remote sensing

MDPI

Article

PAMFPN: Position-Aware Multi-Kernel Feature Pyramid Network with Adaptive Sparse Attention for Robust Object Detection in Remote Sensing Imagery

Xiaofei Yang [1], Suihua Xue [1], Lin Li [1], Sihuan Li [1], Yudong Fang [2,*], Xiaofeng Zhang [3] and Xiaohui Huang [4]

[1] School of Electronics and Communication Engineering, Guangzhou University, Guangzhou 510006, China; xiaofeiyang@gzhu.edu.cn (X.Y.); 1919500073@e.gzhu.edu.cn (S.X.); 2112330037@e.gzhu.edu.cn (L.L.); 2112330082@e.gzhu.edu.cn (S.L.)

[2] Big Data Centre, Ministry of Emergency Management, Beijing 10110, China

[3] Shenzhen Key Laboratory of Internet Information Collaboration, Harbin Institute of Technology (Shenzhen), Shenzhen 518055, China; zhangxiaofeng@hit.edu.cn

[4] School of Information Engineering, East China Jiaotong University, Nanchang 330000, China; 2854@ecjtu.edu.cn

* Correspondence: ydfang2025@gmail.com

Abstract

Deep learning methods have achieved remarkable success in remote sensing object detection. Existing object detection methods focus on integrating convolutional neural networks (CNNs) and Transformer networks to explore local and global representations to improve performance. However, existing methods relying on fixed convolutional kernels and dense global attention mechanisms suffer from computational redundancy and insufficient discriminative feature extraction, particularly for small and rotation-sensitive targets. To address these limitations, we propose a Dynamic Multi-Kernel Position-Aware Feature Pyramid Network (PAMFPN), which integrates adaptive sparse position modeling and multi-kernel dynamic fusion to achieve robust feature representation. Firstly, we design a position-interactive context module (PICM) that incorporates distance-aware sparse attention and dynamic positional encoding. It selectively focuses computation on sparse targets through a decay function that suppresses background noise while enhancing spatial correlations of critical regions. Secondly, we design a dual-kernel adaptive fusion (DKAF) architecture by combining region-sensitive attention (RSA) and reconfigurable context aggregation (RCA). RSA employs orthogonal large-kernel convolutions to capture anisotropic spatial features for arbitrarily oriented targets, while RCA dynamically adjusts the kernel scales based on content complexity, effectively addressing scale variations and intraclass diversity. Extensive experiments on three benchmark datasets (DOTA-v1.0, SSDD, HWPUVHR-10) demonstrate the effectiveness and versatility of the proposed PAMFPN. This work bridges the gap between efficient computation and robust feature fusion in remote sensing detection, offering a universal solution for real-world applications.

Keywords: deep learning; object detection; Transformer; convolution neural network; feature fusion

check for updates

Academic Editor: Zhenwei Shi

Received: 8 May 2025

Revised: 16 June 2025

Accepted: 26 June 2025

Published: 27 June 2025

Citation: Yang, X.; Xue, S.; Li, L.; Li, S.; Fang, Y.; Zhang, X.; Huang, X. PAMFPN: Position-Aware Multi-Kernel Feature Pyramid Network with Adaptive Sparse Attention for Robust Object Detection in Remote Sensing Imagery. *Remote Sens.* **2025**, *17*, 2213. https://doi.org/10.3390/rs17132213

1. Introduction

Remote sensing object detection plays a pivotal role in geospatial intelligence applications, including urban planning, environmental monitoring, and disaster response [1–3]. Despite significant advancements in deep learning, detecting objects in remote sensing images remains challenging due to three inherent characteristics: (1) sparse target distributions (e.g., isolated vehicles in vast landscapes), (2) extreme multi-scale variations (e.g., ships spanning from 10 to 10,000 pixels), and (3) cluttered backgrounds (e.g., buildings and vegetation mimicking target textures) [4–7]. Traditional convolutional neural networks (CNNs) and hybrid CNN–Transformer architectures struggle to address these challenges effectively.

The current remote sensing object detection methods primarily focus on two paradigms: hierarchical feature pyramid networks (FPNs) for multi-scale feature fusion [8,9] and attention mechanisms for contextual modeling [10–13]. While these approaches have achieved notable progress, they exhibit critical limitations when applied to remote sensing scenarios characterized by sparse targets and complex backgrounds.

Hierarchical feature pyramid networks (FPNs) aim to address scale variations by aggregating features from different backbone layers. For example, PANet [8] introduces bidirectional (top-down and bottom-up) paths to enhance the semantic consistency across scales. Similarly, BiFPN [9] employs weighted fusion to prioritize semantically rich features, reducing feature misalignment. However, these methods rely on fixed-size convolutional kernels (e.g., 3×3 in PANet) for cross-scale interactions, struggling to adapt to extreme scale variations in remote sensing imagery. This rigidity stems from their inability to dynamically adjust the kernel scales based on the target size or spatial context.

Attention mechanisms, particularly self-attention in Transformers [14,15] and channel–spatial hybrids like CBAM [16], have been widely adopted to model global dependencies. For example, SwinTransformer [17] achieved a 76.1% mAP50 on DOTA-v1.0 by computing window-based attention, effectively capturing long-range correlations between sparse vehicles. Meanwhile, lightweight variants such as ECA-Net [18] reduce the computational costs through local channel interactions. Despite these advances, two fundamental issues persist. First, dense global attention mechanisms (e.g., in Vision Transformers [14,17]) compute redundant interactions across all spatial positions in DOTA-v1.0 involving background regions, diluting critical sparse features. Second, static attention designs (e.g., fixed window sizes in Swin) fail to prioritize dynamically evolving target distributions.

Although such hybrid architectures perform well in natural image detection, they still exhibit limitations in remote sensing scenarios:

1. Insufficient Sparse Position Modeling: Conventional self-attention mechanisms compute global interactions across all spatial positions, ignoring the inherent sparsity of remote sensing targets. This results in computational redundancy and noise amplification.
2. Coarse-Grained Context Modeling: Existing methods often rely on fixed-scale convolutional kernels, failing to adaptively capture multi-scale local details and long-range contextual relationships.
3. Inefficient Feature Fusion: Cross-scale fusion strategies in feature pyramid networks (FPNs) [9] lack the effective integration of spatial position guidance and multi-kernel feature extraction, limiting their robustness in complex scenes.

To address these limitations and improve the robustness of the model (robustness refers to a model's ability to maintain stable detection performance under challenging remote sensing conditions, including significant scale variations, background clutter, the presence of small objects, and scalability involving different data models), we propose PAMFPN, a novel framework integrating adaptive sparse attention and dynamic multi-kernel fusion to tackle the unique challenges of remote sensing object detection. The framework comprises three core modules and structures: (1) the position-interactive context module (PICM) addresses inefficient sparse modeling with distance-decay sparse attention, dynamically prioritizing critical regions (e.g., sparse vehicles in DOTA-v1.0) via a learnable exponential decay function; (2) the dual-kernel adaptive fusion (DKAF) architecture handles rigid multi-scale perception via region-sensitive attention (RSA) and reconfigurable context aggregation (RCA); (3) lightweight cross-layer interaction optimizes the efficiency via bottleneck restructuring in the C3PAM module. Our key contributions are as follows:

1. We present the position-interactive context module (PICM), a novel plug-and-play module that integrates dynamic position encoding (DEncode) with self-attention to dynamically model sparse spatial correlations in remote sensing imagery;
2. We propose a multi-kernel dynamic feature extraction module, which is a dual-branch architecture combining region-sensitive attention (RSA) to extract local details and reconfigurable context aggregation (RCA) modules for capturing the global dependencies;
3. Rigorous ablation studies and parameter optimizations validate the efficacy of each component; notably, DEncode outperforms traditional positional encoding by 0.6% regarding the mAP50, while the optimal RSA kernel size (k = 5) balances local detail preservation and global context modeling.

The rest of the article is organized as follows. Section 2 presents a historical overview of object detection, remote sensing detection techniques, and attention mechanisms. Section 3 introduces a feature pyramid model for enhanced remote sensing object detection and describes the proposed model architecture and principles in detail. Section 4 describes the dataset and experimental setup. Section 5 presents the experimental results and discussion. Section 6 describes how our approach addresses the challenges encountered in remote sensing object detection. Section 7 highlights the main conclusions of this paper and proposes directions for future research.

2. Related Works

2.1. Object Detection

As one of the fundamental technologies in computer vision, modern object detection algorithms have gradually evolved into paradigms dominated by convolutional neural networks (CNNs) and Transformer architectures, significantly outperforming traditional methods relying on handcrafted features. Existing approaches can be broadly categorized into two-stage and single-stage detectors. Two-stage frameworks, represented by Faster R-CNN [19], first generate region proposals and then perform refined classification and regression. While achieving high detection accuracy, their computational efficiency is constrained by the proposal generation and multi-stage inference processes. In contrast, single-stage detectors such as SSD [20], the YOLO series [21–25], and RT-DETR [26] adopt an end-to-end dense prediction strategy, directly regressing object locations and class probabilities in the original image space. This eliminates the need for complex region proposal steps, resulting in highly efficient detection. Among these, the YOLO series, particularly YOLOv8, has been widely adopted in remote sensing applications due to its strong balance between detection accuracy and computational efficiency [27–29]. However, its native neck structure exhibits notable limitations in remote sensing scenarios, such as insufficient semantic correlation modeling for sparse small objects and overly simplistic receptive fields for multi-scale feature fusion. To address these issues, we propose a neck network compatible with the YOLO series, designed to enhance feature fusion for sparsely distributed targets and improve multi-scale context perception in remote sensing scenarios.

2.2. Remote Sensing Object Detection

The core difference between remote sensing images and natural images is reflected in the high background proportion and multi-angle and multi-scale distribution of targets. For example, Ding et al. [30] introduced a rotation mechanism in the candidate box generation stage, and Xie et al. [31] optimized OBB to enhance OBB representation. The boundary discontinuity problem of rotating box regression makes it difficult for the model to converge, although this can be alleviated via modulation loss (R3Det) [32] or Gaussian distribution modeling. However, the above methods ignore other core differences between remote sensing images and natural images. The existing methods rely on standard backbone networks (such as ResNet [8]) to extract features, and they use fixed-structure bottleneck networks (such as FPNs) for multi-scale fusion, but they fail to include an adaptive feature interaction mechanism for the extreme scale changes and context-sensitive nature of remote sensing targets.

For small object (defined as those occupying fewer than 32 × 32 pixels or less than 1% of the image area) detection, feature pyramid networks (FPNs) and their variants construct multi-scale feature layers and fuse features at different resolutions. Early single-stage detectors like the YOLO series significantly improved small object detection by introducing FPNs, which build top-down feature fusion paths across different backbone layers. However, the conventional FPN's single fusion mode struggles to adapt to the extreme scale differences and complex background interference in remote sensing scenarios. To address this, researchers have proposed enhancements such as the path aggregation network (PAN [33]), which introduces a bottom-up augmentation path to strengthen the semantic representation of shallow features. The bidirectional feature pyramid network (BiFPN [34]) further refines this approach with bidirectional cross-scale connections and adaptive weighting mechanisms, enabling efficient multi-scale feature interaction in YOLO-series models.

To address the unique demands of remote sensing imagery, recent work has explored context-aware feature pyramid designs. For instance, Xu et al. [35] enhanced spatial context modeling through a global attention module and proposed the global feature pyramid network, but its dense computation overhead limits its real-time performance. Yang et al. [36] combined channel and spatial attention mechanisms to suppress background interference and proposed the multi-attention feature pyramid network, but its stacked modules lead to a significant increase in parameters. In the realm of lightweight design, Li et al. [37] reduced the computational costs by pruning redundant connections, but the oversimplified structure reduces its sensitivity to tiny targets. These methods rely on standard feature pyramids for multi-scale fusion, without fully considering the unique characteristics of remote sensing features, such as large intraclass scale variation.

Recent studies on feature extraction backbones, such as that of Li et al. [38], propose selectively expanding the spatial receptive field for larger objects to capture more contextual scene information. However, dilated convolutions risk introducing noise, which is detrimental to small object detection. Cai et al. [39] avoided this issue by leveraging non-dilated multi-kernel depthwise convolutions to extract multi-scale texture features across different receptive fields. While these methods alleviate the challenges in extracting multi-scale features and contextual information for remote sensing targets, they do not specifically address the high background-to-target ratio characteristic of remote sensing images (RSIs). To tackle the challenges posed by high background dominance and multi-angle, multi-scale target distributions in RSIs, we propose PAMFPN, a feature pyramid network that leverages multi-kernel and dynamic kernel convolutions for efficient multi-scale feature fusion and employs adaptive sparse position attention to enhance the feature representation.

2.3. Spatial Feature Correlation Modeling

Attention mechanisms have become a core technology in improving remote sensing object detection performance through dynamic feature enhancement. However, the sparse distribution, arbitrary orientations, and complex backgrounds of remote sensing scenes impose unique requirements on attention design, posing significant challenges in balancing efficiency and effectiveness. Early work, such as the squeeze-and-excitation (SE [40]) module, enhanced the responses of important feature channels through channelwise compression and excitation. However, its static fully connected layers struggle to adapt to the multi-scale nature of remote sensing targets. Subsequent improvements, like Wang et al. [18], adopted lightweight local cross-channel interaction strategies to reduce the computational overhead but neglected spatial dimension correlations.

To integrate spatial and channel information for joint perception, Woo et al. [16] employed parallel channel and spatial attention branches, but they failed to model long-range dependencies, which are crucial for visual tasks. Hou et al. [41] and Xu et al. [42] embedded spatial position information into channel attention. Further advancements, such as the work of Cai et al. [39], leveraged global average pooling and 1D strip convolutions to capture relationships between distant pixels and enhance the features within central regions, guiding attention toward target-dense areas.

While prior work on spatial feature correlation in remote sensing object detection has extensively explored context modeling, none of these methods specifically address the sparse target distribution characteristic of remote sensing scenes. We propose an adaptive sparse encoding mechanism to guide attention toward sparsely distributed targets.

3. Proposed Methodology

3.1. PAMFPN

PAMFPN is a feature pyramid network specifically optimized for object detection in remote sensing scenarios. Given the unique characteristics of RSIs, multi-scale features typically contain substantial background noise (complex textures and visual clutter in non-target regions) and complex target representations, including faint features of small objects and multi-scale target variations. Traditional single-path feature pyramids (e.g., FPNs) often result in the loss of shallow-level details while causing semantic confusion in deeper layers. To address these issues, we employ a dual-path (bottom-up and top-down) multi-scale feature fusion approach. This design leverages high-level semantic information to enhance the feature representation of small targets in low-level layers, while simultaneously injecting detailed features from shallow layers into deep layers to suppress the propagation of background noise in high-level features.

As shown in Figure 1, the input image first passes through a backbone network to extract multi-scale features, which are then fed into our model for efficient multi-scale feature fusion. The fused features are finally delivered to the detection head for category classification and bounding box regression. The PAMFPN architecture consists of two key modules: the C3PAM module and the C2f convolutional module. This design enables seamless integration with various object detectors, such as the YOLO series, to generate the final object detection results for remote sensing imagery.

3.1.1. C3PAM Module

As illustrated in Figure 1, the cross-stage partial module with three PAMs (C3PAM) consists of the Convolution-BatchNorm–SILU (CBS) and position attention module (PAM) (Section 3.2.2) components. Built upon the cross-stage partial with 3 convolutions (C3) architecture [15], it employs a dual-branch feature interaction mechanism to achieve efficient multi-scale feature fusion and noise suppression. For low-level features F_l and high-level features F_h, the processing is as follows:

$$F_{cat} = CAT(F_l, F_h) \tag{1}$$

$$F_{CBS1} = \sigma(Norm(f_{1\times1}(F_{cat}))) \tag{2}$$

$$F_{CBS2} = \sigma(Norm(f_{1\times1}(F_{cat}))) \tag{3}$$

$$F_{PAM} = PAM(F_{CBS1}) \tag{4}$$

$$F_{CBS2} = \sigma(Norm(f_{1\times1}(CAT(F_{PAM}, F_{CBS2})))) \tag{5}$$

Here, σ denotes the SILU activation function, and $f_{k\times k}$ represents convolution operations with kernel size k. $Norm$ stands for normalization operations. CAT represents the splicing of channel dimensions. While maintaining the lightweight design of C3, C3PAM introduces the PAM module to specifically capture multi-scale and semantic feature information in remote sensing scenarios. The PAM enhances cross-position correlations for sparse targets while suppressing irrelevant background features through its adaptive attention mechanism. This design preserves the computational efficiency while significantly improving feature discrimination for objects with varying scales and complex spatial distributions.

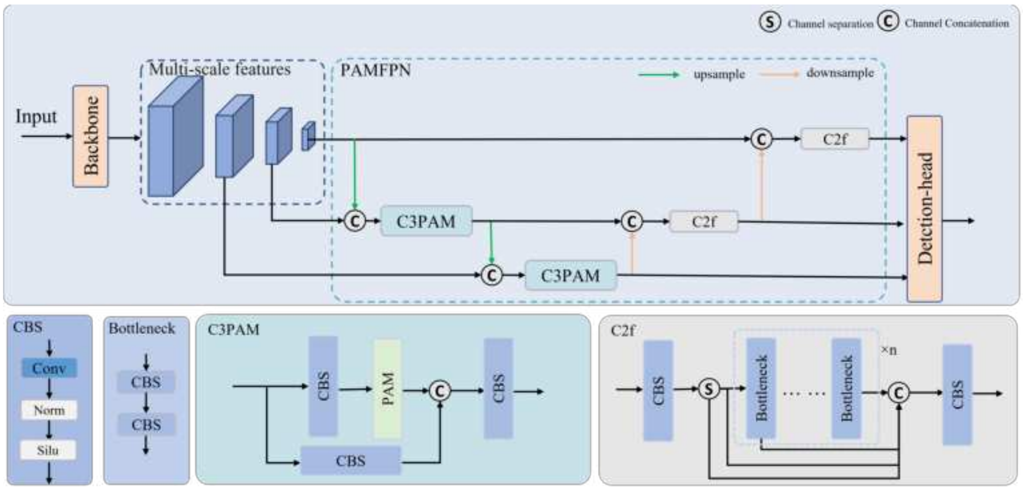

Figure 1. PAMFPM network architecture with C3PAM, C2F, CBS, and bottleneck components.

3.1.2. C2f Module

As illustrated in Figure 1, the C2f module is composed of CBS and bottleneck modules, and the multi-gradient flow feature extraction path is constructed by stacking multiple bottleneck structures, so as to enhance the multi-scale representation ability of the model. The module receives the features from the C3PAM fusion and uses $n + 2$ (3 by default when n = 1) parallel branches to extract features of different granularities, in which the shallow path retains high-resolution details to improve the detection ability for small targets, while the deep path captures the global context information through wide receptive fields to ensure robustness to large targets. Finally, the enhanced features of each branch are fused with concat and CBS to output enhanced features with both local fine features and global semantic information, which significantly improves the detection accuracy for dense small targets and complex large targets in remote sensing scenarios.

3.2. Dual-Kernel Adaptive Fusion Architecture and Position Attention Module

3.2.1. Dual-Kernel Adaptive Fusion Architecture

As shown in Figure 2, the dual-kernel adaptive fusion (DKAF) architecture effectively alleviates the problem of target scale changes in remote sensing through the co-design of the region-sensitive attention (RSA) module and the reconfigurable context aggregation (RCA) module. The architecture uses multi-level residual connections to maintain gradient flow, and, at the same time, multi-branch point convolution is used to implicitly map the channel dimension to learn the long-range dependence between channels. In the spatial dimension, the RCA module realizes multi-scale feature extraction through dynamic kernel adjustment, while the RSA module focuses on key regions through orthogonal convolution, which significantly improves the feature retention ability and positioning accuracy for small targets, especially for multi-scale target detection tasks in complex scenes.

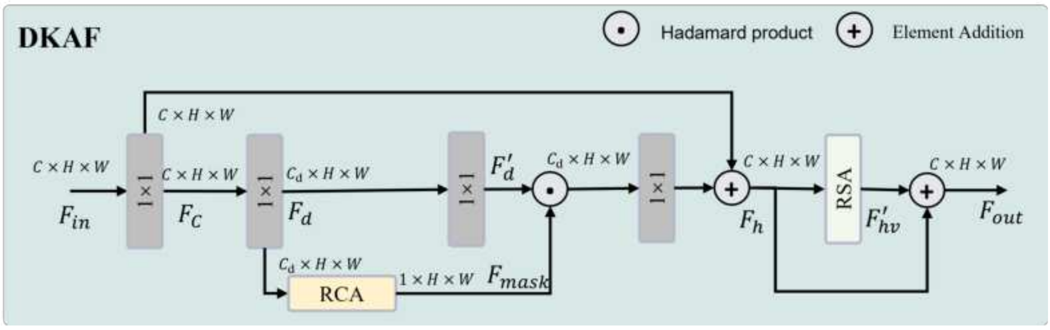

Figure 2. The structure diagram of the dual-kernel adaptive fusion (DKAF) architecture adopts a framework of multi-level residuals to reduce the loss of features.

3.2.2. Position Attention Module

As shown in Figure 3, the position attention module embeds the position-interactive context module (PICM) on the basis of the DKAF architecture to improve the model's association enhancement with sparse targets.

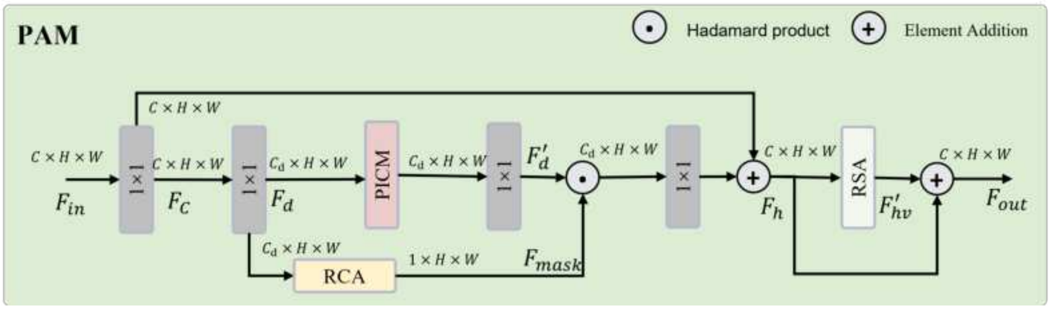

Figure 3. The structure diagram of the PAM module.

For an input feature map F_{in}, the process begins with nonlinear encoding to obtain F_c, followed by channel compression to produce F_d. This compression reduces the computational overhead while maintaining the ability to process high-resolution remote sensing imagery:

$$F_c = \sigma(Norm(f_{1\times1}(F_{in}))) \tag{6}$$

$$F_d = \sigma(Norm(f_{1\times1}(F_c))) \tag{7}$$

The compressed feature F_d is then processed in parallel through two branches. The RCA module extracts rich contextual semantic information via adaptive filtering convolutions:

$$F_{mask} = RCA(F_d) \tag{8}$$

As shown in Figure 4, the PICM performs sparse attention on k potential features F_k. This learns sparse target feature correlations:

$$F_d^{'} = PICM(F_d) \tag{9}$$

$$F_d^{'} = f_{1\times1}(F_d^{'}) \tag{10}$$

$$F_h = \sigma(Norm(f_{1\times1}(F_{mask} \times F_d^{'}))) + F_c \tag{11}$$

Finally, the RSA module performs feature fusion:

$$F_{hv}^{'} = RSA(F_h) \tag{12}$$

$$F_{out} = F_{hv}^{'} + F_h \tag{13}$$

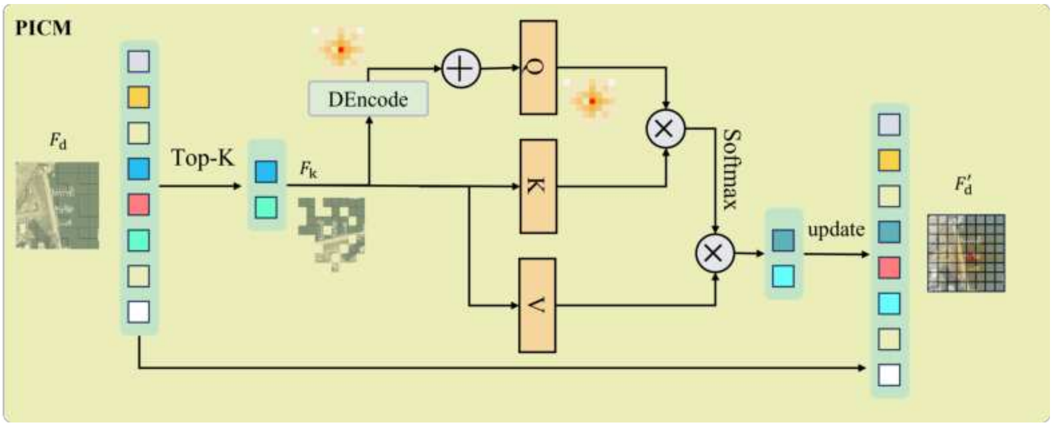

Figure 4. The PICM pre-screens *k* = 64 sparse features, uses DEncode for dynamic position encoding to enhance the semantic information of the query, and then uses self-attention to learn the similarities and differences in sparse features. Finally, the learned information is updated to the original features. For ease of understanding, the figure provides one-dimensional and two-dimensional abstract forms of the features.

3.3. Position-Interactive Context Module

As shown in Figures 3 and 4, the position-interactive context module (PICM) performs sparse attention on k potential features F_k, enhanced by dynamic position encoding (DEncode):

$$F_d^{'} = update(Attention(F_k, F_k, DEncode(F_k) + F_k), F_d) \tag{14}$$

Through Equations (9) and (14), the PICM computes Softmax-based activation scores for F_d; selects the top-K activated features for sparse attention; and replaces the original features with updated sparse representations.

3.3.1. Attention Mechanism and Motivation for Dynamic Position Encoding

The attention mechanism, which originates in the NLP domain, models the input sequence $I \in R^{N \times d_{model}}$ as follows:

$$\text{Attention}(I) = \text{softmax}\left(\frac{QK^{\top}}{\sqrt{d_{\text{model}}}}\right)V \tag{15}$$

$Q \in R^{N \times d_{model}}$ is a query, $K \in R^{N \times d_{model}}$ is a key, $V \in R^{N \times d_{model}}$ is a value, and the mathematical expression is

$$Q = IW_Q \tag{16}$$

$$K = IW_K \tag{17}$$

$$V = IW_V \tag{18}$$

where W_Q, W_K, W_V are the parameters of the model. In order to establish the attention mechanism to model the understanding of feature positions, some researchers have proposed sine and cosine position coding.

In order to realize the modeling of feature position information by attention mechanisms, some studies have proposed static position encoding methods based on sine and cosine functions. However, in the sparse attention mechanism, only focusing on local areas renders the semantic representation vulnerable to a lack of position information, resulting in limited spatial modeling capabilities. To this end, we propose DEncode, a dynamic position encoding method for sparse features, which is used to explicitly model their spatial distribution relationships. Different from traditional static encoding, DEncode can adapt to changes in spatial features and is more suitable for complex and changeable remote sensing environments.

3.3.2. Mechanism of Dynamic Position Encoding

Dynamic position encoding (DEncode) is not a static position code but is dynamically generated based on the selected area of each round of sparse attention. Its core functions are twofold: it provides spatial position information for the attention mechanism, so that the model can understand the relative relationships between sparse features; it promotes contextual connections between targets, especially for small targets with sparse spatial distributions but semantically related in remote sensing images. The position encoding needs to address the uniqueness of the absolute encoding and consistency with the same interval in relative position coding. We characterize the positional information of the feature using binary numbers (x_t, y_t) and use the Euclidean distance to describe the relative positional distance. The specific implementation is as follows:

$$\Delta l_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \tag{19}$$

$$d_i^j = \frac{\Delta l_{ij} - \min \Delta l_{ij}}{\max \Delta l_{ij} - \min \Delta l_{ij}} \tag{20}$$

$$e^{-\sum_j^N d_i^j} \tag{21}$$

In Equation (19), Δl_{ij} represents the relative position, while d_i^j denotes the normalized version of Δl_{ij}. The variable d_i^j encompasses the relative positional information between feature i and other features. In order to mimic the attenuation of the correlation between features with increasing distance, we take the form of an e-exponent.

Considering that each feature I is unique to the spatial radiation of the other features, we modify as follows:

$$I_i^{-\sum_j^N d_i^j} \tag{22}$$

By expanding each term of Equation (22), we get

$$I_i^{-d_i^0} \times I_i^{-d_i^1} \times I_i^{-d_i^i} \times \ldots \times I_i^{-d_i^j} \tag{23}$$

$I_i^{-d_i^j}$ can be expressed as the radiative effect of I_i on the global I_j feature, and it exhibits a range attenuation characteristic. Considering that the value of I is abnormal, so that it does not satisfy the conditions of the exponential function, we perform a sigmoid function on I. In order to preserve its spatial attenuation characteristics, we finally complete the following operations and obtain DEncode:

$$DEncode(I_i) = sigmoid(sigmoid(I_i)^{\sum_j^N d_i^j}) \tag{24}$$

3.4. Region-Sensitive Attention

As shown in Figure 5, the region-sensitive attention (RSA) module is a lightweight contextual attention mechanism specifically designed for detecting arbitrarily oriented targets in remote sensing imagery. At its core, the module employs vertical and horizontal strip-shaped convolutional kernels to extract orientation-aware contextual features and generate spatial attention weights that dynamically enhance the feature responses in target regions.

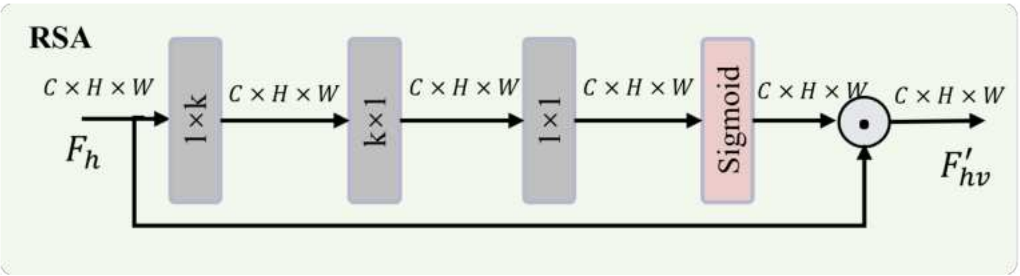

Figure 5. RSA is primarily composed of two orthogonal convolutions that capture the correlation between horizontal and vertical dimensions.

The RSA implementation utilizes two orthogonal sets of convolutions to capture horizontal and vertical contextual information. The horizontal branch processes the input features through a 11 convolution followed by normalization and SILU activation:

$$F_h^{'} = \sigma(Norm(f_{1\times k}(F_h))) \tag{25}$$

$$F_v^{'} = \sigma(Norm(f_{k\times 1}(F_h^{'}))) \tag{26}$$

These orientation-specific features are then fused using a pointwise convolution (11) and passed through a sigmoid function to generate the final spatial attention map. The attention-weighted features are computed as

$$F_{hv}^{'} = Sigmoid(Norm(f_{1\times 1}(F_v^{'}))) \times F_h \tag{27}$$

This design enables the module to effectively model long-range dependencies in complex remote sensing scenes. The orthogonal convolution approach provides comprehensive coverage of spatial orientations, ensuring robust feature extraction regardless of the target rotation angle.

3.5. Reconfigurable Context Aggregation

The reconfigurable context aggregation (RCA) module is a lightweight dynamic filtering structure specifically designed for cross-layer feature fusion in multi-scale remote sensing image analysis. This innovative module dynamically generates adaptive convolutional kernels based on input feature content to enhance feature representation. The key operation involves first processing the input features F_d through a 1 × 1 convolution, followed by pixel shuffle upsampling to reconstruct the spatial dimensions. The module then applies large-kernel average pooling (7 × 7) to extract globally significant regions, generating spatial attention weights that automatically enhance critical target features while suppressing complex background noise in cross-layer concatenated features, as shown in Figure 6. The complete process can be expressed as

$$F_{mask} = \sigma(AvgPool(Reshape(f_{1\times 1}(F_d)))) \tag{28}$$

where σ represents the sigmoid activation function. This design significantly improves the detection robustness in complex remote sensing scenarios by implementing content-aware feature enhancement through its dynamic kernel generation mechanism and global context modeling capability. The RCA module achieves this while maintaining computational efficiency through optimized operations like pixel shuffle and large-kernel pooling, making it particularly suitable for processing high-resolution remote sensing imagery with diverse target scales and complex backgrounds. As illustrated in Figure 6, the module effectively bridges different network layers to enhance multi-scale feature fusion without introducing a substantial computational overhead.

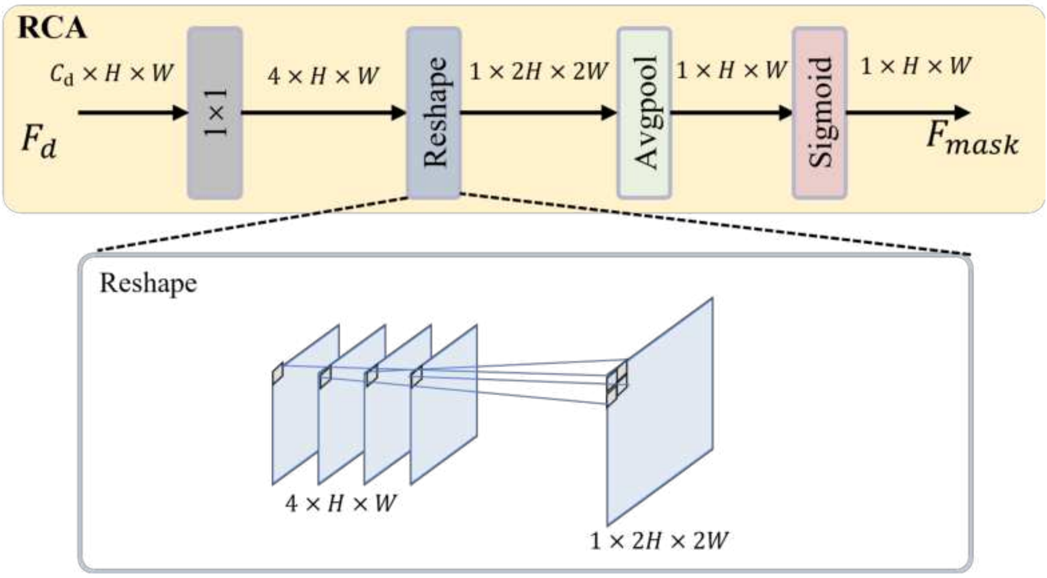

Figure 6. RCA, which expands the information of the feature respatial dimension through pixel rearrangement and filters through the downsampling pooling of large kernels, obtains the adaptive filtering weight of the large receptive field and then strengthens the adaptability of remote sensing scenes.

4. Datasets and Experimental Environment

4.1. Datasets

To validate the effectiveness of our method, we conduct comprehensive experiments on three publicly available benchmark datasets. Figures 7 and 8 depict the distribution of instances of different categories, where the horizontal axis represents the category and the vertical axis represents the number of instances of each category. The distribution of the target size relative to the image size is shown, where the horizontal and vertical axes represent the width and height ratios of the target to the image, respectively.

4.1.1. DOTA-v1.0 Dataset

As a large-scale optical dataset for aerial object detection, DOTA-v1.0 [4] contains 2806 high-resolution images with 188,282 annotated instances across 15 categories: plane (PL), baseball diamond (BD), bridge (BR), ground track field (GTF), small vehicle (SV), large vehicle (LV), ship (SH), tennis court (TC), basketball court (BC), storage tank (ST), soccer-ball field (SBF), roundabout (RA), harbor (HA), swimming pool (SP), and helicopter (HC). The dataset exhibits significant variations in both object orientation and scale. The images are officially split into 1411 training samples, 458 validation samples, and 937 testing samples.

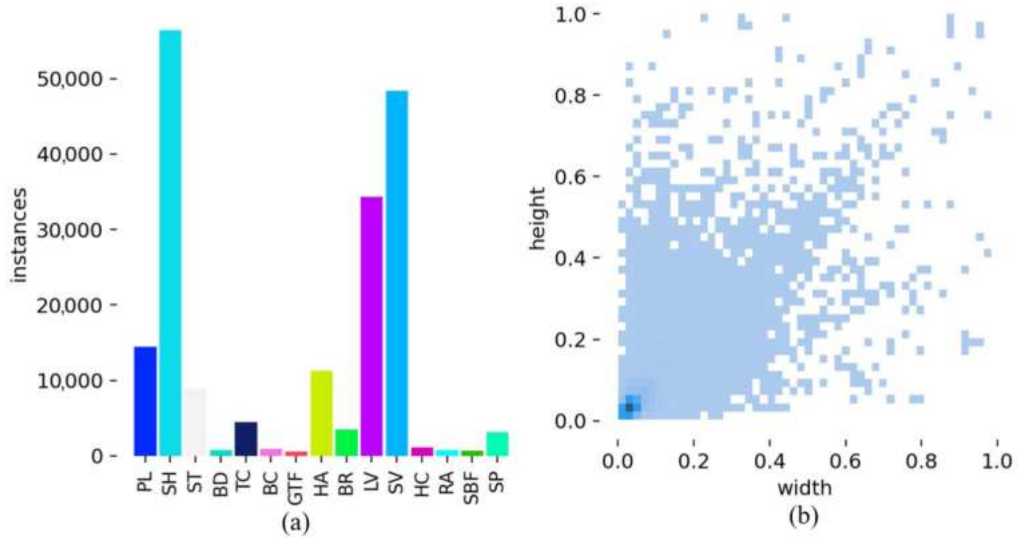

Figure 7. Information about DOTA-v1.0. (**a**) Number of instances per category in the DOTA-v1.0 dataset; (**b**) scale variation (target size to image size ratio) in targets in the DOTA-v1.0 dataset.

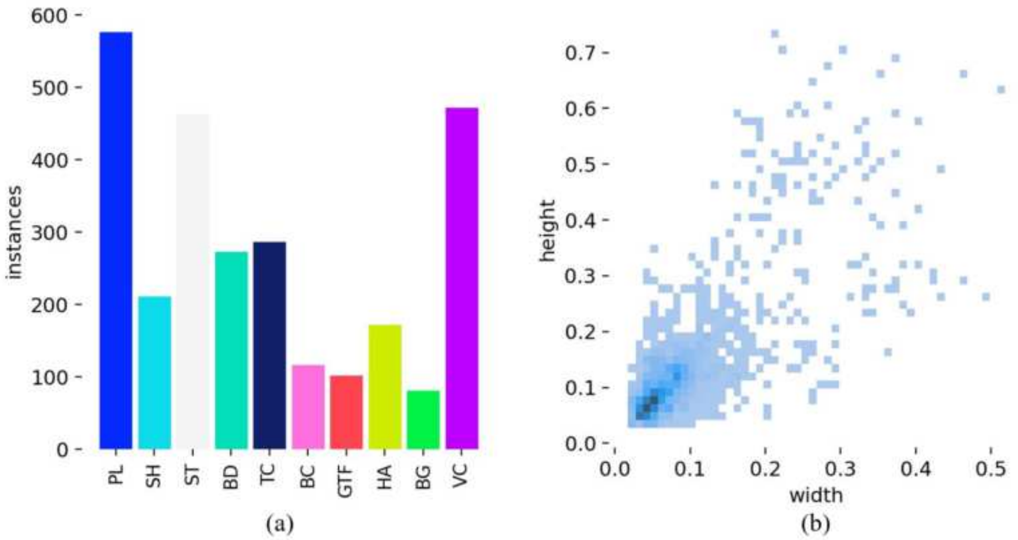

Figure 8. Information about HWPUVHR-10. (**a**) Number of instances per category in the HWPUVHR-10 dataset; (**b**) scale variation (target size to image size ratio) in targets in the HWPUVHR-10 dataset.

4.1.2. SSDD Dataset

This authoritative benchmark dataset for SAR ship detection [6] comprises 1160 synthetic aperture radar images containing 3570 annotated ship instances. Following standard practice, we use 928 images for training and 232 for validation. All ship targets are annotated with horizontal bounding boxes (HBBs).

4.1.3. HWPUVHR-10 Dataset

Designed specifically for high-resolution optical remote sensing object detection [5], this dataset contains 600 images covering 10 typical object categories: airplane (PL), ship (SH), storage tank (ST), baseball diamond (BD), tennis court (TC), basketball court (BC), ground track field (GTF), harbor (HB), bridge (BG), and vehicle (VC). The dataset is divided into 456 training images and 194 validation images, with all objects annotated using horizontal bounding boxes (HBBs).

4.2. Evaluation Indicators

In order to accurately evaluate the effectiveness of the model in multi-view aircraft inspection tasks, we use the average accuracy (mAP) and mAP50 (IOU = 0.5) as key indicators of the detection accuracy. In addition, we consider parameter counting (M) and floating-point arithmetic (GFLOPS) to measure the complexity and computational efficiency of the model.

4.3. Experimental Configuration

The experimental setup is based on Ubuntu 22.04, CUDA 11.8 and PyTorch 2.1.1. The hardware configuration includes an Intel Xeon Silver 4310 processor and two 4090 GPUs. More details of the parameters are provided in Table 1. Additionally, we use data augmentation in the experiment and close the last 10 epochs, as shown on Figure 9. In the experiment, we train on the training set and test and validate on the validation set. Among them, the DOTAv-1.0 dataset uses the OBB box for experiments, and the other datasets use HBB boxes for experiments.

Table 1. Training parameters

Parameter	Value	Parameter	Value
Epochs	300	Optimizer	AdamW
Batch size	8	Learning rate	0.01
Image size	640 × 640	Warmup epochs	3

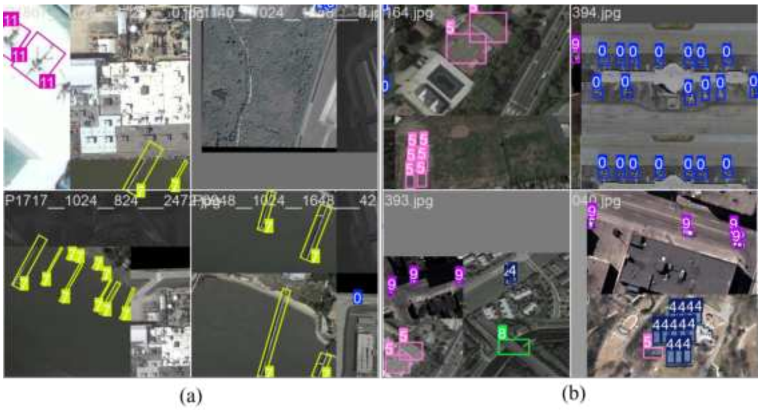

Figure 9. Mosaic is the first innovative data augmentation method proposed for YOLOv4 to greatly improve the performance of small object detection by combining multiple training images. (**a**) is the enhanced effect of the OBB box. (**b**) is the enhanced effect of the HBB box.

5. Experiments

Firstly, we conduct comparative experiments on each module based on YOLOv8, and we carry out ablation experiments, comparison experiments, visualization experiments, and generalization experiments on the whole model.

5.1. Comparison with State-of-the-Art FPN Methods

As shown in Table 2, the experimental results on the HWPUVHR-10 dataset demonstrate the superior performance of our proposed PAMFPN model compared to existing feature pyramid variants. With an impressive 87.7% mAP50, PAMFPN outperforms PAN (87.2%), BiFPN (85.8%), and Slim-Neck (87.1%) while maintaining excellent computational efficiency at just 26.7 GFLOPs. Compared to the baseline PAN architecture, PAMFPN achieves a 0.5% improvement in the mAP50 while simultaneously reducing the computational costs by 6.7%, decreasing from 28.6 GFLOPs to 26.7 GFLOPs. The model shows particularly strong performance on challenging categories, reaching 95.5% accuracy for storage tanks (a 1.9% improvement over PAN) and maintaining 98.9% precision for ground track fields.

Table 2. Performance comparison of different feature pyramid networks integrated with YOLOv8 on object detection tasks (HWPUVHR-10 dataset).

Model	mAP50	GFLOPs	Params	PL	SH	ST	BD	TC	BC	GTF	HB	BG	VC
PANet [33]	87.2	28.6	11	98.8	86.4	93.6	98.8	84.9	60.5	98.8	93.6	69.5	87.3
BIFPN [34]	85.8	46.3	11	99.0	84.5	90.4	97.5	85.0	61.3	97.3	92.5	69.5	80.6
Slim [37]	87.1	24.7	10	99.4	86.5	91.7	98.8	82.0	57.0	98.8	94.7	80.7	81.1
MAFPN [36]	87.1	38.3	11	99.2	85.2	88.7	98.1	85.6	61.6	98.2	88.3	78.0	88.2
AFPM [43]	82.4	33.3	12	99.4	75.5	97.2	97.1	81.0	50.7	98.0	77.4	66.2	81.1
HSFPN [44]	86.6	19.3	6	98.0	81.7	91.6	96.3	85.6	63.0	98.9	95.3	75.8	79.4
GFPN [35]	86.9	35.0	9	99.3	82.9	86.1	98.8	85.5	62.8	98.6	96.3	75.6	83.1
PAMFPN	87.7	26.7	11	99.4	86.0	95.5	98.7	85.1	60.7	98.9	94.1	72.4	86.1

These results highlight PAMFPN's exceptional balance between detection accuracy and computational efficiency, especially when handling the dataset's most demanding scenarios, including extreme scale variations, complex background interference, and high-density object clusters. The model's advantages become even more apparent when compared to other approaches—while BiFPN requires 73% more computational resources (46.3 GFLOPs) and has inferior accuracy, and AFPN suffers from significant accuracy degradation (82.4% mAP50), PAMFPN delivers superior performance without excessive computational demands. This achievement stems from the model's innovative adaptive sparse attention mechanism, which effectively focuses the computation on critical regions, combined with its multi-kernel fusion approach that captures diverse spatial contexts efficiently. The dynamic feature recalibration further ensures consistent performance across targets of varying sizes, making PAMFPN particularly effective for remote sensing applications where both accuracy and efficiency are crucial.

5.2. *Comparison of Attention Modules*

Our comprehensive ablation studies on the DOTA-v1.0 dataset using the YOLOv8 framework provide compelling evidence for the effectiveness of the proposed PAM. As shown in Table 3, the experimental results demonstrate that the PAM significantly enhances the detection performance while maintaining the same model parameters (11 M). The YOLOv8-PAM configuration achieves state-of-the-art performance, with 59.4% mAP and 76.3% mAP50 values, representing improvements of 0.3% and 0.8%, respectively, over the baseline YOLOv8 model.

Table 3. Performance comparison of different attention mechanisms integrated with YOLOv8 on object detection tasks (DOTA-v1.0 dataset).

Model	mAP (%)	mAP50 (%)	GFLOPs	#P	PL	SH	ST	BD	TC	BC	GTF	HA	BR	LV	SV	HC	RA	SBF	SP
YOLOv8	59.1	75.5	28.8	11	96.0	91.0	79.4	84.3	95.8	70.3	65.6	86.6	55.0	88.2	71.0	49.5	71.4	57.0	71.2
Swin [17]	59.2	76.1	30.4	11	96.3	91.2	80.5	84.1	95.8	69.6	66.3	87.4	55.5	89.4	72.4	51.5	71.9	54.9	74.1
ECA [18]	58.9	75.9	27.4	11	96.3	91.2	79.1	82.8	95.6	70.7	64.3	87.3	55.8	88.6	72.7	53.0	69.2	60.4	71.4
ELA [42]	58.7	75.8	27.4	11	96.3	90.8	80.3	82.1	95.9	72.0	62.3	87.0	55.3	88.6	71.5	51.8	71.7	58.3	73.0
CA [41]	59.0	75.8	27.4	11	96.2	91.1	80.0	85.2	96.3	71.8	64.5	86.9	53.4	88.5	71.5	50.3	73.2	55.4	73.0
SE [40]	59.0	75.9	27.4	11	96.3	90.9	79.8	82.5	96.0	69.3	66.3	87.0	55.1	88.6	73.1	53.3	74.4	55.5	70.8
CAA [39]	59.2	75.9	27.6	11	95.9	91.0	79.7	83.5	95.8	71.3	62.4	87.0	54.6	88.2	71.5	56.6	70.5	58.2	72.2
CBMA [16]	59.0	76.0	27.4	11	96.0	91.3	80.3	83.5	95.9	70.9	64.4	87.2	53.4	89.0	73.3	54.1	72.0	54.9	73.4
PAM	59.4	76.3	27.7	11	95.9	91.2	80.1	82.5	96.1	71.0	66.9	86.9	55.3	88.9	72.0	54.1	72.3	59.8	71.5

Notably, YOLOv8-PAM maintains excellent computational efficiency at just 27.7 GFLOPs, comparable to lightweight attention modules like ECA (27.4 GFLOPs) and ELA (27.4 GFLOPs) while delivering superior detection accuracy. Detailed category-wise analysis reveals the PAM's particular advantages in challenging detection scenarios: it achieves 66.9% accuracy on ground track field (GTF), with a 1.3% improvement over the baseline, while maintaining strong accuracy of 71.5% for swimming pool (SP).

The comparative analysis shows that, while other attention mechanisms achieve some performance gains, they suffer from inconsistent improvements across metrics. For instance, SwinTransformer achieves a 76.1% mAP50 but increases the computational cost by 5.6% (30.4 GFLOPs). These results conclusively demonstrate the PAM's unique ability to enhance multi-scale object detection in remote sensing applications without substantially increasing the computational overhead. The module's stable performance improvements across different target categories and scales, coupled with its efficient implementation, make it particularly suitable for practical remote sensing applications where both accuracy and efficiency are critical requirements.

The success of the PAM can be attributed to its innovative design: the sparse position attention mechanism effectively focuses computational resources on critical regions, while the dynamic kernel adaptation ensures robust feature extraction across varying target sizes. This combination enables consistent performance gains without the computational instability observed in alternative attention approaches.

5.3. Component Ablation

Table 4 presents the ablation study results for individual PAM components on the DOTA-v1.0 dataset, systematically demonstrating how each element contributes to the module's overall effectiveness. The complete PAM module (integrating PICM, RSA, and RCA) achieves optimal performance with a 76.3% mAP50—a 0.8% improvement over the baseline—while maintaining computational efficiency at 27.7 GFLOPs.

Table 4. Ablation experimental results for core components in the DOTA-v1.0 dataset. ↑ indicates that the higher the value of the indicator, the better; ↓ indicates that the lower the value of the indicator, the better. ✓ indicates the use of the module.

PICM	RSA	RCA	mAP50 ↑	GFLOPs ↓	Params ↓
			75.5	28.8	11
✓			75.7	27.5	11
	✓		75.5	27.7	11
		✓	75.9	27.5	11
✓	✓		76.0	27.7	11
✓		✓	75.4	27.5	11
	✓	✓	74.7	27.7	11
✓	✓	✓	76.3	27.7	11

The RCA module proves particularly impactful as a standalone component, delivering a 0.4% mAP50 boost through its adaptive filtering mechanism for multi-scale targets. The PICM demonstrates dual benefits, reducing the computational cost by 4.5% while simultaneously improving the accuracy by 0.2% via its position-aware attention mechanism. Meanwhile, the RSA module optimizes the computational efficiency without compromising the accuracy.

A crucial finding emerges in the components' synergistic interaction: the full three-module combination outperforms any dual-component configuration. The PICM+RSA pairing alone achieves a 0.5% performance gain, revealing the complementary strengths of position awareness and rotation-sensitive feature extraction. These results conclusively validate the PAM's innovative three-pronged approach—combining position perception (PICM), rotation adaptation (RSA), and scale adaptability (RCA)—which collectively addresses core challenges in remote sensing object detection. The module's architectural elegance lies in delivering comprehensive performance enhancements without a meaningful computational overhead.

The ablation study further reveals that, while each component offers distinct advantages, their integrated operation creates multiplicative benefits that transcend simple additive improvements. This emergent property stems from the PAM's carefully designed feature interaction pathways, where the PICM focuses the computation on salient regions, RSA maintains orientation robustness, and reconfigurable aggregation (RCA) dynamically adjusts to scale variations.

5.3.1. PICM Ablation

Table 5 compares the performance of different position encoding methods in the PICM on the DOTA-v1.0 dataset. The experimental results show that the PICM using adaptive position coding (DEncode) achieves the best performance with a 76.3% mAP50, which is 0.8% higher than that of the benchmark model, while maintaining the computational efficiency at 27.7 GFLOPs. Specifically, although the no-code scheme can obtain a 75.9% mAP50, it lacks explicit modeling of the spatial distribution of the target. The performance of traditional sinusoidal coding (sincos) is slightly inferior (75.7% mAP50), which verifies the limitations of fixed-mode position coding in remote sensing scenarios. These results fully confirm the importance of adaptive position coding for the spatial modeling of remote sensing targets, and it significantly enhances the position perception ability of the feature pyramid by dynamically adapting to the structural characteristics of the image content.

Table 5. Ablation experimental results for coding in the DOTA-v1.0 dataset.

Method	mAP50	GFLOPs	Params
BASE	75.5	28.821	11.1
no-code	75.9	27.707	10.8
sincos	75.7	27.707	10.8
DEncode	76.3	27.727	10.8

5.3.2. RSA Ablation

In order to determine the optimal size of the cross-convolutional kernel in the RSA module, we carry out systematic ablation experiments on the DOTA-v1.0 dataset and compare the performance of $k = 3, 5, 7, 9, 11$, i.e., five different convolution kernel sizes. The experimental results show that, when $k = 5$, the model shows the best comprehensive performance, reaching a 76.3% mAP50, which is 1% and 0.3% higher than at $k = 3$ and $k = 7$, respectively, while maintaining the computational efficiency at 27.7 GFLOPs, as shown in Table 6. Specifically, the 5 cross-convolutional kernel can achieve a good balance between local detail capture and global context perception, which not only avoids the problem of insufficient modeling of rotational features caused by insufficient receptive fields of the 3 convolutional kernels but also overcomes the effects of excessive background noise introduced by the 7, 9, 11 convolution kernels.

Table 6. Ablation experimental results for orthogonal convolutional kernel in the DOTA-v1.0 dataset.

Kernel Size	mAP	mAP50	GFLOPs	Params
3	58.7	75.3	27.722	10.775
5	59.4	76.3	27.727	10.776
7	59.0	76.0	27.732	10.776
9	59.3	76.2	27.737	10.777
11	58.7	75.2	27.742	11.078

5.4. Generalization Experiments

To comprehensively validate the generalization capability of PAMFPN, we conduct comparative experiments with multiple detectors on two remote sensing datasets (SSDD and HWPUVHR-10), as shown in Tables 7 and 8. The results demonstrate that the improved models equipped with PAMFPN (denoted as "-our") exhibit superior performance in most cases.

On the SSDD dataset, YOLOv5-our achieves the best performance with 76.2% mAP, representing a 1.3% improvement over the original version. YOLOv8-our reaches 75.8% mAP, outperforming the baseline (75.0%). Notably, several improved models achieve performance gains while maintaining or reducing the computational costs, such as YOLOv10-our, which reduces the GFLOPs by 0.4 while increasing the mAP50 by 0.5% .

Table 7. Experimental results for multiple detectors and using a PAMFPN detector on an SSDD dataset, where -our indicates the use of PAMFPN.

Benchmark	GFLOPS	Params	mAP50	mAP
yolov10-our	22.7	7.7	98.0	74.5
yolov10	24.8	8.1	97.5	73.9
yolov9-our	22.7	6.5	98.3	75.3
yolov9	26.7	7.2	98.3	74.9
yolov8	28.6	11.1	98.5	75.0
yolov8-our	26.9	10.8	98.5	75.8
yolov11	22.3	9.7	98.3	74.0
yolov11-our	20.9	9.3	98.2	73.6
yolov6	44.2	16.0	98.1	73.8
yolov5	23.8	9.1	98.9	74.9
yolov5-our	23.3	8.9	98.4	76.2
yolov3	19.1	12.0	95.6	72.1
RT-DETR	130.0	42.0	73.1	49.5

Table 8. Experimental results for multiple detectors and using a PAMFPN detector on the HWPUVHR-10 dataset, where -our indicates the use of PAMFPN.

Benchmark	GFLOPS	Params	mAP50	mAP
yolov10-our	22.7	7.7	84.7	53.7
yolov10	24.8	8.1	83.2	53.0
yolov9-our	22.7	6.5	88.4	55.2
yolov9	26.7	7.2	85.9	54.7
yolov8	28.6	11.1	87.2	54.7
yolov8-our	26.9	10.8	87.7	55.3
yolov11	22.3	9.7	87.9	55.3
yolov11-our	20.9	9.3	88.8	55.0
yolov6	44.2	16.0	86.4	54.7
yolov5	23.8	9.1	87.4	54.4
yolov5-our	23.3	8.9	87.5	54.8
yolov3	19.1	12.0	85.5	50.6
RT-DETR	130.0	42.0	86.5	56.1

On the more challenging HWPUVHR-10 dataset, YOLOv9-our delivers the most outstanding performance, achieving a 88.4% mAP50 and 55.2% mAP, representing improvements of 2.5% and 0.5%, respectively, over the baseline. YOLOv11-our maintains a relatively low computational cost (20.9 GFLOPs) while reaching a 88.8% mAP50, surpassing most comparable models.

Compared with traditional detectors, PAMFPN-enhanced models demonstrate significant advantages. For instance, YOLOv5-our achieves an 87.5% mAP50 on HWPUVHR-10 with only 23.3 GFLOPs, approaching the performance of RT-DETR (86.5%), which requires 130 GFLOPs, confirming PAMFPN's efficiency.

These results fully validate the generalization advantages of the PAMFPN architecture: its position-aware feature fusion mechanism can adapt to different detection frameworks (YOLO series), significantly improving the detection performance in typical remote sensing scenarios (small object detection in SSDD, complex backgrounds in HWPUVHR-10) while maintaining computational efficiency. Particularly when processing datasets like HWPUVHR-10, containing multi-scale targets, PAMFPN demonstrates more pronounced improvements, proving the effectiveness of its rotation-adaptive and scale-adaptive mechanisms.

The experimental results on three representative remote sensing target detection datasets (DOTA, HWPUVHR-10, and SSDD) show that our method has significant advantages in improving model robustness, as shown in Table 9. Robustness here specifically refers to the ability of the model to maintain stable detection performance under complex remote sensing conditions, including significant scale changes, background interference, the presence of small targets, and adaptability to cross-modal data. From the results, it can be seen that, whether integrated into yolov8 or yolov9, the model improved by our method achieves higher detection accuracy (both the mAP and mAP50 have been improved) on both the HWPU and SSDD datasets, accompanied by a slight or even significant reduction in the computational overhead, indicating that it can still maintain stable performance in scenes with dense small targets and complex interference. Especially in the SSDD dataset, the mAP is improved by 0.8%, further verifying its strong perception ability for small targets. In addition, in the DOTA dataset consisting of multi-scale complex scenes, the accuracy of yolov8-our is also slightly improved, further confirming that our method enhances its generalization ability and robustness in diverse remote sensing scenarios while keeping the model lightweight.

Table 9. Detection performance comparison on multiple datasets.

Model	GFLOPs	Params	DOTA		HWPU		SSDD	
			mAP	mAP50	mAP	mAP50	mAP	mAP50
yolov9	26.7	7.2	48.6	71.7	54.7	85.9	74.9	98.3
yolov9-our	22.7	6.5	48.7	71.7	55.2	88.4	75.3	98.3
yolov8	28.8	11.1	59.1	75.5	54.7	87.2	75.0	98.5
yolov8-our	27.7	10.8	59.4	76.3	55.3	87.7	75.8	98.5

5.5. Visualizing Experiments

To demonstrate the performance advantages of PAMFPN in complex remote sensing scenarios more intuitively, we conduct comparative experiments with the baseline model, YOLOv8, on three representative scenarios from the DOTA-v1.0 dataset, as shown in Figures 10 and 11. The results indicate that PAMFPN achieves significant performance improvements under conditions of extreme scale variation, severe occlusion, and extreme rotation.

As illustrated in Figure 10, the upper group of airport scenes presents three key challenges: (1) extreme scale variation among aircraft targets, (2) approximately half of the targets being occluded, and (3) annotations including only fully visible targets. PAMFPN successfully detects all 17 aircraft (including 10 occluded ones), while YOLOv8 identifies only 13 complete aircraft. In the lower group of court scene comparisons, we specifically examine PAMFPN's detection performance under tree shadow occlusion conditions. PAMFPN successfully detects all tennis courts obscured by tree shadows, whereas the baseline model misses three instances.

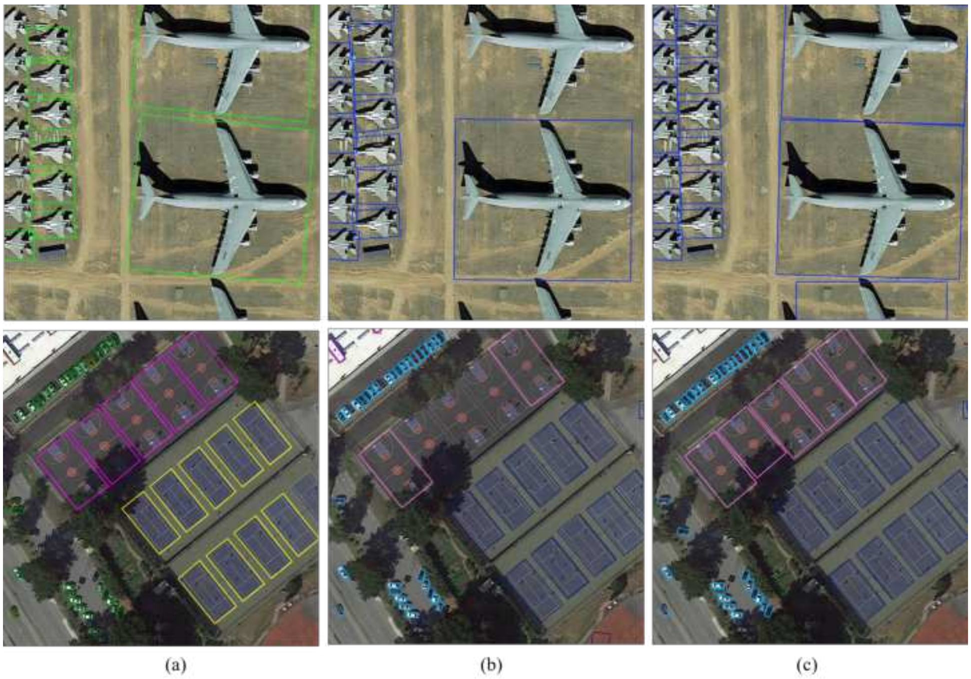

Figure 10. (**a**) is the true label, (**b**) is the benchmark model prediction, and (**c**) is our prediction.

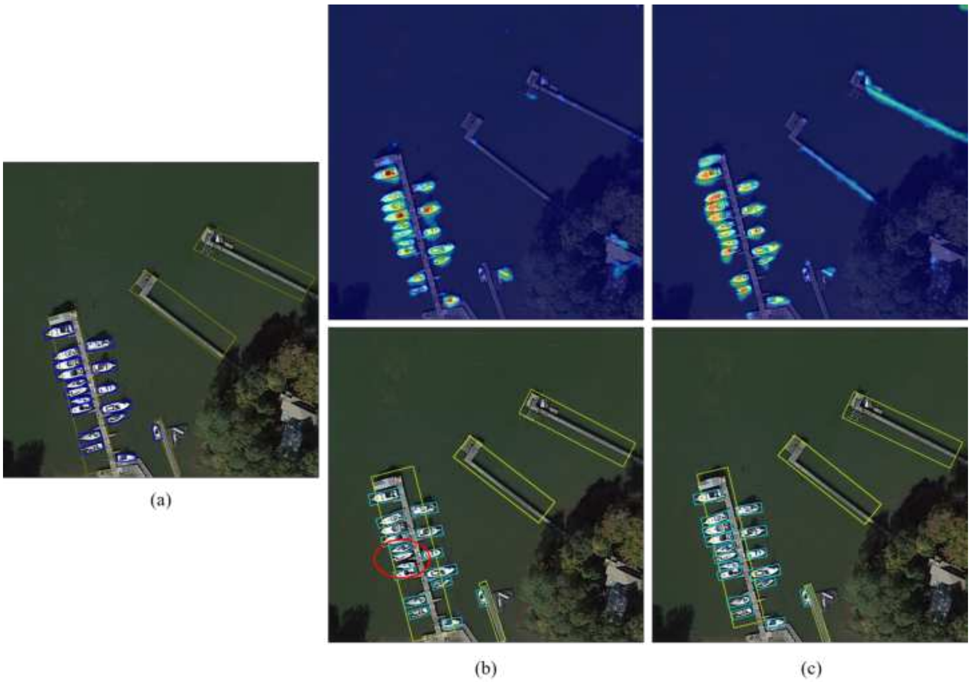

Figure 11. (**a**) is the true label, (**b**) is the benchmark model prediction, and (**c**) is our prediction. The heat map is visualized using GradCAM [45].

As shown in Figure 11, we further validate PAMFPN's capability in detecting arbitrarily oriented targets through the visualization of the detection results in a typical harbor scene. The red-marked areas highlight two vessels sailing in opposite directions. The experimental results demonstrate that PAMFPN successfully detects both opposite-direction vessels in the marked region, with the heatmaps precisely focusing on the ships' bows. In contrast, the baseline model shows missed detections while primarily focusing on darker-colored hull sections, failing to capture key target features and ultimately missing three vessels.

These visualization experiments comprehensively verify PAMFPN's advantages: the PICM module enables precise position awareness, the RSA module extracts rotation-sensitive features, and the RCA module achieves multi-scale feature adaptation. Their synergistic operation effectively addresses various challenges in remote sensing object detection, including scale variation, occlusion, illumination changes, and arbitrary target orientations. Compared to the baseline, PAMFPN demonstrates significant performance improvements in these complex scenarios, providing an effective solution for high-precision remote sensing target detection.

6. Discussion

Through systematic experimental validation, we demonstrate that PAMFPN's core advantage lies in its innovative feature fusion mechanism, which fully considers the unique characteristics of remote sensing scenarios: small target proportions, arbitrary orientations, and significant scale variations. The carefully designed PICM employs adaptive position encoding (DEncode) to achieve the sparse modeling of target spatial distributions, thereby enhancing the feature correlation for sparse targets. The RSA component captures horizontal and vertical positional relationships, strengthening the multi-dimensional perception of rotated targets. The RCA adaptive filtering module dynamically adjusts the kernels to extract contextual information, effectively addressing three key challenges in remote sensing object detection: multi-scale target detection, complex background interference, and arbitrary orientation target recognition.

Extensive experiments validate PAMFPN's generalization capabilities. On both the SSDD and HWPUVHR-10 remote sensing datasets, improved models equipped with PAMFPN demonstrate consistent performance gains. Notably, while maintaining or reducing the computational costs, the models achieve improved detection accuracy across multiple YOLO variants. For instance, YOLOv5-our achieves 76.2% mAP on the SSDD dataset, representing a 1.3% improvement over the original version, fully demonstrating PAMFPN's universality and efficiency.

The visualization analysis further confirms the model's advantages. In various typical scenarios, including airports, tennis courts, and harbors, PAMFPN consistently demonstrates strong robustness against scale variations, occlusions, and orientation changes. Particularly in extreme cases like detecting vessels sailing in opposite directions, the model maintains accurate detection and orientation estimation, which holds significant practical importance.

In conclusion, through innovative module design and hierarchical feature processing, PAMFPN establishes an efficient and robust framework for remote sensing object detection. Extensive experiments validate its advantages in terms of accuracy, efficiency, and generalization, providing a reliable solution for high-precision remote sensing object detection. Future work may further explore this framework's potential in other remote sensing tasks, such as scene classification and change detection.

7. Conclusions

This paper presents PAMFPN, a novel model tailored to addressing the unique challenges of remote sensing object detection, including sparse target distributions, multi-scale variations, and complex background interference. By integrating three key innovations—the C3PAM module for adaptive sparse position modeling, the RSA module for rotation-sensitive feature extraction, and the RCA module for dynamic multi-scale context aggregation—PAMFPN achieves robust and efficient feature fusion across diverse remote sensing scenarios. Extensive experiments on three benchmark datasets (DOTA-v1.0, SSDD, and HWPUVHR-10) demonstrate the superiority of our approach, with PAMFPN achieving state-of-the-art performance (76.3% mAP50 on DOTA-v1.0 and 87.7% mAP50 on HWPUVHR-10) while maintaining computational efficiency (26.7–27.7 GFLOPs). The modular design ensures compatibility with mainstream detectors, as evidenced by the consistent performance gains across YOLO-series variants. Visualization analyses further validate the model's capability to handle extreme scale variations, occlusions, and arbitrary orientations through the synergistic operation of its attention mechanisms.

Future work will focus on lightweight deployment for edge devices, temporal modeling for video-based detection, and cross-modal fusion strategies. PAMFPN establishes a versatile framework that bridges the gap between theoretical innovation and practical application in remote sensing interpretation, offering significant potential for intelligent geospatial analysis systems.

Author Contributions: Conceptualization, X.Y.; formal analysis, S.X.; methodology, X.Y. and S.X.; investigation, L.L. and S.X.; visualization, S.X.; validation, X.Y.; data curation, L.L. and S.L.; writing—original draft, S.X., L.L. and S.L.; writing—review & editing, X.Y., Y.F. and X.Z.; project administration, X.Z.; supervision, Y.F.; funding acquisition, X.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Natural Science Foundation of China (62301174, 62462031), Guangzhou basic and applied basic research topics (2024A04J2081, 2025A04J3375) and Natural Science Foundation of Jiangxi Province (20242BAB26023).

Data Availability Statement: The codes and parameters of our model are publicly available at https://github.com/SuiHuaXue/PAMFPN (accessed on 7 May 2025).

Acknowledgments: This work was supported in part by the National Natural Science Foundation of China under Grant No. 62462031 and in part by the Natural Science Foundation of Jiangxi Province under Grant 20242BAB26023.

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

References

1. Zhao, Y.; Yang, T.; Wang, S.; Su, H.; Sun, H. Adaptive Dual-Domain Dynamic Interactive Network for Oriented Object Detection in Remote Sensing Images. *Remote Sens.* **2025**, *17*, 950. [CrossRef]
2. Cheng, G.; Wang, J.; Li, K.; Xie, X.; Lang, C.; Yao, Y.; Han, J. Anchor-Free Oriented Proposal Generator for Object Detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–11. [CrossRef]
3. Ding, J.; Xue, N.; Xia, G.S.; Bai, X.; Yang, W.; Yang, M.Y.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; et al. Object Detection in Aerial Images: A Large-Scale Benchmark and Challenges. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 7778–7796. [CrossRef]
4. 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 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3974–3983. [CrossRef]
5. Cheng, G.; Han, J.; Zhou, P.; Guo, L. Multi-class geospatial object detection and geographic image classification based on collection of part detectors. *ISPRS J. Photogramm. Remote Sens.* **2014**, *98*, 119–132. [CrossRef]
6. Zhang, T.; Zhang, X.; Li, J.; Xu, X.; Wang, B.; Zhan, X.; Xu, Y.; Ke, X.; Zeng, T.; Su, H. SAR Ship Detection Dataset (SSDD): Official Release and Comprehensive Data Analysis. *Remote Sens.* **2021**, *13*, 3690. [CrossRef]
7. Wang, X.; Han, C.; Huang, L.; Nie, T.; Liu, X.; Liu, H.; Li, M. AG-Yolo: Attention-Guided Yolo for Efficient Remote Sensing Oriented Object Detection. *Remote Sens.* **2025**, *17*, 1027. [CrossRef]
8. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]

9. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 936–944. [CrossRef]
10. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. *arXiv* **2016**, arXiv:1506.02640.
11. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6517–6525. [CrossRef]
12. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
13. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004.10934.
14. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *arXiv* **2023**, arXiv:1706.03762.
15. Zhu, X.; Lyu, S.; Wang, X.; Zhao, Q. TPH-YOLOv5: Improved YOLOv5 Based on Transformer Prediction Head for Object Detection on Drone-captured Scenarios. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Montreal, BC, Canada, 11–17 October 2021; pp. 2778–2788. [CrossRef]
16. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. *arXiv* **2018**, arXiv:1807.06521.
17. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October 2021; pp. 9992–10002. [CrossRef]
18. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 11531–11539. [CrossRef]
19. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
20. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. In Proceedings of the Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016.
21. Li, C.; Li, L.; Jiang, H.; Weng, K.; Geng, Y.; Li, L.; Ke, Z.; Li, Q.; Cheng, M.; Nie, W.; et al. YOLOv6: A Single-Stage Object Detection Framework for Industrial Applications. *arXiv* **2022**, arXiv:2209.02976.
22. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 7464–7475. [CrossRef]
23. Varghese, R.; M., S. YOLOv8: A Novel Object Detection Algorithm with Enhanced Performance and Robustness. In Proceedings of the 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS), Chennai, India. 18–19 April 2024; pp. 1–6. [CrossRef]
24. Wang, C.Y.; Yeh, I.H.; Liao, H.Y.M. YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information. *arXiv* **2024**, arXiv:2402.13616.
25. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J.; Ding, G. YOLOv10: Real-Time End-to-End Object Detection. *arXiv* **2024**, arXiv:2405.14458
26. Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; Chen, J. DETRs Beat YOLOs on Real-time Object Detection. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; pp. 16965–16974. [CrossRef]
27. Li, X.; Duan, W.; Fu, X.; Lv, X. R-SABMNet: A YOLOv8-Based Model for Oriented SAR Ship Detection with Spatial Adaptive Aggregation. *Remote Sens.* **2025**, *17*, 551. [CrossRef]
28. Li, K.; Zheng, X.; Bi, J.; Zhang, G.; Cui, Y.; Lei, T. RMVAD-YOLO: A Robust Multi-View Aircraft Detection Model for Imbalanced and Similar Classes. *Remote Sens.* **2025**, *17*, 1001. [CrossRef]
29. Liu, S.; Shao, F.; Chu, W.; Dai, J.; Zhang, H. An Improved YOLOv8-Based Lightweight Attention Mechanism for Cross-Scale Feature Fusion. *Remote Sens.* **2025**, *17*, 1044. [CrossRef]
30. 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 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
31. Xie, X.; Cheng, G.; Wang, J.; Yao, X.; Han, J. Oriented R-CNN for Object Detection. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October 2021; pp. 3500–3509. [CrossRef]
32. Yang, X.; Yan, J.; Feng, Z.; He, T. R3Det: Refined Single-Stage Detector with Feature Refinement for Rotating Object. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 11–15 October 2021; Volume 35, pp. 3163–3171. [CrossRef]
33. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path Aggregation Network for Instance Segmentation. *arXiv* **2018**, arXiv:1803.01534.

34. Tan, M.; Pang, R.; Le, Q.V. EfficientDet: Scalable and Efficient Object Detection. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 10778–10787. [CrossRef]
35. Xu, X.; Jiang, Y.; Chen, W.; Huang, Y.; Zhang, Y.; Sun, X. DAMO-YOLO: A Report on Real-Time Object Detection Design. *arXiv* **2023**, arXiv:2211.15444.
36. Yang, Z.; Guan, Q.; Zhao, K.; Yang, J.; Xu, X.; Long, H.; Tang, Y. Multi-Branch Auxiliary Fusion YOLO with Re-parameterization Heterogeneous Convolutional for accurate object detection. *arXiv* **2024**, arXiv:2407.04381.
37. Li, H.; Li, J.; Wei, H.; Liu, Z.; Zhan, Z.; Ren, Q. Slim-neck by GSConv: A lightweight-design for real-time detector architectures. *J.-Real-Time Image Process.* **2024**, *21*, 62. [CrossRef]
38. Li, Y.; Hou, Q.; Zheng, Z.; Cheng, M.M.; Yang, J.; Li, X. Large Selective Kernel Network for Remote Sensing Object Detection. In Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; pp. 16748–16759. [CrossRef]
39. Cai, X.; Lai, Q.; Wang, Y.; Wang, W.; Sun, Z.; Yao, Y. Poly Kernel Inception Network for Remote Sensing Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; pp. 27706–27716.
40. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141. [CrossRef]
41. 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, Nashville, TN, USA, 20–25 June 2021.
42. Xu, W.; Wan, Y. ELA: Efficient Local Attention for Deep Convolutional Neural Networks. *arXiv* **2024**, arXiv:2403.01123.
43. Yang, G.; Lei, J.; Zhu, Z.; Cheng, S.; Feng, Z.; Liang, R. AFPN: Asymptotic Feature Pyramid Network for Object Detection. *arXiv* **2023**, arXiv:2306.15988.
44. Chen, Y.; Zhang, C.; Chen, B.; Huang, Y.; Sun, Y.; Wang, C.; Fu, X.; Dai, Y.; Qin, F.; Peng, Y.; et al. Accurate leukocyte detection based on deformable-DETR and multi-level feature fusion for aiding diagnosis of blood diseases. *Comput. Biol. Med.* **2024**, *170*, 107917. [CrossRef]
45. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Int. J. Comput. Vis.* **2019**, *128*, 336–359. [CrossRef]

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.