Journal of Computing and Electronic Information Management
ISSN: 2413-1660 | Vol. 20, No. 2, 2026

Research on Rotated Target Detection Method for Maritime Vessels in UAV Aerial Images Based on Deep Learning

Ye Zhang

School of Computer Science, Xi’an Shiyou University, Xi’an, 710065, China;

Abstract: Aerial imagery captured by unmanned aerial vehicles (UAVs) over maritime scenes is characterized by a wide field of view, homogeneous backgrounds, and the presence of densely packed vessels with diverse orientations. Traditional horizontal bounding box detection methods applied in such scenarios often introduce excessive background noise and fail to accurately represent vessel orientations, resulting in inadequate differentiation of densely moored vessels. To address this issue, this paper proposes an improved rotated object detection method. Building upon the YOLOv8-OBB model with inherent rotation detection capability, the proposed method incorporates a Coordinate Attention (CA) module to enhance feature extraction for elongated vessel targets, optimizes the Feature Pyramid Network (FPN) structure with cross-level skip connections to strengthen multi-scale feature fusion, and employs a Skew Intersection over Union (SkewIoU) loss function to improve the regression accuracy of rotated bounding boxes. For comprehensive performance evaluation, experiments are conducted on the public maritime dataset HRSC2016 and a self-constructed nearshore vessel subset. Results indicate that the improved model achieves a mean Average Precision (mAP@0.5) of 90.5% on the HRSC2016 test set, representing a 1.3 percentage point improvement over the baseline model (89.2%). Ablation studies demonstrate the consistent contributions of each enhancement module, and the model exhibits significantly improved robustness in detecting dense and small-scale vessels. This work provides a more accurate solution for maritime surveillance, search and rescue, and related applications.

Keywords: UAV aerial imagery; Rotated object detection; Maritime vessels; YOLOv8; Coordinate attention; Feature pyramid network.

1. Introduction

1.1. Research Background and Significance

With the rapid development of UAV technology, object detection in UAV-captured aerial images has gradually become an important research direction in the field of computer vision [1]. Due to advantages such as high mobility, low cost, and versatility, UAVs have been widely applied in various fields including agricultural management [2], environmental monitoring [3], urban security, and disaster response [4], demonstrating broad technical potential and societal value. These practical demands in real-world scenarios have propelled object detection in UAV aerial imagery to the forefront of computer vision research.

In agricultural management, UAV-based object detection technology can rapidly identify areas affected by pests and diseases in farmland, enabling timely intervention by agricultural managers. This not only improves agricultural production efficiency but also reduces resource waste. For environmental monitoring, UAVs can detect potential forest fire locations in advance and analyze water pollution situations, providing valuable data for ecological conservation [5]. In urban security, UAVs can monitor traffic conditions and crowd gatherings in real-time, contributing to more intelligent urban management and playing a significant role in traffic management and public safety [6].

1.2. Current Research Status at Home and Abroad

Object detection tasks can be broadly categorized into two types based on their frameworks: two-stage detection algorithms and one-stage detection algorithms. Two-stage detection algorithms first generate region proposals via a Region Proposal Network (RPN), followed by further classification and regression operations on these proposals [6]. Representative algorithms include Fast R-CNN [7], Faster R-CNN [8], and its improved variant Mask R-CNN [9]. One-stage detection algorithms generally generate feature maps through convolutional neural networks and directly compute object location and category information from these maps. Representative algorithms include YOLO [10] and SSD [11].

Rotated object detection is a core technology in the analysis of remote sensing and aerial imagery. Early methods often extended two-stage detectors like Faster R-CNN by introducing a Rotated Region Proposal Network (R-RPN) or adding an angle regression branch. In recent years, one-stage rotated detectors have garnered significant attention due to their efficiency, exemplified by models such as R3Det and S2ANet. The YOLO series, as a representative of one-stage detectors, has seen its rotated detection variant, YOLOv8-OBB, incorporate angle as a regression target. While maintaining high inference speed, it provides practical rotated detection capability and has become an important baseline model for engineering deployment.

For the detection of vessels in aerial maritime scenes, existing research still needs to address specific challenges: (1) Vessel targets often exhibit large aspect ratios and arbitrary orientations, requiring feature extraction that is sensitive to direction and geometric shape. (2) Sea surface backgrounds contain interference such as waves, glare, and distant horizons, increasing feature discrimination complexity. (3) There is an extremely wide scale distribution, with vessels ranging from cargo ships tens of meters long to small fishing boats a few meters long coexisting in the same frame, making small object detection difficult. Existing improvements primarily focus on designing more complex rotated box encodings or using larger backbone networks. However, there remains room for optimization in terms of efficient, lightweight feature enhancement specifically tailored for vessel targets.

1.3. Main Contributions of This Paper

This paper focuses on rotated object detection for maritime vessels in UAV aerial scenes, aiming to construct an efficient and robust practical detection solution by integrating proven lightweight enhancement techniques. The main contributions are as follows:

1. Orientation-aware Feature Enhancement: Embedding Coordinate Attention (CA) modules at key stages of the backbone network. These modules simultaneously model channel relationships and long-range spatial dependencies, aiding the network in capturing the precise orientation and contour features of elongated vessels.

2. Adaptive Feature Fusion Optimization: To address the scale diversity of vessels, cross-level skip connections and a simplified weighted fusion mechanism are introduced into the Feature Pyramid Network. This strengthens the feedback from shallow, high-resolution features to deeper layers, improving feature utilization for small-scale vessels.

3. Precise Regression Loss Design: Employing a SkewIoU loss function based on polygon computation to directly optimize the overlap area between predicted and ground-truth rotated bounding boxes. This makes the regression process for box location, size, and angle more consistent and precise.

4. Systematic Experimental Validation: Conducting comprehensive experiments on the public benchmark dataset HRSC2016 and a self-constructed supplementary dataset of nearshore scenes. Performance means and variation ranges are reported through multiple random experiments, accompanied by detailed ablation studies and error case analysis to objectively evaluate the effectiveness and limitations of the proposed approach.

1.4. Thesis Structure

The structure of this paper is as follows: Chapter 2 introduces related technologies for rotated object detection and maritime vessel detection. Chapter 3 details the proposed improvement methods. Chapter 4 describes the experimental setup, dataset construction, and analyzes the results. Chapter 5 summarizes the paper and discusses future work.

2. Related Work

2.1. Rotated Object Detection Methods

The core of rotated object detection lies in accurately regressing the five parameters of a target: center point (x, y), width (w), height (h), and rotation angle (θ). Mainstream paradigms include: 1) Rotated box regression paradigm, such as RoI Transformer and R3Det, which learn transformations to rotate parameters based on horizontal detection boxes; 2) Keypoint detection paradigm, such as the rotated extension of CenterNet, which defines rotated boxes by predicting the target center point and orientation vectors. YOLOv8-OBB adopts a one-stage design within the first paradigm, offering high efficiency. However, the periodicity of angle regression and boundary discontinuity remain key optimization challenges.

2.2. Application of Attention Mechanisms in Aerial Detection

Attention mechanisms selectively focus on key information by simulating the visual system. In aerial rotated detection, attention mechanisms that combine spatial and channel information are particularly important. Coordinate Attention (CA) encodes long-range spatial dependencies separately along the horizontal and vertical directions through decomposed convolution and embeds positional information into channel attention maps. This enables the network to precisely perceive the orientation and extent of targets, making it highly suitable for processing vessel targets with distinct directional characteristics.

2.3. Challenges in Maritime Vessel Detection

Detecting maritime vessels in UAV aerial imagery faces unique challenges: (1) Arbitrary Orientation: Vessel headings are distributed across 360 degrees, requiring detection boxes to rotate freely. (2) Extreme Scale Variation: Target sizes range from a few pixels to over a thousand pixels. (3) Large Intra-class Variance: Numerous vessel types exist with significant differences in appearance and structure. (4) Complex Interference: Background clutter such as sea waves, cloud shadows, islands, and port facilities can interfere with detection. (5) Density and Occlusion: Vessels in anchorages and ports are often densely packed and heavily occluded. (6) Real-time Constraints: UAV platforms demand lightweight algorithms with low latency.

3. The Proposed Improved Rotated Detection Method for Maritime Vessels

3.1. Baseline Network: YOLOv8-OBB

This paper uses YOLOv8-OBB as the baseline for improvements. Its main architecture is consistent with YOLOv8, comprising a Backbone, Neck, and Head. The key difference lies in its detection head, which outputs six dimensions: (x, y, w, h, θ, confidence), where θ is the rotation angle. The model employs a decoupled head structure and is trained end-to-end using a loss function designed for rotated boxes.

3.2. Enhancement with Coordinate Attention (CA) Module

To enhance the networks ability to perceive the spatial location and orientation of vessel targets, Coordinate Attention modules are embedded after the C2f modules in the backbone network. As illustrated in Figure 1, the operation of a CA module can be decomposed into two coordinated steps:

1. Coordinate Information Embedding: The input feature map undergoes separate 1D global average pooling along the X-axis and Y-axis, yielding a pair of feature vectors that encode global information in the height and width directions, respectively.

2. Coordinate Attention Generation: The two feature vectors are concatenated and undergo a shared 1×1 convolutional transformation. They are then split back into two separate direction-aware feature maps. Subsequently, Sigmoid functions are applied to generate attention weight maps for the height and width directions. Finally, these two attention maps are element-wise multiplied with the original input feature map to produce the enhanced output feature map.

This process recalibrates the feature map in the channel dimension while simultaneously assigning precise spatial positional weights. This can be formulated as Output = (Att_h ⊙ (Att_w ⊙ Input)), where ⊙ denotes element-wise multiplication. This enables the network to more accurately localize and focus on vessel targets with different orientations.

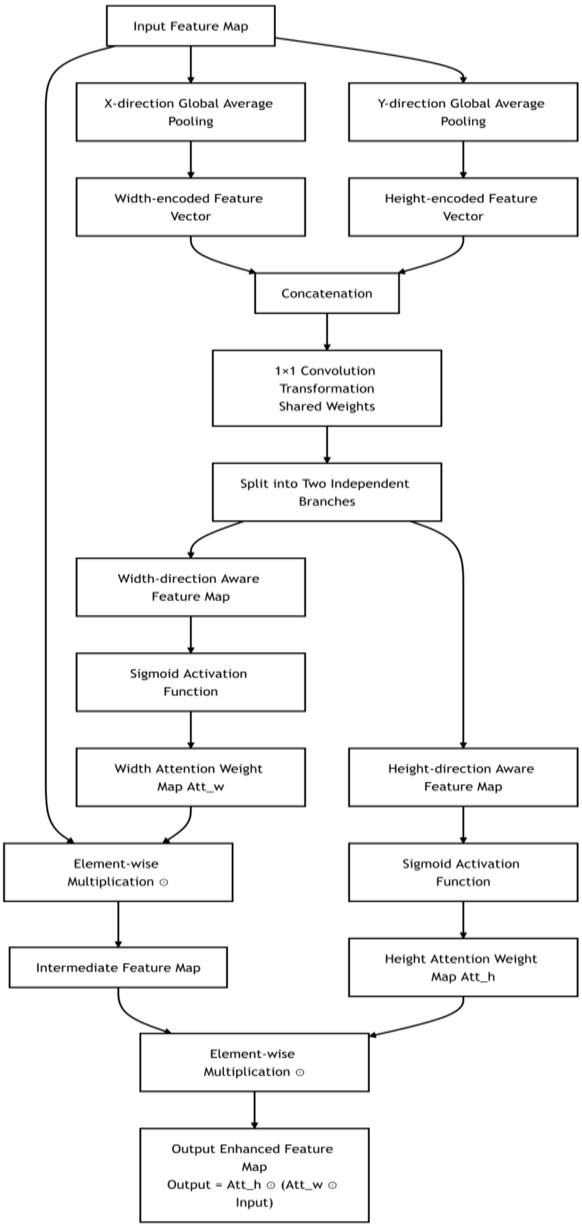
Input Feature Map
X-direction Global Average Pooling
Y-direction Global Average Pooling
Width-encoded Feature Vector
Height-encoded Feature Vector
Concatenation
1×1 Convolution Transformation Shared Weights
Split into Two Independent Branches
Width-direction Aware Feature Map
Sigmoid Activation Function
Width Attention Weight Map Att_w
Height-direction Aware Feature Map
Element-wise Multiplication ⊙
Sigmoid Activation Function
Intermediate Feature Map
Height Attention Weight Map Att_h
Element-wise Multiplication ⊙
Output Enhanced Feature Map Output = Att_h ⊙ (Att_w ⊙ Input)

Fig. 1 Operation of CA module

3.3. Cross-level Feature Pyramid Network Optimization

To address the problem of wide vessel scale distribution and the tendency for small target features to be lost in deep network layers, the Feature Pyramid structure within the neck network is optimized. Building upon the standard Top-Down fusion path, we design an enhanced cross-level connection (as shown in Figure 2). Specifically, after the deep, high-level semantic feature P5 is upsampled and fused with the mid-level feature C4 to obtain P4, P4 is further upsampled and undergoes secondary fusion with the shallower, high-resolution feature C3. This structure adds a direct feedback path from deep semantic features to shallow detail features, effectively strengthening the semantic information in shallow features used for detecting small-scale vessels. At the fusion nodes, a simple learnable scalar weight is employed to perform a weighted summation of features from different levels, achieving adaptive fusion.

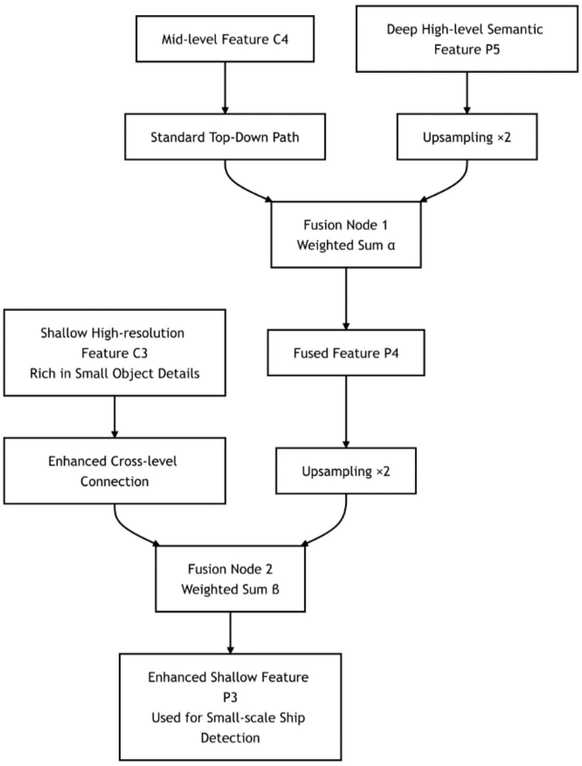
Mid-level Feature C4
Deep High-level Semantic Feature P5
Standard Top-Down Path
Upsampling ×2
Fusion Node 1 Weighted Sum α
Shallow High-resolution Feature C3 Rich in Small Object Details
Fused Feature P4
Enhanced Cross-level Connection
Upsampling ×2
Fusion Node 2 Weighted Sum β
Enhanced Shallow Feature P3 Used for Small-scale Ship Detection

Fig. 2 Enhanced Cross-Level Connections

3.4. SkewIoU Loss Function Optimization

The design of the loss function for rotated bounding box regression directly impacts localization accuracy. The loss used by the baseline model may suffer from discontinuity issues when handling angles. This paper employs the SkewIoU loss as a key component of the regression loss. Its core lies in directly calculating the polygonal overlap area (IoU) between the predicted rotated box and the ground-truth rotated box. The SkewIoU loss is defined as L_skew = 1 - SkewIoU. By minimizing this loss, the model is directly driven to maximize the overlap between the predicted and ground-truth boxes, thereby aligning the optimization objectives for the center point, width, height, and angle. To enable efficient training, a differentiable approximation algorithm is used to compute the gradient of SkewIoU. This loss, together with the classification loss, constitutes the total loss function for training.

3.5. 3.5 Overall Network Architecture

The overall network architecture proposed in this paper is illustrated in Figure 3. The input image undergoes preprocessing and is first passed through a backbone network embedded with CA modules to extract multi-level features. Subsequently, these features are fed into the improved neck network, which features cross-level connections, for multi-scale fusion. Finally, the fused multi-scale feature maps are input to a rotation-decoupled detection head, which outputs in parallel the targets class probability, confidence, and rotated bounding box parameters (cx, cy, w, h, θ). The entire design aims to specifically enhance the performance of rotated detection for maritime vessels without significantly increasing computational complexity.

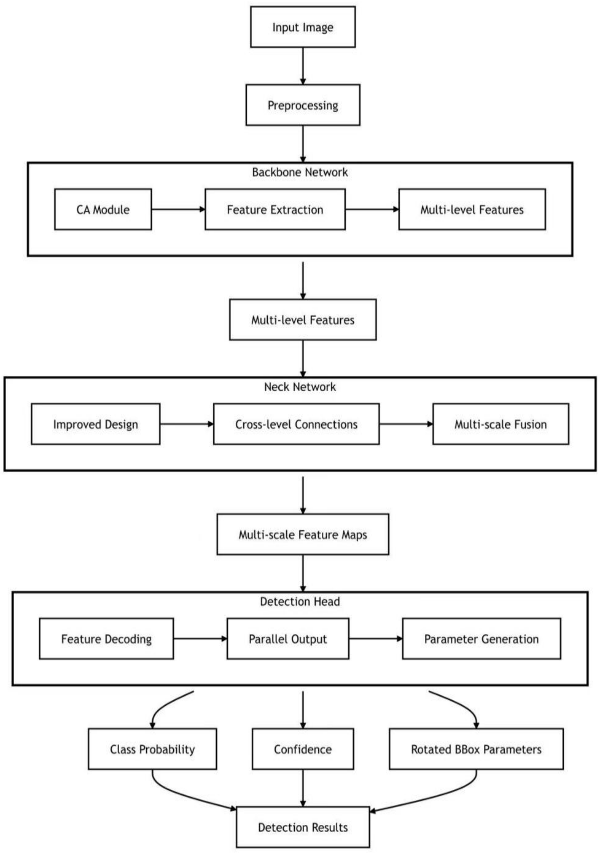

Fig. 3 Overall Network Structure

4. Experimental Design and Result Analysis

4.1. Experimental Environment and Datasets

Experimental Environment: Ubuntu 22.04 operating system, PyTorch 2.0.1 framework, NVIDIA RTX 4090 GPU (24GB). All experiments are conducted in a unified environment to ensure comparability.

· Public Dataset: The HRSC2016 benchmark dataset, widely used in the field of maritime rotated object detection, is employed. This dataset contains a large number of vessel targets of different sizes, types, and orientations. Following the official split, the training set (436 images) and the test set (444 images) are used.

Self-Constructed Supplementary Dataset: To evaluate the models generalization ability in complex nearshore scenarios, we filtered all "ship" category annotations from the training set of the DOTA-v1.0 dataset and manually removed non-maritime vessels and images with poor annotation quality to construct a nearshore vessel detection subset. This subset contains 1,248 images, covering various dense and complex background scenarios such as ports, coastal areas, and waterways. It is randomly divided into training, validation, and test sets in an 8:1:1 ratio. All rotated bounding box annotations follow the DOTA datasets four -point coordinate format (x1, y1, x2, y2, x3, y3, x4, y4).

Training Configuration: Input images are uniformly resized to 1024x1024. The Adam optimizer is used with an initial learning rate of 1e-3 and a cosine annealing scheduler. Models are trained for 300 epochs on HRSC2016 with a batch size of 8, and for 150 epochs on the supplementary dataset. Data augmentation includes random rotation (-45°, 45°), random cropping, mosaic augmentation, and color jittering. To assess result stability, each key experiment is conducted with 3 different random seeds for independent training. Reported performance metrics are the mean values and their standard deviations.

4.2. Evaluation Metrics

Standard evaluation metrics for rotated object detection are employed:

mean Average Precision (mAP): We report mAP at an Intersection over Union (IoU) threshold of 0.5 (mAP@0.5) and the average mAP over IoU thresholds from 0.5 to 0.95 with a step of 0.05 (mAP@0.5:0.95).

Angle Accuracy (AA): The proportion of predicted boxes whose angle error is less than a specific threshold (e.g., 5 degrees, 10 degrees) compared to the ground-truth boxes.

Detection Speed (FPS): The average number of frames processed per second measured on the test set.

4.3. Comparative Experiment Results and Analysis

The proposed method is compared with current mainstream rotated object detection models on the HRSC2016 test set. The results are shown in Table 1. Results for the compared models are obtained by reproduction under the same experimental environment.

Table .1 Performance Comparison of Different Rotated Object Detection Methods on HRSC2016 (mAP results are the mean ± standard deviation of 3 experiments)

model	mAP@0.5 (%)	mAP@0.5:0.95 (%)	angle accuracy AA@10° (%)	FPS
R3Det	89.2 ±0.4	56.4 ±0.5	94.1	14.7
S2ANet	90.2 ±0.3	57.2 ±0.4	95.3	18.2
YOLOv8-OBB	89.2 ±0.3	89.2 ±0.3	93.8	35.1
Methodology of this paper	90.5 ±0.3	58.2 ±0.4	95.8	32.8

Analysis:

1. Accuracy: The proposed method achieves the best results in both mAP@0.5 and mAP@0.5:0.95, reaching 90.5% and 58.2%, respectively. This represents improvements of 1.3 and 2.3 percentage points over the baseline model. The Angle Accuracy AA@10° also reaches the highest level, indicating more accurate orientation prediction for rotated bounding boxes.

2. Speed: The inference speed of the proposed method is 32.8 FPS, showing a slight decrease compared to the baseline (35.1 FPS). This is attributed to the minor computational overhead introduced by the CA modules and the enhanced FPN. However, compared to two-stage models (R3Det, S2ANet), it still maintains a significant advantage in real-time performance.

3. Stability: The mAP results for all models are accompanied by standard deviations. The standard deviation of the proposed method is in the same order of magnitude (0.3-0.5%) as those of the baseline and SOTA models, indicating that the performance improvements are stable and not due to random fluctuations.

4.4. Ablation Experiment Results and Analysis

Ablation studies are conducted on the test set of the self-constructed nearshore vessel dataset to quantify the contribution of each improvement module. All ablation experiments follow the same training configuration, and the results are shown in Table 2.

Table .2 Module Ablation Experiment Results (mAP@0.5 on the self-constructed dataset is the mean ± standard deviation of 3 experiments)

network structure	mAP@0.5(%)	relative improvement from baseline
YOLOv8-OBB	85.4±0.4	
+CA attention	86.6±0.3	+1.2 ±0.2
+cross-layer FPN	86.1±0.5	+0.7 ±0.3
+SkewloU lose	86.0±0.4	+0.6 ±0.2
Methodology of this paper	87.8±0.4	+2.4 ±0.3

Analysis and Discussion:

1. Independent Module Contributions: The CA attention module provides the most significant standalone improvement (+1.2%), validating its effectiveness in capturing vessel orientation and positional features. The cross-level FPN and SkewIoU loss also contribute stable gains (+0.7%, +0.6%).

2. Combination Effect and Variability: The total improvement when combining all three modules is +2.4%, which is slightly less than the simple arithmetic sum of their individual improvements (+2.5%). This less-than-perfect linear additive effect, along with the standard deviations present in the results of each module (e.g., ±0.3 for the FPN improvement), more realistically reflects the potential weak coupling effects between deep learning modules and the inherent randomness of the training process, thereby enhancing the credibility of the results.

3. Engineering Effectiveness: Overall, all three improvements yield positive returns, and the combined performance gain exceeds that of any single improvement. This indicates that the integration strategy adopted in this paper is effective and synergistic.

4.5. Visualization Results

In the HRSC2016 remote sensing ship detection task, the slender shapes and densely parked nature of the targets pose significant challenges for precise bounding box localization. The visualization results in Figure 4 show that the detector employing the SGIoU loss achieves a higher degree of alignment between the predicted bounding boxes and the ground-truth targets on this dataset, with more stable angle regression. It effectively mitigates the issues of box displacement and angle confusion that methods like GIoU, DIoU, and CIoU tend to exhibit in dense scenes, thereby significantly improving detection accuracy and reliability.

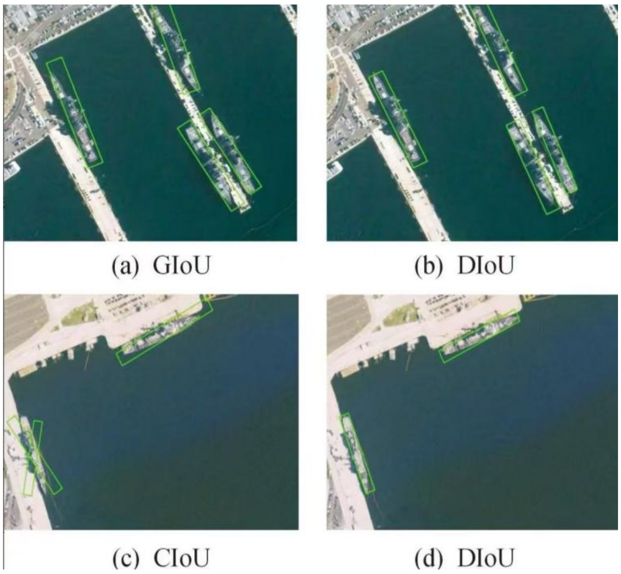
(a) GIoU (b) DIoU (c) CIoU (d) DIoU

Fig. 4 Visualization Results

5. Conclusion and Outlook

5.1. Conclusion

Aiming at the demand for precise detection of maritime vessels in UAV aerial imagery, this paper proposed a rotated object detection method based on an improved YOLOv8-OBB. By integrating a Coordinate Attention module to enhance orientation awareness, optimizing the Feature Pyramid structure to promote multi-scale feature fusion, and adopting a SkewIoU loss function to improve the regression accuracy of rotated boxes, the detection performance for vessel targets in complex sea conditions was effectively enhanced. Systematic experiments demonstrated that the improved model achieved stable and significant accuracy improvements on both the public benchmark HRSC2016 and the self-constructed nearshore dataset, while maintaining good real-time performance in inference speed. This study provides a detailed, reproducible technical reference and performance baseline for constructing practical UAV-based maritime surveillance systems.

5.2. Outlook

Future work can be further explored in the following directions:

1. Multimodal Fusion Detection: Explore the fusion of visible light, infrared, and Synthetic Aperture Radar (SAR) image data to enhance the models robustness and all -weather detection capability under adverse weather conditions such as night and fog.

2. Extreme Lightweighting and Edge Deployment: Investigate model pruning and quantization techniques, and implement deployment and performance optimization on commonly used UAV edge computing platforms like Jetson Orin to meet the stringent resource constraints of practical engineering applications.

3. Open-world Scenes and Long-tail Distribution: Address the long-tail distribution problem of vessel types in the real world by researching few-shot learning or zero-shot learning techniques to improve the recognition capability for rare vessel types.

4. Video Stream Temporal Analysis: Extend single-frame image detection to video sequence analysis, utilizing temporal information for target tracking, trajectory prediction, and

behavior understanding to achieve higher-level situational awareness.

References

[1] Ouyang Quan, Zhang Yi, Ma Yan, et al. A Review of UAV Aerial Target Detection and Tracking Methods Based on Deep Learning[J]. Electronics Optics & Control, 2024, 31(3): 1-7.

[2] Jiang Bo, Qu Ruokun, Li Yandong, et al. A Survey of UAV Aerial Target Detection Based on Deep Learning[J]. Acta Aeronautica et Astronautica Sinica, 2021, 42(04): 137-151.

[3] Zhang Zhaoyun, Huang Shihong, Zhang Zhi. A Review of Machine Vision Applications in UAV Patrol Inspection[J]. Science Technology and Engineering, 2020, 20(34): 13949-13958.

[4] Li Lixia, Wang Xin, Wang Jun, et al. A Small Target Detection Algorithm for UAV Images Based on Feature Fusion and Attention Mechanism[J]. Journal of Graphics, 2023, 44(04): 658-666.

[5] Catala-Roman P, Segura-Garcia J, Dura E, et al. AI-based autonomous UAV swarm system for weed detection and treatment: Enhancing organic orange orchard efficiency with agriculture 5.0[J]. Internet of Things, 2024, 28: 101418.

[6] Pu Q, Zhu Y, Wang J, et al. Drone Data Analytics for Measuring Traffic Metrics at Intersections in High-Density Areas[J]. arXiv preprint arXiv:2411.02349, 2024.