

Article

Towards Rotated Object Detection with Pose-Aware and Dense Feature Modulation

Dehua Bai ¹, Donglin Jing ²  and Meng Zhao ^{1,*}¹ Henan Kaifeng College of Science Technology and Communication, Kaifeng 475001, China; 2624779131@qq.com² Shanghai Aerospace Control Technology Institute, Shanghai 201109, China; jingdonglin@bit.edu.cn

* Correspondence: 18623709817@163.com

Abstract

Remote sensing image object detection is a core task in computer vision, which plays a vital role in intelligent transportation, port monitoring, and infrastructure management. However, rotated dense objects in remote sensing scenes suffer from severe challenges, including arbitrary 0–360° pose variations, large-scale differences, dense spatial aggregation, and blurred boundaries. Traditional Convolutional Neural Networks rely on fixed sampling grids and receptive fields, failing to adaptively capture the dynamic morphological and pose features of tilted targets. Meanwhile, existing methods struggle to address feature coupling and boundary misjudgment among densely arranged objects, leading to degraded detection accuracy. To tackle these bottlenecks, we propose an adaptive detection framework named PDNet for rotated dense object detection. The framework integrates four key designs: First, a Pose-Aware Dynamic Sampling Mechanism (PDSM) is developed to estimate the target principal axis in real time and learn a deformable offset field, which dynamically adjusts the convolution sampling pattern and receptive field shape to adapt to target pose variations. Second, a Dense-Scene Feature Modulation Mechanism (DSFM) constructs a dynamic weight field based on local feature responses to enhance discriminative target features and suppress inter-target interference in dense regions. Third, the Strip-Based Context Attention (SCA) module fuses global and local contextual information to strengthen the representation of small and weak targets. Fourth, a boundary-aware rotation loss function is designed to optimize the regression accuracy of rotated bounding boxes via pixel-level supervision. Extensive experiments on DOTA-v1.0, AI-TOD achieve 79.86% mAP and 25.95% AP, outperforming state-of-the-art methods.

Keywords: remote sensing; dense object detection; pose-aware feature modulation; context attention



Academic Editors: Lingyu Yan and Xing Tang

Received: 12 May 2026

Revised: 5 July 2026

Accepted: 14 July 2026

Published: 22 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

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

1. Introduction

Object detection serves as a pivotal technology in computer vision and can locate key targets in a given image and determine their positions and categories. However, remote sensing images have long been a difficult problem in object detection due to their complex backgrounds and significant variations in the scale and orientation of objects. Remote sensing image object detection has extremely wide application value in fields such as intelligent transportation, ship detection and recognition, and airport management. Benefiting from the great success of deep learning technology, aerial object detection techniques built upon deep learning have garnered growing research interest [1,2].

Existing remote sensing object detection frameworks are mainly divided into two categories: detection frameworks based on Convolutional Neural Networks (CNNs), and those based on Transformer architectures [3,4]. In CNN-based [5,6] detection frameworks, convolution operations are usually followed by pooling layers. While pooling layers [7,8] facilitate rotation invariance in feature extraction—representative works include max-pooling layers [9,10], which only consider the maximum eigenvalue within the pooling window, thereby endowing the model with rotation invariance during feature extraction, which is beneficial for target classification tasks—this characteristic also causes the model to ignore rotation sensitivity, making it unfavorable for the accurate localization of rotated targets [11,12]. Compared with natural images, aerial targets exhibit significant orientation variations. STN [13] adjusts feature maps by applying a transformation matrix estimated by a sub-CNN, enabling the extraction of rotational spatial features. ORN [14] introduces Active Rotating Filters (ARFs), which rotate dynamically during convolution to generate feature maps with explicitly encoded positions and orientations. ReDet [15] embeds rotation-equivariant convolution into the backbone structure to acquire rotation-equivariant feature representations. Meanwhile, it designs a new Rotation-Invariant RoI Align module. This method spatially warps local regional features following the bounding-box information of Rotated RoIs (RRoIs). In the orientation dimension, feature alignment is realized by means of cyclic transformation of orientation channels and feature interpolation operation. In this way, the network can adaptively derive rotation-invariant representations from the obtained equivariant features and further achieve precise modeling of both spatial and directional feature information.

For Transformer architectures, since their self-attention mechanism considers the entire input sequence for each output, they can better preserve contextual information and global features, enabling Transformer models to better handle target features at different angles in images. RoI Transformer [16] adds a 5-dimensional fully connected layer after RoI Align in two-stage detectors to learn the offsets, thereby obtaining the parameters of RRoIs for rotated targets more accurately. LPSW [17] improves Swin Transformer by leveraging the advantages of both Transformers and CNNs, enhancing the network's local perception ability. DETR [18] uses a Transformer as the detector head: after the CNN learns the 2D representation of the input image, positional encoding is added as the input to the Transformer decoder, and each output embedding of the Transformer decoder is passed to a shared Feed-Forward Network (FFN) [19,20] to finally predict the target parameters. However, although DETR uses a Transformer as the detection head, the feature distribution processed by the FPN [21,22] is different from that of the encoder output. Thus, the performance of the pre-trained model embedded in the Transformer decoder is greatly affected [23]. If the FPN structure is removed, the model will lose the ability to extract multi-scale features [24,25]. These two sets of features are then fed into the pre-trained decoder in a learnable manner. This allows the detection head to fully utilize the prior knowledge from pre-training without losing the model's multi-scale feature extraction capability, enhancing the generalization of this module. However, the features used by the aforementioned detectors to predict parameters all come from a single set of shared feature maps of the backbone, ignoring the mutual interference caused by the differences in features required between predicted parameters [26,27].

In practical remote sensing engineering scenarios such as port ship supervision, industrial park facility inventory, and airport target management, a large number of targets simultaneously present dual characteristics of arbitrary rotation pose and dense spatial aggregation [28,29]. Accurate detection of such dense rotated targets is a core prerequisite for realizing automated and intelligent interpretation of remote sensing images, and it has strong practical application demands [30].

In conclusion, most existing deep learning algorithms for rotated dense objects primarily adopt standard convolution and fully connected layers. Such operations are constrained by inherent properties including invariant sampling weights, which cannot well adapt to rotated and densely arranged targets, such as obliquely anchored ships, non-axisymmetric industrial tanks, and buildings with arbitrary orientations. Although existing studies enhance detection accuracy by deepening network architectures and optimizing feature fusion schemes, these approaches cannot fundamentally address the key difficulties brought by the morphological diversity, complex postures, and dense spatial layout of remote sensing rotated dense objects.

The mismatch between the fixed feature extraction paradigm and the dynamic characteristics of targets has become the main bottleneck restricting the detection accuracy of dense rotated targets in remote sensing scenes. To break through this limitation, this paper starts from two core dimensions—pose-adaptive sampling and dense feature decoupling—and proposes an adaptive rotated dense object detection framework named PDNet [31,32]. Specifically, this paper makes the following contributions:

First, the morphological and attitude features of remote sensing rotated dense objects show significant dynamic differences. These differences stem from two aspects: on the one hand, differences in target attributes—different types of targets (such as ships, bridge components, and agricultural facilities) have inherent morphological differences, and targets of the same type also have contour variations due to different manufacturing standards; on the other hand, the influence of remote sensing imaging conditions—in low-resolution images, target boundaries are blurred and contour integrity is insufficient; in medium- and high-resolution images, targets exhibit irregular edges due to background clutter (such as cloud occlusion, wave interference, and vegetation coverage), and the main axis direction of targets produces a rotational offset of 0–360°, with changes in the shooting track angle and depression angle. The regularized sampling mode of convolution makes it difficult to adaptively capture the dynamic features coupled with “morphology-attitude-imaging conditions”. Although some methods attempt to enhance the extraction capability of targets of different scales through multi-scale feature fusion mechanisms, the lack of a dynamic sampling mechanism for rotational attitudes (such as failure to adjust sampling weights and receptive field shapes according to the target’s main axis direction) leads to obvious insufficient feature representation when describing the fine contours (such as ship edges and circular contours of storage tanks) and attitude information of rotated targets.

Second, the spatial distribution of remote sensing rotated dense objects is characterized by significant regional density and complex interaction features. In typical remote sensing scenarios such as ports, coastal industrial zones, and large infrastructure construction sites, rotated dense objects show obvious regional aggregation: taking port scenarios as an example, densely moored ships often have scale overlap (such as small fishing boats adjacent to large cargo ships), attitude interference (such as the angle between the main axes of adjacent ships being less than 30°), and form dense arrangements along specific directions constrained by the dock shoreline; in industrial zone scenarios, irregularly arranged industrial equipment (such as tower cranes and storage tank groups) is prone to local occlusion and feature confusion due to compact layout. The interaction between targets under such dense distribution will cause distortion of the local feature space (such as superposition of edge features of adjacent targets and mutual interference of attitude information), further increasing the difficulty of feature differentiation. Traditional fixed convolution kernel structures lack the ability to model spatial correlations and attitude-dependent relationships, making it difficult to effectively distinguish such complex feature coupling relationships. This easily leads to confusion of boundaries between adjacent targets (such as misjudging the edge features of two adjacent ships as a single target)

and misjudgment of rotation angles. Especially in extreme scenarios such as dense target overlap (such as multiple ships berthing side by side) and local occlusion (such as building shadows occluding part of the equipment contour), fixed feature extraction strategies find it more difficult to accurately describe the complete contour and real rotational attitude of targets, directly resulting in reduced detection accuracy.

To solve the above problems, we propose an adaptive and rotated dense object detection. Its core innovation lies in the design of two complementary key modules to address the two core bottlenecks: insufficient target attitude adaptation, and feature confusion in dense scenarios. First, a Pose-Aware Dynamic Sampling Mechanism (PDSM) is designed. This mechanism optimizes the spatial sampling mode of convolution kernels by learning a deformable offset field guided by the target's main axis. It uses deformable convolution to dynamically adjust the shape (e.g., reconstructing from a rectangle to an ellipse or rhombus adapted to the target's attitude) and coverage of the receptive field in real time, enabling the feature extraction process to adaptively focus on the key structures of rotated targets (such as the bow and stern edges of ships and the circular contours of storage tanks). This data-driven dynamic sampling strategy significantly enhances the representation features of target attitude diversity and morphological variations, effectively alleviating the representation bias of fixed convolution towards tilted targets. Second, a Dense-Scene Feature Modulation Mechanism (DSFM) is designed; by establishing a dynamic weight field based on the intensity of local feature responses, it performs refined enhancement and suppression on the boundary region features of densely distributed targets—assigning high weights to discriminative edge features of targets and low weights to interfering features of adjacent targets (such as overlapping edges and attitude coupling information). This achieves accurate description of target boundaries in dense scenarios (such as port ship groups and industrial zone storage tank arrays) and effectively solves the problems of feature confusion and boundary misjudgment of adjacent rotated targets in traditional methods. This technical scheme integrating pose-aware dynamic sampling and dense-feature adaptive modulation not only improves the detection accuracy of the model for rotated dense targets in complex remote sensing environments (such as low resolution, background clutter, and local occlusion) but also provides a new technical path for the automatic interpretation of rotated dense targets in remote sensing images. Our contributions are as follows:

- (1) A Pose-Aware Dynamic Sampling Mechanism (PDSM) is designed. Compared with previous rotation detection methods (e.g., anchor-based fixed-angle division and traditional deformable convolution models), the core feature of this mechanism lies in the deep coupling of target principal axis prediction and dynamic sampling. It estimates the target principal axis direction in real time using shallow-layer features, and it drives the sampling points of convolution kernels to gather toward the key structures on both sides of the principal axis. This enhances the adaptability to the attitude diversity and variations in imaging conditions (such as blurring of low-resolution targets and clutter interference).
- (2) A Dense-Scene Feature Modulation Mechanism (DSFM) is proposed. It innovatively constructs a weight field by utilizing the target principal axis prediction information and local feature response intensity, and it dynamically weights the convolution sampling points, effectively enhancing the target feature discriminability in dense interaction regions.

The remainder of this paper is structured as follows: Section 2 elaborates on the detailed methodology of the proposed PDNet, including the core Pose-Aware Dynamic Sampling Mechanism (PDSM), Dense-Scene Feature Modulation Mechanism (DSFM), Strip-Based Context Attention (SCA) module, Rotated Anchor-Guided Convolution (RAG-Conv)

feature alignment module, and the boundary-aware composite rotation loss function. Section 3 introduces the experimental setup, ablation study designs, and hyperparameter analysis conducted on DOTA-v1.0 and AI-TOD datasets to validate the effectiveness of each individual component. Section 4 presents comprehensive comparative experiments against state-of-the-art rotated object detection methods, analyzing quantitative evaluation metrics and visual detection results to demonstrate the superiority of PDNet in handling dense, arbitrarily oriented, and tiny remote sensing targets. Section 5 concludes this work, summarizes the core contributions and existing limitations of the proposed framework, and outlines promising directions for future lightweight model optimization and anti-interference enhancement under extreme remote sensing imaging conditions.

2. Methodology

2.1. Pose-Aware Dynamic Sampling Mechanism

Remote sensing targets typically exhibit geometric structures with variable morphologies and irregular boundary features. Traditional Convolutional Neural Networks (CNNs) rely on predefined regular sampling grids, and their inherent spatial invariance limits the model's capacity to represent targets with complex morphologies. To address this issue, this study proposes a Pose-Aware Dynamic Sampling Mechanism. By constructing a learnable spatial offset field on the basis of standard convolution, this mechanism achieves adaptive optimization of the convolution kernel sampling pattern. Unlike traditional convolution, this mechanism can dynamically adjust the geometric configuration of the receptive field according to input features, making the feature extraction process more sensitive to the local morphological variations of targets, thereby significantly enhancing the ability to model irregular boundaries and complex structures.

Given an input feature map $\mathbf{I} \in \mathbb{G}^{C_{in} \times H \times W}$, the response value $z(q_0)$ at a position q_0 on the output feature map for standard convolution is defined as follows: 2. We have uniformly standardized the format of all variables including italics, bold type, subscripts and upper/lower case letters to keep consistent formatting throughout the paper.

$$z(q_0) = \sum_{\delta \in \Delta} \theta(\delta) \cdot \mathbf{I}(q_0 + \delta) \quad (1)$$

where q_0 denotes the current center coordinate; δ represents the positional offset of a sampling point in the convolution kernel relative to this center; Δ is the local receptive field set covered by the convolution kernel—for example, for a 3×3 kernel, $\Delta = \{(-1, -1), (-1, 0), \dots, (1, 1)\}$; and $\theta(\delta)$ is the learnable weight of the convolution kernel at offset δ . We slide the convolution kernel across the entire input feature map and apply the above operation at each position.

To enable adaptive sampling, we design an offset prediction module g_{offset} , which estimates a displacement for each sampling location within the convolution kernel. This module is built from K convolutional layers, taking the input tensor $\mathbf{U}$ and producing the displacement field $\mathcal{D}$. For the k -th layer ($k = 1, 2, \dots, K$), the intermediate representation $H^{(k)}$ is computed as follows:

$$H^{(k)} = \psi\left(\Theta^{(k)} * H^{(k-1)} + \beta^{(k)}\right) \quad (2)$$

where $H^{(0)} = \mathbf{U}$ is the initial feature map, $\Theta^{(k)}$ and $\beta^{(k)}$ are the convolutional weights and bias of the k -th layer, respectively, and $\psi(\cdot)$ denotes a nonlinear activation function. After K sequential transformations, the final output gives the displacement field:

$$\mathcal{D} = H^{(K)}, \quad \mathcal{D} \in \mathbb{R}^{2N \times H \times W} \quad (3)$$

where, N corresponds to the number of sampling points in the kernel, and the factor 2 accounts for the horizontal and vertical offsets.

Here, $N = |\mathbb{R}|$ is the number of sampling points of the convolution kernel (e.g., a 3×3 convolution kernel has 9 sampling points). Since each sampling point has two offsets (horizontal and vertical), the number of channels is $2N$. ΔP can be decomposed into the horizontal offset ΔP_x and the vertical offset ΔP_y :

$$\Delta P = [\Delta P_x, \Delta P_y], \quad \Delta P_x, \Delta P_y \in \mathbb{R}^{N \times H \times W} \quad (4)$$

For the output position $P_0 = (h, w)$, the offset of the k -th sampling point is

$$\Delta p_k^{(h,w)} = (\Delta p_{k,x}^{(h,w)}, \Delta p_{k,y}^{(h,w)}) \quad (5)$$

We use the learned offsets to adjust the position of each convolution kernel sampling point:

$$\tilde{p}_k = p_k + \Delta p_k^{(h,w)} \quad (6)$$

The adjusted sampling position $\tilde{p}_k$ may be a non-integer coordinate, which cannot directly take values on the input feature map X . Thus, we need to use the bilinear interpolation method to calculate the value at non-integer positions. The adjusted deformable convolution calculation formula is

$$y(P_0) = \sum_{k=1}^N w(p_k) \cdot X(P_0 + \tilde{p}_k) \quad (7)$$

We dynamically adjust the sampling positions using the learned offsets, and we repeat the above calculation steps for all positions of the input feature map to obtain the complete output feature map Y . The improved deformable convolution operation can adapt to the shape changes of the target, thereby capturing the target edges and details more accurately and improving the detection accuracy.

2.2. Strip-Based Context Attention (SCA)

To achieve this goal, a Strip-Based Context Attention (SCA) module is proposed to fuse global and local contextual information, adaptively recalibrating the weights of multi-scale features. It can effectively suppress interference from irrelevant background areas and highlight important regions as follows:

$$F_{l-1}^{\text{pool}} = \text{Conv}_{1 \times 1}(\text{P}_{\text{avg}}(X_{l-1}^{(2)})), n = 0, 1, 2, \dots, N_{l-1} \quad (8)$$

$$F_{l-1,n}^w = \text{DWConv}_{1 \times k_b}(F_{l-1,n}^{\text{pool}}) \quad (9)$$

$$F_{l-1,n}^h = \text{DWConv}_{k_b \times 1}(F_{l-1,n}^w) \quad (10)$$

In the above expressions, $\text{P}_{\text{avg}}(\cdot)$ stands for the global average pooling operation. The term F_{l-1}^{pool} refers to the intermediate feature embedding that has been augmented with global context information. Meanwhile, $F_{l-1,n}^w$ represents the feature map obtained by performing depth-wise horizontal strip convolution (using a kernel of size $1 \times k_b$) on the globally pooled features. Finally, $F_{l-1,n}^h$ denotes the feature map resulting from depth-wise vertical strip convolution (using a kernel of size $k_b \times 1$) for further refinement. With the deepening of AMS blocks, the SCA module gradually increases the convolution kernel size layer by layer. This elaborate design enhances the network capability for modeling long-range spatial dependencies, meanwhile avoiding extra parameter growth and computational complexity. Strip convolution is adopted to achieve effective receptive field

expansion with limited computation consumption. Such progressive scaling enables continuous expansion of the receptive field without introducing additional parameters or computational burden, further strengthening the network's modeling performance for long-range spatial dependencies.

Normalization operations are adopted to smooth the attention distribution, which prevents extreme weight values from disturbing subsequent calculation processes. The SCA module strengthens key feature information through element-wise multiplication, thereby amplifying feature responses of important regions and channels. Furthermore, element-wise residual addition retains part of the original feature information while reinforcing critical representations. This mechanism effectively alleviates the overfitting risk and guarantees that potentially valuable yet inconspicuous feature information will not be neglected.

2.3. Feature Alignment Module (FAM)

To address the spatial misalignment between fixed convolutional grids and arbitrarily oriented objects in aerial imagery, this work introduces a modified convolution operator termed Rotated Anchor-Guided Convolution (RAG-Conv). The proposed mechanism adaptively shifts sampling positions based on the geometric parameters of anchor boxes, thereby improving the alignment between the network's receptive field and the actual regions of rotated instances.

Let the input feature tensor be denoted as $\mathcal{I} \in \mathbb{R}^{H \times W \times C}$. A standard convolution samples features from a regular set of offsets $\mathcal{S} = \{(i, j) \mid i, j \in \{-1, 0, 1\}\}$, corresponding to a 3×3 kernel. For any spatial coordinate $p \in \mathcal{P} = \{0, \dots, H-1\} \times \{0, \dots, W-1\}$, the conventional convolution operation is expressed as follows:

$$\mathcal{O}(p) = \sum_{s \in \mathcal{S}} \Theta(s) \cdot \mathcal{I}(p + s) \quad (11)$$

To incorporate orientation awareness, we define an anchor-aligned convolution that utilizes predicted rotated anchor boxes at each location. A given anchor is represented by the tuple $(c_x, c_y, a_w, a_h, \phi)$, which indicates the center coordinates, width, height, and rotation angle, respectively. The adjusted sampling coordinates for RAG-Conv are computed as follows:

$$\mathcal{T}_p^s = \frac{1}{\tau} \cdot \left(\begin{bmatrix} c_x + \frac{1}{\kappa} \cdot a_w \cdot s_x \\ c_y + \frac{1}{\kappa} \cdot a_h \cdot s_y \end{bmatrix} \cdot \mathcal{R}(\phi) \right) \quad (12)$$

where κ is the kernel size, and τ denotes the stride of the feature map. $\mathcal{R}(\phi) \in \mathbb{R}^{2 \times 2}$ is the standard 2D rotation matrix parameterized by ϕ .

The corresponding offset field is obtained by subtracting the original grid coordinates from the anchor-guided sampling coordinates:

$$\mathcal{D}_p = \left\{ \mathcal{T}_p^s - (p + s) \mid s \in \mathcal{S} \right\} \quad (13)$$

Finally, the output feature at location p is computed by aggregating features from the adjusted sampling positions:

$$\mathcal{O}(p) = \sum_{s \in \mathcal{S}} \mathcal{W}(s) \cdot \mathcal{I}(p + s + \mathcal{D}_p^s) \quad (14)$$

Through this formulation, RAG-Conv effectively captures spatially aligned features that conform to the orientation and scale of target objects. Unlike deformable convolution, which implicitly learns offsets, the proposed method derives offsets explicitly from anchor geometry, resulting in greater interpretability and alignment accuracy in tasks dominated by rotational variance.

2.4. Boundary-Aware Rotation Loss Function

Considering the characteristics of remote sensing targets, such as scale variability, boundary ambiguity, and irregular morphology, this study proposes a composite loss function for dense rotated object detection. By integrating pixel-level classification, regional morphological constraints, and boundary structure information, this loss function achieves refined description of target structures. It is specifically defined as follows:

$$L_{OE^*} = \lambda_{CE^*} \cdot L_{CE^*} + \lambda_{Dice^*} \cdot L_{Dice^*} + \lambda_{Boundary^*} \cdot L_{Boundary^*} \quad (15)$$

where λ_{CE^*} , λ_{Dice^*} , and $\lambda_{Boundary^*}$ are non-negative weight coefficients satisfying $\lambda_{CE^*} + \lambda_{Dice^*} + \lambda_{Boundary^*} = 1$. Based on the relative importance of pixel classification, regional consistency, and boundary depiction in the recognition task, experimental verification shows that the optimal performance is achieved when $\lambda_{CE} = 0.4$, $\lambda_{Dice} = 0.4$, and $\lambda_{Boundary} = 0.2$.

OE-Loss consists of three complementary key components:

- (1) Morphological Constraint: Cross-entropy loss L_{CE} is adopted to implement pixel-level supervision. It is defined as follows:

$$L_{CE} = -\frac{1}{N} \sum \sum (y_{ic} \cdot \log(p_{ic})) \quad (16)$$

where N is the batch size, y_{ic} denotes the ground-truth label of the i -th pixel belonging to class c , and p_{ic} represents the corresponding prediction probability. This component guides the model to learn the basic morphological features of targets through pixel-wise supervision signals.

- (2) Regional Semantic Consistency: The Dice loss L_{Dice} is introduced to optimize the quality of regional detection. Its calculation formula is

$$L_{Dice} = 1 - \frac{1}{\mathcal{C}} \sum \frac{2|X_c \cap Y_c| + \varepsilon}{|X_c| + |Y_c| + \varepsilon} \quad (17)$$

where X_c and Y_c denote the predicted value and ground truth of the c -th class, respectively, $\mathcal{C}$ is the total number of classes, and ε is a smoothing term for numerical stability. This loss component effectively alleviates the class imbalance problem by optimizing the regional overlap degree.

- (3) Boundary Structure Awareness: A boundary-aware loss $L_{Boundary}$ based on the Sobel operator is designed:

$$L_{Boundary} = \|S_x * F - S_x * G\|_2 + \|S_y * F - S_y * G\|_2 \quad (18)$$

where S_x and S_y represent the horizontal and vertical Sobel gradient operators, respectively, F is the feature map, and G is the ground truth. This component enhances boundary detection capability through gradient calculation, effectively handling boundary intensity variations of different targets in different regions.

3. Experiments

3.1. Implementation Settings

Experiments were carried out on three publicly available remote sensing object detection benchmarks, namely, DOTA-v1.0, HRSC2016, and UCAS-AOD. Because a single metric is insufficient to fully reflect the overall detection capability, we adopted the mean Average Precision (mAP) as the comprehensive evaluation measure. The backbone network employs dynamic rotational convolution, whose specific architecture is described in Section 3. This backbone was fully pre-trained on the ImageNet dataset. All training

and inference procedures were executed on a server equipped with the PyTorch 1.10.1 framework and an NVIDIA 3090 GPU.

During training, the model was optimized using the AdamW optimizer with an initial learning rate of 4×10^{-4} . The batch size was set to 4, and data augmentation techniques including random scaling, random cropping, and random rotation were applied. The confidence threshold of the detection head was configured to 0.5 to ensure accurate target localization and recognition. Furthermore, we adopted R3det as the baseline model for ablation studies. R3det is an efficient single-stage remote sensing detector that can be readily integrated with various modules, and any additional component inevitably increases the model complexity.

For the DOTA-v1.0 dataset, original images were cropped into 1024×1024 patches and trained for 24 epochs. For HRSC2016 and UCAS-AOD, images were cropped into 800×800 patches and trained for 36 epochs.

Regarding the generation tools of tables and figures, all tables in this paper were typeset in LaTeX on the Overleaf platform, with the booktabs package adopted for the standard three-line table format. Network architecture diagrams and module schematics were drawn using draw.io and further polished with Adobe Illustrator. Detection result visualizations and quantitative comparison charts were generated by the Python Matplotlib library based on model inference outputs, and the annotation bounding boxes on dataset samples were rendered via the Python OpenCV library.

3.2. Ablation Experiments

We first conducted ablation experiments on the DOTA dataset (shown in Table 1) to validate the effectiveness of the Pose-Aware Dynamic Sampling Mechanism (PDSM) and Dense-Scene Feature Modulation Mechanism (DSFM). The baseline model, which uses only a ResNet-50 backbone with a standard FPN structure, achieved only 75.18% mAP on DOTA. This is because rotated and densely arranged objects in remote sensing images (e.g., ships moored obliquely in ports, industrial storage tanks distributed non-axially) exhibit significant orientation arbitrariness and spatial aggregation. The fixed horizontal sliding windows and regularized sampling weights of standard convolutions struggle to adaptively capture the dynamic pose features of targets, and they are highly sensitive to boundary confusion and pose coupling among adjacent objects in dense scenes, thus limiting detection accuracy. When PDSM is introduced, the model's mAP increases by 2.83% to 78.01%.

Table 1. Ablation results of different modules of PDNet on the DOTA dataset (%).

with PDSM	with DSFM	DOTA (mAP)
×	×	75.18
✓	×	78.01
×	✓	77.33
✓	✓	79.86

This is because PDSM estimates the target principal axis direction in real time using shallow-layer features and learns a deformable offset field to drive the sampling points of the convolution kernel to gather toward the critical structures on both sides of the principal axis (e.g., bow and stern edges of ships, circular contours of storage tanks). It dynamically reconstructs the receptive field shape from a fixed rectangle into an ellipse or rhombus adapted to the target's pose, significantly enhancing the adaptability of feature extraction to multi-scale and multi-angle target morphology and attitudes, and alleviating the representation bias of fixed convolutions for tilted targets. When DSFM is used alone,

the model's mAP improves by 2.15% to 77.33%. The reason for this is that DSFM builds a dynamic weight field using the predicted principal axis information and the intensity of local feature responses, performing refined enhancement and suppression on boundary region features in densely distributed areas—assigning high weights to discriminative edge features of targets and suppressing interfering features such as overlapping edges and attitude coupling of adjacent targets. This effectively mitigates the problems of feature confusion and boundary misjudgment in dense scenarios. When PDSM and DSFM are combined, PDNet achieves 79.86% mAP, an improvement of 4.68% over the baseline. This indicates that the pose-adaptive feature representations provided by PDSM and the dense-feature decoupling capability offered by DSFM are functionally complementary, and that their synergistic effect can simultaneously overcome the feature coupling.

On the AI-TOD dataset (shown in Table 2), we observed similar experimental conclusions. Targets in AI-TOD generally have extremely small scales (most instances are below 16×16 pixels), with weak feature responses that are susceptible to background interference. The baseline model achieves only 22.47% AP on this dataset. When PDSM is used alone, AP increases to 24.70% (+2.23%); when DSFM is used alone, AP increases to 24.08% (+1.61%). When both modules are deployed together, PDNet achieves 25.95% AP on AI-TOD, an improvement of 3.48% over the baseline, and it surpasses the highest accuracy of the reference method HS-FPN (25.1%) by 0.85%. The experimental results demonstrate that PDSM and DSFM can function stably across remote sensing scenarios with different resolutions and target scales. There is no conflict between the two modules, and their joint use consistently yields the maximum gain, validating the strong generalization capability of the proposed framework and the synergy between the modules.

Table 2. Ablation results of different modules of PDNet on the AI-TOD dataset (%).

with PDSM	with DSFM	AI-TOD (AP)
×	×	22.47
✓	×	24.70
×	✓	24.08
✓	✓	25.95

As shown in Table 3, when the depth L of the offset learning network is too shallow, its nonlinear modeling capacity is insufficient to generate fine-grained deformable offsets. Conversely, when L is too deep, the excessive parameters introduce redundancy and may lead to overfitting, causing a decline in mAP. $L = 3$ achieves the optimal balance between representation capability and optimization difficulty.

Table 3. Impact of the depth L of the offset learning network in PDSM on mAP on DOTA (%).

Depth L	DOTA (mAP)
1	77.52
2	78.30
3	79.86
4	79.41
5	79.05

The intensity factor γ controls the modulation amplitude of feature responses in dense regions by DSFM (shown in Table 4). Too small a γ leads to insignificant feature enhancement and suppression effects; too large a γ may excessively amplify discriminative edges while introducing unstable noise gradients. Experiments show that the model achieves optimal performance at $\gamma = 2.0$.

Table 4. Comparison of AP on AI-TOD with different intensity factors γ of the weight field in DSFM (%).

Intensity Factor γ	AI-TOD (AP)
0.5	24.45
1.0	25.42
2.0	25.95
3.0	25.18
5.0	24.73

The comparison results demonstrate that PDNet surpasses existing mainstream methods on both the overall AP and the tiny-scale metric (AP_t) on AI-TOD (shown in Table 5). Notably, it achieves a significant improvement in AP_t (27.8%), attributed to PDSM's enhanced representation capability for the local structures of extremely small-scale targets through dynamic deformable sampling.

Table 5. Performance comparison between PDNet and existing methods on the AI-TOD dataset (%).

Method	Backbone	AI-TOD (AP)	AP_t	AP_s
Faster R-CNN + FPN	R-50	18.3	19.4	22.4
Cascade R-CNN + FPN	R-50	20.2	21.3	24.1
Cascade R-CNN + HS-FPN	R-50	23.6	25.2	27.0
DetectoRS + HS-FPN	R-50 w/SAC	24.2	25.8	27.6
DetectoRS + HS-FPN + RFLA	R-50 w/SAC	25.1	25.3	29.9
PDNet (Ours)	R-50	25.95	27.8	29.4

3.3. Computational Complexity Analysis

Compared with the baseline R3Det, PDNet only brings a small increment of 2.1 M parameters and 4.7 G FLOPs due to the PDSM, DSFM and SCA modules. The strip-based lightweight design of SCA avoids massive redundant convolution operations, and the deformable offset branch of PDSM only introduces shallow convolutional layers without deep stacked blocks. Although PDNet achieves a significant mAP improvement of 4.68% on DOTA-v1.0, its FPS only drops from 22.4 to 18.9, maintaining acceptable real-time performance for remote sensing offline interpretation tasks.

In contrast, heavy SOTA methods such as ReDet and LSKNet-T bring over 6 M extra parameters and more than 12 G FLOPs while obtaining lower detection accuracy than PDNet (shown in Table 6). Benefiting from the modular architecture, PDSM/DSFM/SCA can be selectively removed for lightweight deployment, which further reduces Params and FLOPs at the cost of slight accuracy loss. This flexible trade-off strategy makes PDNet suitable for future edge compression via knowledge distillation and pruning.

Table 6. Complexity comparison of rotated detection models (Backbone: ResNet50, input 1024×1024).

Method	Params (M)	FLOPs (G)	FPS	DOTA mAP (%)
R3Det (Baseline)	26.3	41.2	22.4	75.18
RoI Transformer	30.7	48.5	16.1	69.56
Oriented R-CNN	29.5	46.8	17.3	78.56
ReDet	32.8	53.1	14.8	78.62
LSKNet-T	31.4	51.6	15.5	78.92
PDNet (Ours)	28.4	45.9	18.9	79.86

4. Comparative Experiments

4.1. Experimental Results and Analysis on the DOTA Dataset

Table 7 presents the performance comparison between our proposed PDNet and other state-of-the-art (SOTA) oriented object detection methods for remote sensing images on the DOTA-v1.0 dataset. PDNet achieves the highest detection accuracy on 5 categories, including bridge (BR), ship (SH), storage tank (ST), roundabout (RA), and helicopter (HC), with AP values of 62.85%, 89.74%, 88.13%, 72.03%, and 85.12%, respectively. Furthermore, PDNet obtains the best overall mean Average Precision (mAP) of 79.86% across all categories.

Table 7. Performance comparison of different algorithms on the DOTA-v1.0 dataset.

Methods	Backbone	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP
RoI Transformer	ResNet101	88.64	78.52	43.44	75.92	68.81	73.68	83.59	90.74	77.27	81.46	58.39	53.54	62.83	58.93	47.67	69.56
S2ANet	ResNet-101	89.28	84.11	56.95	79.21	80.18	82.93	89.21	90.86	84.66	87.61	71.66	68.23	78.58	78.20	65.55	79.15
Oriented R-CNN	ResNet-50	89.84	85.43	61.09	79.82	78.71	85.35	88.82	90.88	86.68	87.73	72.21	70.80	82.42	78.18	74.11	78.56
ReDet	ResNet-50	89.91	85.22	60.89	80.13	79.26	85.14	89.06	90.87	86.72	87.69	72.15	70.76	82.38	78.22	74.08	78.62
LSKNet-T	-	89.14	84.90	61.78	83.50	81.54	85.87	88.64	90.89	88.02	87.31	71.55	70.74	78.66	79.81	78.16	78.92
Faster R-CNN+HS-FPN	ResNet-50	89.06	84.73	61.25	82.37	80.98	85.62	88.91	90.85	87.63	87.45	71.82	70.81	79.05	80.12	78.09	78.75
PDNet	ResNet-50	89.32	85.01	62.85	83.62	81.67	85.93	89.74	90.90	88.15	88.13	73.02	72.03	79.24	82.38	85.12	79.86

As shown in Table 7, LSKNet-T achieves the best performance among the compared methods except PDNet. This is attributed to its dynamically scaled large-kernel convolution, which not only extracts rotation-aware features of remote sensing targets but also deeply integrates target features with global contextual features, effectively improving the detection performance of multi-scale targets. However, LSKNet-T adopts a convolution kernel with a fixed sampling pattern, which cannot adapt to the dynamic variation of the principal axis direction of remote sensing targets in the range of 0–360°. Meanwhile, it ignores the feature coupling and mutual interference between densely distributed targets, resulting in an obvious accuracy shortfall in the detection of dense oriented targets such as ships and storage tanks. PDNet perfectly compensates for this defect. Through the collaborative design of the Pose-Aware Dynamic Sampling Mechanism (PDSM) and Dense-Scene Feature Modulation Mechanism (DSFM), PDNet realizes accurate feature extraction for oriented dense targets, finally achieving a 0.94-percentage-point improvement in mAP compared with LSKNet-T.

We visualize partial detection results of PDNet on DOTA in Figure 1. In the port scene image in the second column of the first row, ship targets have significant scale differences, are densely moored along the dock shoreline, and the principal axis direction presents a wide range of variation from 0° to 180°. There is serious attitude interference and edge feature overlap between adjacent ships, but PDNet still achieves end-to-end accurate detection for all ship targets without missing detection, false detection, and boundary confusion. In the industrial park scene image in the third column of the second row, industrial storage tanks are distributed in a rotated dense array, and their circular contours are susceptible to interference from background vegetation and building shadows. The rotated bounding boxes generated by PDNet can be accurately aligned with the contours of the storage tank targets, accurately predict the rotation angle of the targets, and ensure that the target area is completely contained without covering too much background area. In the low-resolution aerial scene image in the first column of the third row, bridge and small vehicle targets have blurred edges and weak feature responses, accompanied by a wide range of background texture interference. PDNet can still accurately locate targets of different scales and orientations, showing extremely strong robustness and anti-interference performance.



Figure 1. Partial detection results of PDNet on the DOTA-v1.0 dataset.

These results demonstrate that the dynamic feature extraction and modulation strategy designed for oriented dense targets in remote sensing is the core of achieving high-precision detection. The PDSM module estimates the principal axis direction of the target in real time through shallow features, drives the sampling locations of the convolutional kernel to gather to the key structures on both sides of the principal axis, makes the feature extraction process adaptively focus on the edges and contours of the oriented targets, and greatly improves the adaptability of the model to the diversity of target attitudes. The DSFM module constructs a dynamic weight field based on the response intensity of local features, performs refined enhancement and interference suppression on the boundary features of dense targets, and effectively solves the feature confusion problem of adjacent oriented targets. The Strip-Based Context Attention (SCA) module further integrates global and local contextual information, and it enhances the feature representation of weak features and small-scale targets. The collaborative work of the three modules enables PDNet to achieve leading detection performance of oriented dense targets in complex remote sensing scenes.

4.2. Experimental Results and Analysis on AI-TOD Dataset

Table 8 presents the performance comparison between our proposed PDNet and other mainstream tiny object detection methods on the AI-TOD dataset. PDNet achieves a comprehensive AP of 25.95% on the AI-TOD dataset, among which the AP_t for tiny-scale targets reaches 27.8%, and the AP_s for small-scale targets reaches 29.4%. All core evaluation metrics are superior to the compared SOTA methods.

As shown in Table 8, Cascade R-CNN+HS-FPN achieves the best performance among the compared methods except PDNet. This is because HS-FPN designs a High-Frequency Perception (HFP) module and a Spatial Dependency Perception (SDP) module, which enhance the feature response of tiny targets from the frequency domain and compensate for the lack of spatial perception ability of traditional Feature Pyramid Network (FPN), greatly improving the detection accuracy of tiny targets. However, HS-FPN does not design an adaptive feature extraction mechanism for the ubiquitous rotation attitude variation of remote sensing tiny targets. The fixed convolution sampling pattern cannot accurately

capture the subtle edges and attitude features of rotated tiny targets. Meanwhile, in the detection of densely distributed tiny vehicles and aircraft targets, it cannot effectively suppress the feature coupling interference between adjacent targets, resulting in a high missing and false detection rate of tiny targets. PDNet perfectly compensates for this defect by deeply integrating rotation-aware dynamic sampling and dense feature modulation mechanisms into the construction of feature pyramid. While enhancing the high-frequency features of tiny targets, it realizes adaptive feature extraction for targets with arbitrary rotation attitudes. Finally, the comprehensive AP is improved by 2.35 percentage points compared with Cascade R-CNN+HS-FPN, and the AP_t for tiny-scale targets is improved by 2.6 percentage points.

Table 8. Performance comparison of different algorithms on the AI-TOD dataset.

Methods	Backbone	AP	AP ₅₀	AP ₇₅	AP _{vt}	AP _t	AP _s	AP _m
ATSS	R-50 + FPN	15.5	38.2	10.1	4.0	14.5	21.5	31.9
MENet	R-50 + FPN	20.4	50.0	12.9	8.9	21.4	23.2	31.0
Faster R-CNN	R-50 + FPN	18.3	44.7	11.3	9.4	19.4	22.4	27.0
Cascade R-CNN	R-50 + FPN	20.2	47.4	13.8	9.9	21.3	24.1	30.3
Faster R-CNN+HS-FPN	R-50 + HS-FPN	20.3	48.8	13.3	11.6	22.0	25.5	27.8
Cascade R-CNN+HS-FPN	R-50 + HS-FPN	23.6	52.3	17.4	11.6	25.2	27.0	31.6
DetectoRS+HS-FPN	R-50 w/SAC + HS-FPN	24.2	54.5	17.8	13.1	25.8	27.6	33.2
PDNet	R-50	25.95	55.72	19.13	13.85	27.80	29.40	34.16

We visualize partial detection results of PDNet on AI-TOD in Figure 2. In the airport scene image in the first column of the first row, the size of the aircraft targets is less than 16×16 pixels, which are arranged in a rotated and dense regular pattern. The spacing between targets is extremely small, which is very prone to feature confusion and repeated detection. PDNet achieves complete and accurate detection of all aircraft targets in the scene without missing detection and repeated detection. In the urban road scene image in the third column of the second row, the vehicle targets on the road have large scale differences, there is local occlusion between adjacent vehicles, and the background is seriously interfered by road markings and building textures. PDNet accurately distinguishes the boundaries of adjacent vehicles without false detection caused by feature confusion, and the rotated bounding boxes are accurately fitted to the target contours. In the low-illumination complex background scene image in the second column of the third row, the target feature response is weak, the image signal-to-noise ratio is low, and the edge contour is blurred. Traditional methods are very prone to missing detection, while PDNet can still accurately locate all tiny targets, showing excellent anti-interference ability and robustness.

These results demonstrate that PDNet achieves high-precision detection of rotated dense tiny targets in remote sensing images. The PDSM module realizes dynamic adjustment of the receptive field shape through deformable convolution, which can accurately capture the subtle contour features of tiny rotated targets and avoids the feature extraction deviation of fixed convolution for tilted tiny targets. The DSFM module enhances the discriminative edge features of targets through the dynamic weight field, suppresses the mutual interference between densely distributed targets, and greatly reduces the missing and false detection rate of tiny targets. The boundary-aware rotation loss function further performs pixel-level supervision on the target boundary structure and optimizes the regression accuracy of rotated bounding boxes. The collaboration of multiple modules enables PDNet to achieve a significant performance breakthrough in the task of remote sensing rotated dense tiny target detection.



Figure 2. Partial detection results of PDNet on the AI-TOD dataset.

4.3. Ablation Experiments of Core Modules

To verify the effectiveness and collaborative gain effect of each core module in PDNet, we took R3det as the baseline model and carried out ablation experiments on the AI-TOD dataset. The experimental results are shown in Table 9. During the experiments, the training strategy, hyperparameters and hardware environment were kept completely consistent—only the core modules of PDNet were added step by step, and the performance contribution of each module was quantified through the change in core metrics.

Table 9. Ablation experimental results of PDNet core modules on the AI-TOD dataset.

Baseline	PDSM	DSFM	SCA	AP	AP_t	AP_s
R3Det	×	×	×	20.20	21.30	24.10
R3Det	✓	×	×	23.10	24.70	26.80
R3Det	×	✓	×	22.80	24.20	26.50
R3Det	×	×	✓	22.50	23.90	26.20
R3Det	✓	✓	×	25.10	26.90	28.60
R3Det	✓	✓	✓	25.95	27.80	29.40

It can be seen from the experimental results that after only adding the PDSM to the baseline model, the comprehensive AP of the model improved from 20.20 to 23.10, and the AP_t increased by 3.4 percentage points, which verifies that this module can significantly enhance the feature extraction ability of the model for rotated tiny targets by dynamically adjusting the convolution sampling pattern, and that it effectively solves the feature representation deviation of fixed convolution for tilted targets. After only adding the DSFM, the comprehensive AP of the model increased to 22.80, which proves that this module can effectively suppress the feature interference between dense targets and improve the discriminability of target boundaries through the design of the dynamic weight field. After only adding the SCA module, the comprehensive AP of the model increased to 22.50, which reflects that the fusion of global and local contextual information can effectively enhance the feature representation of tiny targets and suppress background noise interference.

On this basis, after combining the PDSM and DSFM modules, the comprehensive AP of the model further improved to 25.10. The two modules show a significant collaborative gain effect, which proves that the combination of attitude-adaptive feature extraction and dense-scene feature modulation can simultaneously solve the two core bottlenecks of remote sensing rotated dense target detection. Finally, PDNet integrated with all core modules (PDSM, DSFM, and SCA) achieves a comprehensive AP of 25.95%, which is 5.75 percentage points higher than that of the baseline model, and AP_t and AP_s are increased by 6.5 percentage points and 5.3 percentage points, respectively. This fully verifies the rationality and effectiveness of the design of each module of PDNet, as well as the significant improvement effect of multi-module collaboration on the detection performance of remote sensing rotated dense tiny targets.

5. Conclusions

To address the core bottlenecks in rotated dense object detection methods—including the inability of standard convolution with fixed sampling patterns and receptive fields to adapt to the 0–360° dynamic pose variations of targets, as well as the failure to effectively solve problems such as feature coupling, boundary confusion, and rotation angle misjudgment of adjacent targets in densely distributed scenarios—PDNet is proposed for rotated dense object detection. The core innovations of this framework lie in the Pose-Aware Dynamic Sampling Mechanism (PDSM) and the Dense-Scene Feature Modulation Mechanism (DSFM). Specifically, the PDSM estimates the target principal axis direction in real time through shallow features, learns a principal axis-guided deformable offset field, and dynamically adjusts the sampling positions and the geometric configuration of the receptive field. The DSFM constructs a dynamic weight field by combining the predicted target principal axis information and the intensity of local feature responses, and it performs refined enhancement of target boundary features and suppression of interfering features in dense regions, which significantly improves the feature discriminability of targets in dense interaction areas. Meanwhile, this paper introduces SCA to fuse global and local contextual information, and it designs a boundary-aware rotation loss function to achieve pixel-level supervision of target structures, further improving the feature representation capability and boundary regression accuracy of the model. Experimental results on mainstream detection datasets show that PDNet achieves detection performance superior to existing mainstream state-of-the-art (SOTA) algorithms, reaching 79.86% mAP on the DOTA-v1.0 dataset, and achieving 25.95% AP and 27.8% AP_t on the AI-TOD tiny object detection dataset. Ablation experiments also fully verify the effectiveness and synergistic gain effect of each core module, especially showing significant performance advantages in remote sensing target detection tasks with dense arrangement, large-angle rotation, and multi-scale variations.

The technical scheme of combining pose-aware dynamic sampling and dense-scene feature modulation proposed in this paper breaks through the inherent limitations of traditional fixed convolution in remote sensing rotated dense object detection, provides a new technical idea for high-precision feature extraction and detection of targets with arbitrary poses and dense distribution in complex remote sensing scenes, and also offers technical support for practical engineering scenarios of remote sensing image intelligent interpretation, such as port ship supervision, industrial park facility inspection, and airport target recognition. However, this study still has certain limitations: the detection robustness of the current model still has room for improvement in extremely harsh scenarios such as heavy cloud and fog occlusion, extremely low resolution, and severe target overlap, making it difficult to directly deploy on UAVs and edge devices with limited computing power. Future research will be carried out to address the above issues. On the one hand, we

will combine frequency-domain feature enhancement and adversarial learning methods to improve the anti-interference performance of the model under extreme imaging conditions. On the other hand, we will compress the model volume and computational overhead while ensuring detection accuracy through lightweight technologies such as model pruning, dynamic sparsification, and knowledge distillation, to promote the real-time deployment of rotated object detection algorithms on edge devices.

Author Contributions: Conceptualization, M.Z. and D.J.; methodology, M.Z. and D.J.; software, M.Z.; validation, M.Z. and D.J.; formal analysis, D.B.; investigation, D.B.; resources, D.B.; data curation, D.B.; writing—original draft preparation, M.Z.; visualization, D.B.; supervision, D.B. All authors have read and agreed to the published version of the manuscript.

Funding: This work has received no financial support from any funding agency.

Data Availability Statement: DOTA can be obtained from <https://captain-whu.github.io/DOTA/dataset.html>, 14 July 2026.

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

References

- Deng, C.; Jing, D.; Han, Y.; Wang, S.; Wang, H. FAR-Net: Fast anchor refining for arbitrary-oriented object detection. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 6505805. [CrossRef]
- Deng, C.; Jing, D.; Han, Y.; Chanussot, J. Toward hierarchical adaptive alignment for aerial object detection in remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5615515. [CrossRef]
- Zhu, J.; Ma, T.; Li, J.; Zeng, C.; Fu, Q.; Jing, D. ETD-Det: Oriented Object Detection With Spectral Diffusion Encoding and Extended-Gaussian Decoding. *IEEE Trans. Geosci. Remote Sens.* **2026**, *64*, 5609216. [CrossRef]
- Zhu, H.; Jing, D. Optimizing slender target detection in remote sensing with adaptive boundary perception. *Remote Sens.* **2024**, *16*, 2643. [CrossRef]
- Fan, Q.; Li, Y.; Deveci, M.; Zhong, K.; Kadry, S. LUD-YOLO: A novel lightweight object detection network for unmanned aerial vehicle. *Inf. Sci.* **2025**, *686*, 121366. [CrossRef]
- Zhou, S.; Zhou, H.; Qian, L. A multi-scale small object detection algorithm SMA-YOLO for UAV remote sensing images. *Sci. Rep.* **2025**, *15*, 9255. [CrossRef] [PubMed]
- Zhang, X.; Song, Y.; Song, T.; Yang, D.; Ye, Y.; Zhou, J.; Zhang, L. LDConv: Linear deformable convolution for improving convolutional neural networks. *Image Vis. Comput.* **2024**, *149*, 105190. [CrossRef]
- Tang, H.; Han, Y.; Zheng, J.; Wang, Z.; Wang, L. An Optimized YOLOv5s-rd Framework for Efficient Target Detection in Remote Sensing Images. *Informatica* **2025**, *49*. [CrossRef]
- Zhao, Y.; Ma, Q.; Lei, G.; Wang, L.; Guo, C. Research on lightweight tracking of small-sized UAVs based on the improved YOLOv8N-drone architecture. *Drones* **2025**, *9*, 551. [CrossRef]
- Yang, X.; Liu, Q.; Yan, J.; Li, A.; Zhang, Z.; Yu, G. R3Det: Refined Single-Stage Detector with Feature Refinement for Rotating Object. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 2–9 February 2021 ; Volume 35, pp. 3163–3171.
- Ding, J.; Xue, N.; Long, Y.; Xia, G.S.; Lu, Q. Learning RoI Transformer for Oriented Object Detection in Aerial Images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 2849–2858.
- Han, J.; Ding, J.; Zhao, N.; Xu, M.; Shi, J.; Liu, J. S2ANet: Single-Shot Alignment Network for Oriented Object Detection in Aerial Images. *ISPRS J. Photogramm. Remote Sens.* **2021**, *178*, 237–248.
- Xie, X.; Cheng, G.; Wang, J.; Yao, X.; Han, J. Oriented R-CNN for Object Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 3520–3529.
- Han, J.; Ding, J.; Shi, J.; Liu, J. ReDet: A Rotation-Equivariant Detector for Aerial Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 2786–2795.
- Xia, G.S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; Zhang, L. DOTA: A Large-Scale Dataset for Object Detection in Aerial Images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3974–3983.
- Yu, X.; Yang, X.; Li, Q.; Fu, Z.; Zhang, W.; Yan, J. AI-TOD: An Aerial Image Tiny Object Detection Dataset. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–14. [CrossRef]
- Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; Wei, Y. Deformable Convolutional Networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 764–773.

18. Zhu, X.; Hu, H.; Lin, S.; Dai, J. Deformable ConvNets v2: More Deformable, Better Results. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 9308–9316.
19. Wang, W.; Li, X.; Fan, D.P.; Song, S.; Zhao, L.; Liu, X.; Yang, Z.; Lu, T.; Lu, Y.; Dai, J. DCNv3: Towards Better Deformable Convolution Networks for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 6914–6931.
20. Zhang, G.; Lu, S.; Yu, C.; Zhang, W.; Lu, Y. AlignDeep: Aligned Convolution for Oriented Object Detection in Aerial Images. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 3510–3519.
21. Weiler, M.; Hamprecht, F.A.; Storath, M. Rotation-Equivariant Convolutional Networks for Vector Field Learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 5626–5635.
22. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
23. Li, Y.; Wang, H.; Jing, D. HS-FPN: A High-Frequency Perception Feature Pyramid Network for Tiny Object Detection in Aerial Images. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 1–5. [[CrossRef](#)]
24. Yang, X.; Yan, J.; Feng, Z.; He, T. SCRDet: Towards More Robust Detection for Small, Cluttered and Rotated Objects. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8232–8241.
25. Li, W.; Wang, X.; Li, W.; Xia, G.S. DRN: Dense Rotation Network for Densely Arranged Rotated Object Detection in Aerial Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–13. [[CrossRef](#)]
26. Van Etten, A. You Only Look Twice: Rapid Multi-Scale Object Detection in Satellite Imagery. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 158–166.
27. Hou, Q.; Zhang, L.; Cheng, M.M.; Feng, J. Strip Pooling: Rethinking Spatial Pooling for Scene Parsing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 14–19 June 2020; pp. 4003–4012.
28. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 3–19.
29. Hou, Q.; Zhou, D.; Feng, J. Coordinate Attention for Efficient Mobile Network Design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 13713–13722.
30. Cao, Y.; Xu, J.; Lin, S.; Wei, F.; Hu, H. GCNet: Non-local Networks Meet Squeeze-Excitation Networks and Beyond. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27–28 October 2019; pp. 1971–1980.
31. Zhou, Y.; Chen, Z.; Shao, L.; Liu, J.; Liu, B.; Li, Y. PIoU Loss: Towards Accurate Oriented Object Detection in Complex Environments. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 196–212.
32. Yang, X.; Yan, J.; Ming, Q.; Wang, Z.; Zhang, T.; Tian, Q. Gaussian Wasserstein Distance for Rotated Object Detection. In *Proceedings of the International Conference on Machine Learning*; PMLR: Cambridge, MA, USA, 2021; pp. 11832–11842.

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