Orientation-Aware Feature Fusion for Accurate Rotated Object Detection in Remote Sensing Images

Jing Zhang[1], Mas Rina Mustaffa[2]*, Fatimah Khalid[3], Zainal Abdul Kahar[4]
Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Serdang, Malaysia[1, 2, 3, 4]
Department of General Education, Anhui Vocational College of City Management, Hefei, China[1]

***Abstract*—Conventional feature fusion mechanisms largely overlook orientation information, making it difficult to effectively represent objects with diverse rotational patterns. To address this issue, we propose YOLO-RSL, a lightweight rotated object detector that introduces orientation awareness into feature representation, feature fusion, and localization optimization. The proposed Rotation-Oriented Iterative Attentional Feature Fusion (RO-iAFF) module incorporates implicit orientation information derived from orthogonal asymmetric depthwise convolutions into a two-stage channel attention process, enabling orientation-sensitive multi-scale feature aggregation. In addition, a Swin Transformer block and a large-kernel feature branch are employed to enhance global context modeling and improve small-object representation. To achieve more accurate localization, an enhanced regression loss (LFFE Loss) is designed to jointly optimize overlap quality, rotation angle, center offset, and aspect ratio. Experimental results on the DIOR-R and UCAS-AOD datasets show that YOLO-RSL achieves $mAP_{0.5}$ scores of 84.17% and 98.44%, surpassing the baseline by 3.15% and 2.91%, respectively. Moreover, the performance gains become more pronounced as target aspect ratios increase, demonstrating the effectiveness of the proposed orientation-aware design.**

***Keywords*—Rotated object detection; oriented bounding box; feature fusion; remote sensing image**

I. Introduction

Remote sensing platforms provide a wide range of observation data with high space-time resolution, which makes them irreplaceable and play an important role in the fields of urban planning, disaster assessment, and traffic monitoring. In remote sensing image (RSI) analysis, rotating target detection is a basic but difficult task. In this task, most targets, such as aircraft, ships, and vehicles, will appear at any angle in a complex context, and there are obvious scale changes. Using directional enclosure frames (OBBs) can give more compact geometric representations and effectively reduce background interference, so rotating target detection has become a key task in the field of remote sensing analysis [1].

The target detector based on deep learning has greatly improved the performance of general target detection. Representative frameworks such as Faster R-CNN [2] and SSD [3] have laid a solid foundation for the modern target detection architecture. After that, the YOLO-based detector has made significant progress in balancing detection accuracy and inference efficiency. However, because of the limited perception ability of target orientation, it has not been able to fully model global context information. There are still many problems in the direct use of traditional general-purpose detectors in remote sensing image analysis. Existing rotary target detectors, such as SCRDet++ [4] and R3Det [5], effectively improve the accuracy of target positioning by introducing rotary perception regression and refined feature learning mechanisms. Even so, when encountering the limitations of lightweight deployment, these methods still have some shortcomings in dealing with small targets and complex background problems.

In recent years, Transformer-based architectures have shown strong capabilities in long-range dependency modeling and global feature representation. Vision Transformer (ViT) took the lead in introducing the self-attention mechanism into computer vision tasks, while Pyramid Vision Transformer (PVT) [6] and MobileViT [7] improved computing efficiency through hierarchical and lightweight design. Deformable DETR [8] and OrientedFormer [9] extend the Transformer mechanism to direction detection in RSI. However, these methods address global context modeling rather than the orientation sensitivity of the fusion stage itself, and their computational overhead continues to limit deployment in lightweight detectors.

Feature fusion is the stage at which multi-scale feature maps from different backbone levels are combined before being passed to detection heads. Existing channel attention modules employed in this stage—SE [10], CBAM [11], and even the iterative iAFF [12]—compute channel weighting solely from global average pooling or local convolutional statistics of the feature maps, without any encoding of target orientation. This means the same channel weights are assigned irrespective of whether a target appears at 0°, 45°, or 90°. In RSIs, where targets span the full 180° orientation range and object categories such as aircraft and ships have aspect ratios exceeding 6:1, orientation-agnostic fusion leads to a systematic failure mode: features discriminative for one orientation are suppressed when those features happen to score poorly on a channel statistic computed over a spatially mixed feature map. The result is missed detections for densely arranged, arbitrarily oriented objects and imprecise OBB regression for elongated targets—precisely the objects that most frequently appear in RSI benchmarks.

A further challenge lies in rotated bounding-box regression. Existing methods, such as circular smoothing label (CSL) [13] and H2RBox-v2 [14], solve the problems of periodic discontinuity and angle blur in the process of angle prediction. At the same time, ProbIoU [15] optimizes the rotating boundary frame by modeling geometric uncertainty. However,

*Corresponding author.

the existing loss function is still unable to fully model the joint relationship between the overlapping area, center offset, aspect ratio, and rotation angle at the same time, resulting in the unstable positioning of slender and densely arranged objects. Second, multi-scale feature fusion in the neck is typically orientation-agnostic, producing channel-wise attention weights that are insensitive to the rotation state of targets. Third, conventional IoU-based losses provide incomplete geometric supervision for OBBs because they neglect center displacement, aspect-ratio mismatch, and angle residual simultaneously. The present work addresses all three deficiencies within a single, parameter-efficient framework. The main contributions of this study are as follows:

1) A novel Rotation-Oriented iterative Attentional Feature Fusion (RO-iAFF) module is proposed. RO-iAFF extends the standard iAFF framework by introducing orientation encoding angle embeddings generated from the cosine and sine values of the predicted rotation angle into the dual-stage MS-CAM attention computation, thereby enabling channel-wise feature weighting to become sensitive to object orientation. In addition, an extra 10×10 deep feature extraction layer is incorporated into the neck to enlarge the effective receptive field for small targets.

2) Swin Transformer blocks replace selected C3K2 modules in the YOLOv11s-OBB backbone to enrich global contextual representation, providing high-quality multi-scale features on which orientation-aware fusion can act most effectively.

3) A Four-Factor Loss Function for Enhanced regression is proposed to jointly constrain overlap, angle, center offset, and aspect ratio, thereby improving rotated bounding-box localization stability and accuracy.

II. Related Work

A. YOLO-Based Rotated Object Detection

YOLO-based detectors have achieved significant progress in balancing detection accuracy and inference efficiency. Compared with two-stage frameworks such as Faster R-CNN [2], the YOLO architecture provides more efficient end-to-end interference, making it very suitable for real-time RSI applications. SCRDet++ [4] improves rotation detection through instance-level feature denoising and rotation loss smoothing, while R³Det [5] enhances positioning performance through refined feature optimization. CSL [13] re-expresses angle regression as a classification problem to alleviate periodic discontinuity, while H2RBox-v2 [14] improves direction estimation through symmetric perception self-supervised learning. YOLOv11-OBB [16] expands the direction detection head for OBB regression on the basis of YOLOv11, achieving high positioning accuracy with lightweight overhead. Despite these advances, these detectors have not solved the problem of direction insensitivity in their feature fusion stage: no matter what the modeling effect of the detection head is on the rotation, the attention weight of the neck channel is still independent of the direction, resulting in a continuous characterization bottleneck, and YOLO-RSL is directly aimed at this bottleneck.

B. Attention and Feature Fusion Mechanisms

Attention mechanisms are the dominant tool for feature fusion in modern detectors. SE [10] enhances channel-wise recalibration through global average pooling, yet its scalar channel weights are derived from orientation-invariant global statistics, making it blind to target rotation. CBAM [11] combines channel and spatial attention for refined enhancement, but generates identical orientation-agnostic weights regardless of whether a target is oriented at 0° or 90°. To improve multi-scale feature interaction, iAFF [12] introduces iterative attention weighting to dynamically fuse heterogeneous semantic features across different layers. Compared with traditional linear fusion methods, iterative feature fusion achieves more effective semantic alignment and feature refinement. In addition, BoTNet [17] incorporates multi-head self-attention into convolutional bottleneck structures, enabling stronger global dependency modeling while maintaining computational efficiency.

However, most existing feature fusion strategies mainly focus on channel or spatial attention and neglect explicit orientation information during feature aggregation. Consequently, their performance remains limited for densely distributed and arbitrarily rotated objects in remote sensing imagery.

C. Transformer-Based Object Detection

The Transformer architecture shows a strong ability in modeling long-range dependencies and global context information. PVT [6] improves efficiency through hierarchical feature extraction, while MobileViT [7] combines lightweight convolution with a Transformer module. For object detection, Deformable DETR [8] reduces global attention cost through sparse deformable sampling, ARS-DETR [18] improves aspect-ratio-sensitive feature learning for aerial detection, and OrientedFormer [9] introduces orientation-aware positional encoding for rotated RSI detection. Although these methods effectively improve contextual perception, they address representation quality rather than fusion-stage orientation sensitivity, and their computational overhead continues to hinder lightweight deployment. We adopt Swin Transformer blocks [19] selectively in the backbone to gain their hierarchical global context benefits at moderate cost, complementing rather than relying on them for the orientation-awareness provided by RO-iAFF.

D. Research Gap and Motivation

The survey above reveals a consistent and unaddressed gap: while rotation-aware regression losses and backbone representations have been extensively studied, the feature fusion stage of lightweight YOLO-series detectors has remained universally orientation-agnostic. The three literature streams reviewed above address complementary aspects—orientation-aware regression (Section II-A), feature representation (Section II-C), and channel attention quality (Section II-B)—but none provides a mechanism for making multi-scale channel fusion weights sensitive to target orientation. We fill this gap by proposing RO-iAFF, which introduces orientation awareness directly into the iterative attention fusion computation, and validate that this targeted intervention yields the largest performance gain among all

proposed components, particularly for elongated, high-aspect-ratio targets.

III. Proposed Method

A. Overall Architecture

The proposed detector is designed around a three-level orientation-aware strategy applied to YOLOv11s-OBB (Fig. 1). At the representation level, selected C3K2 modules at backbone stages 2 and 3 are replaced with Swin Transformer blocks [19] to enrich global semantic representation. At the fusion level—the central innovation—each neck output branch is connected to the detection head via the proposed RO-iAFF module, which replaces the standard direct skip connection and provides orientation-sensitive channel weighting. Additional convolutional and C3K2 layers before the SPPF stage generate a 10×10 deep feature map that is upsampled and fused with the 20×20 PANet output to form a fourth detection branch, enlarging the effective receptive field for small targets. At the supervision level, the original regression loss is replaced with LFFE, which jointly constrains four geometric factors of predicted OBBs. These three levels are mutually reinforcing: the Swin backbone provides richer multi-scale features, RO-iAFF fuses them in an orientation-sensitive manner, and LFFE ensures that the resulting predictions are geometrically precise. Together, they introduce only 0.56 additional GFLOPs and 0.21 M additional parameters over the baseline.

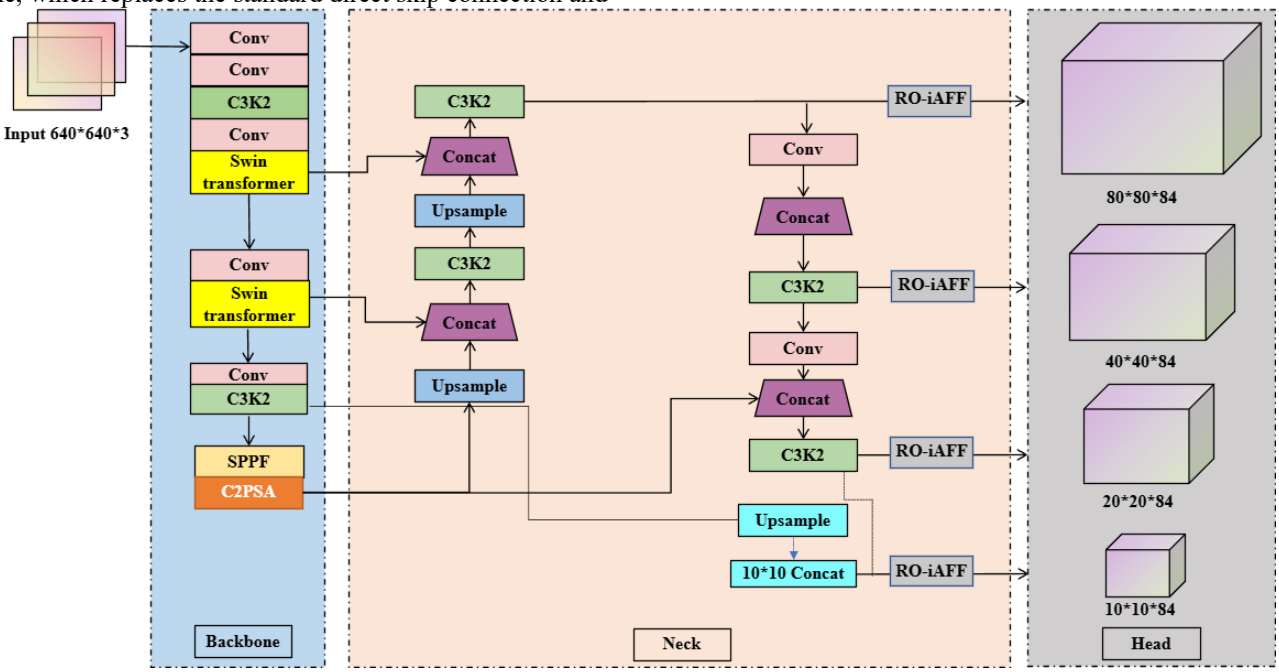

Fig. 1. Overall network structure.

B. Rotation-Oriented Iterative Attention Feature Fusion

Standard multi-scale feature fusion modules, including the original iAFF, compute channel attention weights solely from the feature maps themselves, rendering the fusion process rotation-agnostic. In RSIs, where targets such as aircraft, ships, and bridges exhibit wide orientation variation, orientation-agnostic fusion leads to suboptimal channel selection. The original iAFF framework consists of two core components: a Multi-Scale Channel Attention Mechanism (MS-CAM), which extracts complementary global and local channel descriptors to generate soft attention weights, and a two-stage iterative fusion strategy that progressively refines the fused representation. Building upon the standard iAFF framework, the proposed RO-iAFF introduces a new component called the Directional Channel Modulation (DCM) descriptor. Unlike conventional channel attention modules that derive weights purely from pooling statistics, DCM pulls directional information directly out of the feature maps themselves. This is done through a pair of asymmetric depthwise convolutions applied to the element-wise sum of the two input feature maps. The resulting descriptor is then injected into the MS-CAM global branch at both fusion stages, allowing the channel weights to shift in response to the dominant orientation present in the fused features.

A key design decision in DCM is that it does not take a rotation angle as an external input. Instead, it recovers directional information implicitly from the feature maps. Given the element-wise sum S of the two feature maps to be fused, two directional feature maps are computed below:

$$\boldsymbol{f}_h = \boldsymbol{DWConv}_{1\times k}(\boldsymbol{S}) \in ¡^{H\times W\times C} \quad (1)$$

$$\boldsymbol{f}_v = \boldsymbol{DWConv}_{k\times 1}(\boldsymbol{S}) \in ¡^{H\times W\times C} \quad (2)$$

$\boldsymbol{DWConv}_{1\times k}$ uses a 1×k depthwise convolution to capture horizontal structure, and $\boldsymbol{DWConv}_{k\times 1}$ uses a k×1 depthwise convolution to capture vertical structure, where k is set to 7. A 1×k kernel responds preferentially to intensity patterns that extend along the horizontal direction, while a k×1 kernel responds preferentially to vertically oriented patterns. For targets with a spindle tilt angle of θ, the activation value of the horizontal feature diagram $\boldsymbol{f}_h$ is proportional to cos(θ), while the activation value of the vertical feature diagram $\boldsymbol{f}_v$ is

proportional to sin(θ). The relative magnitude of the two maps therefore encodes the dominant orientation of the scene in a continuous and differentiable manner. This encoding spans the full 180-degree orientation range and requires no explicit angle label or supervision signal from the detection head, making DCM entirely self-contained within the neck. This continuous, self-supervised orientation encoding is what differentiates DCM from all prior channel attention methods, which discard directional spatial structure by pooling over entire feature maps.

Global average pooling is then applied to each directional map, and the resulting vectors are concatenated and compressed to a C-dimensional directional descriptor via a 1×1 convolution:

$$\boldsymbol{d} = \boldsymbol{Conv}_{1\times1}\left(\left[\boldsymbol{GAP}(\boldsymbol{f}_h) \| \boldsymbol{GAP}(\boldsymbol{f}_v)\right]\right) \in ¡^{C} \tag{3}$$

The parameter overhead of DCM is $2\boldsymbol{kC} + \boldsymbol{C}^2$, negligible relative to the backbone. Since $\boldsymbol{d}$ is derived entirely from the feature maps, DCM is fully self-contained within the neck and introduces no dependency on detection head outputs.

The directional descriptor $\boldsymbol{d}$ is injected additively into the MS-CAM global attention branch (Fig. 2) prior to Sigmoid normalization at both fusion stages. For Stage 1, given the global descriptor $\boldsymbol{g} = \boldsymbol{GAP}(\boldsymbol{S})$ and local attention map $\boldsymbol{a}_{loc} = \boldsymbol{Conv}_{1\times1}(\boldsymbol{S})$, the rotation-oriented attention weight is computed as:

$$\boldsymbol{a}_{glob}^{dir} = \boldsymbol{\sigma}\left(\boldsymbol{W}_2 \cdot \boldsymbol{BN}(\mathrm{Re}\,\boldsymbol{LU}(\boldsymbol{W}_1\boldsymbol{g})) + \boldsymbol{d}\right) \tag{4}$$

$$\boldsymbol{a}_1 = \boldsymbol{\sigma}\left(\boldsymbol{a}_{glob}^{dir} + \boldsymbol{a}_{loc}\right) \tag{5}$$

The intermediate fused feature $\boldsymbol{Z}_0$ is then obtained as:

$$\boldsymbol{Z}_0 = \boldsymbol{a}_1 \,\mathrm{e}\, \boldsymbol{X} + (1-\boldsymbol{a}_1)\,\mathrm{e}\, \boldsymbol{Y} \tag{6}$$

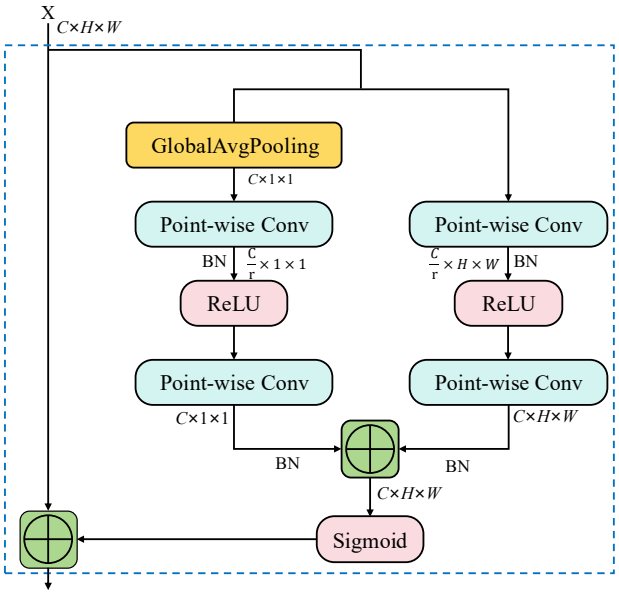

Fig. 2. Architecture of MS-CAM.

Stage 2 refines $\boldsymbol{Z}_0$ by applying a second rotation-oriented MS-CAM block under the same directional descriptor d:

$$\boldsymbol{a}_{glob,0}^{dir} = \boldsymbol{\sigma}\left(\boldsymbol{W}_4 \cdot \boldsymbol{BN}(\mathrm{Re}\,\boldsymbol{LU}(\boldsymbol{W}_3\boldsymbol{g}_0)) + \boldsymbol{d}\right) \tag{7}$$

$$\boldsymbol{a}_2 = \boldsymbol{\sigma}\left(\boldsymbol{a}_{glob,0}^{dir} + \boldsymbol{a}_{loc,0}\right) \tag{8}$$

$$\boldsymbol{Z} = \boldsymbol{a}_2 \,\mathrm{e}\, \boldsymbol{Z}_0 + (1-\boldsymbol{a}_2)\,\mathrm{e}\, \boldsymbol{Y} \tag{9}$$

where, $\boldsymbol{g}_0 = \boldsymbol{GAP}(\boldsymbol{Z}_0)$ and $\boldsymbol{a}_{loc,0} = \boldsymbol{Conv}_{1\times1}(\boldsymbol{Z}_0)$. Sharing $\boldsymbol{d}$ across both stages ensures consistent directional guidance without doubling the DCM overhead. where $\boldsymbol{W}_i (\boldsymbol{i} = 1,...,4)$ are the learnable weight matrices of the two-layer bottleneck MLPs in Stage 1 and Stage 2, respectively, with reduction ratio r = 4. Stage 1 performs coarse channel selection guided by the directional prior, while Stage 2 refines the fused representation by re-examining $\boldsymbol{Z}_0$ under the same directional context. When the feature maps contain no directional asymmetry, $\boldsymbol{d}$ approaches a uniform vector and RO-iAFF degenerates to standard iAFF, providing stable behavior on isotropic targets. The DCM weights are initialized with near-zero values to ensure that training begins from a well-understood iAFF baseline. The complete forward pass is summarized in Algorithm 1.

Algorithm 1: RO-iAFF Forward Pass

Input: Feature maps X, Y $\in ¡^{H\times W\times C}$

Directional Channel Modulation

1. $\boldsymbol{S} \leftarrow \boldsymbol{X} + \boldsymbol{Y}$
2. $\boldsymbol{f}_h \leftarrow \boldsymbol{DWConv}_{1\times k}(\boldsymbol{S})$ //horizontal directional feature
   $\boldsymbol{f}_v \leftarrow \boldsymbol{DWConv}_{k\times1}(\boldsymbol{S})$ //vertical directional feature
3. $\boldsymbol{d} \leftarrow \boldsymbol{Conv}_{1\times1}\left(\left[\boldsymbol{GAP}(\boldsymbol{f}_h) \| \boldsymbol{GAP}(\boldsymbol{f}_v)\right]\right) \in ¡^{C}$

Stage 1: Coarse fusion

4. $\boldsymbol{g} \leftarrow \boldsymbol{GAP}(\boldsymbol{S})$
5. $\boldsymbol{a}_{glob} \leftarrow \boldsymbol{\sigma}\left(\boldsymbol{W}_2 \cdot \boldsymbol{BN}(\mathrm{Re}\,\boldsymbol{LU}(\boldsymbol{W}_1\boldsymbol{g})) + \boldsymbol{d}\right)$
   $\boldsymbol{a}_{loc} \leftarrow \boldsymbol{Conv}_{1\times1}(\boldsymbol{S})$, $\boldsymbol{a}_1 \leftarrow \boldsymbol{\sigma}\left(\boldsymbol{a}_{glob} + \boldsymbol{a}_{loc}\right)$
6. $\boldsymbol{Z}_0 \leftarrow \boldsymbol{a}_1 \,\mathrm{e}\, \boldsymbol{X} + (1-\boldsymbol{a}_1)\,\mathrm{e}\, \boldsymbol{Y}$

Stage 2: Refined fusion

7. $\boldsymbol{g}_0 \leftarrow \boldsymbol{GAP}(\boldsymbol{Z}_0)$
8. $\boldsymbol{a}_{glob,0} \leftarrow \boldsymbol{\sigma}\left(\boldsymbol{W}_4 \cdot \boldsymbol{BN}(\mathrm{Re}\,\boldsymbol{LU}(\boldsymbol{W}_3\boldsymbol{g})) + \boldsymbol{d}\right)$
   $\boldsymbol{a}_{loc,0} \leftarrow \boldsymbol{Conv}_{1\times1}(\boldsymbol{Z}_0)$, $\boldsymbol{a}_2 \leftarrow \boldsymbol{\sigma}\left(\boldsymbol{a}_{glob,0} + \boldsymbol{a}_{loc,0}\right)$
9. $\boldsymbol{Z} \leftarrow \boldsymbol{a}_2 \,\mathrm{e}\, \boldsymbol{Z}_0 + (1-\boldsymbol{a}_2)\,\mathrm{e}\, \boldsymbol{Y}$

Output: Fused feature Z $\in ¡^{H\times W\times C}$

C. Swin Transformer Block

While RO-iAFF provides orientation-aware fusion, the quality of the fused representations depends on the richness of the backbone features. The C3K2 module extracts local features through small volumes, but its receptive field is limited. In order to enrich the global context information, we replaced some C3K2 modules in stages 2 and 3 of the backbone network with the Swin Transformer module [19].

Swin Transformer adopts a hierarchical framework, including image block segmentation, feature extraction, and image block merging. Each module replaces the standard global multi-head self-attention mechanism (MSA) with the window-based multi-head self-attention mechanism (W-MSA) and the shifted window multi-head self-attention mechanism (SW-MSA), as shown in Fig. 3. W-MSA calculates self-attention in non-overlapping local windows; SW-MSA loops to move window boundaries to achieve cross-window information exchange. After the attention sublayer, layer normalization, multi-layer perceptron (MLP), and residual connection are applied in turn, and finally the semantic feature diagram is generated.

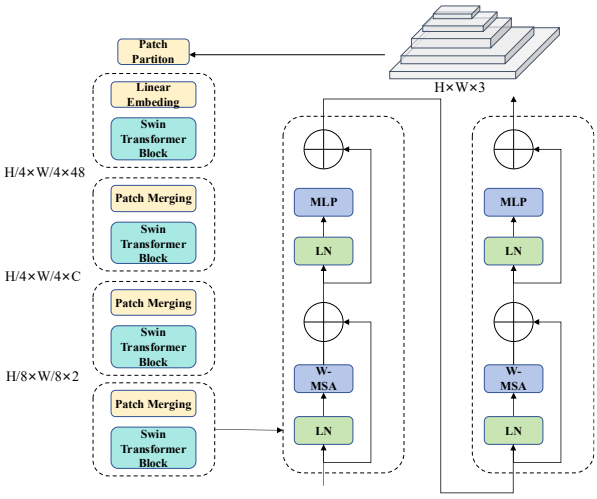

Fig. 3. Structure of Swin Transformer.

D. Four-Factor Bounding-Box Loss

Standard IoU-based losses present a known difficulty for rotated object detection, as OBB intersections become non-differentiable under certain parameter configurations. While orientation-aware fusion improves representational quality, accurate OBB regression still demands a loss function that captures all relevant geometric attributes jointly. LFFE addresses this need by optimizing four orthogonal geometric factors: overlap area, rotation angle residual, center-point offset, and aspect-ratio discrepancy. The total training loss is:

$$L_{total} = \frac{\lambda_1}{N}\sum_{n=1}^{N} L_{obj} + \frac{\lambda_2}{N}\sum_{n=1}^{N} L_{cls} + \frac{\lambda_3}{N}\sum_{n=1}^{N} L_{box} \tag{10}$$

In the loss formulation L_{obj} 、 L_{cls} and L_{box} correspond to the confidence loss, classification loss, and bounding box loss functions $\lambda_1 = \lambda_2 = 1$ and $\lambda_3 = 0.05$ adopts the default setting. The regression loss is:

$$L_{box} = L_{SkewIoU} + \lambda_4 L_{center} + \lambda_5 (\theta_p - \theta_t)^2 \tag{11}$$

Here, SkewIoU is the rotated IoU indicator. L_{center} is a center-related constraint. To achieve a r reasonable balance between positioning accuracy and angle alignment, the weighting factors λ_4 and λ_5 are given values of 1.0 and 0.5, respectively. By imposing constraints on four geometric attributes, LFFE provides more reliable supervision for OBB and further enhances the direction-sensitive characteristics generated by RO-iAFF.

IV. Experiments

A. Datasets and Metrics

DIOR-R is a rotated variant of the DIOR dataset comprising over 23,000 high-resolution RSIs annotated with OBBs across 20 object categories, sourced from diverse sensors at varying spatial resolutions and imaging angles. Large images are partitioned into 1024×1024 patches using a sliding window, and Mosaic data augmentation is applied during training to enrich the spatial distribution of objects across scales and orientations. The data augmentation examples are illustrated in Fig. 4. The training and test sets are partitioned at an 8:2 ratio. The training and test sets are partitioned at an 8:2 ratio. The statistical overview includes instances of object categories, the size and quantity of ground-truth bounding boxes, the center positions and aspect ratios of objects, as shown in Fig. 5.

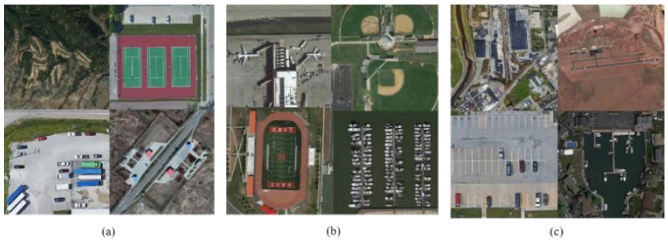

Fig. 4. Mosaic data augmentation on the DIOR dataset.

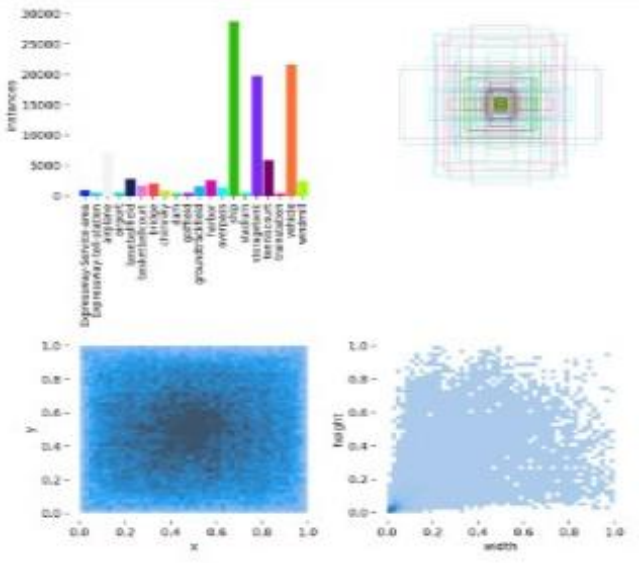

Fig. 5. Statistics of the DIOR dataset.

UCAS-AOD contains high-resolution aerial images with OBB annotations for two categories (airplane and vehicle). Versions 1.0 and 2.0 together provide 1,000 airplane images (7,482 instances), 510 car images (7,114 instances), and 910 background images. Maximum annotated aspect ratios reach 8:1. Random rotation, angle jitter, and brightness variation are applied for data augmentation. Training and test sets comprise 1,110 and 400 images, respectively.

B. Implementation Details

All experiments are conducted on Ubuntu 20.04 with an NVIDIA RTX 3080 GPU, PyTorch 1.10.0, CUDA 11.3, and cuDNN 8.3.4. The SGD optimizer uses a learning rate of 0.01, momentum of 0.937, and weight decay of 0.0005. A 3-epoch linear warm-up phase is followed by cosine annealing decay. The batch size is 8, and all models are trained for 200 epochs. Training results on the UCAS-AOD dataset are shown in Fig. 6.

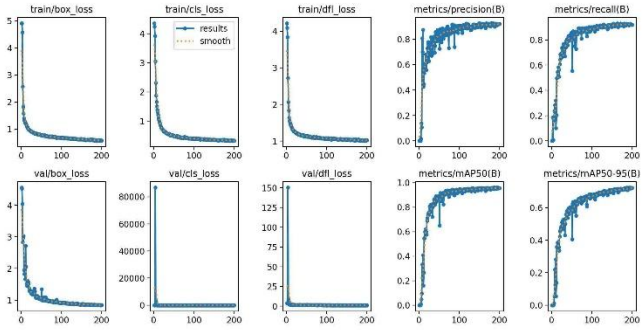

Fig. 6. YOLOv11s-OBB baseline training results on the UCAS-AOD dataset.

C. Ablation Experiment

To justify the selection of RO-iAFF over alternative attention modules, SE [10], CBAM [11], Coordinate Attention (CA) [20], the standard iAFF [12], and the proposed RO-iAFF are each inserted at the four neck output branches and evaluated on UCAS-AOD under identical training conditions (Table I). RO-iAFF achieves 96.64% mAP0.5, surpassing standard iAFF by 0.61% and CA by 0.66%, while adding only 0.03 M parameters over the baseline. The consistent performance ordering (Baseline < SE ≤ CBAM < CA < iAFF < RO-iAFF) mirrors the degree to which each method incorporates spatial or directional information, strongly supporting the interpretation that orientation-awareness is the decisive factor.

TABLE I. COMPARISON OF THE IMPACT OF DIFFERENT ATTENTION MECHANISMS

Method	$mAP_{0.5}$	Param(M)
Baseline	95.53	7.72
SE	95.56	7.77
CBAM	95.57	7.82
CA	95.98	7.74
RO-iAFF	**96.64**	**7.75**

The heat maps (Fig. 7) on DIOR further show that RO-iAFF can maintain a clearer attention to the target area at different scales and in complex contexts. Placing RO-iAFF at multiple levels of the neck can achieve finer features across scales, so as to obtain better positioning and classification effects.

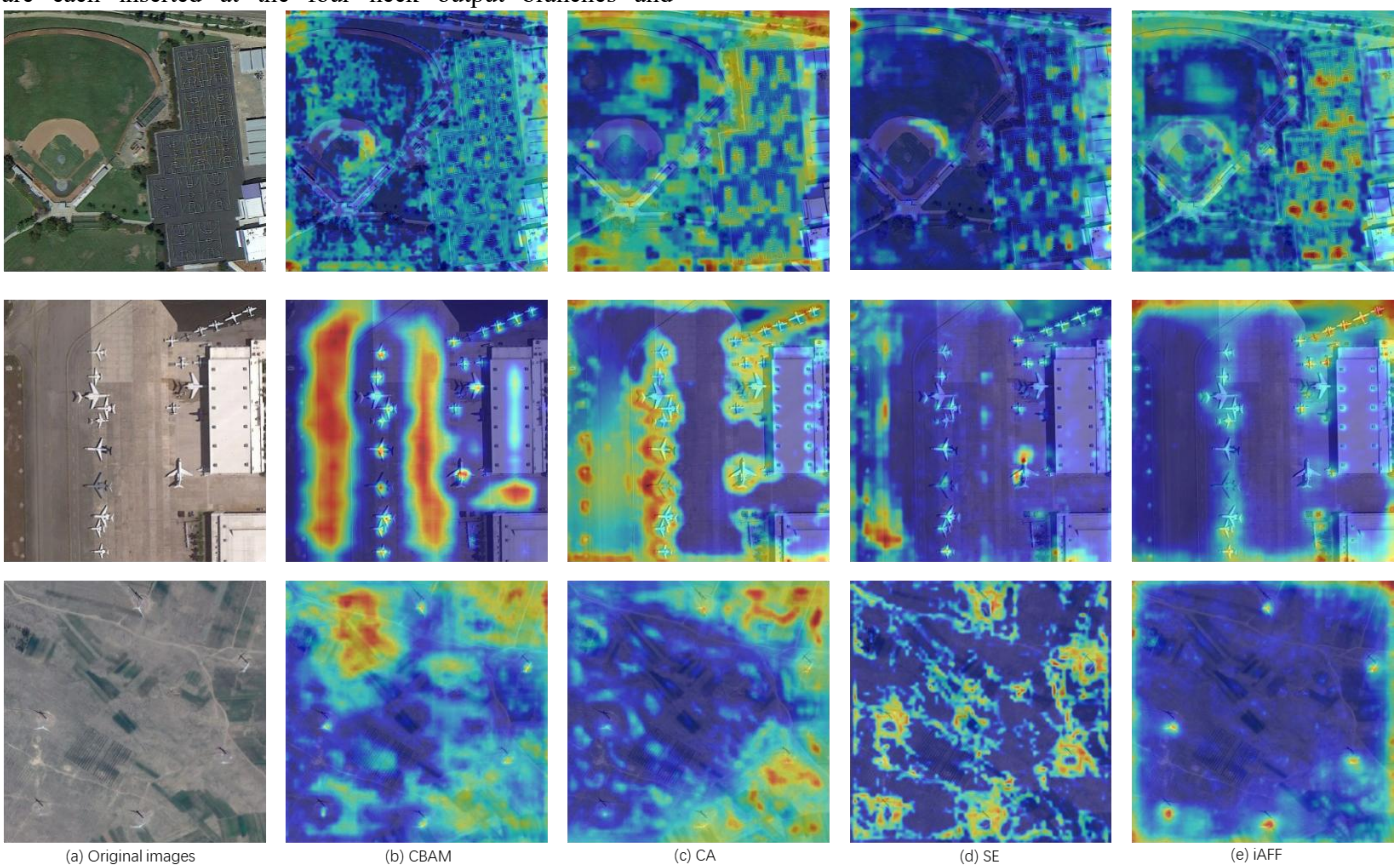

Fig. 7. Heatmap images by different neck designs on the DIOR dataset.

Table II shows that each proposed component brings steady improvements on both DIOR-R and UCAS-AOD. On DIOR-R, the baseline YOLOv11s-OBB achieves 81.02% mAP0.5 and 49.12% mAP0.5:0.95. After the introduction of the Swin Transformer module, these scores increased to 81.56% and 50.54%, respectively, confirming that the self-attention mechanism based on hierarchical windows can enhance the global context representation in messy remote sensing scenarios. After replacing the standard iAFF with our proposed RO-iAFF (also integrating the 10×10 depth feature branch), the single-component gain is the largest, mAP0.5 increased to 83.24%, and mAP0.5:0.95 increased to 51.73%, indicating that directional perception fusion can significantly improve the multi-scale feature quality of slender and densely arranged targets. The subsequent integration of LFFE achieved the best performance, with mAP0.5 at 84.17% and mAP0.5:0.95 at 52.98%, an increase of 3.15% and 3.86% respectively over the benchmark, while maintaining only 16.54 GFLOPs of computing cost and 8.12 M parameters.

On the UCAS-AOD data set of small densely distributed targets, each proposed component has made similar continuous improvements. The baseline model achieved mAP0.5 at 95.53% and mAP0.5:0.95 at 71.35%. After the introduction of the Swin Transformer module, the score increased to 96.64% and 71.93%, respectively, verifying that global context modeling can improve target detection performance even on a relatively compact two-category benchmark dataset. RO-iAFF further increased mAP0.5 to 97.25% and mAP0.5:0.95 to 72.84%. This result shows that directional perception fusion is particularly effective for slender targets such as aircraft and vehicles. After adding LFFE, the mAP0.5 of the complete model reached 98.44%, and the mAP0.5:0.95 reached 74.19%, which was 2.91% and 2.84% higher than the baseline model, respectively. Although there are differences in the number of categories, object scale, and annotation density, the model performance on the two datasets has improved, indicating that the three components can work well together and apply to different remote sensing conditions.

TABLE II. ABLATION EXPERIMENT ON THE DIOR AND UCAS-AOD DATASETS

Model	Dataset	Swin Transformer	RO-iAFF	LFFE	$mAP_{0.5}$	#Param(M)	GFLOPs
YOLOv11s-OBB	DIOR	×	×	×	81.02	7.91	16.45
		✓	×	×	81.56	7.95	15.91
		✓	✓	×	83.24	8.08	16.23
		✓	✓	✓	84.17	8.12	**16.54**
	UCAS-AOD	×	×	×	95.53	7.72	16.47
		✓	×	×	96.64	7.75	16.55
		✓	✓	×	97.25	7.78	16.63
		✓	✓	✓	98.44	7.93	**16.65**

TABLE III. PERFORMANCE GAIN STRATIFIED BY TARGET ASPECT RATIO ON DIOR-R

Aspect Ratio Group	Target Examples	Baseline $mAP_{0.5}$(%)	YOLO-RSL $mAP_{0.5}$(%)	Gain
AR < 2	Vehicle, Dam	86.5	88.3	+1.8
2 ≤ AR < 4	Bridge, Harbor	83.3	86.7	+3.4
4 ≤ AR < 6	Ship, Airplane	79.5	84.2	+4.7
AR ≥ 6	Aircraft	74.7	82.1	**+7.4**

To verify the directional perception of YOLO-RSL, we group according to the target aspect ratio (AR) to analyze the performance of $mAP_{0.5}$ on the DIOR-R dataset (Table III). Four groups are defined: AR < 2 (near-isotropic targets), 2 ⩽ AR < 4 (moderately elongated targets), 4 ⩽ AR < 6 (elongated targets), and AR ≥ 6 (highly elongated targets).

The performance of YOLO-RSL in each group is better than that of the baseline model, but the increase is not uniform. It is worth mentioning that the performance improvement gradually increases with the increase of the target aspect ratio. For targets whose aspect ratio is less than 2, the performance improvement is limited. The visual appearance of such targets does not change much with rotation. In this case, the direction-sensitive channel weighting of RO-iAFF is also limited. With the increase in aspect ratio, the performance improvement continues to increase and peaks in the AR 6 group. The appearance of the clearer target of the spindle changes more at the rotation angle, and the advantage of RO-iAFF integrating direction information into channel attention calculation is becoming more and more obvious.

D. Comparison with State-of-the-Art Methods

To evaluate how the proposed method compares against existing work, extensive experiments were carried out on the DIOR dataset. The comparison covers a variety of detector architectures. On the two-stage side, Faster R-CNN [2] and Mask R-CNN [21] are included. On the one-stage side, the comparison includes RetinaNet [22], YOLOv8 [23], YOLOv10 [24], and YOLOv11 [16]. Several specialized models are also included, namely CANet [25], CE-FPN [26], SCRDet [4], and YOLOP [27]. The full results are reported in Table IV.

TABLE IV. COMPARISON OF STATE-OF-THE-ART MODELS ON THE DIOR DATASET

Model	airplane	ship	bridge	dam	vehicle	harbor	airport	chimney	$mAP_{0.5}$	Param	GFLOPs
Faster-RCNN	88.1	77.3	81.2	72.3	84.5	75.4	79.5	74.7	79.4	82.4	33.1
Mask-RCNN	88.3	78.1	81.4	70.9	85.0	75.7	80.0	75.3	79.7	80.5	28.6
RetinaNet	82.4	82.7	81.9	82.0	81.8	82.7	83.1	81.6	82.5	7.0	15.8
YOLOv5	87.8	76.9	80.6	71.5	83.1	74.1	78.3	73.5	78.5	7.1	17.4
YOLOv8s	89.4	78.4	82.3	73.6	85.3	76.3	80.6	75.8	80.2	6.98	13.7
YOLOv10	89.2	79.2	82.3	74.3	86.2	77.0	81.0	76.4	80.9	16.2	16.8
YOLOP	87.1	75.3	79.1	70.8	82.4	73.3	79.5	72.6	77.3	18.9	19.7
YOLOv11s -OBB	90.3	79.5	83.3	74.7	86.5	77.9	81.9	77.2	81.02	7.91	16.45
CANet	91.2	81.2	84.4	76.2	87.1	79.0	83.3	78.0	82.9	37.1	12.3
CE-FPN	91.0	81.3	83.7	75.4	84.4	82.3	82.0	77.5	81.5	34.8	15.5
SCRDet	91.7	81.7	85.1	76.6	88.2	79.4	83.4	79.1	83.7	30.3	14.7
Ours	92.0	82.5	82.9	83.2	81.1	86.3	84.3	85.9	**84.17**	**8.12**	**16.5**

As demonstrated in Table IV, our model exhibits superior performance across the majority of object categories in the DIOR dataset. Specifically, the proposed method achieves the highest AP across the categories in the DIOR dataset. Furthermore, the model attains a robust mAP0.5 of 84.17%, surpassing both traditional models and recent variants of YOLO. Notably, while maintaining competitive GFLOPs at 16.5 and a relatively lightweight parameter size of 8.12 million, our model outperforms heavier architectures like Faster R-CNN and Mask R-CNN, which achieve mAP0.5 scores of 79.4% and 79.7%, respectively, as well as specialized models like SCRDet and CANet. These findings underscore the efficacy and efficiency of the proposed method in large-scale RSOD tasks.

To rigorously evaluate the performance of the proposed model in the target detection task, we perform a comparison with several state-of-the-art models on the UCAS-AOD dataset, as summarized in Table IV. Compared with Faster R-CNN, NAS-FPN [28], RT-DETR [29], Zhang et al. [30], Li et al. [31], S²Anet [32], ARS-DETR [33], Oriented-Former [9], and SASM [34], our model achieves better accuracy and efficiency.

On UCAS-AOD (Table V), our method achieves an impressive mAP0.5 of 98.44%, surpassing all other models in this category. This indicates that our model achieves higher detection accuracy while also maintaining an effective precision-recall balance. Furthermore, the model's mAP0.5:0.95 score of 74.19% demonstrates its consistent performance across a range of intersection-over-union thresholds, further highlighting its robustness. The proposed model operates with a relatively low parameter count of 7.93M and a GFLOP value of 16.65. While the parameter count is slightly higher than some of the other models, its superior detection performance, as evidenced by the mAP scores, justifies this trade-off. Overall, this result demonstrates that explicit orientation-aware feature fusion at the neck level is more parameter-efficient than relying on Transformer-based global attention for orientation modeling. Show more8:07 PM Claude responded: The strong performance on UCAS-AOD is consistent with the aspect-ratio analysis reported earlier. Aircraft dominate this dataset, with aspect ratios reaching up to 8:1. These are exactly the targets for which orientation-aware channel weighting delivers the largest gains.

TABLE V. COMPARISON OF STATE-OF-THE-ART MODELS ON THE UCAS-AOD DATASET

Model	$mAP_{0.5}$	$mAP_{0.5:0.95}$	Params(M)	GFLOPs
Faster-RCNN	89.43	72.47	38.43	11.3
NAS-FPN	90.64	73.93	34.20	9.49
RT-DETR	90.85	71.25	31.27	16.1
Zhang et al.	91.46	72.79	28.73	7.2
Li et al.	92.86	73.73	24.56	7.3
S^2Anet	89.87	72.55	49.14	15.6
SASM	86.54	69.63	41.58	15.8
Oriented RepPoints	85.72	68.16	36.67	20.4
Gliding Vertex	91.65	73.23	41.32	15.56
ARS-DETR	90.13	72.13	61.60	15.84
Oriented-Former	92.25	72.98	58.47	16.38
YOLOv11s-OBB	95.53	73.57	7.72	16.47
Ours	**98.44**	**74.19**	**7.93**	**16.65**

E. Visualization

To evaluate the performance of YOLO-RSL, representative samples were selected from both DIOR and UCAS-AOD for visual inspection and quantitative testing. Fig. 8 compares detection results of YOLOv11s-OBB and our method on DIOR-R. The orientation-agnostic baseline frequently misses densely arranged targets at non-canonical orientations and generates false positives in cluttered backgrounds with repeated texture patterns. YOLO-RSL can accurately identify and locate small targets (aircraft and vehicles less than 30 × 30 pixels). YOLO-RSL extracts the local characteristics of small objects through its improved feature enhancement module, thus improving the detection accuracy. In scenarios involving large

objects, the recognition accuracy of the baseline algorithm is also low. The most significant differences are in the ship and aircraft categories, which have the highest aspect ratio. These remarkable results are due to the improvement of the YOLO-RSL feature fusion path and backbone network.

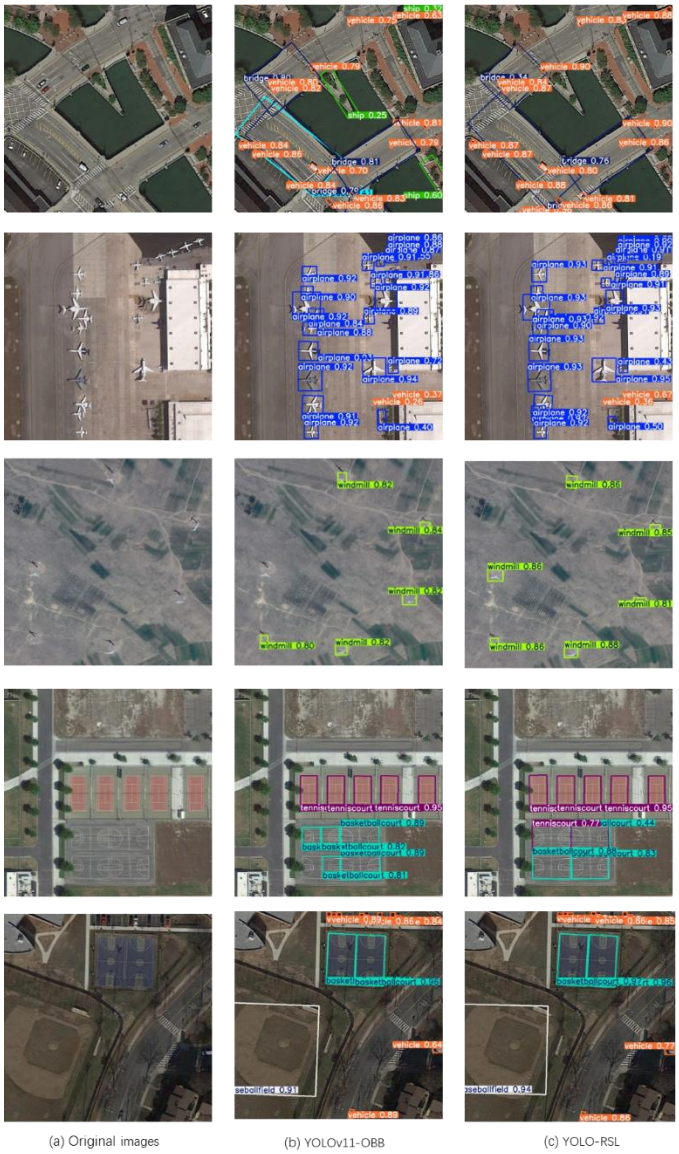

Fig. 8. Comparison of detection results on the DIOR dataset.

Fig. 9 shows the test results of YOLO-RSL on the UCAS-AOD dataset. Our method shows high accuracy in target detection, especially in aircraft detection, and the mAP value of the model is relatively high. The mAP0.5 value of automobile detection is slightly lower than that of aircraft detection, but the overall score is still very high. This result shows that the model maintains good robustness for different types of objects. In high-density scenarios such as airport aprons, the direction-sensitive channel weighting mechanism of RO-iAFF can also be handled well. However, the confidence score of some test results is lower. These low-trust situations are concentrated in two specific scenarios: the target is blocked, and the imaging platform is different from the perspective of the training data. At this time, the shape and proportion of the target may be very different from the content recognized by the model, resulting in a decrease in detection confidence. The feature extraction component designed to be more stable under changes in viewing angle and occlusion is a far-reaching research direction in the future.

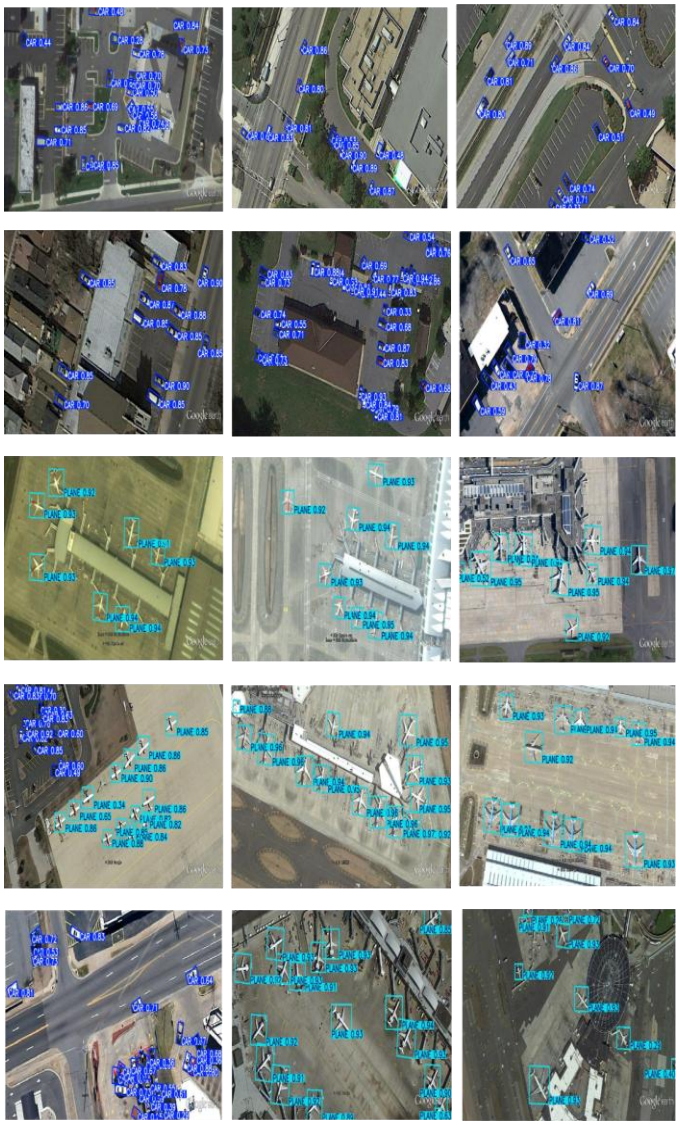

Fig. 9. The visualization of detection on the UCAS-AOD dataset.

V. Conclusion

This study presents YOLO-RSL, a unified framework that operates across three synergistic levels: backbone representation, neck-level feature fusion, and bounding-box regression supervision. First, the novel RO-iAFF module injects rotation-angle embeddings into the dual-stage iterative attention fusion, making cross-scale channel weighting explicitly orientation-aware; this represents the first known integration of rotation prior into the multi-scale feature fusion stage of a YOLO-series detector. Second, Swin Transformer blocks embedded in the YOLOv11s-OBB backbone enrich global feature representation while preserving inference speed. Third, the LFFE loss jointly constrains four geometric

attributes of predicted OBBs, yielding more stable and complete regression supervision. Experimental results on DIOR-R and UCAS-AOD validate the effectiveness of each component and demonstrate that YOLO-RSL achieves state-of-the-art accuracy with fewer parameters than transformer-heavy competitors. Future work will focus on extending orientation awareness to the backbone stage, incorporating multi-modal optical and SAR fusion, and applying knowledge distillation for edge-device deployment.

REFERENCES

[1] T. Zhang, P. Shen, and L. Wang, "A review of remote sensing applications in urban planning and environmental monitoring," Remote Sensing, vol. 15, no. 8, p. 2053, 2023.

[2] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 28, 2015, pp. 91–99.

[3] W. Liu, D. Anguelov, D. Erhan et al., "SSD: Single shot multibox detector," in Proc. Eur. Conf. Comput. Vis. (ECCV), 2016, pp. 21–37.

[4] X. Yang, J. Yan, W. Liao et al., "SCRDet++: Detecting small, cluttered and rotated objects via instance-level feature denoising and rotation loss smoothing," IEEE Trans. Pattern Anal. Mach. Intell., vol. 45, no. 2, pp. 2384–2399, 2023.

[5] G. Cheng, Y. Zhu, Y. Lin, S. Wang et al., "R³Det: Refined single-stage detector with feature refinement for rotating object detection," IEEE Trans. Image Process., vol. 31, pp. 1213–1226, 2022.

[6] W. Wang, E. Xie, X. Li, D. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2021, pp. 568–578.

[7] S. Mehta and M. Rastegari, "MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops, 2022, pp. 109–119.

[8] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," in Proc. Int. Conf. Learn. Representations (ICLR), 2021.

[9] J. Zhao, Z. Ding, Y. Zhou, H. Zhu, W. Du, R. Yao, and A. El Saddik, "OrientedFormer: An end-to-end transformer-based oriented object detector in remote sensing images," IEEE Trans. Geosci. Remote Sens., vol. 62, pp. 1–16, 2024, doi: 10.1109/TGRS.2024.3456240.

[10] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018, pp. 7132–7141.

[11] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 3–19.

[12] Y. Dai, F. Gieseke, S. Oehmcke, Y. Wu, and K. Barnard, "Attentional feature fusion," in Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV), 2021, pp. 3560–3569.

[13] Y. Yang, H. Liu et al., "CSL-YOLO: A YOLO-based remote sensing-oriented object detector with circular smooth label," IEEE Geosci. Remote Sens. Lett., 2022.

[14] Y. Yu et al., "H2RBox-v2: Incorporating symmetry for oriented object detection," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 36, 2023.

[15] J. M. Llerena, L. F. Zeni, L. N. Kristen et al., "Gaussian bounding boxes and probabilistic intersection-over-union for object detection," arXiv preprint arXiv:2106.06072, 2021.

[16] Ultralytics, "YOLO11 Documentation," 2024. [Online]. Available: https://docs.ultralytics.com/models/yolo11/

[17] A. Srinivas, T.-Y. Lin, N. Parmar, J. Shlens, P. Abbeel, and A. Vaswani, "Bottleneck transformers for visual recognition," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2021, pp. 16519–16529.

[18] Y. Zeng et al., "ARS-DETR: Aspect ratio-sensitive detection transformer for aerial oriented object detection," IEEE Trans. Geosci. Remote Sens., vol. 62, pp. 1–15, 2024, doi: 10.1109/TGRS.2024.3303814.

[19] A. Maharek, A. Abozeid, R. Orban et al., "SwinVid: Enhancing video object detection using Swin Transformer," Comput. Syst. Sci. Eng., vol. 48, no. 2, pp. 305–320, 2024.

[20] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2021, pp. 13713–13722.

[21] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2961–2969.

[22] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2980–2988.

[23] G. Jocher, A. Chaurasia, and J. Qiu, "YOLO by Ultralytics," GitHub repository, 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[24] A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, and J. Han, "YOLOv10: Real-time end-to-end object detection," arXiv preprint arXiv:2405.14458, 2024.

[25] X. Hou, M. Liu, S. Zhang et al., "CANet: Contextual information and spatial attention-based network for detecting small defects in manufacturing industry," Pattern Recognit., vol. 140, p. 109558, 2023.

[26] Y. Luo, X. Cao, J. Zhang et al., "CE-FPN: Enhancing channel information for object detection," Multimed. Tools Appl., vol. 81, no. 21, pp. 30685–30704, 2022.

[27] J. Zhan et al., "YOLOPX: Anchor-free multi-task learning network for panoptic driving perception," Pattern Recognit., vol. 148, p. 110152, 2024.

[28] N. Wang, Y. Gao, H. Chen et al., "NAS-FCOS: Efficient search for object detection architectures," Int. J. Comput. Vis., vol. 129, pp. 3299–3312, 2021.

[29] Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang, Y. Liu, and J. Chen, "DETRs beat YOLOs on real-time object detection," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2024, pp. 16965–16974.

[30] X. Zhang, Z. Gong, H. Guo et al., "Adaptive adjacent layer feature fusion for object detection in remote sensing images," Remote Sensing, vol. 15, no. 17, p. 4224, 2023.

[31] J. Li, M. Chen, S. Hou et al., "An improved S2A-Net algorithm for ship object detection in optical remote sensing images," Remote Sensing, vol. 15, no. 18, p. 4559, 2023.

[32] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

[33] Y. Zeng et al., "ARS-DETR: Aspect ratio-sensitive detection transformer for aerial oriented object detection," IEEE Trans. Geosci. Remote Sens., vol. 62, pp. 1–15, 2024, doi: 10.1109/TGRS.2024.3303814.

[34] L. Hou, K. Lu, J. Xue, and Y. Li, "Shape-adaptive selection and measurement for oriented object detection," in Proc. AAAI Conf. Artif. Intell., 2022, pp. 923–932.