Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.058952

ARTICLE

# ASL-OOD: Hierarchical Contextual Feature Fusion with Angle-Sensitive Loss for Oriented Object Detection

**Kexin Wang[1,#], Jiancheng Liu[1,#,*], Yuqing Lin[2,*], Tuo Wang[1], Zhipeng Zhang[1], Wanlong Qi[1], Xingye Han[1] and Runyuan Wen[3]**

[1]Northwest Institute of Mechanical and Electrical Engineering, Xianyang, 712099, China

[2]School of Information Engineering, Chang'an University, Xi'an, 710064, China

[3]School of Computer Science and Technology, Xidian University, Xi'an, 710071, China

*Corresponding Authors: Jiancheng Liu. Email: jianchengliu@njust.edu.cn; Yuqing Lin. Email: yuqinglin@chd.edu.cn

[#]Kexin Wang and Jiancheng Liu contributed equally to this work

Received: 24 September 2024 Accepted: 03 December 2024 Published: 17 February 2025

## ABSTRACT

Detecting oriented targets in remote sensing images amidst complex and heterogeneous backgrounds remains a formidable challenge in the field of object detection. Current frameworks for oriented detection modules are constrained by intrinsic limitations, including excessive computational and memory overheads, discrepancies between predefined anchors and ground truth bounding boxes, intricate training processes, and feature alignment inconsistencies. To overcome these challenges, we present ASL-OOD (Angle-based SIOU Loss for Oriented Object Detection), a novel, efficient, and robust one-stage framework tailored for oriented object detection. The ASL-OOD framework comprises three core components: the Transformer-based Backbone (TB), the Transformer-based Neck (TN), and the Angle-SIOU (Scylla Intersection over Union) based Decoupled Head (ASDH). By leveraging the Swin Transformer, the TB and TN modules offer several key advantages, such as the capacity to model long-range dependencies, preserve high-resolution feature representations, seamlessly integrate multi-scale features, and enhance parameter efficiency. These improvements empower the model to accurately detect objects across varying scales. The ASDH module further enhances detection performance by incorporating angle-aware optimization based on SIOU, ensuring precise angular consistency and bounding box coherence. This approach effectively harmonizes shape loss and distance loss during the optimization process, thereby significantly boosting detection accuracy. Comprehensive evaluations and ablation studies on standard benchmark datasets such as DOTA with an mAP (mean Average Precision) of 80.16 percent, HRSC2016 with an mAP of 91.07 percent, MAR20 with an mAP of 85.45 percent, and UAVDT with an mAP of 39.7 percent demonstrate the clear superiority of ASL-OOD over state-of-the-art oriented object detection models. These findings underscore the model's efficacy as an advanced solution for challenging remote sensing object detection tasks.

## KEYWORDS

Oriented object detection; transformer; deep learning

## 1 Introduction

The rapid advancement of computer vision and optical remote sensing technologies has brought significant challenges to the forefront of target detection, particularly in addressing complex image backgrounds, varying resolutions, and the inherently multi-scale nature of military targets. Object detection techniques can be broadly divided into traditional methods that depend on manually engineered features and deep learning approaches that leverage automatically learned representations. As the complexity of image backgrounds and resolutions continues to increase, manually designed features have proven insufficient in capturing the variability and fine-grained details of intricate targets, leading to degraded detection performance. Conversely, the emergence of deep learning-based object detection methodologies, powered by convolutional neural networks and other sophisticated architectures, has enabled the autonomous extraction of image features. These innovations have demonstrated superior efficacy in overcoming the challenges posed by complex scenarios, establishing themselves as state-of-the-art solutions in this field [1].

Deep learning-based target detection methodologies have been refined into three principal paradigms. The traditional two-stage framework, typified by Region-Based Convolutional Neural Network (R-CNN) [2] and Faster R-CNN [3], employs region proposal algorithms to generate candidate bounding boxes, which are then refined and classified using deep neural networks. Conversely, one-stage methods, such as the You Only Look Once (YOLO) series [4,5], recast the detection task as a regression problem, enabling the concurrent prediction of bounding box coordinates and object categories with heightened computational efficiency. In recent advancements, Transformer-based approaches, exemplified by the Detection Transformer [6] and Meta's Segment Anything Model [7], have emerged as a groundbreaking paradigm. These models harness sophisticated attention mechanisms to model global dependencies among image features, achieving state-of-the-art performance in intricate scenarios featuring densely distributed or overlapping objects.

Despite considerable progress, object detection methods that utilize horizontal bounding boxes as anchors exhibit critical shortcomings when addressing complex backgrounds and the detection of multi-scale military targets. Firstly, when the size of a horizontal bounding box exceeds the ground truth boundaries of smaller objects, precise localization becomes increasingly difficult, leading to deviations that degrade detection performance. Secondly, in cases where a horizontal bounding box encloses multiple densely packed small objects, the conflation of feature information often results in feature loss or blurring, thereby hindering accurate feature extraction and classification. Lastly, in intricate scenarios such as maritime traffic involving ships, horizontal bounding boxes frequently encompass substantial background regions, heightening the risk of background interference and increasing the likelihood of misclassification.

In recent years, oriented object detection has undergone significant advancements, driven by innovative methodologies and architectural developments. Early methods, such as oriented region proposal networks (ORPN) [8], employed predefined anchors with multiple angles and scales (e.g., 54 angles per position), yet faced severe limitations due to their substantial computational and memory overheads. Furthermore, the reliance on manually designed anchors often resulted in suboptimal alignment with ground truth, thereby constraining detection accuracy, especially in complex and cluttered scenes. To address these challenges, the Region of Interest (RoI) Transformer [9] was introduced, which converts horizontal bounding boxes into oriented ones using a learned RoI transformation strategy. Although the RoI Transformer achieved commendable results, its dependence on horizontal box transformations inherently limited its effectiveness for densely distributed and multi-scale objects. SANet (Structure-aware network) [10] pushed the boundaries further by integrating Convolutional Neural Networks

(CNNs) to extract discriminative shape features and Recurrent Neural Networks (RNNs) to capture contextual dependencies. However, the combined use of CNN and RNN architectures introduced issues such as feature bias and orientation misalignment, which remained prevalent in scenarios with occlusion or high-density object distributions, ultimately reducing the accuracy of oriented bounding box predictions.

To tackle these challenges, we present ASL-OOD, a cutting-edge one-stage framework meticulously crafted to enhance the detection of oriented objects in complex, multi-scale environments. ASL-OOD introduces several pivotal innovations, notably the integration of TB and TN architecture, complemented by ASDH. The TB module combines the c2f module from YOLOv8 with the Swin Transformer, facilitating advanced multi-scale feature extraction and efficient modeling of global dependencies. This hybrid design not only supports the detection of objects across varying scales but also enhances the representation of intricate, fine-grained details. The TN module further refines detection capabilities, particularly for small and densely packed objects, by incorporating long-range dependencies and generating high-resolution feature representations. The ASDH module leverages an angle-sensitive SIOU loss function to address the shortcomings of conventional horizontal bounding box approaches. By ensuring precise alignment between predicted and ground truth angles, this component significantly enhances bounding box orientation consistency, leading to superior detection performance in demanding scenarios. Our work introduces key innovations that distinguish it from existing methods:

1. The integration of the Swin Transformer into Backbone and Neck modules enhances multi-scale feature representation and leverages attention mechanisms to capture long-range dependencies, achieving superior detection accuracy in complex, occluded, and multi-scale scenarios.

2. By introducing an angle offset representation and a novel Angle-SIOU loss, we address the limitations of horizontal box-based methods, ensuring precise orientation alignment and improving the detection of symmetric objects like vehicles and aircraft in military and remote sensing applications.

3. Extensive evaluations on DOTA, HRSC2016, MAR20, and UAVDT demonstrate ASL-OOD's mAP superiority, showcasing its robustness and generalization in handling dense, occluded, and symmetric targets where traditional models falter.

## 2 Related Works

In recent years, object detection has undergone transformative advancements, propelled by the emergence of convolutional networks and Transformer-based architectures [11]. Despite these innovations, traditional horizontal detectors remain fundamentally constrained, limiting their applicability in complex scenarios. To overcome these shortcomings, oriented object detection has gained prominence as a promising paradigm with diverse applications [12]. To enhance oriented bounding box regression, SCRDet (Detection for Small, Cluttered and Rotated) [13] introduces a feature fusion architecture that leverages anchor sampling angles and refines smooth L1 loss with an Intersection Over Union (IOU) constant factor, effectively addressing boundary inconsistencies.

Additionally, SCRDet incorporates a supervised multi-dimensional attention network to mitigate the influence of background noise. Tackling the challenge of densely packed small targets in remote sensing images, Ding et al. propose the RoI Transformer module [14], which significantly improves detection accuracy within a two-stage detection framework. The RoI Transformer utilizes an oriented RoI module to transform horizontal RoIs into oriented RoIs, facilitating the extraction of orientation-invariant features through RoI Warping for downstream classification and regression tasks. Xu et al. advance the field by introducing a gliding vertices representation [15], enabling arbitrary quadrilateral object detection through vertex offset learning within the head regression branch of Faster R-CNN. However, these approaches continue to depend on horizontal RoIs for classification and regression, limiting their ability to fully address the challenges associated with horizontal bounding boxes in complex scenarios.

In one-stage object detection networks, the removal of region proposal generation and RoI alignment operations substantially enhances the efficiency and simplicity of oriented object detection [16,17]. Cheng et al. proposed $C^2$-YOLO [18], which incorporates angle prediction and an angle loss mechanism into the YOLOv5 framework, enabling accurate and streamlined detection of oriented objects. Similarly, Han et al. introduced the SANet framework [19], which leverages a CNN for inter-class feature discrimination and an RNN for structural modeling, effectively addressing discrepancies between classification confidence and positional precision. To resolve challenges related to spatial and feature alignment as well as regression uncertainty in label assignment, Ming et al. proposed the dynamic anchor learning method [20]. This approach employs a novel matching degree metric to comprehensively evaluate anchor positioning capabilities, ensuring optimal label allocation. Pan et al. advanced the field with the dynamic refinement network (DRN) [21], which dynamically adjusts neuron receptive fields based on the shape and orientation of target objects while refining predictions in a target-aware manner, significantly improving detection accuracy.

This study introduces ASL-OOD, an advanced one-stage framework for oriented object detection. By moving beyond the constrained contextual representation of traditional YOLO models reliant on convolutional neural networks, ASL-OOD employs the Swin-Transformer to deliver comprehensive multi-scale feature representation, markedly enhancing detection performance across varying target sizes. Additionally, an effective angle offset representation model, integrated with an angle-based SIOU loss, is proposed to improve detection precision and robustness, particularly for military targets in complex environments.

## 3 Method

The architecture of ASL-OOD, shown in Fig. 1, comprises three core components: the Transformer-based Backbone, the Transformer-based Bidirectional Feature Pyramid Network (BiFPN) Neck, and the Angle-SIOU-based Decoupled Head. Traditional YOLO models, constrained by the limited receptive field of convolutional operations, face challenges in capturing global context from ground images with complex scenes, varying resolutions, and small-scale targets. ASL-OOD addresses these limitations by incorporating the Swin-Transformer within the Backbone and BiFPN Neck, enabling efficient multi-level feature extraction and fusion. The Angle-SIOU technique extends SIOU to accommodate targets with extreme aspect ratios, while the decoupled head ensures accurate generation of oriented bounding boxes.

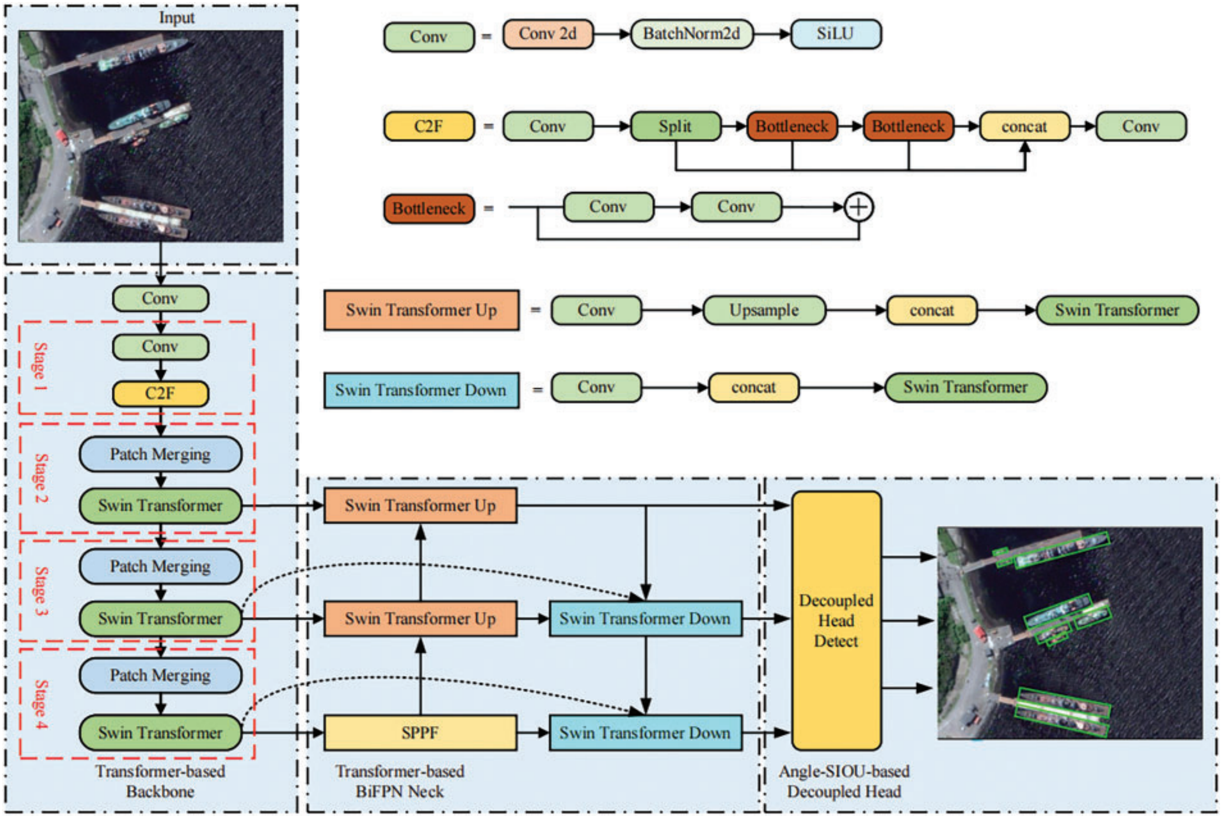

Input
Conv = Conv 2d → BatchNorm2d → SiLU
C2F = Conv → Split → Bottleneck → Bottleneck → concat → Conv
Bottleneck = Conv → Conv → ⊕
Swin Transformer Up = Conv → Upsample → concat → Swin Transformer
Swin Transformer Down = Conv → concat → Swin Transformer
Conv
Stage 1
Conv
C2F
Stage 2
Patch Merging
Swin Transformer
Stage 3
Patch Merging
Swin Transformer
Stage 4
Patch Merging
Swin Transformer
Transformer-based Backbone
Swin Transformer Up
Swin Transformer Up
Swin Transformer Down
SPPF
Swin Transformer Down
Transformer-based BiFPN Neck
Decoupled Head Detect
Angle-SIOU-based Decoupled Head

**Figure 1:** Architecture of ASL-OOD. Features are extracted via a transformer-based backbone and fused through bidirectional neck to enhance multi-scale representation. The decoupled head, utilizing the proposed oriented definition, generates high-quality detection boxes with precise alignment

### *3.1 Transformer-Based Backbone (TB)*

In complex scenarios, objects vary widely in scale and aspect ratio, with small objects like cars and bridges occupying only a few pixels, while larger objects, such as planes and ships, cover extensive areas. Capturing both shallow local features and deep global features is essential. However, traditional YOLO networks, based on CNNs, struggle to model global contextual information and feature interdependencies due to the inherent limitations of convolutional kernels. To overcome these constraints, we propose an enhanced Backbone leveraging the Transformer mechanism, as shown in Fig. 2. By utilizing the self-attention mechanism of Transformers, the model effectively captures global feature interactions while improving the extraction of local target features. Specifically, we integrate the Swin-Transformer [22], which employs W-MSA (Window Multi-Head Self-Attention) to parallelize computations and reduce complexity. SW-MSA (Shifted Window Multi-Head Self Attention) enhances feature communication across windows, while the patch merging module downsamples feature maps to optimize resolution and channel dimensions, enabling a hierarchical design that boosts computational efficiency.

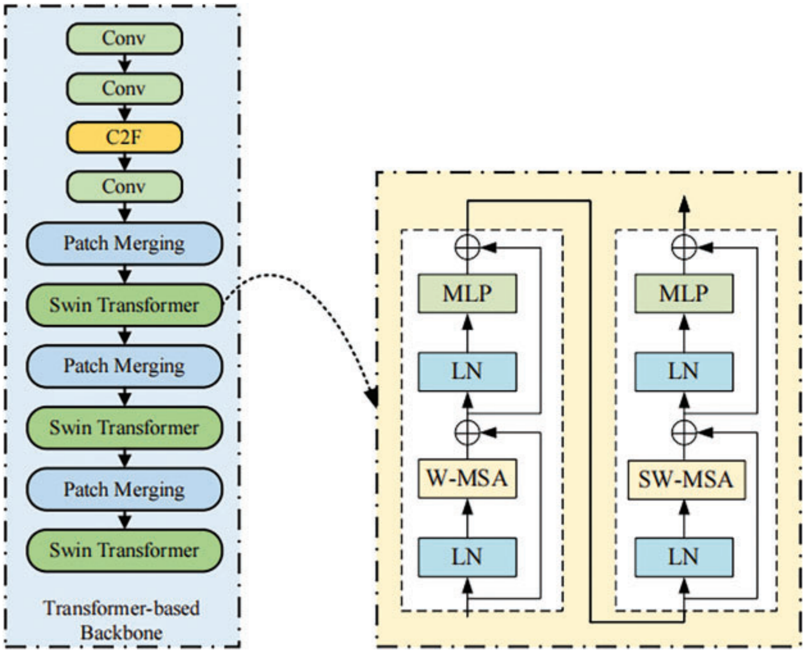


**Figure 2:** Architecture of transformer-based backbone

Initially, we replaced all convolutional layers in the Backbone with Swin-Transformer Blocks during experimentation. However, this led to a loss of essential contextual features, negatively impacting detection accuracy. To address this, we introduced the c2f structure from YOLOv8 at the Backbone's front end to enhance preliminary feature extraction. Additionally, Swin-Transformer Blocks were deployed in stages 2, 3, and 4 to capture low-resolution features efficiently. This design achieves a balance between performance and resource consumption, making the model more viable for industrial applications.

### *3.2 Transformer-Based Neck (TN)*

Within YOLO network architectures, the Neck module, as described in [23], serves as a vital component, linking the backbone network to the detection head. Its core objective is to refine up-down sampling and feature fusion processes across backbone features at multiple stages, thereby enhancing the model's capacity to detect targets of varying scales. The original YOLOv5 framework employs PANet (Path Aggregation Network) for feature fusion, utilizing both top-down and bottom-up pathways to generate multi-scale feature maps, achieving notable performance gains. Building upon this, BiFPN, introduced in [24], offers superior performance by integrating bidirectional connections and additional edges within the PANet framework, enabling more efficient cross-scale fusion. Furthermore, BiFPN incorporates learnable parameters to dynamically optimize feature weights across different levels during the fusion process.

However, accurately detecting minuscule objects remains a persistent challenge in our dataset due to their dependency on high-resolution features. BiFPN's top-down and bottom-up fusion mechanisms often struggle to retain or amplify the fine-grained details necessary for small object detection. Moreover, the sparse representation of these tiny objects within feature maps exacerbates information loss and dilution during the fusion process, further impairing detection accuracy.

To address these challenges, we present the TN module, a novel integration of BiFPN and Swin-Transformer. TN leverages localized window interactions and global path dependency modeling with weighted edges, enabling the effective capture of long-range dependencies critical for detecting tiny objects within expansive receptive fields. Additionally, TN adopts a hierarchical feature resolution reduction and rescaling mechanism, harnessing the strengths of the Swin-Transformer. This ensures the preservation and enhancement of feature resolution at multiple levels, facilitating precise localization and the retention of fine-grained details essential for small object detection. Lastly, TN employs an FPN-inspired multi-scale feature fusion framework, ensuring efficient information transfer across feature map levels through bidirectional sampling pathways. This design significantly improves the representation of small-scale targets across diverse contexts, enhancing detection accuracy and robustness in complex scenarios.

When fusing features of different scales from the TB module using our TN module, we employ fast normalized fusion to assign weights to the additional edges. The calculation of the normalized weight, denoted as $O$, is defined as follows:

$$O = \sum\nolimits_i \frac{\omega_i}{\epsilon + \sum\nolimits_j \omega_j} \cdot I_i, \tag{1}$$

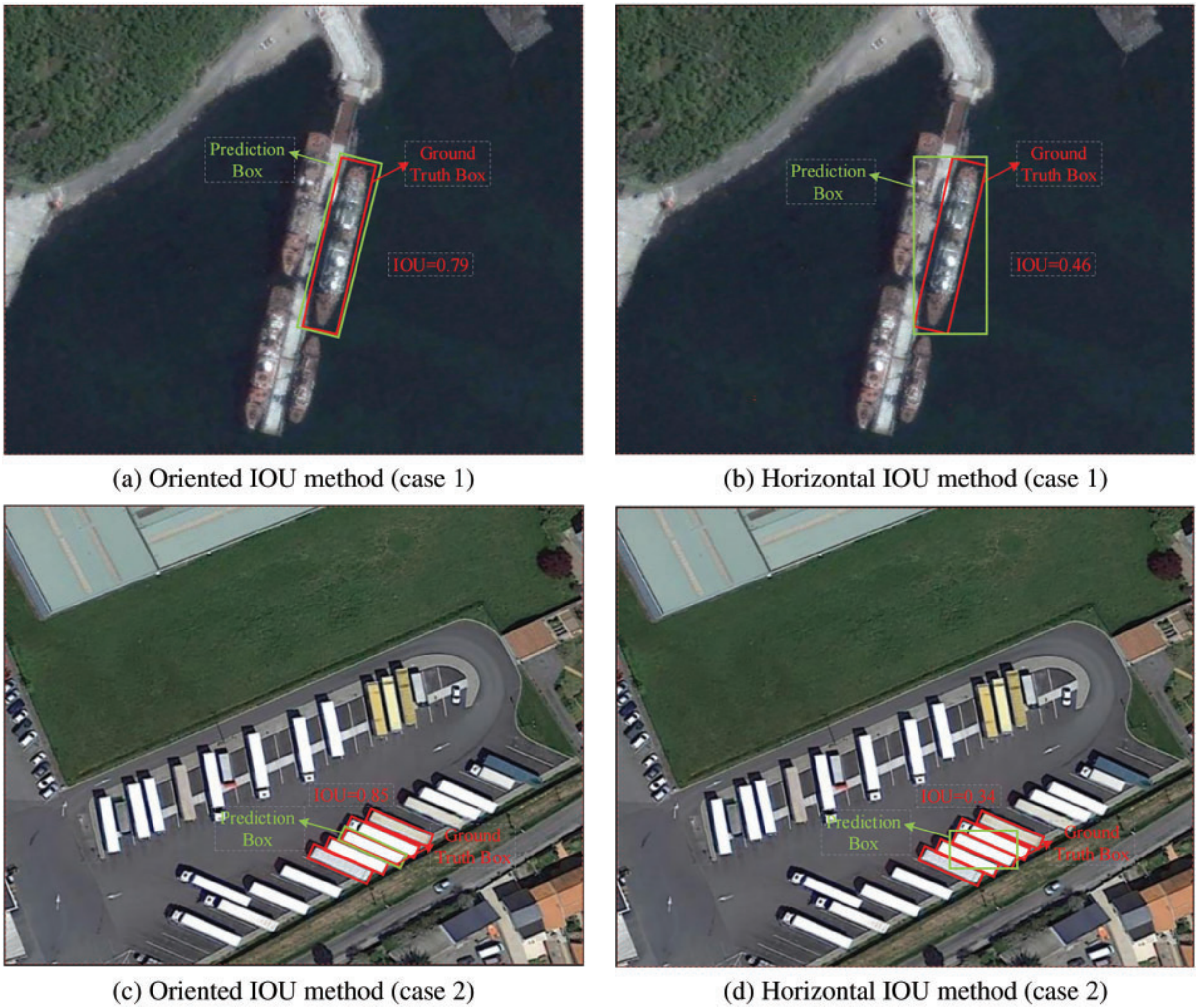
$$\beta = \begin{cases} \beta_{v1}, & if \sigma_v > \tau_v \\ 1, & otherwise \end{cases}, \tag{2}$$

where $\omega_i$ represents a learnable weight matrix, while $\epsilon = 0.0001$ is a small value introduced to mitigate numerical instability. This approach enables the learning of weights $\omega_i$ and performs a normalization operation, ensuring that the contribution of the additional edge is appropriately adjusted to enhance the effect of feature fusion.

### *3.3 Angle-SIOU-Based Decoupled Head*

In complex scenes, targets such as vehicles, buildings, or ships often possess distinct orientations or tilts, rendering horizontal bounding boxes inadequate for accurately delineating their shapes and boundaries, thus leading to unreliable detection. Moreover, horizontal boxes often encompass densely packed small objects and irrelevant background regions, injecting noise into feature computation and significantly degrading network performance.

To illustrate the limitations of traditional horizontal boxes, Fig. 3 provides examples of various scenarios. In Fig. 3a,b, horizontal boxes include significant irrelevant background information, increasing the risk of misclassifying background regions as objects or expanding bounding boxes to encompass extraneous areas. Fig. 3c,d reveals another challenge: horizontal boxes often enclose densely packed small objects, leading to feature merging that causes loss or blurring of target-specific details, hindering the accurate extraction and classification of individual small targets. In contrast, as shown in Fig. 3, oriented boxes achieve consistently higher IOU values, underscoring their ability to better constrain target positions and deliver more precise localization, effectively addressing the shortcomings of horizontal boxes.



(a) Oriented IOU method (case 1)

(b) Horizontal IOU method (case 1)

(c) Oriented IOU method (case 2)

(d) Horizontal IOU method (case 2)

**Figure 3:** Examples of different types of horizontal boxes and oriented boxes under the same ground truth

### *3.3.1 Angle-Offset Representation*

In this section, we propose a novel object representation scheme called angle-offset representation, which accurately represents a rotated bounding box using its minimum bounding rectangle and angle. Fig. 4 illustrates this representation, where the red box represents the ground truth oriented box of the object, the blue box represents the minimum bounding rectangle, and the black point denotes the midpoint of each side of the minimum bounding rectangle.

Let $(x, y)$ be the center coordinates of the ground truth oriented box, $\theta$ be the angle with respect to the $x-axis$, and $m$ and $n$ be the length and width of the minimum bounding rectangle, respectively. To obtain the offsets $\Delta x$ and $\Delta y$ of the oriented box relative to the bounding rectangle based on the angle $\theta$, we calculate:

$$\Delta x = \frac{m - nsin2\theta}{2cos2\theta} \text{ and } \Delta y = \frac{msin2\theta - n}{2cos2\theta}. \qquad (3)$$

By predicting the five values $(x, y, m, n, \theta)$, we can obtain the coordinate representation $(v_1, v_2, v_3, v_4)$ of the oriented box:

$$\begin{cases} v_1 = \left(x + \Delta x, y - \dfrac{n}{2}\right), v_2 = \left(x + \dfrac{m}{2}, y - \Delta y\right) \\ v_3 = \left(x - \Delta x, y + \dfrac{n}{2}\right), v_4 = \left(x - \dfrac{m}{2}, y + \Delta y\right) \end{cases}. \tag{4}$$

The angle-offset representation effectively encodes the angular deviation of the oriented box relative to its bounding rectangle, simplifying the interpretation of its orientation and shape by utilizing their spatial relationship. This approach facilitates direct prediction and decoding of the oriented box's center coordinates, orientation angle, and dimensions (length and width) of the bounding rectangle, obviating the need for complex post-processing steps. Furthermore, it offers a more compact and efficient representation of the position and shape of the oriented box compared to existing methods [25]. Based on this representation, our loss function integrates three components: angle-based distance loss, angle-based shape loss, and IOU loss.

### *3.3.2 Angle-Based Distance Loss*

Conventional distance loss focuses solely on the Euclidean distance between ground truth and predicted boxes, disregarding angular discrepancies that critically affect alignment. To resolve this limitation, we introduce angle-guided distance loss, which integrates 2D angular differences into the distance computation. This refinement mitigates the impact of distance-independent variables, enhancing the accuracy and robustness of the loss function.

In Fig. 5, the angle $\alpha$ represents the angle between the line connecting the predicted box and the ground truth box and the horizontal line. The center points of the predicted box and the ground truth box are denoted as $(x, y)$ and $(x^{gt}, y^{gt})$, respectively. The horizontal and vertical differences are denoted as $c_w$ and $c_h$, respectively. Inspired by the Selective IOU (SIOU), we introduce an angle coefficient $\lambda$:

$$\lambda = 1 + 2sin^2\left(arcsin\frac{c_h}{d} - \frac{\pi}{4}\right), d = \sqrt{(y - y^{gt})^2 + (x - x^{gt})^2}, \tag{5}$$

$$c_w = \max\left(x, x^{gt}\right) - \min\left(x, x^{gt}\right), c_h = \max\left(y, y^{gt}\right) - \min(y, y^{gt}). \tag{6}$$

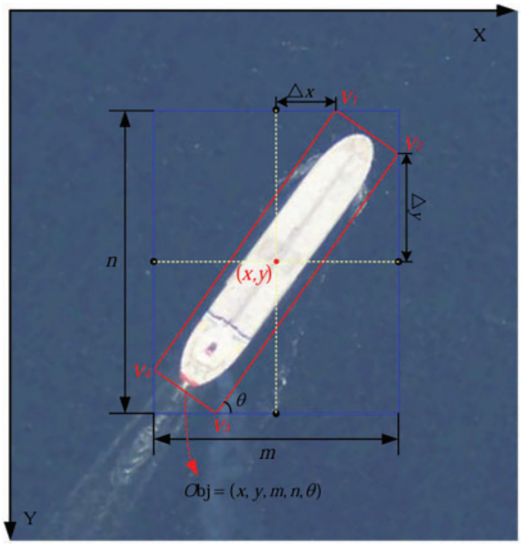


**Figure 4:** Examples of angle-offset representation

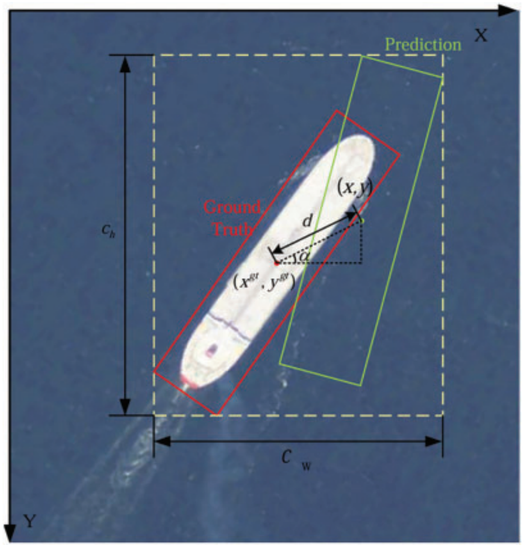


**Figure 5:** Examples of angle-based distance loss

The angle-based distance loss $\mathcal{L}_D$ can be defined as:

$$\mathcal{L}_D = \sum\nolimits_{t=(x,y)} (1 - e^{-\lambda\rho_t}), \tag{7}$$

$$\rho_x = \left(\frac{x^{gt} - x}{c_w}\right)^2, \rho_y = \left(\frac{y^{gt} - y}{c_h}\right)^2. \tag{8}$$

More specifically, when the angle $\alpha$ is 0 and the coefficient $\lambda$ is 2, the distance cost follows a more conventional formulation. Conversely, when $\alpha$ is $\frac{\pi}{4}$ and $\lambda$ is 1, the contribution of the distance cost is amplified, allowing the model to prioritize the consistency of distances between the ground truth box and the predicted box.

### *3.3.3 Angle-Based Shape Loss*

Detecting oriented objects with pronounced aspect ratios requires more than Euclidean distance, as it fails to fully capture their geometric and orientation-specific properties. Moreover, conventional shape loss is highly susceptible to pose variations and tilts, adversely affecting detection accuracy. To address these limitations, we incorporate an angle-aware component, $\gamma$, into the Angle-based Shape Loss. This enhancement improves the model's sensitivity to orientation changes, accurately captures the similarity between oriented objects, and substantially enhances detection precision for such targets.

As shown in the Fig. 6, we consider the calculation of the shape loss, where the ground truth oriented box has a length of $w^{gt}$ and a width of $h^{gt}$, while the predicted box has a length of $w$ and a width of $h$. The offset angles of the ground truth and predicted boxes are denoted as $\theta^{gt}$ and $\theta$, respectively.

The angle-aware component $\gamma$ can be defined as:

$$\gamma = 1 + 3sin^2\left(\theta^{gt} - \theta - \frac{\pi}{2}\right). \tag{9}$$

The angle-based shape loss $\mathcal{L}_S$ can be defined as:

$$\mathcal{L}_S = \sum\nolimits_{t=(w,h)} (1 - e^{-\mu_t})^{\gamma}, \tag{10}$$

$$\mu_w = \frac{|w - w^{gt}|}{\max(w, w^{gt})}, \mu_h = \frac{|h - h^{gt}|}{\max(h, h^{gt})}. \tag{11}$$

When the difference between the angles of the ground truth box and the predicted box, $\theta^{gt}$ and $\theta$, is 0, it indicates a small angle disparity between the oriented box and the real box. In such cases, we set $\gamma = 4$, which effectively reduces the shape cost. Conversely, when the difference $\theta^{gt} - \theta = \frac{\pi}{2}$ represents the largest angle difference between the oriented box and the real box, we assign $\gamma = 1$, thereby increasing the shape cost.

Our method outperforms SIOU in angle consistency by integrating variations in oriented angles, increasing the sensitivity of shape loss to orientation changes. Significant alterations in the object box's orientation result in a proportional increase in shape loss, effectively penalizing shape inconsistencies and improving the accuracy and stability of the box representation. Additionally, the Angle-based Shape Loss further enhances angle consistency by explicitly accounting for orientation variations. Its heightened responsiveness to substantial angular deviations ensures effective penalization of shape inconsistencies, significantly boosting the precision and stability of object box shape representation.

### *3.3.4 Angle-Based Loss Function*

Subsequent to the incorporation of the oriented angle component to guide both shape and distance loss, we introduce an IOU loss to comprehensively assess the position, shape, and overlap of the object bounding box. As shown in Fig. 7, the IOU loss, denoted as $\mathcal{L}_{IOU}$, quantifies the degree of overlap between the predicted object box and the ground truth object box, providing an evaluation of the accuracy of the object box. A low IOU loss indicates a small overlap between the predicted target box and the real target box, suggesting potential errors in the position or shape of the target box. The IOU loss is computed using the following formula:

$$\mathcal{L}_{IOU} = 1 - \frac{A \cap B}{A \cup B}, \tag{12}$$

where $A$ represents the ground truth box and $B$ represents the predicted box.

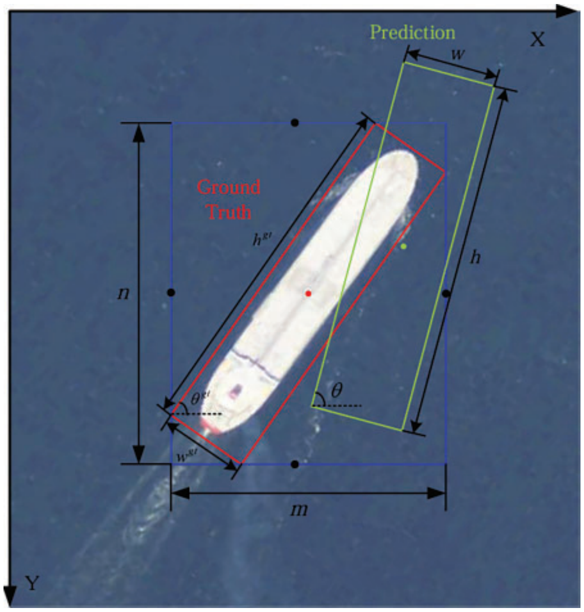


**Figure 6:** Examples of angle-based shape loss

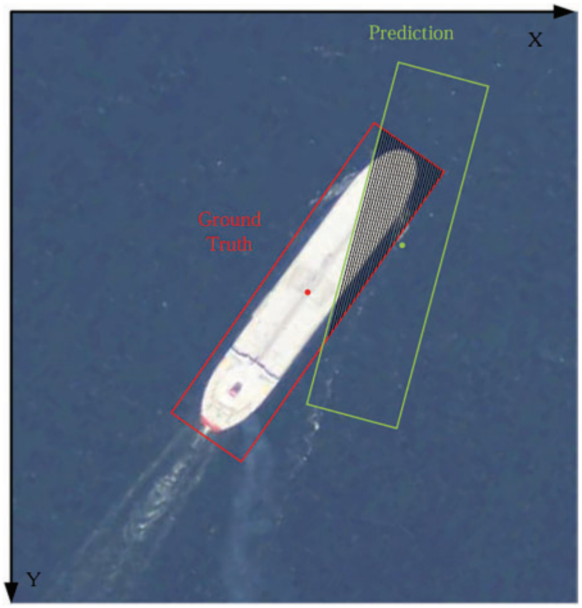


**Figure 7:** Examples of IOU loss

Optimizing the IOU loss ensures precise alignment between the predicted object box and the ground truth, effectively reducing errors in shape and position. This process guides the model during training to iteratively refine the object box's position and shape, thereby enhancing detection accuracy and localization precision. To holistically address the shape, position, and overlap of targets in ASL-OOD, we combine $\mathcal{L}_D$, $\mathcal{L}_S$, and $\mathcal{L}_{IOU}$. This comprehensive optimization approach enables the model to better adapt to objects with different shapes and oriented angles, thereby improving detection performance and robustness. The overall loss, denoted as $\mathcal{L}_{TAS}$, is defined as follows:

$$\mathcal{L}_{TAS} = \mathcal{L}_D + \mathcal{L}_S + \mathcal{L}_{IOU}. \tag{13}$$

Traditional YOLO architectures utilize a unified head module to simultaneously predict target location and category information. In this work, we adopt the decoupled head design from YOLOX [26], which separates these predictions into independent branches. This decoupling enables separate training and optimization of location and category components, enhancing model flexibility and improving the overall efficiency and effectiveness of the training process.

## 4 Experiments

To rigorously evaluate the performance of the proposed ASL-OOD model across a diverse range of object categories and scales, we conducted extensive experiments on multiple benchmark datasets. Specifically, we utilized DOTA [27], HRSC2016 [28], MAR20 [29], and UAVDT [30], each presenting distinct and challenging object detection scenarios. These datasets were carefully selected to provide a comprehensive evaluation of ASL-OOD's capabilities and robustness in handling complex detection tasks.

### *4.1 Datasets*

**DOTA Dataset:** The DOTA dataset is a comprehensive benchmark for object detection in aerial imagery, featuring 2806 high-resolution images and 188,282 labeled object instances spanning 15 diverse classes, including vehicles, infrastructure, and recreational areas such as small vehicles (SV), tennis courts (TC), and planes (PL). Image resolutions range from 800 × 800 to 4000 × 4000 pixels. For training preparation, the images were cropped into 608 × 608 sub-images with a 100-pixel stride and split into training, validation, and test sets in an 8:1:1 ratio. Multi-scale training and testing involved resizing images to scales of 0.5×, 1.0×, and 1.5×, followed by cropping into 1024 × 1024 patches with strides of 824 and 524 pixels. The ASL-OOD model was trained over 45 epochs with an initial learning rate of 0.005, which was decreased by a factor of 10 at epochs 24 and 33.

**HRSC2016 Dataset:** The HRSC2016 dataset, specifically designed for ship detection in maritime surveillance, consists of 1070 images capturing ships in a variety of orientations and scales. Image resolutions range from 300 × 300 to 1500 × 900 pixels. To maintain aspect ratios, images were resized such that the shorter side was scaled to 800 pixels, with the longer side constrained to a maximum of 1333 pixels. The dataset was divided into training, validation, and test subsets following an 8:1:1 split, consistent with the DOTA dataset. The ASL-OOD model was trained on HRSC2016 over 40 epochs, with an initial learning rate of 0.005, which was systematically reduced by a factor of 10 at epochs 24 and 33.

**UAV Datasets (UAVDT and MAR20):** To evaluate the generalization ability of our model in UAV (Unmanned Aerial Vehicle)-based target detection, we utilized two UAV-specific datasets: UAVDT and MAR20. While both focus on detecting targets in UAV imagery, they pose distinct challenges, and experiments were conducted separately to highlight their unique attributes.

The UAVDT dataset [30] comprises 100 video sequences with over 80,000 annotated frames, capturing urban environments at varying altitudes and under diverse weather conditions. It includes three primary detection categories (car, bus, and truck) with significant variations in scale, occlusion, and viewpoint. Frames extracted from the videos were resized to 1024 × 1024 pixels and divided into training, validation, and test sets with an 8:1:1 split. The model was trained on UAVDT for 50 epochs, initialized with a learning rate of 0.005, which was reduced by a factor of 10 at epochs 30 and 40. The mAP was employed as the primary evaluation metric, with IOU thresholds of 0.5 and 0.75 to assess detection performance under different precision requirements.

The MAR20 dataset targets the detection of 20 distinct aircraft types using remote sensing imagery. It contains 3800 images with a resolution of 800 × 800 pixels, divided into training, validation, and test sets in an 8:1:1 ratio, consistent with the UAVDT dataset. The model was trained on MAR20 for 75 epochs, with an initial learning rate of 0.005, reduced by a factor of 10 at epochs 50 and 65. The mAP was employed as the primary evaluation metric, with performance assessed on both small- and large-scale targets to gauge the model's robustness in handling diverse target types.

### *4.2 Evaluation Metrics*

In our experiments, mAP serves as the primary evaluation metric, using a fixed IOU threshold of 0.5, referred to as mAP@0.5. This metric evaluates the detection performance by quantifying the overlap between predicted bounding boxes and ground truth, with predictions deemed correct if the IOU is equal to or greater than 0.5. The average precision (AP) for each class is determined as the area under the precision-recall curve, defined as:

$$AP = \int_{0}^{1} p(r)dr, \tag{14}$$

where $p(r)$ is the precision at recall level $r$. The mean of $AP$ values across all object classes gives the $mAP$:

$$mAP@0.5 = \frac{1}{N}\sum\nolimits_{i=1}^{N} AP_i, \tag{15}$$

where $N$ is the number of object classes, and $AP_i$ is the AP for class $i$ at $IOU = 0.5$. In all subsequent experiments, $mAP$ refers exclusively to $mAP@0.5$, meaning that the performance is evaluated based on an IOU threshold of 0.5. This metric reflects the model's ability to detect objects with a moderate overlap requirement, making it a reliable measure for general object detection tasks.

### *4.3 Comparison with State-of-the-Art Methods*

**Results on the DOTA Dataset:** The experimental results on the DOTA dataset are summarized in Table 1, detailing the category-wise AP for each class and the overall mAP achieved by the ASL-OOD model. For comparison, existing object detection methods are categorized into one-stage and two-stage models. As shown in Table 1, traditional one-stage models like SSD (Single Shot MultiBox Detector), which are limited to horizontal detection, struggle to address the oriented characteristics of targets in complex scenes, leading to suboptimal performance. In contrast, ASL-OOD consistently outperforms other oriented detection models, demonstrating superior performance for dense, oriented small objects and large objects. On the test set, ASL-OOD achieves an impressive mAP of 80.16%. These results highlight ASL-OOD's ability to effectively model the oriented relationships between ground-truth and predicted boxes by incorporating angle-aware components and optimizing the loss function. This approach mitigates the influence of distance-independent variables, enhances detection accuracy, and improves the modeling of oriented objects, ultimately bolstering the model's robustness across diverse detection scenarios.

**Table 1:** Comparison with state-of-the-art methods on DOTA dataset

| Methods | PL | BD | BR | GTF | SV | LV | SH | TC | BC | ST | SBF | RA | HA | SP | HC | mAP |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **One-stage** | | | | | | | | | | | | | | | | |
| PIOU [31] | 80.90 | 69.70 | 24.10 | 60.20 | 38.30 | 64.40 | 64.80 | 90.90 | 77.20 | 70.40 | 46.50 | 37.10 | 57.10 | 61.90 | 64.00 | 60.50 |
| DRN [21] | 88.91 | 80.22 | 43.52 | 63.35 | 73.48 | 70.69 | 84.94 | 90.14 | 83.85 | 84.11 | 50.12 | 58.41 | 67.62 | 68.60 | 52.50 | 70.70 |
| RSDet [25] | 89.80 | 82.90 | 48.60 | 65.20 | 69.50 | 70.10 | 70.20 | 90.50 | 85.60 | 83.40 | 62.50 | 63.90 | 65.60 | 67.20 | 68.00 | 72.20 |
| R3Det [16] | 88.76 | 83.09 | 50.91 | 67.27 | 76.23 | 80.39 | 86.72 | 90.78 | 84.68 | 83.24 | 61.98 | 61.35 | 66.91 | 70.63 | 53.94 | 73.79 |
| SANet [19] | 89.11 | 82.84 | 48.37 | 71.11 | 78.11 | 78.39 | 87.25 | 90.83 | 84.90 | 85.64 | 60.36 | 62.60 | 65.26 | 69.13 | 57.94 | 74.12 |
| YOLOv5 [32] | 88.15 | 84.04 | 49.89 | 54.13 | 69.63 | 77.91 | 87.52 | 90.80 | 80.17 | 86.68 | 59.93 | 54.64 | 68.81 | 72.66 | 44.23 | 71.21 |
| C2-YOLO [18] | 87.19 | 83.16 | 48.77 | 51.83 | 78.88 | 85.07 | 88.12 | 90.07 | 82.83 | 87.62 | 57.65 | 65.57 | 75.95 | 80.69 | 49.72 | 74.20 |
| R-YOLO [33] | 90.15 | 84.51 | 54.27 | 68.45 | 78.86 | 86.97 | 89.32 | 90.84 | 74.26 | 89.06 | 66.78 | 67.84 | 74.54 | 74.21 | 65.13 | 77.01 |
| K-CBST-YOLO [34] | 89.50 | 83.30 | 65.90 | 67.60 | 80.10 | 91.20 | 80.90 | 65.50 | 64.30 | 83.60 | 93.30 | 81.80 | 77.20 | 80.90 | 69.50 | 78.40 |

(Continued)

**Table 1 (continued)**

| Methods | PL | BD | BR | GTF | SV | LV | SH | TC | BC | ST | SBF | RA | HA | SP | HC | mAP |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Two-stage** | | | | | | | | | | | | | | | | |
| SCRDet [13] | 89.98 | 80.65 | 52.09 | 68.36 | 68.36 | 60.32 | 72.41 | 90.85 | 87.94 | 86.86 | 65.02 | 66.68 | 66.25 | 68.24 | 65.21 | 72.61 |
| Vertex [15] | 89.64 | 85.00 | 52.26 | 77.34 | 73.01 | 73.14 | 86.82 | 90.74 | 79.02 | 86.81 | 59.55 | 70.91 | 72.94 | 70.86 | 57.32 | 75.02 |
| MaskOBB [35] | 89.61 | 85.09 | 51.85 | 72.90 | 75.28 | 73.23 | 85.57 | 90.37 | 82.08 | 85.05 | 55.73 | 68.39 | 71.61 | 69.87 | 66.33 | 74.86 |
| FAOD [36] | 90.21 | 79.58 | 45.49 | 76.41 | 73.18 | 68.27 | 79.56 | 90.83 | 83.40 | 84.68 | 53.40 | 65.42 | 74.17 | 69.69 | 64.86 | 73.28 |
| FR-Est [37] | 89.63 | 81.17 | 50.44 | 70.19 | 73.52 | 77.98 | 86.44 | 90.82 | 84.13 | 83.56 | 60.64 | 66.59 | 70.59 | 66.72 | 60.55 | 74.20 |
| **Ours** | | | | | | | | | | | | | | | | |
| **ASL-OOD** | 91.23 | 86.42 | 63.99 | 65.89 | 82.98 | 88.03 | 90.19 | 91.87 | 87.99 | 89.31 | 65.99 | 69.37 | 81.66 | 83.13 | 64.39 | 80.16 |

We provide visualizations of representative detection results from the DOTA dataset using ASL-OOD, depicted in Fig. 8. These examples showcase the model's precision in capturing fine details of small objects and its robustness in handling targets of varying scales. The visualizations substantiate the efficacy of the proposed TN and TB modules, which leverage long-range dependency modeling, high-resolution feature representation, and multi-scale feature fusion. These advancements significantly improve detection accuracy and performance in complex object detection scenarios.

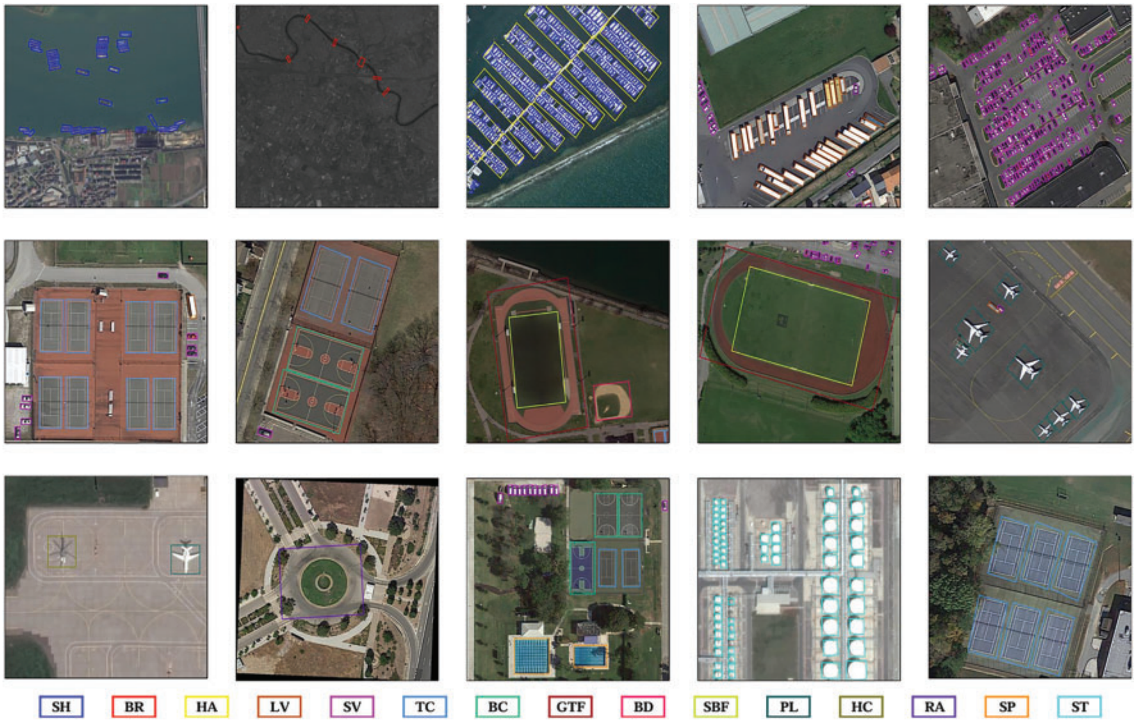


**Figure 8:** Examples of detection results on the DOTA dataset. Boxes of different colors correspond to objects of different colors

**Results on the HRSC2016 Dataset:** The performance of ASL-OOD on the HRSC2016 dataset is detailed in Table 2, with evaluations conducted using both the PASCAL VOC07 and VOC12 metrics. The mAP for VOC07 is computed using the 11-point interpolation method, which assesses precision at 11 evenly spaced recall levels (0, 0.1, 0.2, . . . , 1.0), providing a coarse yet consistent measure of detection performance across various recall levels. In contrast, VOC12 employs a continuous interpolation method, calculating precision at every point along the precision-recall curve for a more granular and comprehensive evaluation. As shown in Table 2, ASL-OOD achieves outstanding mAP scores of 91.07% on VOC07 and 98.13% on VOC12, outperforming all other listed methods. Fig. 9 presents visualizations of selected detection results from the HRSC2016 dataset, further illustrating ASL-OOD's superior accuracy and robustness in detecting oriented objects.

**Table 2:** Comparison with state-of-the-art methods on MAR20, UAVDT, and HRSC2016 datasets

| MAR20 | | UAVDT | | HRSC2016 | | |
|---|---|---|---|---|---|---|
| Methods | mAP | Methods | mAP | Methods | VOC 07 | VOC 12 |
| FCOS-O [38] | 70.69 | YOLOv3-tiny [39] | 26.90 | PIOU [31] | 89.20 | – |
| RetinaNet [40] | 73.43 | YOLOv4-tiny [41] | 27.70 | DRN [21] | – | 92.70 |
| SANet [19] | 81.10 | YOLOv5-S [42] | 29.80 | DAL [20] | 89.77 | – |
| ATSS [43] | 72.20 | YOLOX-tiny [26] | 29.10 | SANet [19] | 90.17 | 95.01 |
| Faster R-CNN [3] | 81.35 | PP-PicoDet-L [44] | 31.10 | C2-YOLO [18] | 89.77 | – |
| RoI-tran [14] | 82.72 | YOLOv7-tiny [45] | 31.20 | RoI-tran [14] | 86.20 | – |
| Vertex [15] | 81.48 | YOLOv8-S [46] | 31.90 | Vertex [15] | 88.20 | – |
| Oriented R-CNN [47] | 81.92 | LWUAVDet-S [48] | 34.10 | HLA [49] | 90.42 | – |
| HLA [49] | 83.97 | YOLOv10 [5] | 36.00 | Oriented R-CNN [47] | 90.50 | 97.60 |
| **ASL-OOD** | **85.45** | **ASL-OOD** | **39.70** | **ASL-OOD** | **91.07** | **98.13** |

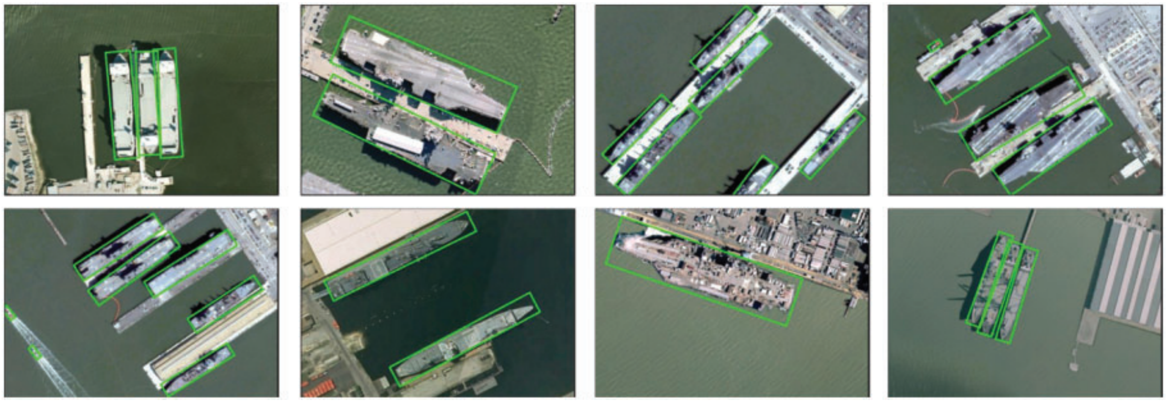


**Figure 9:** Examples of detection results on the HRSC2016 dataset

**Results on the MAR20 & UAVDT Dataset:** To further validate ASL-OOD's performance in specific domains, we conducted experiments on the MAR20 dataset, with results detailed in Table 2. ASL-OOD achieved an impressive mAP of 85.45% across 15 aircraft categories, significantly surpassing all comparison models. Visualizations in Fig. 10 demonstrate ASL-OOD's ability to leverage its angle-offset scheme, precisely capturing the position and shape of symmetric targets such as airplanes, thereby ensuring more reliable detection results. For evaluating ASL-OOD's rotational detection capabilities, particularly for highly symmetric objects, we also conducted experiments on the UAVDT dataset. As presented in Table 2, ASL-OOD attained a remarkable mAP of 39.7%, outperforming all state-of-the-art models on this challenging dataset. UAVDT introduces distinct challenges, including small, densely packed, and heavily occluded targets within complex urban settings. Many of these targets, such as vehicles, exhibit high symmetry, often leading to detection ambiguities in conventional models. ASL-OOD effectively addresses these challenges, demonstrating enhanced robustness and precision in these demanding scenarios.

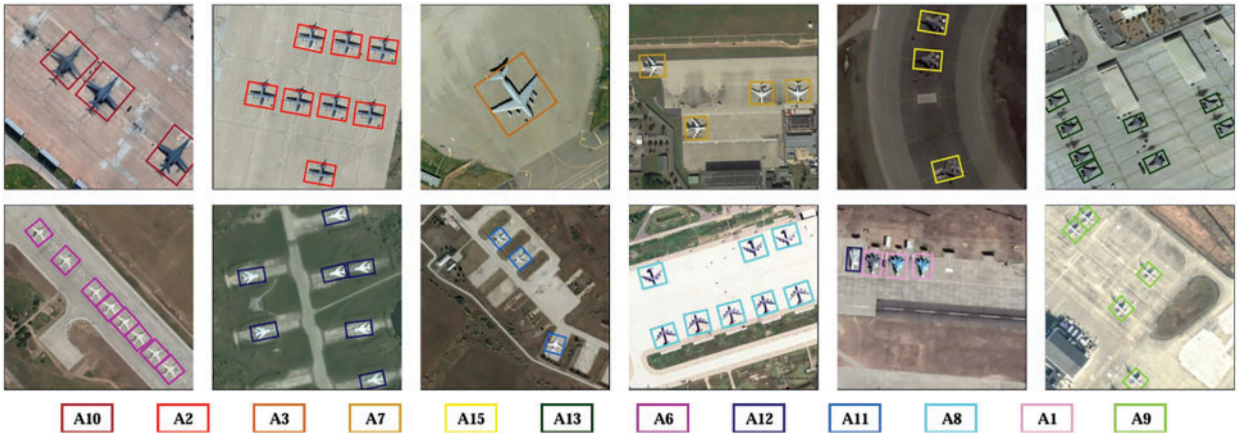


**Figure 10:** Examples of detection results on the MAR20 dataset. Boxes of different colors correspond to objects of different colors

ASL-OOD, powered by its TB and TNmodules, effectively overcomes these challenges with its sophisticated rotational detection mechanism. By integrating an angle offset scheme, ASL-OOD precisely captures the orientation and alignment of symmetric objects, achieving notable improvements in detection accuracy and localization precision. This method proves particularly effective for highly symmetric objects like cars and buses, which are prone to misclassification in traditional detection frameworks.

### *4.4 Ablation Experiment*

To evaluate the impact of the TB and TN modules in the ASL-OOD model, we performed ablation experiments on the DOTA, HRSC2016, and MAR20 datasets. The results, measured in terms of mAP, are summarized in Table 3.

**Table 3:** Ablation study on ASL-OOD modules across datasets

| Models | DOTA | HRSC2016 | MAR20 |
|---|---|---|---|
| ASL-OOD-TB | 78.18 (−1.98) | 89.66 (−1.41) | 83.32 (−2.13) |
| ASL-OOD-TN | 78.58 (−1.58) | 89.58 (−1.49) | 83.73 (−1.72) |
| ASL-OOD-ASDH | 72.37 (−7.79) | 85.82 (−5.25) | 81.27 (−4.18) |
| ASL-OOD | **80.16** | **91.07** | **85.45** |

**ASL-OOD-TB (Transformer-Based Backbone):** ASL-OOD-TB denotes a variant of ASL-OOD where the TB module is replaced with the standard YOLOv5 Backbone, allowing us to assess the TB module's contribution to detecting multi-scale targets in complex backgrounds. As reported in Table 3, removing the TB module causes an absolute mAP decrease of 1.98 on DOTA, 1.41 on HRSC2016, and 2.13 on MAR20. These reductions highlight the critical role of the Transformer-based Backbone in enhancing ASL-OOD's performance under challenging conditions, including complex scenes and varying target scales.

**ASL-OOD-TN (Transformer-Based Neck):** The ASL-OOD-TN variant replaces the Transformer-based Neck (TN) module with the standard PANet architecture to assess the impact of multi-scale feature fusion strategies on detection performance. As presented in Table 3, this modification leads to a reduction in mAP by 1.58 on DOTA, 1.49 on HRSC2016, and 1.72 on MAR20, relative to the full ASL-OOD model. These results highlight the critical role of the Transformer-based Neck module in improving detection accuracy for multi-scale and densely distributed targets. By enhancing feature fusion across scales, the TN module enables the model to better capture fine-grained details in complex scenes, thereby bolstering overall detection performance.

**ASL-OOD-ASDH (Angle-SIOU-Based Decoupled Head):** The removal of ASDH leads to significant mAP drops of 7.79 on DOTA, 5.25 on HRSC2016, and 4.18 on MAR20, highlighting its critical role in optimizing angle consistency and bounding box alignment. The impact is most pronounced on DOTA, a dataset with large aspect ratio variations and complex orientations, where precise angle predictions are crucial for detecting diverse, densely packed targets. On HRSC2016, ASDH proves essential for elongated maritime targets, ensuring accurate orientation alignment. The smaller drop on MAR20 reflects its relatively simpler aspect ratio and orientation characteristics.

These ablation experiments underscore the pivotal roles of the TB, TN, and ASDH modules in boosting ASL-OOD's detection capabilities. The TB module enhances overall performance by delivering a robust, context-sensitive backbone, while the TN module facilitates accurate localization and the detailed extraction of multi-scale target features through sophisticated multi-scale fusion strategies and Transformer-based methodologies. Furthermore, the ASDH module plays a critical role in refining angle predictions and bounding box alignment, significantly improving detection performance for oriented and elongated objects across datasets with diverse aspect ratios and orientations.

### *4.5 Model Complexity Analysis*

To augment the performance evaluation, we present an in-depth analysis of the computational complexity of the proposed ASL-OOD model in comparison with state-of-the-art object detection models such as YOLOv4 and YOLOv5. This analysis considers essential metrics, including the number of parameters, floating-point operations (FLOPs), and training/inference speeds, providing a holistic assessment of the models' complexity, efficiency, and suitability for real-world applications.

As detailed in Table 4, ASL-OOD comprises 21.89 MB of parameters, slightly more than YOLOv5 (21.25 MB) while remaining significantly lighter than YOLOv4 (27.6 MB). Despite this marginal increase in parameter count, ASL-OOD delivers a substantial boost in detection accuracy, achieving a mAP of 80.16 compared to 68.18 for YOLOv5 and 67.23 for YOLOv4. In terms of computational cost, ASL-OOD requires 51.38G FLOPs, which is marginally higher than YOLOv5 but still lower than YOLOv4, further underscoring its computational efficiency. The model also demonstrates competitive training and inference speeds, with a training speed of 327 ms per image and an inference speed of 34.21 ms, showing only a slight increase compared to YOLOv5. These results underscore ASL-OOD's ability to achieve an excellent trade-off between complexity and performance. The minimal increase in parameters and FLOPs relative to YOLOv5 leads to a significant improvement in detection accuracy, solidifying ASL-OOD as a highly efficient and effective model for object detection tasks.

**Table 4:** Model complexity comparison in terms of FLOPs, number of parameters, and speed (DOTA)

| Model | Parameters (MB) | FLOPs (G) | Train speed (ms) | Test speed (ms) | mAP |
|---|---|---|---|---|---|
| YOLOv4 | 27.6 | 52.21 | 323 | 33.88 | 67.23 |
| YOLOv5 | 21.25 | 49.71 | 316 | 32.33 | 68.18 |
| **ASL-OOD** | 21.89 | 51.38 | 327 | 34.21 | 80.16 |

## 5 Conclusion

In this paper, we introduce ASL-OOD, a one-stage oriented object detection framework that overcomes the limitations of conventional YOLO-based approaches, particularly their reliance on horizontal bounding boxes. By incorporating Transformer-based modules in both the Backbone and Neck, ASL-OOD enhances multi-scale feature extraction and achieves effective information fusion, particularly for small and low-resolution targets. Additionally, we propose an innovative angle offset representation and an angle-based IOU loss, which significantly enhance detection accuracy for oriented objects across diverse scales and orientations. Extensive experiments on four military target datasets demonstrate the superior performance of ASL-OOD, achieving state-of-the-art mAP and confirming its robustness in oriented object detection tasks. In future work, we plan to expand the framework's capabilities by addressing challenges associated with boundary blurring in dynamic military targets. This may include exploring advanced data augmentation techniques and employing image restoration networks to tackle these complexities effectively.

**Acknowledgement:** The authors wish to express their appreciation to the reviewers for their helpful suggestions which greatly improved the presentation of this paper.

**Funding Statement:** This research was supported by the Key Research and Development Program of Shaanxi Province (2024GX-YBXM-010).

**Author Contributions:** The authors confirm contribution to the paper as follows: Study conception and design: Kexin Wang, Yuqing Lin, Jiancheng Liu; Data collection: Yuqing Lin, Wanlong Qi, Xingye Han; Analysis and interpretation of results: Kexin Wang, Runyuan Wen, Zhipeng Zhang; Draft manuscript preparation: Kexin Wang, Tuo Wang. All authors reviewed the results and approved the final version of the manuscript.

**Availability of Data and Materials:** The data that support the findings of this study are openly available in DOTA at https://captain-whu.github.io/DOTA (accessed on 25 February 2021), in HRSC2016 at http://www.escience.cn/people/liuzikun/DataSet.html (accessed on 02 December 2024), in UAVDT at https://sites.google.com/site/daviddo0323/ (accessed on 02 December 2024), and in MAR20 at https://gcheng-nwpu.github.io/ (accessed on 02 September 2024).

**Ethics Approval:** Not applicable.

**Conflicts of Interest:** The authors declare no conflicts of interest to report regarding the present study.

## References

[1] X. Yang, J. Yan, W. Liao, X. Yang, J. Tang and T. He, "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, 2022. doi: 10.1109/TPAMI.2022.3166956.

[2] Y. Jiang *et al.*, "R2CNN: Rotational region CNN for orientation robust scene text detection," 2017, *arXiv:1706.09579*.

[3] R. Girshick, "Fast R-CNN," 2015, *arXiv:1504.08083*.

[4] J. Redmon, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016.

[5] A. Wang *et al.*, "YOLOv10: Real-time end-to-end object detection," 2024, *arXiv:2405.14458*.

[6] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov and S. Zagoruyko, "End-to-end object detection with transformers," in *Eur. Conf. Comput. Vis.*, Glasgow, UK, Springer International Publishing, 2020, pp. 213–229.

[7] A. Kirillov *et al.*, "Segment anything," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 4015–4026.

[8] J. Ma *et al.*, "Arbitrary-oriented scene text detection via rotation proposals," *IEEE Trans. Multimed.*, vol. 20, no. 11, pp. 3111–3122, 2018. doi: 10.1109/TMM.2018.2818020.

[9] J. Ding, N. Xue, Y. Long, G. -S. Xia, and Q. Lu, "Learning RoI transformer for detecting oriented objects in aerial images," 2018, *arXiv:1812.00155*.

[10] H. Fan and H. Ling, "SANet: Structure-aware network for visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2017, pp. 42–49.

[11] Z. Yang, S. Liu, H. Hu, L. Wang, and S. Lin, "RepPoints: Point set representation for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9657–9666.

[12] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," *Proc. IEEE*, vol. 111, no. 3, pp. 257–276, 2023. doi: 10.1109/JPROC.2023.3238524.

[13] X. Yang *et al.*, "SCRDet: Towards more robust detection for small, cluttered and rotated objects," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 8232–8241.

[14] J. Ding, N. Xue, Y. Long, G. -S. Xia, and Q. Lu, "Learning RoI transformer for oriented object detection in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2849–2858.

[15] Y. Xu *et al.*, "Gliding vertex on the horizontal bounding box for multi-oriented object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 4, pp. 1452–1459, 2020. doi: 10.1109/TPAMI.2020.2974745.

[16] X. Yang, J. Yan, Z. Feng, and T. He, "R3Det: Refined single-stage detector with feature refinement for rotating object," *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 4, pp. 3163–3171, 2021.

[17] J. Yi, P. Wu, B. Liu, Q. Huang, H. Qu, and D. Metaxas, "Oriented object detection in aerial images with box boundary-aware vectors," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2021, pp. 2150–2159.

[18] X. Cheng and C. Zhang, "C2-YOLO: Rotating object detection network for remote sensing images with complex backgrounds," in *2022 Int. Joint Conf. Neural Netw. (IJCNN)*, IEEE, 2022, pp. 1–8.

[19] J. Han, J. Ding, J. Li, and G. -S. Xia, "Align deep features for oriented object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–11, 2021.

[20] Q. Ming, Z. Zhou, L. Miao, H. Zhang, and L. Li, "Dynamic anchor learning for arbitrary-oriented object detection," *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 3, pp. 2355–2363, 2021.

[21] X. Pan *et al.*, "Dynamic refinement network for oriented and densely packed object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11207–11216.

[22] Z. Liu *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10012–10022.

[23] T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2117–2125.

[24] P. Y. Chen, M. C. Chang, J. W. Hsieh, and Y. S. Chen, "Parallel residual bi-fusion feature pyramid network for accurate single-shot object detection," *IEEE Trans. Image Process.*, vol. 30, pp. 9099–9111, 2021. doi: 10.1109/TIP.2021.3118953.

[25] W. Qian, X. Yang, S. Peng, J. Yan, and Y. Guo, "Learning modulated loss for rotated object detection," *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 3, pp. 2458–2466, 2021.

[26] Z. Ge, "YOLOX: Exceeding YOLO series in 2021," 2021, *arXiv:2107.08430*.

[27] G. S. Xia *et al.*, "DOTA: A large-scale dataset for object detection in aerial images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3974–3983.

[28] Z. Liu, H. Wang, L. Weng, and Y. Yang, "Ship rotated bounding box space for ship extraction from high-resolution optical satellite images with complex backgrounds," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 8, pp. 1074–1078, 2016. doi: 10.1109/LGRS.2016.2565705.

[29] Y. U. Wenqi *et al.*, "MAR20: A benchmark for military aircraft recognition in remote sensing images," *Natl. Remote Sens. Bull.*, vol. 27, no. 12, pp. 2688–2696, 2024.

[30] D. Du *et al.*, "The unmanned aerial vehicle benchmark: Object detection and tracking," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 370–386.

[31] Z. Chen *et al.*, "PIoU loss: Towards accurate oriented object detection in complex environments," in *Comput. Vis.–ECCV 2020: 16th Eur. Conf.*, Glasgow, UK: Springer International Publishing, 2020, pp. 195–211.

[32] G. Jocher *et al.*, "ultralytics/yolov5: v5.0-YOLOv5-P6 1280 models, AWS, Supervise.ly and YouTube integrations," *Zenodo*, 2021. doi: 10.5281/zenodo.4679653.

[33] Y. Hou *et al.*, "R-YOLO: A YOLO-based method for arbitrary-oriented target detection in high-resolution remote sensing images," *Sensors*, vol. 22, no. 15, 2022, Art. no. 5716. doi: 10.3390/s22155716.

[34] A. Cheng, J. Xiao, Y. Li, Y. Sun, Y. Ren and J. Liu, "Enhancing remote sensing object detection with K-CBST YOLO: Integrating CBAM and swin-transformer," *Remote Sens.*, vol. 16, no. 16, 2024, Art. no. 2885. doi: 10.3390/rs16162885.

[35] J. Wang, J. Ding, H. Guo, W. Cheng, T. Pan and W. Yang, "Mask OBB: A semantic attention-based mask oriented bounding box representation for multi-category object detection in aerial images," *Remote Sens.*, vol. 11, no. 24, 2019, Art. no. 2930. doi: 10.3390/rs11242930.

[36] W. Qian, C. Zhou, and D. Zhang, "FAOD-Net: A fast AOD-Net for dehazing single image," *Math. Probl. Eng.*, vol. 2020, no. 1, 2020, Art. no. 4945214. doi: 10.1155/2020/4945214.

[37] K. Fu, Z. Chang, Y. Zhang, and X. Sun, "Point-based estimator for arbitrary-oriented object detection in aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4370–4387, 2020. doi: 10.1109/TGRS.2020.3020165.

[38] Z. Tian, C. Shen, H. Chen, and T. He, "Fully convolutional one-stage object detection," 2019, *arXiv:1904.01355*.

[39] J. Redmon, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.

[40] T. Lin, "Focal loss for dense object detection," 2017, *arXiv:1708.02002*.

[41] A. Bochkovskiy, C. Y. Wang, and H. Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.

[42] L. Cao, P. Song, Y. Wang, Y. Yang, and B. Peng, "An improved lightweight real-time detection algorithm based on the edge computing platform for UAV images," *Electronics*, vol. 12, no. 10, pp. 2274, 2023. doi: 10.3390/electronics12102274.

[43] S. Zhang, C. Chi, Y. Yao, Z. Lei, and S. Z. Li, "Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9759–9768.

[44] G. Yu *et al.*, "PP-PicoDet: A better real-time object detector on mobile devices," 2021, *arXiv:2111.00902*.

[45] C. Y. Wang, A. Bochkovskiy, and H. Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 7464–7475.

[46] G. Wang, Y. Chen, P. An, H. Hong, J. Hu and T. Huang, "UAV-YOLOv8: A small-object-detection model based on improved YOLOv8 for UAV aerial photography scenarios," *Sensors*, vol. 23, no. 16, 2023, Art. no. 7190. doi: 10.3390/s23167190.

[47] X. Xie, G. Cheng, J. Wang, X. Yao, and J. Han, "Oriented R-CNN for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 3520–3529.

[48] X. Min, W. Zhou, R. Hu, Y. Wu, Y. Pang and J. Yi, "LWUAVDet: A lightweight UAV object detection network on edge devices," *IEEE Internet Things J.*, vol. 11, no. 13, pp. 24013–24023, 2024. doi: 10.1109/JIOT.2024.3388045.

[49] Q. Chen, T. Zheng, L. Liu, L. Yu, and Z. Chen, "HLA: Harmonized label assigner for two-stage oriented object detection," in *2022 Int. Conf. Autom., Robot. Comput. Eng. (ICARCE)*, IEEE, 2022, pp. 1–5.