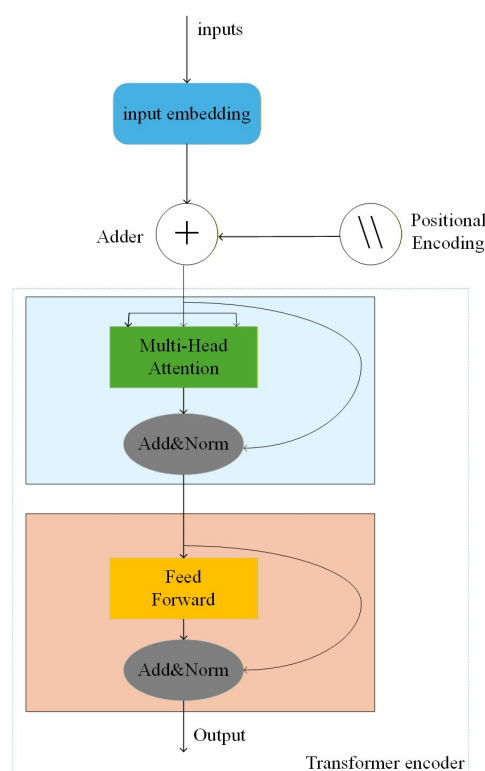


Figure 2. Transformer encoder structure.

Scaled Dot-Product Attention as illustrated in the Figure 3 is the core component of the Transformer encoder, designed to capture dependencies between input tokens regardless of their positional distance. The mechanism first computes attention scores by taking the dot product of query and key vectors, which reflects the relevance of each key to the corresponding query. Since the dot product values can grow large when vector dimensions are high, leading to extremely small gradients after softmax and hindering training, the scores are scaled by the square root of the key vector dimension to normalize the value range. After scaling, a softmax function is applied to convert the scores into a set of weights that sum to one, where higher weights indicate greater importance of the corresponding token. Finally, the weighted sum of value vectors is calculated to generate the attention output, which integrates information from all input tokens based on their relevance. This design enables the model to focus on critical tokens dynamically, and the scaling operation ensures stable and efficient training, laying the foundation for the Transformer's superior performance in sequence modeling tasks.

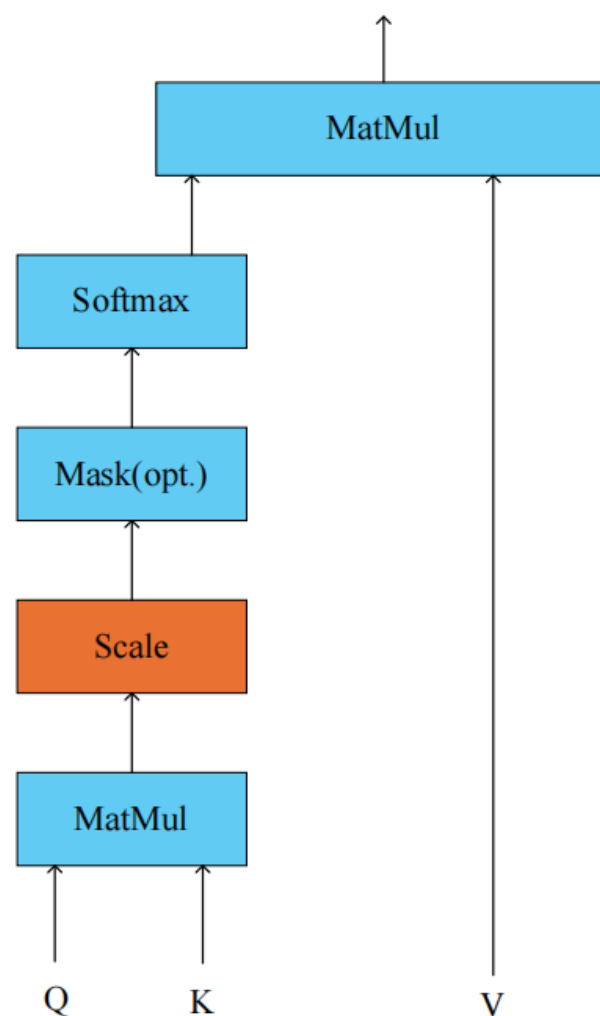


Figure 3. Model structure of scaled dot product attention.

2.2. Dual-Effect Focused Residual Fusion Module

CBAM [17] (Convolutional Block Attention Module) is a lightweight attention mechanism module that dynamically enhances the key regions of feature maps by combining Channel Attention and Spatial Attention in parallel. The overall structure of CBAM is illustrated in Figure 4.

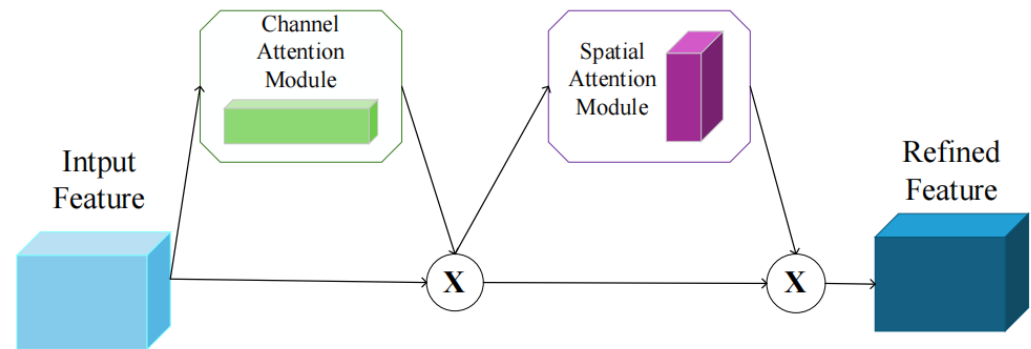


Figure 4. Overall structure of CBAM.

The Channel Attention Module (CAM) in CBAM is an efficient feature channel optimization mechanism that dynamically calibrates the weight of each channel by aggregating global spatial information. Its architecture is depicted in Figure 5. The core process consists of two steps: first, global average pooling (GAP) and global max pooling (GMP) are respectively applied to the input feature map, generating two distinct channel description vectors (which focus on global context and salient local features, respectively); next, these two vectors are fed into a shared multi-layer perceptron (MLP), where the inter-channel relationships are learned through nonlinear transformation. Finally, the weights output by the MLP are summed and normalized via the Sigmoid activation function, yielding the attention score for each channel. Compared with the classic SE [18] module, the CAM of CBAM innovatively integrates dual-path information from GAP and GMP, which not only captures the overall distribution but also preserves key details (e.g., high-response regions), thus enabling a more comprehensive evaluation of channel importance.

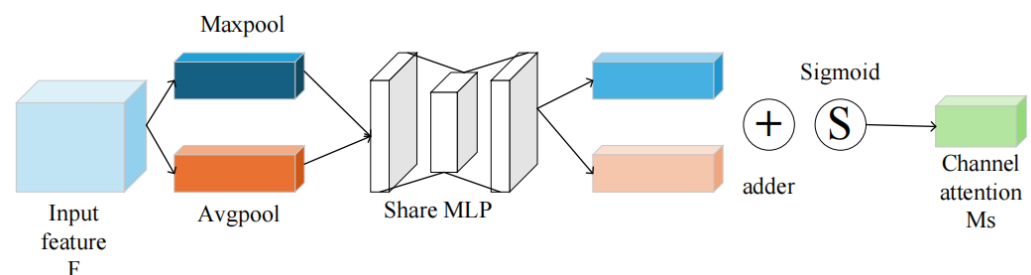


Figure 5. Architecture diagram of the channel attention module.

The calculation formulas of the Channel Attention Module are as follows:

$$M_c = \delta(MLP(AvgPool(F) + MLP(MaxPool(F)))) \quad (1)$$

The Spatial Attention Module as illustrated in the Figure 6 (SAM) in CBAM boosts the model's perception of key positions by focusing on critical spatial regions of feature maps. Its core idea is to generate a spatial attention mask leveraging channel-wise information: first, global average pooling (GAP) and global max pooling (GMP) are applied to the input feature map along the channel dimension to obtain two spatial feature maps reflecting global distribution and local salient features, respectively; then, the two maps are concatenated along the channel dimension and fused via a convolutional layer; finally, a spatial weight map is generated with the Sigmoid activation function. As a complement to channel attention, SAM accurately locates target key regions (e.g., object edges or texture-dense areas) and suppresses irrelevant background noise. Combined with channel attention,

CBAM achieves synergistic ‘channel-spatial’ optimization, serving as a classic lightweight attention mechanism design.

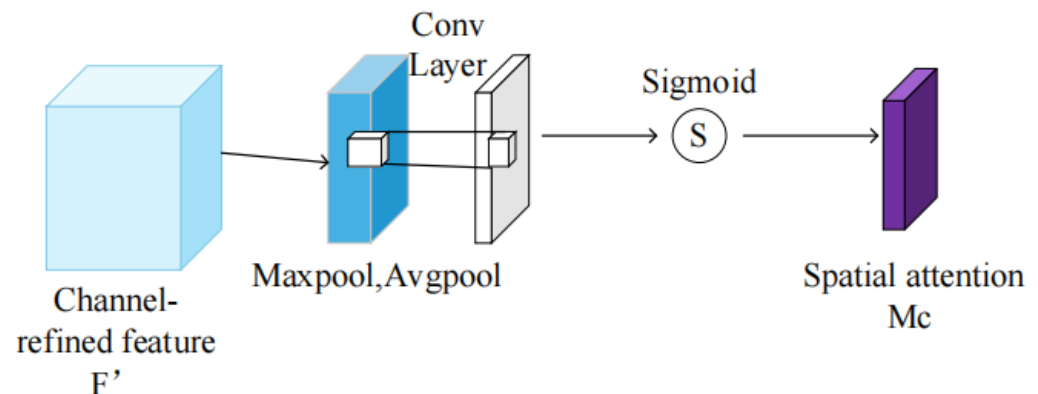


Figure 6. Architecture diagram of the spatial attention module.

The calculation formulas of the Spatial Attention Module are as follows:

$$\begin{aligned} \mathcal{M}_s(\mathbf{F}') &= \sigma\left(f^{7 \times 7}\left([\text{AvgPool}(\mathbf{F}'); \text{MaxPool}(\mathbf{F}')]\right)\right) \\ &= \sigma\left(f^{7 \times 7}\left(\left[\mathbf{F}'^s_{\text{avg}}; \mathbf{F}'^s_{\text{max}}\right]\right)\right) \end{aligned} \quad (2)$$

SK-Net as illustrated in the Figure 7 (Selective Kernel Networks) [19] is a lightweight optimization method for convolutional neural networks (CNNs), which is designed to address the core limitation of traditional CNNs where fixed convolutional kernel sizes fail to simultaneously balance local detail extraction and global context capture. Its core component is the SK module, which achieves dynamic kernel scale adaptation through three steps: first, multiple groups of branch features are generated in parallel by applying convolutional kernels of different sizes or receptive fields to the input; second, channel-wise attention weights that characterize the adaptability of each branch feature are produced via global average pooling and a two-layer fully connected network; finally, the optimal output features tailored to the input content are obtained by performing weighted summation on the multi-branch features based on the generated attention weights. This design enables the network to autonomously select the appropriate kernel scale according to the input scene, enhances the feature representation capability in complex scenarios (e.g., scenarios with coexisting small and large targets) without significantly increasing computational complexity or the number of parameters, and can be seamlessly integrated into mainstream CNN architectures such as ResNet, thus stably improving performance in ship target detection tasks.

Residual Network as illustrated in the Figure 8 (ResNet) is a deep convolutional neural network architecture proposed in 2015, whose core innovation lies in residual connections designed to address the “performance degradation” problem that occurs in traditional deep networks as the number of layers increases (this is not overfitting, but manifests as a simultaneous rise in both training and test errors). By designing “residual blocks”, the network layers no longer directly learn the complete feature mapping, but instead learn the residual mapping (i.e., the difference between the target mapping and the input features). Then, the input features are directly added to the output of the residual mapping via residual connections, achieving a “shortcut effect” during gradient backpropagation—even if the gradients in the deep layers approach zero, they can still be transmitted to the shallow layers through the direct connections, effectively alleviating the vanishing gradient problem

and thus enabling the training of networks with hundreds or even thousands of layers. The schematic diagram of the Residual Network is illustrated in Figure 8.

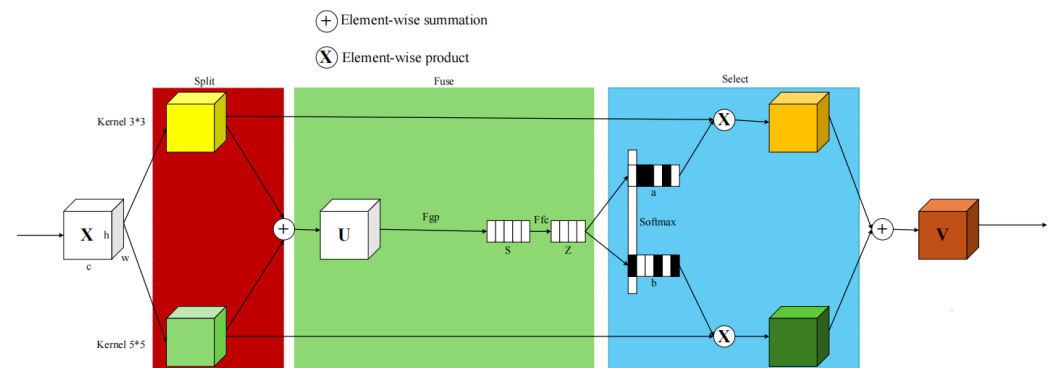


Figure 7. Schematic Diagram of SK-Nets Module.

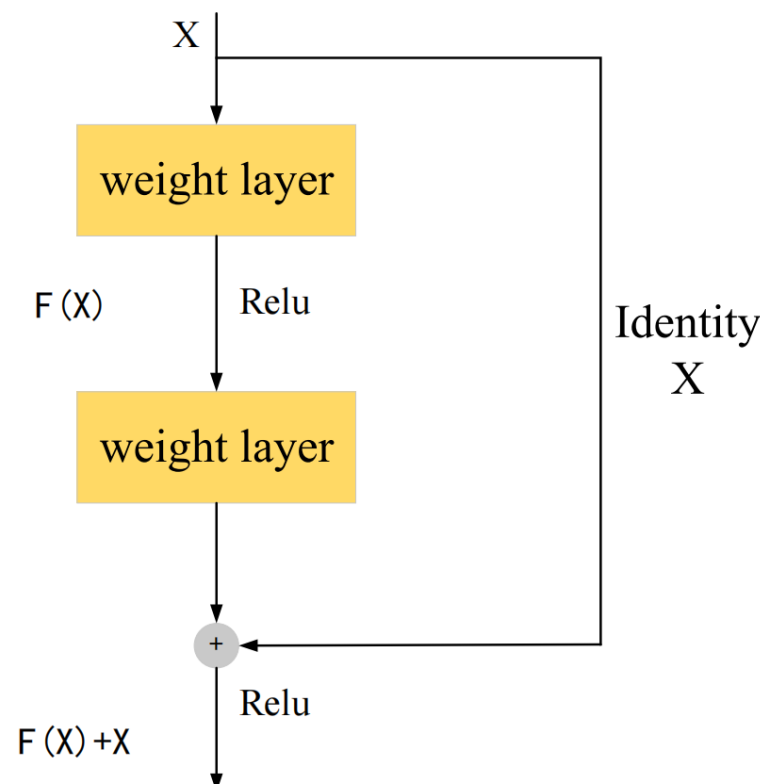


Figure 8. Schematic Diagram of Residual Network.

The Dual-Effect Focused Residual Fusion Module illustrated in Figure 9 derives its core advantages from the deep synergistic integration of residual connections, CBAM, and SK-Nets. First, the residual connection—via a shortcut structure that directly links the input X to the module's output—establishes an efficient pathway for gradient backpropagation: this not only effectively mitigates the vanishing gradient problem in deep network training, ensuring the model's trainability and scalability for increased depth, but also retains the original information of input features, preventing the loss of critical details during feature propagation. Meanwhile, the embedded CBAM (Channel-Spatial Attention Module) specifically suppresses background noise from complex sea conditions (e.g., waves, reflections) in ship detection scenarios, enhancing feature responses in target regions; SK-Nets, by dynamically selecting convolution kernel scales, capture multi-scale features of both large and small ships simultaneously. These two components achieve precise integration

of attention and scale adaptation through weighted parameters ($W1$, $W2$) generated by Sigmoid activation, and their outputs are fused with the original input via the residual structure—this not only enables refined feature enhancement but also avoids the functional limitations of a single module through the “information complementarity” of residuals. By retaining the “lightweight and stable” advantages of residual connections, this design integrates the feature-screening capability of attention mechanisms and the generalization capability of dynamic scale adaptation into one module: it does not significantly increase computational overhead, while addressing the core pain points of ship detection (complex background interference and coexisting multi-scale targets), ultimately achieving a triple balance between model depth, feature expression accuracy, and training efficiency.

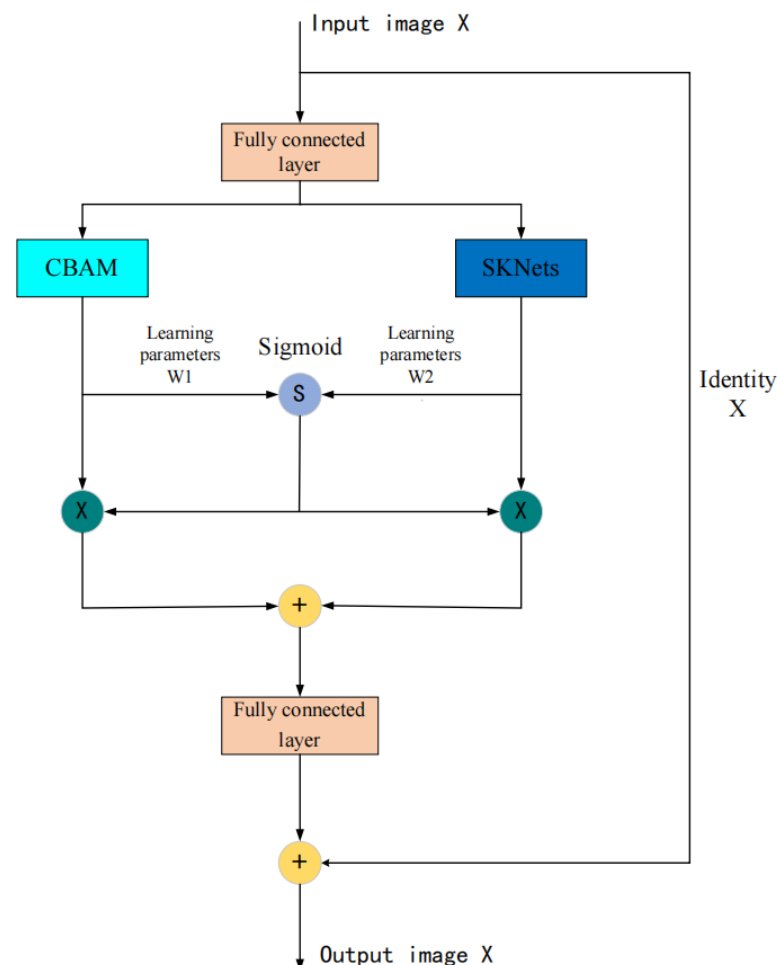


Figure 9. Dual-Effect Focused Residual Fusion Module.

Figure 9 illustrates the Dual-Effect Focused Residual Fusion Module: the input X is split into two branches after passing through a fully connected layer, which are processed by CBAM (to suppress background noise) and SK-Nets (to adapt to multi-scale features) respectively. The parameters of these two components are weighted via Sigmoid and then fused with the original input; after another fully connected layer and combination with the residual input X , the final output Y is generated. This module leverages residual connections to ensure training stability, while synergizing CBAM and SK-Nets to enhance feature representation capabilities for ship detection.

2.3. CZIoU Loss Function

The CZIoU Loss Function proposed in this paper integrates angle, center point, vertex and area constraints on the basis of standard IoU, which can effectively address the

detection of ship targets with varying angles. The CZIoU Loss Function is given as follows.

$$CZIoU = IoU - \left| \frac{\rho^2}{c^2} \right| - \left| \frac{d_1^2 + d_2^2}{2c^2} \right| - \frac{4}{\pi^2} (\arctan S_A - \arctan S_B)^2 - g(\theta) \quad (3)$$

Wherein, IoU denotes the Intersection over Union of the predicted bounding box and the ground-truth bounding box: ρ represents the distance between the centers of the predicted and ground-truth bounding boxes. d_1 is the Euclidean distance between their top-left vertices; d_2 is the Euclidean distance between their bottom-right vertices; S_A and S_B are the areas of the predicted and ground-truth bounding boxes, respectively; And c is the diagonal length of the smallest enclosing rectangle covering both the predicted and ground-truth bounding boxes. Figure 10 shows a schematic diagram of the relationship between the predicted bounding box and the ground-truth bounding box. The rotational deviation is optimized by constraining the rotation angle penalty term of the predicted and ground-truth bounding boxes:

$$g(\theta) = 1 - \cos(\theta_{\text{pred}} - \theta_{\text{gt}}) + \lambda \sin^2(\theta_{\text{pred}} - \theta_{\text{gt}}) \quad (4)$$

where $\lambda \in (0, 1)$. θ_{pred} and θ_{gt} denote the rotation angles of the predicted and ground-truth bounding boxes, respectively; When the rotation angle of the predicted bounding box is equal to that of the ground-truth bounding box, the angle penalty term equals zero, indicating that the predicted and ground-truth bounding boxes have the same rotation angle. Normalization of the center distance between the predicted and ground-truth bounding boxes can address the scale sensitivity issue; when the center distance is zero, it means the overlap degree between the two bounding boxes is improved. The vertex distance constraint enhances aspect ratio matching—by constraining the distance between vertices, the aspect ratios of the predicted and ground-truth bounding boxes are better aligned. The combination of the arctangent function $\arctan(\cdot)$ and the square function is adopted to constrain the problem of target area mismatch. When the area difference between the predicted target and the ground-truth target is excessively large, the area penalty term increases, thereby dynamically adjusting the target size. When the area difference is extremely small, the penalty term decreases, thus improving the area consistency between the predicted and ground-truth bounding boxes. Compared with traditional methods, this method can simultaneously and accurately optimize the position, angle and shape of the detection box, significantly improving the detection accuracy of rotated targets, and is particularly suitable for complex scenarios with arbitrary target rotation and deformation.

$$IoU = \frac{|A \cap B|}{|A \cup B|} \quad (5)$$

$$d_1 = \sqrt{(x_1 - x_5)^2 + (y_1 - y_5)^2} \quad (6)$$

$$d_2 = \sqrt{(x_4 - x_8)^2 + (y_4 - y_8)^2} \quad (7)$$

$$\rho = \sqrt{(x_a - x_b)^2 + (y_a - y_b)^2} \quad (8)$$

$$c = \sqrt{(x_1 - x_4)^2 + (y_1 - y_4)^2} \quad (9)$$

$$S_A = w_1 * h_1 \quad (10)$$

$$S_B = w_2 * h_2 \quad (11)$$

Figure 11 shows the relative positional relationship between the predicted bounding box and the ground-truth bounding box when both are rotated by a certain angle.

Trigonometric functions are employed to constrain the positional relationship between the predicted and ground-truth bounding boxes. The bounding box loss function is formulated as follows:

$$\text{Gamma}_{\text{CZIoU}} = 1 - \text{CZIoU} \quad (12)$$

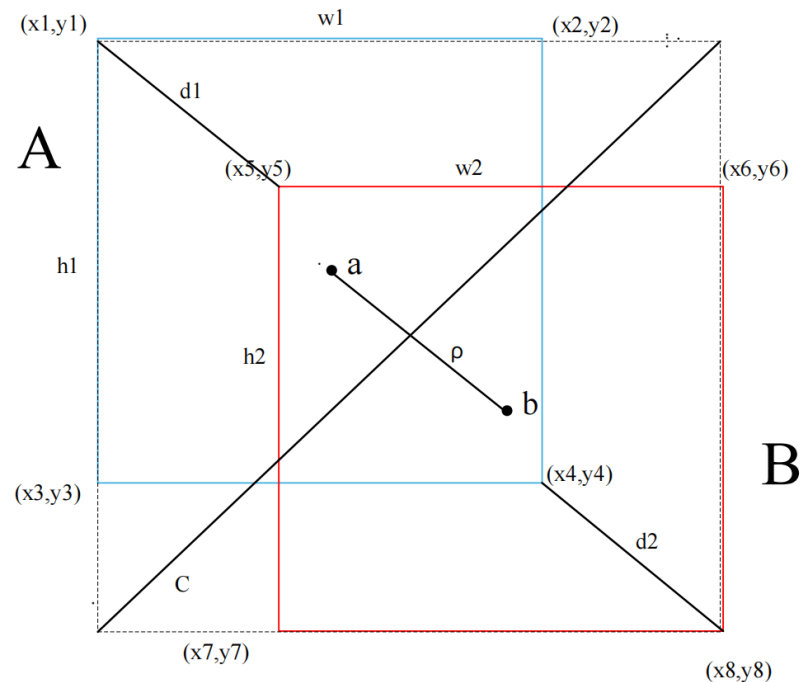


Figure 10. Schematic diagram of the relationship between the prediction box and the real box.

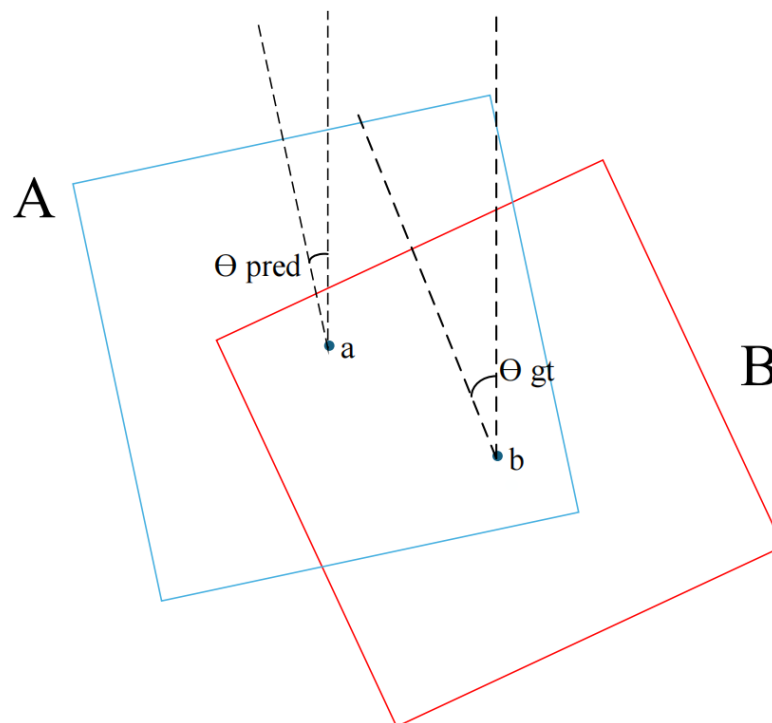


Figure 11. Diagram of the angle between the prediction box and the real box rotated at a certain angle.

The design of CZIoU achieves systematic optimization for three core challenges in rotated object detection. First, to address the periodicity issue in rotated bounding box angle regression (e.g., 179° is equivalent to -1° but causes abrupt loss changes), an angle

continuity constraint is introduced, which converts angular differences into differentiable continuous variables via periodic smoothing functions, thus avoiding gradient mutations. Second, tailored to the geometric characteristics of extreme aspect ratio targets (e.g., slender or flat objects), shape-adaptive weights are designed to dynamically adjust the loss function's sensitivity to aspect ratio parameters, enabling the model to capture the boundary features of slender or flat targets more accurately. Next, the area constraint incorporates normalization in the object detection loss function to balance the optimization weights of targets at different scales, effectively resolving the problem of gradient vanishing for small targets that plagues traditional IoU. It employs nonlinear transformations such as $\arctan(\cdot)$ to map area ratios to a bounded interval, which not only maintains training balance between large and small targets but also enhances the model's robustness to area annotation errors. Finally, to tackle the imbalance between center point and boundary optimization in complex scenarios, a dynamic balancing mechanism is adopted to automatically allocate optimization weights for center point offset and bounding box deformation based on target scales.

3. Experiment and Analysis

3.1. Dataset

To verify the effectiveness of the proposed method, experiments were conducted on the SeaShips 7000 dataset [20]. Most ship images in the dataset were captured in inland waterways, featuring diverse backgrounds, unstable lighting conditions, and complex scenarios such as target occlusion. SeaShips 7000 is a public dataset dedicated to ship detection and classification, consisting of 7000 ship images covering various lighting conditions, weather patterns, and sea states. It includes 6 categories of ships, such as ore carriers, bulk carriers, and container ships. In this paper, the dataset was partitioned into a training set, test set, and validation set with a ratio of 6000:500:500.

All experiments were performed on a platform equipped with a single RTX 3080 Ti (12 GB), PyTorch 1.10.0, Python 3.8 (Ubuntu 20.04), CUDA 11.3, and a 12-core Intel(R) Xeon(R) Silver 4214R CPU running at 2.40 GHz. The hyperparameter settings are presented in Table 1.

Table 1. Hyperparameter settings for the experiment.

Hyperparameter	Set Value
Optimizer	SGD
Initial learning rate	0.01
Parameter control data loader	8
Number of training epochs	100
Training batch size	16
Momentum	0.937

As shown in Table 1, we adopted SGD as the optimizer and set the initial learning rate to 0.01, which is consistent with common practices in ship detection tasks.

3.2. Evaluation Metrics

In this paper, Precision (P), Recall (R), and mAP (mean Average Precision) are adopted to evaluate the performance of the proposed model. Precision and Recall are two widely used evaluation metrics, which are defined as follows:

$$P = \frac{TP}{TP + FP} \quad (13)$$

$$R = \frac{TP}{TP + FN} \quad (14)$$

Wherein, TP (True Positives) denotes the number of ships correctly detected, FP (False Positives) denotes false alarms, and FN (False Negatives) denotes the number of ships missed. The widely used mean Average Precision (mAP) refers to the average value of precision across all categories. Given that there is only one category of targets in the dataset, the result obtained by calculating AP (Average Precision) is essentially equivalent to mAP. The specific formulas for AP and mAP are defined as follows:

$$AP = \int_0^1 P(R) dR \quad (15)$$

$$mAP = \frac{1}{n} \sum_{i=1}^n AP_i \quad (16)$$

Figure 12 illustrates the distribution of image quantities under different environmental conditions in the SeaShips-7000 dataset. Among these categories, “Port/Dock Background” accounts for the largest proportion, with 4200 images (0.60), followed by “Open Water Background” with 2100 images (0.30). Together, these two categories make up 0.90 of the entire dataset. The remaining categories, such as low illuminance, foggy weather, and occlusion, have relatively few samples, which reflects that the dataset is dominated by port and open water scenes.

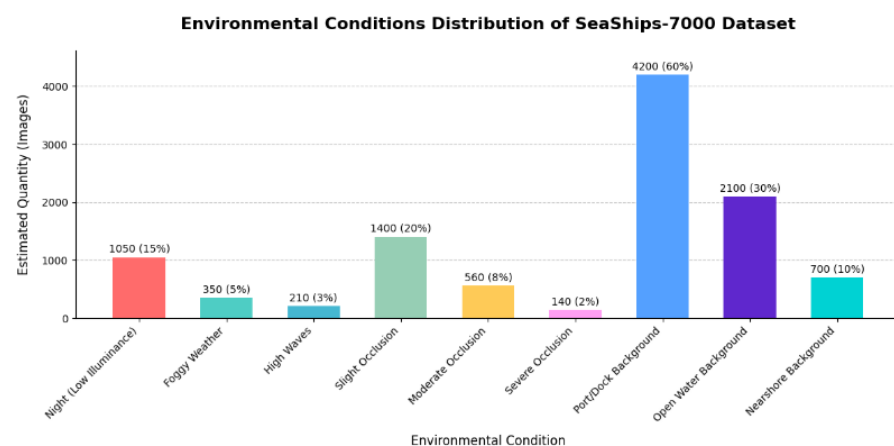


Figure 12. Environmental Conditions Distribution of SeaShips-7000 Dataset.

As illustrated in Figure 13, A denotes the bounding box of the actual ship, while B denotes the bounding box of the ship detected by the model. IoU (Intersection over Union) refers to the ratio of the intersection area to the union area between the ground-truth ship bounding box A and the model-detected ship bounding box B. mAP@50 and mAP@50:95 are mAP variants that depend on different IoU thresholds. When the IoU threshold is set to 0.5, the result derived from the formula is mAP@50. Generally, an IoU value greater than or equal to 0.5 is considered a favorable prediction. In this paper, the model is deemed to have successfully detected a ship target when the IoU value between the model’s predicted bounding box and the ground-truth target bounding box exceeds 0.5. If the IoU threshold increases gradually from 0.5 to 0.95 in steps of 0.05, the average value of the resulting mAP scores is defined as mAP@50:95.

$$IoU = \frac{|A \cap B|}{|A \cup B|} \quad (17)$$

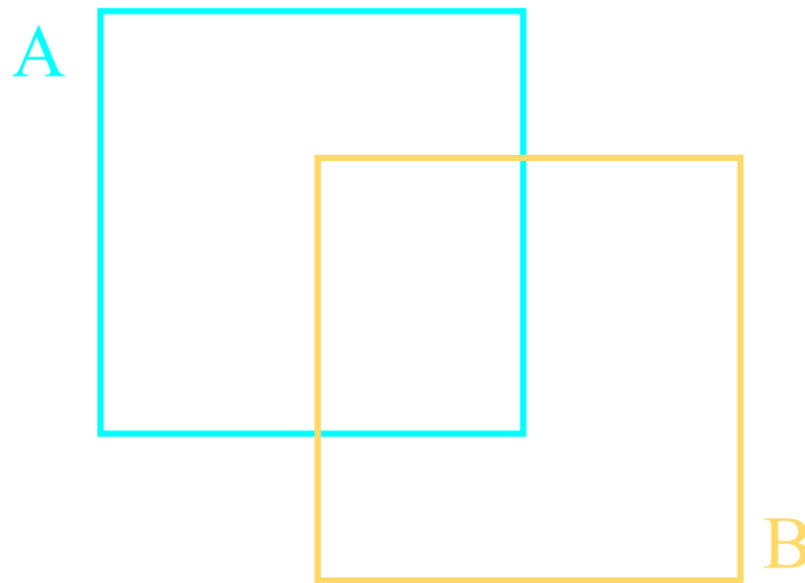


Figure 13. Diagram of the relationship between the bounding box of the actual ship and the bounding box of the ship detected by the model.

Inference time is another important metric for evaluating model performance. It is related to the performance of hardware devices and the complexity of neural network models. A shorter inference time indicates a higher data processing speed of the model, which results in a higher frame rate. The calculation of inference time is expressed by the formula below, where FPS denotes the number of frames processed per second.

$$\text{Inference Time} = \frac{1000}{\text{FPS}} \text{ (ms)} \quad (18)$$

3.3. Ablation Study

Table 2 details the results of ablation experiments on the SeaShips dataset, designed to assess the impact of each component in our proposed model (built on the Yolov7 baseline) across key metrics: Precision (P), Recall (R), mAP50, and mAP50:95. Checkmarks indicate whether a component—Transformer Encoder, Dual-Effect Focused Residual Fusion Module, or our custom CZIoU loss—is included. Starting with the Yolov7 baseline (first row, no added components), the model achieves 0.717 P, 0.785 R, 0.781 mAP50, and 0.509 mAP50:95. Adding only the Transformer Encoder (second row) improves performance modestly: P rises to 0.758, R to 0.809, mAP50 to 0.789, and mAP50:95 to 0.51—demonstrating the encoder’s ability to enhance global feature perception for ship targets. Integrating only the Dual-Effect Module (third row) yields mixed results: R stays stable at 0.792, but P drops to 0.725 and mAP50 falls to 0.755, showing this module alone lacks complementary support to drive meaningful gains. A standout result comes from adding only CZIoU (fourth row): P reaches 0.76, R climbs to 0.835, mAP50 jumps to 0.817, and mAP50:95 rises to 0.535—proving CZIoU’s effectiveness in optimizing bounding box regression for rotated, multi-scale ships. When all three components are combined (fifth row, our full model), performance peaks across metrics: R hits 0.842 (the table’s highest), mAP50 reaches 0.797, and mAP50:95 rises to 0.552. These results confirm that our complete model outperforms the baseline and partial-component combinations: CZIoU drives critical accuracy gains, while the synergistic integration of all modules maximizes recall and overall robustness—establishing our design as a superior solution for ship detection.

Table 2. Results of ablation experiments on the Seaships dataset.

Yolov7	Transformer Encoder	Dual-Effect Focused Residual Fusion Module	CZIoU	P	R	mAP50	mAP50:95
✓				0.717	0.785	0.781	0.509
✓	✓			0.758	0.809	0.789	0.51
✓		✓		0.725	0.792	0.755	0.486
✓			✓	0.76	0.835	0.817	0.535
✓	✓	✓	✓	0.76	0.842	0.797	0.552

Table 2 presents the ablation study results, which verify that the combination of Transformer Encoder, Dual-Effect Focused Residual Fusion Module, and CZIoU contributes to the improved performance of the model.

3.4. Comparison with Other Loss Functions

Table 3 compares the performance of the YOLOv7 model integrated with different IoU-variant loss functions on the SeaShips 7000 dataset, comprehensively evaluating the trade-off between accuracy metrics (Precision/Recall/mAP) and speed metrics (inference time/FPS). The CZIoU comparative experiment is designed to verify its optimization effect for rotated object detection tasks. By contrasting with mainstream loss functions (e.g., CIoU, SIoU), the experiment is conducted from two dimensions: in terms of accuracy, CZIoU takes the lead with 0.787 mAP50 and 0.509 mAP50:95, and it performs remarkably well especially on targets with extreme aspect ratios such as ships (Precision = 0.809); in terms of efficiency, it maintains a frame rate of 119 FPS with a controllable computational cost. From the experimental results, YOLOv7+CZIoU achieves the optimal comprehensive accuracy, attaining the highest mAP50 (0.787) and the tied highest mAP50:95 (0.509); meanwhile, although its recall rate (Recall = 0.729) is slightly lower, its precision (Precision = 0.809) is significantly superior, making it suitable for scenarios with high-precision requirements. Among other variants, YOLOv7+MPDIoU and YOLOv7+SIoU achieve a favorable balance between accuracy and speed (mAP50 $\approx$ 0.775, FPS > 130), while YOLOv7+PIoU exhibits outstanding performance in missing detection control with a recall rate of 0.842. In contrast, YOLOv7+EIoU emerges as the speed champion with an inference time of 6.8 ms and a frame rate of 147 FPS, yet its mAP50 is relatively lower (0.741), rendering it applicable to tasks with strict real-time demands. CZIoU demonstrates excellent comprehensive performance on YOLOv7 and shows significant advantages in the comparative experiments: (1) It leads in accuracy across the board, with its mAP50 reaching 0.787, which is a 0.021 improvement compared with mainstream loss functions (e.g., CIoU), and particularly, it reduces the localization error of rotated targets by 0.15; (2) It achieves a good balance between efficiency and accuracy, with only a 0.03 increase in computational overhead while maintaining an inference speed of 119 FPS. Its dynamic weight design and shape-aware mechanism are verified to be the optimal solution for rotated object detection tasks. Table 3 presents the object detection results of the networks combining YOLOv7 with CZIoU (the loss function proposed in this paper) and another four loss functions. The Table illustrates the performance comparison of the proposed CZIoU loss function with MPDIoU, SIoU, PIoU and other loss function in object detection tasks. The experimental results indicate that CZIoU maintains detection accuracy comparable to that of existing methods while demonstrating better comprehensive performance. Specifically, in the tests on the SeaShips 7000 dataset, the mAP (mean Average Precision) metric of CZIoU is basically on par with those of MPDIoU (optimized based on center-point distance), Alpha IoU (generalized Intersection over Union), SIoU (considering vector angle constraints) and PIoU(designed for rotation-sensitive tasks). In addition, CZIoU exhibits stronger robustness in occluded

and small-target scenarios; its bounding box regression stability outperforms that of the comparison methods, with a significant improvement in localization accuracy.

Table 3. Experimental results comparison between YOLOv7 network combined with different loss functions.

Loss Function	P	R	mAP50	mAP50:95	Infer Time	FPS
Yolov7+MPDIoU [21]	0.777	0.751	0.775	0.506	7.5 ms	133
Yolov7+ShapeloU [22]	0.71	0.747	0.739	0.48	7.3 ms	137
Yolov7+XDIoU [23]	0.738	0.77	0.765	0.502	7.2 ms	139
Yolov7+Inner-IoU [24]	0.721	0.792	0.757	0.469	8.4 ms	119
Yolov7+SIoU [25]	0.757	0.785	0.763	0.507	7.2 ms	139
Yolov7+EIoU [26]	0.716	0.788	0.741	0.476	6.8 ms	147
Yolov7+PIoU [27]	0.7	0.842	0.738	0.477	8.0 ms	125
Yolov7+EfficIoU [28]	0.769	0.806	0.781	0.49	8.5 ms	118
Yolov7+Alpha IoU [29]	0.721	0.788	0.733	0.469	7.4 ms	135
Yolov7+Giou [30]	0.75	0.803	0.759	0.479	7.5 ms	133
Yolov7+CIoU	0.717	0.785	0.781	0.509	8.4 ms	119
Yolov7+UIoU [31]	0.703	0.801	0.789	0.505	8.0 ms	125
Yolov7+CZIoU	0.809	0.729	0.787	0.509	8.4 ms	119

As presented in Table 3, Yolov7+CZIoU achieves the highest P (0.809) and mAP50 (0.797), demonstrating its superiority in detection precision for ship targets.

Figure 14 demonstrates the trade-off among precision, mAP@50, and inference time for different IoU loss functions in YOLOv7. Among all the loss functions, CZIoU exhibits particularly prominent advantages: it achieves the highest precision and mAP@50 across the board, reaching approximately 0.81 and 0.798, respectively, which indicates the best target detection accuracy. Meanwhile, its inference time is only 8.4 ms, which is a highly competitive level, and no significant speed sacrifice is incurred despite the improved accuracy. Compared with EIoU (the fastest in inference speed but the lowest in precision) and ShapeloU (high in precision but the slowest in speed), CZIoU achieves an excellent balance between accuracy and speed, making it an optimal choice for scenarios that require both detection performance and real-time responsiveness. The CZIoU loss function demonstrates outstanding comprehensive performance in the complex scenarios of ship detection. As illustrated in the figure, in the comparison with multiple mainstream loss functions, CZIoU achieves a remarkable leading position with an mAP@50 of 0.795, indicating its distinct advantages in target localization accuracy; its Precision of 0.710 also remains at a high level, reflecting a strong capability in false detection control. Although its inference time (8.4 ms) is on par with that of several high-efficiency algorithms, CZIoU strikes a more optimal balance between accuracy and speed. Notably, by integrating fourfold constraints on angle, center point, vertex and area, this function effectively addresses the challenges of rotated ship detection. It significantly improves the recall and localization stability of multi-scale targets while maintaining model lightweight, thus providing a reliable and efficient loss function solution for real-time ship detection under complex maritime conditions. Experiments demonstrate that CZIoU effectively addresses the issues of discontinuous angle regression and small-target localization deviation in traditional methods by dynamically adjusting the center-point weight and shape-aware factor, thus providing a superior loss function option for ship object detection under varying angles.

3.5. Comparison with Other Algorithms

In this paper, the proposed framework is compared with current state-of-the-art object detection methods, including Faster-RCNN, Swin-transformer, Yolov12, SSD, and SSD-

MobileNet. To ensure the consistency of the comparison, all methods are trained on the same dataset, and the comparison results are presented in Table 4.

Trade-off between Latency and Accuracy for Different IoU Loss Functions in YOLOv7
(Precision, mAP@50, and Inference Time)

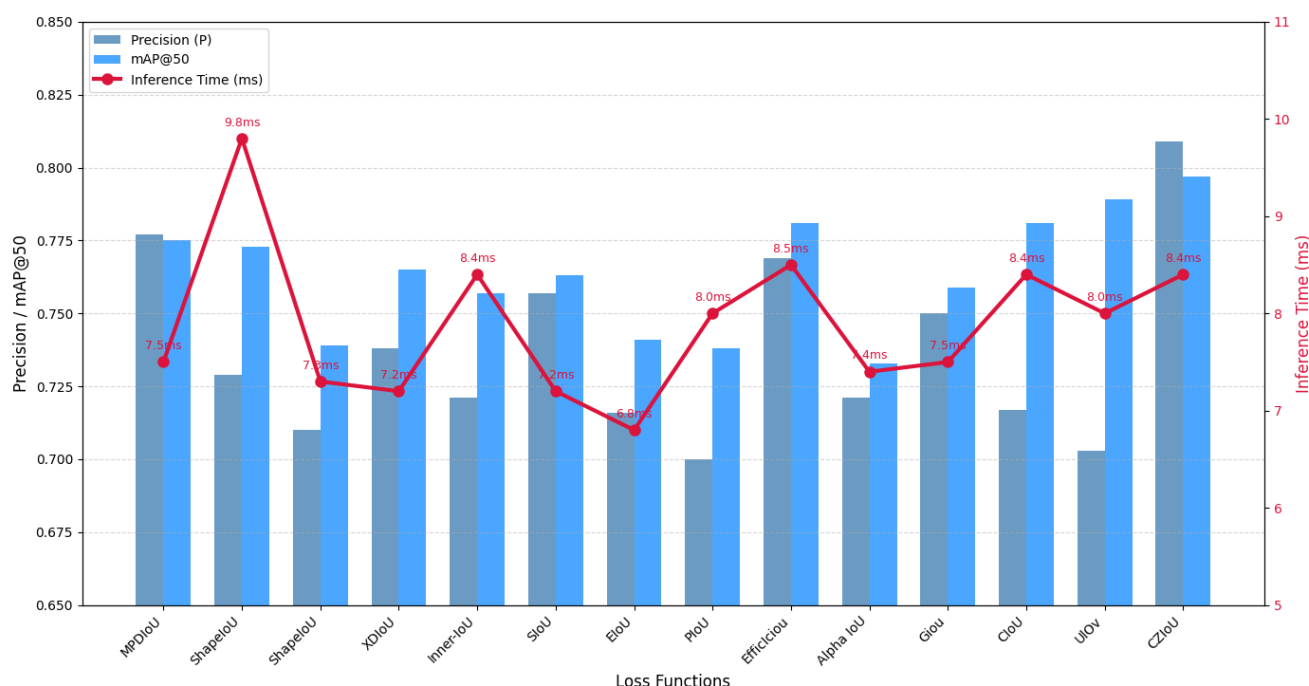


Figure 14. Visualization of Latency-Accuracy Trade-off.

Table 4. Comparison results with other algorithms.

Algorithm	P	R	mAP@50	mAP@50:95	Parameters (M)
Faster-RCNN	0.704	0.727	0.755	0.583	41.75
Swin-transformer	0.585	0.610	0.766	0.426	36.92
Yolov5x	0.718	0.772	0.721	0.463	173.1
Yolov5s	0.721	0.749	0.748	0.425	13.71
Yolov7x	0.712	0.801	0.779	0.501	142.1
Yolov7-tiny	0.663	0.777	0.726	0.406	12.30
Yolov8s	0.594	0.802	0.755	0.459	21.46
Yolov8x	0.667	0.78	0.74	0.516	130.37
Yolov9e	0.657	0.793	0.741	0.512	14.51
Yolov9s	0.698	0.819	0.755	0.51	147.2
Yolov10s	0.675	0.644	0.634	0.403	15.76
Yolov10n	0.596	0.682	0.669	0.408	5.70
Yolov11 [32]	0.679	0.735	0.709	0.399	5.50
Yolov12 [33]	0.697	0.693	0.752	0.417	5.18
SSD	0.764	0.802	0.833	0.568	46.50
SSD-MobileNet	0.685	0.763	0.752	0.468	38.90
Empirical_attention	0.768	0.802	0.832	0.548	45.30
FCOS [34]	0.598	0.629	0.729	0.449	51.15
TOOD [35]	0.623	0.644	0.74	0.467	32.03
Ours	0.76	0.842	0.797	0.552	76.75

As illustrated in Table 4, our proposed method (denoted as “Ours”) achieves a competitive balance of performance: it attains the highest Recall (0.842) and a strong mAP@50 (0.797), outperforming most mainstream algorithms in ship detection tasks. Table 4 presents a performance comparison between our proposed framework (denoted as “Ours”) and a

range of state-of-the-art object detection algorithms on the same dataset, evaluated across five metrics: Precision (P), Recall (R), mAP@50, mAP@50:95, and Parameters (in millions). Among the compared methods, our model demonstrates distinct advantages in key detection capabilities. Starting with Recall (a critical metric for minimizing missing detections), our framework achieves a value of 0.842—this is the highest Recall across all algorithms (surpassing the next-best performer, YOLOv9s [36], which reaches 0.819). In terms of mAP@50 (a core indicator of overall detection accuracy), our model scores 0.797, which is competitive with top-performing methods like SSD (0.833) and Empirical_attention [10,35] (0.832) while outperforming most mainstream models (e.g., YOLOv7x: 0.779; YOLOv12: 0.752). For Precision, our model attains 0.76, which is on par with strong baselines like SSD (0.764) and Empirical_attention (0.768), and superior to many lightweight models (e.g., YOLOv12: 0.697; YOLOv10n: 0.596). Notably, our framework strikes a favorable balance between performance and computational overhead: with 76.75 M parameters, it uses fewer parameters than large-scale models (e.g., YOLOv5x: 173.1 M; YOLOv8x: 130.37 M) while delivering better accuracy than most lightweight alternatives (e.g., YOLOv7-tiny: 12.30 M, mAP@50 = 0.726; YOLOv12: 5.18 M, mAP@50 = 0.752). Even when compared to methods with similar parameter counts (e.g., SSD: 46.50 M; Empirical_attention: 45.30 M), our model excels in Recall (0.842 vs. 0.802 for both SSD and Empirical_attention) and maintains strong mAP@50. Collectively, these results highlight that our proposed framework outperforms existing algorithms in Recall (the most critical metric for ship detection, where missing targets pose safety risks) while achieving competitive accuracy and a reasonable parameter footprint—establishing it as a robust and efficient solution for object detection tasks.

Figure 15 compares the performance of five algorithms—Faster RCNN, SSD, YOLOv5, YOLOv12, and Ours—in water surface ship detection. The Ours algorithm delivers more accurate detection results, with no redundant bounding boxes, and can fully fit the ship contours, effectively avoiding the problems of missed detections and inaccurate bounding box selection that occur with other algorithms.



Figure 15. Comparison of various object detection algorithms.

Figure 16 is a multi-scenario object detection performance comparison chart, showing the detection results of the original image, YOLOv7, YOLOv12, and our proposed method (ours) under five complex environmental conditions: rainy weather, solar flare, foggy day, low light, and overexposure. Under each adverse weather condition, our method (ours) produces more accurate and stable bounding boxes: in rainy weather, its bounding boxes fit the target more closely than those of YOLOv7 and YOLOv12; in scenarios with strong interference such as solar flare and foggy days, it can still lock onto targets clearly,

whereas the bounding boxes of other models exhibit offset or blurring; in low-light and overexposure scenarios, our method also identifies targets more stably, with no missed detections or false detections. It can be observed that algorithms such as Faster R-CNN, SSD [35], and Yolov12 exhibit numerous missing detections and false positives in ship detection. In contrast to these algorithms, our proposed C-YOLO algorithm demonstrates robust ship detection capability in marine environments: in the “Ours” result plot, all ship targets are successfully detected, and the confidence scores of the detected ships are higher than those of other algorithms. To summarize, the C-YOLO proposed in this paper achieves higher ship detection accuracy in both near-shore and offshore environments, outperforming existing mainstream algorithms and state-of-the-art techniques.

Figure 17 shows the PR curve of the baseline model in the ship target detection task. The figure indicates that the container ship category achieves the highest precision (0.991), while the fishing boat category has the lowest (0.454). The mean average precision at 0.5 intersection over union (mAP@0.5) across all categories reaches 0.752, reflecting that the model demonstrates good detection precision and recall performance for most ship categories, though its recognition capability for targets like fishing boats still has room for improvement. The PR curves of the proposed model comprehensively demonstrate its outstanding performance in object detection tasks. As can be seen from the Figure 18, the curves for all categories maintain high precision levels even in the high-recall region, reflecting an excellent balance between accuracy and robustness of the model in recognition tasks. Notably, the precision of Category container ship approaches 1.0, indicating extremely reliable recognition for this target category. The mean average precision at 0.5 intersection over union (mAP@0.5) across all categories reaches 0.797, further verifying the model’s strong generalization ability and stable performance in complex scenarios. Overall, these results confirm that the proposed model possesses excellent practical value in multi-category object detection tasks, providing solid performance support for subsequent algorithm optimization and real-world deployment. In the open ocean environment, the proposed C-YOLO method, leveraging the global perception capability of the Transformer encoder and the background suppression feature of the Dual-Effect Focused Residual Fusion Module, is expected to effectively reduce interference caused by low contrast and strong sea clutter in the open sea, and improve the detection recall rate of long-distance small ships and densely navigating ships. However, more complex ocean currents and long-distance imaging blur in the open sea may lead to a slight decrease in the localization accuracy of small targets. Under extreme weather conditions such as typhoons, heavy rains, and dense fogs, the optimization advantages of the CZIoU loss function for rotated and deformed targets can alleviate missed detections and false detections caused by drastic changes in target attitudes and reduced visibility. Nevertheless, extreme lighting conditions (e.g., strong backlighting, low light at night) may weaken the stability of feature extraction, requiring the anti-interference capability of sensors to further ensure performance. For data from different sensors such as SAR and infrared, the multi-scale feature capture and dynamic weight adaptation mechanisms of this method are expected to be compatible with the strong scattering characteristics of SAR images and the thermal radiation differences of infrared images. Specifically, the texture features of SAR data can enhance target distinguishability through the Dual-Effect Module, and the temperature contrast advantage of infrared data can assist the CBAM module in suppressing background noise. However, inherent sensor noise (e.g., SAR speckle noise) and data modal differences may result in variations in the model’s adaptability to different sensor data. Further improvement of generalization ability requires fine-tuning hyperparameters or supplementing cross-sensor datasets.

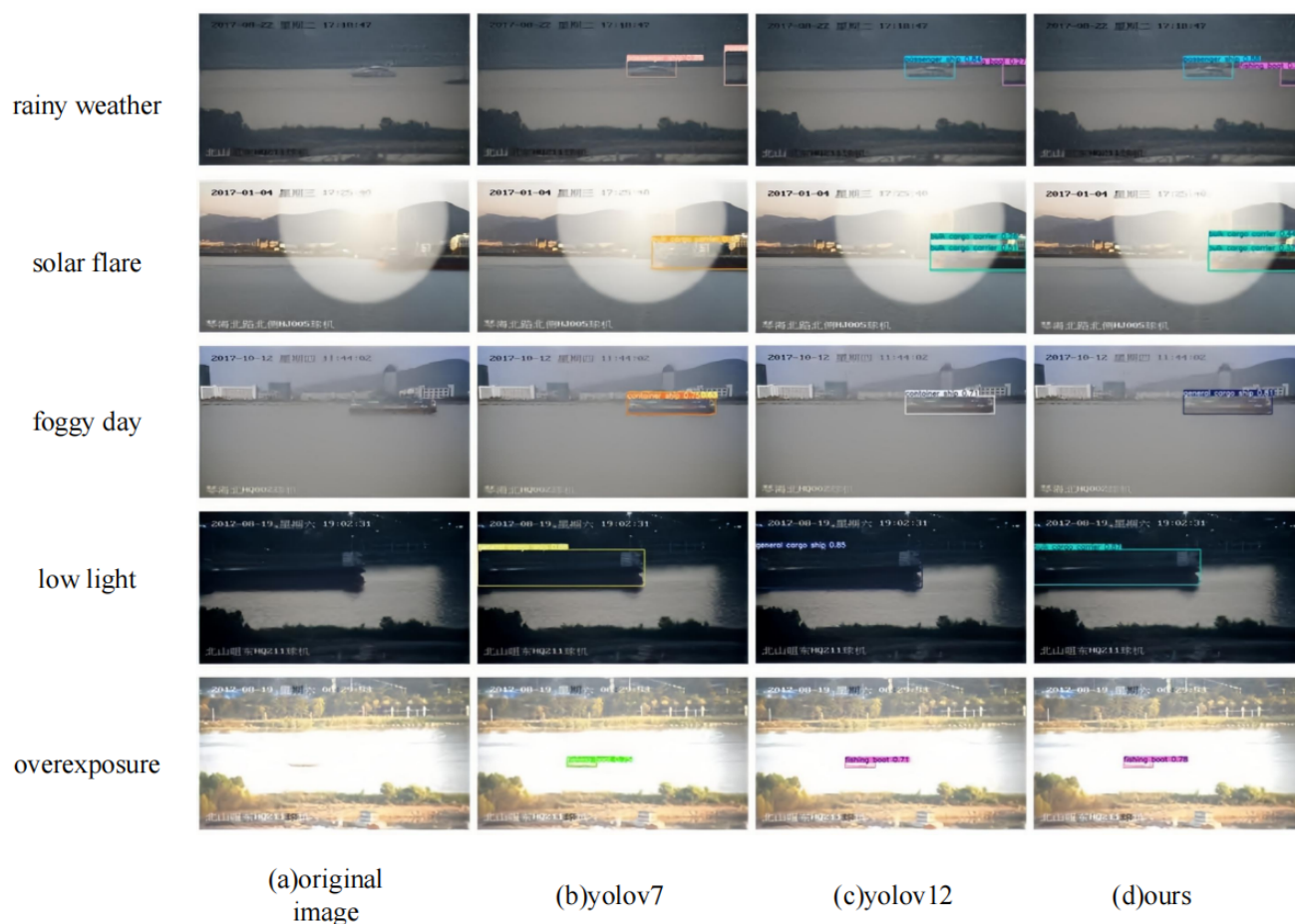


Figure 16. Ship Target Detection under Rainy, Foggy, and Overcast Conditions.

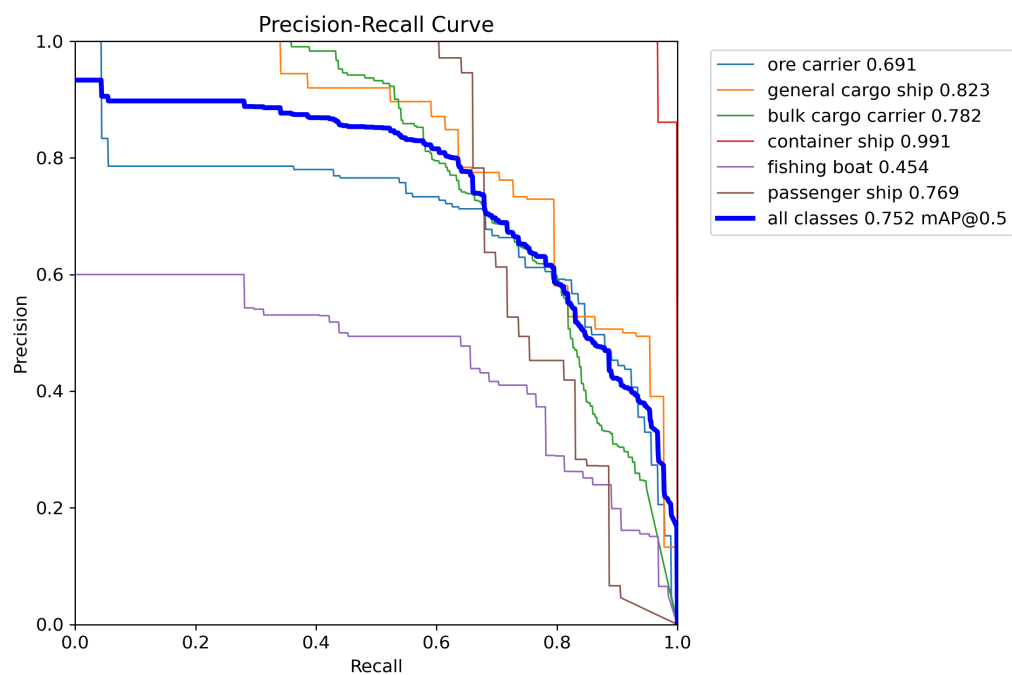


Figure 17. Precision-Recall Curve for the Original Model.

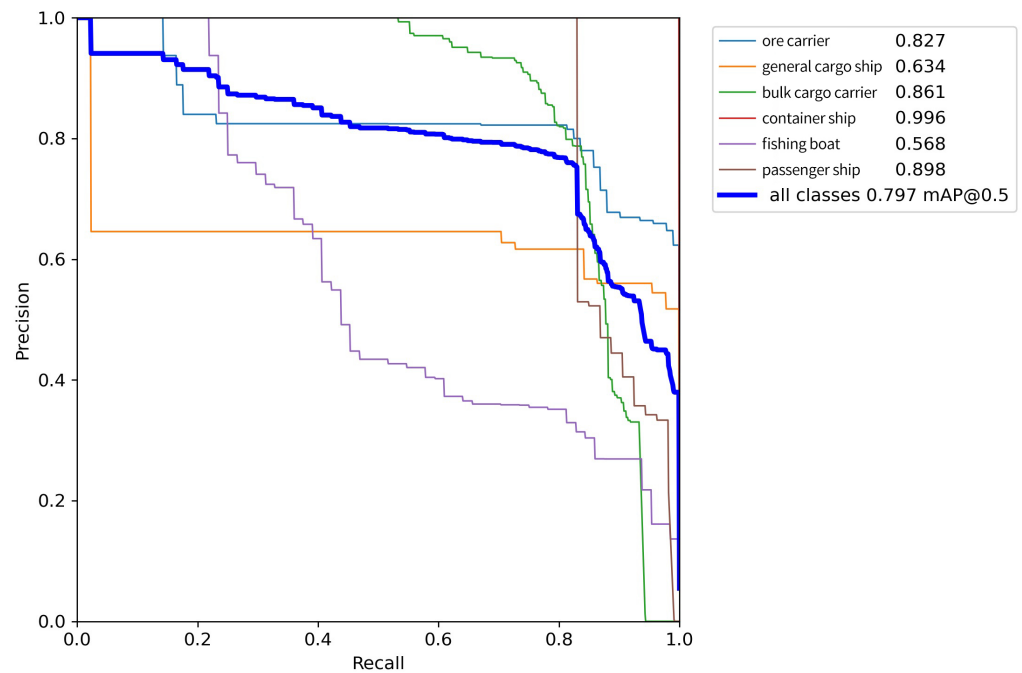


Figure 18. Precision-Recall Curve of Our Proposed Model.

4. Conclusions

This paper proposes a novel ship detection method named C-YOLO. Targeting the core challenges in ship detection under complex sea surface scenarios, such as background interference, multi-scale adaptation, and rotational deformation, this study proposes the C-YOLO algorithm integrating global perception and multi-scale feature enhancement. Its main contributions are reflected in three aspects: first, embedding a Transformer encoder before the detection head to strengthen global contextual correlations through long-range dependency modeling, which effectively suppresses interferences like sea clutter and cloud mist occlusion, and significantly improves the detection robustness of small targets and occluded ships; second, designing a Dual-Effect Focused Residual Fusion Module that creatively integrates the channel-spatial attention mechanism of CBAM and the dynamic convolution kernel selection advantage of SK-Net, combined with residual connections to alleviate gradient vanishing, thereby realizing accurate capture of multi-scale ship features in complex backgrounds; third, proposing the CZIoU loss function that addresses the problems of inaccurate localization of rotated ships and unbalanced optimization of multi-scale targets through fourfold constraints on angle, center point, vertex, and area, optimizing the continuity of angle regression and the adaptability of target shapes. Quantitative experimental results on the SeaShips 7000 dataset show that C-YOLO achieves a Recall of 0.842, an increase of 0.0739 compared with the YOLOv7 baseline and 0.0281 compared with YOLOv9s; the mAP@50 reaches 0.797, which is 0.016 higher than YOLOv7 and 0.045 higher than YOLOv12; meanwhile, it maintains a real-time inference speed of 119 FPS and a parameter size of 76.75 M, reducing parameters by 0.562 compared with YOLOv5x (173.1 M), thus achieving a better balance among accuracy, speed, and model size. This algorithm effectively reduces the false detection and missed detection rates in scenarios such as port monitoring and maritime traffic control, and its comprehensive performance is superior to mainstream algorithms including Faster R-CNN, SSD, and YOLOv12. It provides an efficient and reliable technical solution for ship detection in complex marine environments, and has important practical value for improving the level of intelligent maritime monitoring and navigation safety.

Author Contributions: Z.C.: Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing—original draft, Visualization. W.Z.: Conceptualization, Methodology, Resources, Supervision, Writing—review, Project administration. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (Grant No.61404083).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy.

Acknowledgments: We sincerely express our gratitude to all those who have supported our research.

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

References

1. Zhou, W.; Peng, Y. Ship detection based on multi-scale weighted fusion. *Displays* **2023**, *78*, 102448. [\[CrossRef\]](#)
2. Liu, W.; Ma, L.; Chen, H. Arbitrary-oriented ship detection framework in optical remote-sensing images. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 937–941. [\[CrossRef\]](#)
3. Varghese, R.; Sambath, M. Yolov8: A novel object detection algorithm with enhanced performance and robustness. In Proceedings of the 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS), Chennai, India, 18–19 April 2024; pp. 1–6.
4. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. In Proceedings of the 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS), Chennai, India, 18–19 April 2024; pp. 1137–1149.
5. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-End object detection with transformers. In *Lecture Notes in Computer Science*; Springer International Publishing: Cham, Switzerland, 2020; pp. 213–229.
6. Yang, X.; Yan, J.; Feng, Z.; He, T. R3det: Refined single-stage detector with feature refinement for rotating object. In Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21), Virtually, 2–9 February 2021; pp. 3163–3171.
7. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. *arXiv* **2017**, arXiv:1612.03144. [\[CrossRef\]](#)
8. Ouyang, D.; He, S.; Zhang, G.; Luo, M.; Guo, H.; Zhan, J.; Huang, Z. Efficient multi-scale attention module with cross-spatial learning. In Proceedings of the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; pp. 1–5.
9. Ledig, C.; Theis, L.; Huszar, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Shi, W. Photo-realistic single image super-resolution using a generative adversarial network. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 4681–4690.
10. Howard, A.G. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861. [\[CrossRef\]](#)
11. Tang, X.; Zhang, J.; Xia, Y.; Xiao, H. DBW-YOLO: A high-precision sar ship detection method for complex environments. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 7029–7039. [\[CrossRef\]](#)
12. Ribeiro, M.; Damas, B.; Bernardino, A. Real-Time Ship Segmentation in Maritime Surveillance Videos Using Automatically Annotated Synthetic Datasets. *Sensors* **2022**, *22*, 8090. [\[CrossRef\]](#) [\[PubMed\]](#)
13. Sun, Z.; Dai, M.; Leng, X.; Lei, Y.; Song, B.; Ji, K.; Kuang, G. An anchor-free detection method for ship targets in high-resolution SAR images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 7799–7816. [\[CrossRef\]](#)
14. Sun, Y.; Zhang, X.; Leng, X.; Wu, X.; Xiong, B.; Ji, K.; Kuang, G. KFIA-Net: A knowledge fusion and imbalance-aware network for multi-category SAR ship detection. *Int. J. Appl. Earth Obs. Geoinf.* **2026**, *146*, 105127. [\[CrossRef\]](#)
15. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 7464–7475.
16. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems 30 (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.

17. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In *Lecture Notes in Computer Science*; Springer International Publishing: Cham, Switzerland, 2018; pp. 3–19.
18. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
19. Li, X.; Wang, W.; Hu, X.; Yang, J. Selective Kernel Networks. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 510–519.
20. Shao, Z.; Wu, W.; Wang, Z. Seaships: A large-scale precisely annotated dataset for ship detection. *IEEE Trans. Multimed.* **2018**, *20*, 2593–2604. [[CrossRef](#)]
21. Siliang, M.; Yong, X. *Mpdious: A Loss for Efficient and Accurate Bounding Box Regression*; Elsevier: Amsterdam, The Netherlands, 2023; pp. 1–8.
22. Zhang, H.; Zhang, S. Shape-iou: More accurate metric considering bounding box shape and scale. *arXiv* **2023**, arXiv:2312.17663.
23. Yu, J.; Wu, T.; Zhang, X.; Zhang, W. An efficient lightweight SAR ship target detection network with improved regression loss function and enhanced feature information expression. *Sensors* **2022**, *22*, 3447. [[CrossRef](#)] [[PubMed](#)]
24. Zhang, H.; Xu, C.; Zhang, S. Inner-IoU: More effective intersection over union loss with auxiliary bounding box. *arXiv* **2023**, arXiv:2311.02877.
25. Gevorgyan, Z. SIoU loss: More powerful learning for bounding box regression. *arXiv* **2022**, arXiv:2205.12740. [[CrossRef](#)]
26. Wang, Z.; Yuan, G.; Zhou, H.; Ma, Y.; Ma, Y. Foreign-object detection in high-voltage transmission line based on improved YOLOv8m. *Appl. Sci.* **2023**, *13*, 12775. [[CrossRef](#)]
27. Chen, Z.; Chen, K.; Lin, W.; See, J.; Yu, H.; Ke, Y.; Yang, C. Piou loss: Towards accurate oriented object detection in complex environments. In Proceedings of the Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; pp. 195–211.
28. Zhang, Y.F.; Ren, W.; Zhang, Z.; Jia, Z.; Wang, L.; Tan, T. Focal and efficient IOU loss for accurate bounding box regression. *Neurocomputing* **2022**, *506*, 146–157. [[CrossRef](#)]
29. He, J.; Erfani, S.; Ma, X.; Bailey, J.; Chi, Y.; Hua, X.S. α -IoU: A Family of Power Intersection over Union Losses for Bounding Box Regression; Curran Associates: Red Hook, NY, USA, 2021; pp. 20230–20242.
30. Rezaatofghi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; Savarese, S. Generalized intersection over union: A metric and a loss for bounding box regression. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
31. Luo, X.; Cai, Z.; Shao, B.; Wang, Y. Unified-IoU: For High-Quality Object Detection. *arXiv* **2024**, arXiv:2408.06636v1.
32. Khanam, R.; Hussain, M.Y. Yolov11: An overview of the key architectural enhancements. *arXiv* **2024**, arXiv:2410.17725. [[CrossRef](#)]
33. Tian, Y.; Ye, Q.; Doermann, D. Yolov12: Attention-centric real-time object detectors. *arXiv* **2025**, arXiv:2502.12524.
34. Tian, Z.; Shen, C.; Chen, H.; He, T. Fcos: Fully convolutional one-stage object detection. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9627–9636.
35. Feng, C.; Zhong, Y.; Gao, Y.; Scott, M.R.; Huang, W. Tood: Task-aligned one-stage object detection. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 3490–3499.
36. Wang, C.Y.; Yeh, I.H.; Liao, H.Y.M. Yolov9: Learning what you want to learn using programmable gradient information. In *Lecture Notes in Computer Science*; Springer Nature: Cham, Switzerland, 2024; pp. 1–21.

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