remote sensing

MDPI

Article

HSF-DETR: Hyper Scale Fusion Detection Transformer for Multi-Perspective UAV Object Detection

Yi Mao [1], Haowei Zhang [1], Rui Li [1], Feng Zhu [2,3], Rui Sun [2] and Pingping Ji [1,*]

[1] College of Information Science and Engineering, Key Laboratory of Maritime Intelligent Cyberspace Technology of Ministry of Education, Hohai University, Nanjing 210098, China; maoyi@hhu.edu.cn (Y.M.); 231623010079@hhu.edu.cn (H.Z.); lirui628@hhu.edu.cn (R.L.)
[2] College of Civil Aviation, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China; rui.sun@nuaa.edu.cn (R.S.)
[3] Nanjing Research Institute of Electronic Engineering, Nanjing 210007, China
* Correspondence: 20241619@hhu.edu.cn

Abstract: Unmanned aerial vehicle (UAV) imagery detection faces challenges in preserving small object features during multi-level downsampling, handling angle and altitude-dependent variations in aerial scenes, achieving accurate localization in dense environments, and performing real-time detection. To address these limitations, we propose HSF-DETR, a lightweight transformer-based detector specifically designed for UAV imagery. First, we design a hybrid progressive fusion network (HPFNet) as the backbone, which adaptively modulates receptive fields to capture multi-scale information while preserving fine-grained details critical for small object detection. Second, building upon features extracted by HPFNet, we develop MultiScaleNet, which enhances feature representation through dual-layer optimization and cross-domain feature learning, significantly improving the model's capability to handle complex aerial scenarios with diverse object orientations. Finally, to address spatial–semantic alignment challenges, we devise a position-aware align context and spatial tuning (PACST) module that ensures effective feature calibration through precise alignment and adaptive fusion across scales. This hierarchical architecture is complemented by our novel AdaptDist-IoU loss with dynamic weight allocation, which enhances localization accuracy, particularly in dense environments. Extensive experiments using standard detection metrics (mAP50 and mAP50:95) on the VisDrone2019 test dataset demonstrate that HSF-DETR achieves superior performance with 0.428 mAP50 (+5.4%) and 0.253 mAP50:95 (+4%) when compared with RT-DETR, while maintaining real-time inference (69.3 FPS) on an NVIDIA RTX 4090D GPU with only 15.24M parameters and 63.6 GFLOPs. Further validation across multiple public remote sensing datasets confirms the robust generalization capability of HSF-DETR in diverse aerial scenarios, offering a practical solution for resource-constrained UAV applications where both detection quality and processing speed are crucial.

Keywords: UAV imagery; object detection; hyper scale fusion detection transformer (HSF-DETR); real time; multi-domain feature fusion; hierarchical feature optimization

Academic Editor: Konstantinos Topouzelis

Received: 8 April 2025
Revised: 22 May 2025
Accepted: 3 June 2025
Published: 9 June 2025

Citation: Mao, Y.; Zhang, H.; Li, R.; Zhu, F.; Sun, R.; Ji, P. HSF-DETR: Hyper Scale Fusion Detection Transformer for Multi-Perspective UAV Object Detection. *Remote Sens.* **2025**, *17*, 1997. https://doi.org/10.3390/rs17121997

1. Introduction

With the convergence of computational innovation and aerial technology, the unmanned aerial vehicle (UAV) has evolved from being not only a simple aircraft but also into a sophisticated edge computing system. These aerial systems need to undertake the onboard processing of complex algorithms, enabling real-time decisions without constant

ground station communication. Such capabilities have propelled UAVs into critical applications spanning traffic management, disaster response coordination, defense reconnaissance, and intelligent surveillance networks [1–4]. While these applications demonstrate remarkable potential, they also reveal fundamental challenges in aerial vision systems, particularly when operating in densely populated regions where target identification demands both precision and speed. The unique operational context of urban monitoring presents a distinctive set of technical hurdles that conventional computer vision frameworks struggle to overcome effectively.

In complex urban environments, UAVs have been equipped with high-resolution cameras and embedded detection systems [5,6] to enhance their capability for detecting subtle changes and small objects. Although UAVs offer reduced operational expenses and accelerated data acquisition compared with satellite remote sensing [7], the inherent characteristics of aerial imagery present substantial obstacles for conventional object detectors. These impediments manifest as feature representation issues, where inconsistent distributions are generated by motion blur from high-speed flight and multi-angle imaging. Due to variations in flight altitude and perspective, identical objects may appear at diverse scales, whereby traditional networks struggle to process such variations, resulting in inconsistent detection performance. To overcome this altitude and angle-dependent variation challenge, our proposed MultiScaleNet employs a dual-layer optimization approach combined with cross-domain feature learning, enabling robust feature representation across diverse viewing angles and scales in aerial scenarios. Conventional frameworks predominantly employ fixed-size receptive fields, failing to simultaneously attend to objects of varying dimensions and consequently neglecting small object features. Despite the integration of high-precision hardware, embedded detection models remain inadequately optimized for preserving critical details of small objects during feature extraction, leading to significant information degradation. To address this critical challenge of small object feature preservation, we propose a hybrid progressive fusion network (HPFNet), which employs progressive kernel expansion and partial convolutions to specifically retain fine-grained details of small objects throughout the feature extraction process. In urban scenarios, building occlusions and background complexity further compromise detection systems' ability to differentiate foreground targets from background elements, particularly diminishing localization precision for minimal-area objects [8–10]. We specifically address this dense environment localization challenge through our position-aware align context and spatial tuning (PACST) module, which ensures effective feature calibration through precise alignment and adaptive fusion across scales, significantly enhancing localization accuracy in cluttered urban environments. While deep learning methodologies currently dominate this domain [11], the rapid advancement of UAV technology has generated an urgent requirement for computational frameworks that concurrently address fixed receptive field limitations, multi-scale feature fusion challenges, detail preservation, and complex background interference, while maintaining both high accuracy and real-time performance.

Several categories of object detection approaches have emerged from the evolution of deep learning. Traditional methods primarily fall into two categories based on convolutional neural networks: one-stage and two-stage detectors. Without explicit region proposal generation, direct object detection and location prediction are performed by one-stage detectors like RetinaNet [12] and the You Only Look Once series [13–15]; conversely, candidate regions before classification and location regression are first generated by two-stage detectors such as Faster R-CNN [16]. While these CNN-based approaches have achieved remarkable success in general object detection, they face significant challenges in UAV-based scenarios, particularly in preserving small object features, handling multi-scale variations, and achieving precise localization in dense environments.

More recently, transformer-based detectors have emerged as a promising direction. DETR [17] marked a paradigm shift by pioneering transformers and self-attention mechanisms [18] into object detection, reformulating it as a set prediction problem and eliminating traditional post-processing steps. Building upon this foundation, RT-DETR [19] and other variants have further improved efficiency and accuracy. However, most existing detectors, whether CNN-based or transformer-based, are primarily designed for general object detection scenarios and struggle to address the unique challenges presented by aerial imagery. The distinctive characteristics of UAV-captured data—including extreme scale variations, dense object distributions, and complex viewpoint changes—demand specialized architectural designs that can effectively preserve critical spatial details while maintaining semantic understanding across multiple scales.

Despite recent advances in object detection, UAV-perspective scenarios present unique challenges that existing models struggle to address. A fundamental challenge lies in preserving small object features during feature extraction, as conventional backbone networks with fixed receptive fields often fail to capture critical details [20]. While attention mechanisms have shown promise in global modeling, they prove less effective in early backbone stages [21,22], and traditional CNN feature maps contain substantial redundancy that increases computational overhead. To address this challenge, we propose a hybrid progressive fusion network (HPFNet), which is a multi-stage optimization framework specifically designed for UAV detection. HPFNet employs multi-scale partial convolution block (MPCBlock) in shallow layers to preserve fine-grained details during downsampling, while integrating multi-scale partial attention fusion block (MSPAFBlock) in deeper layers to enhance global context modeling while maintaining computational efficiency [23].

Another critical challenge in UAV object detection is maintaining consistent feature representation across varying scale structures in aerial imagery, where objects of identical categories may appear at dramatically different scales due to altitude variations [24,25]. Our proposed MultiScaleNet systematically addresses this challenge by integrating high-resolution P2 features enhanced through space-to-depth convolution (SPDConv) [26] for efficient spatial information reorganization, while implementing cascaded shuffle convolution tuning (CSCT) at the P3 layer. The multi-domain fusion stage introduces a Domain-ScaleKernel structure that synergizes spatial and frequency domain modeling, optimized through a cross-stage partial (CSP) structure [27] to ensure precise feature extraction while maintaining computational efficiency.

The balance between semantic richness and spatial precision presents a third significant hurdle in UAV imagery analysis. Research indicates that deeper network levels excel in semantic understanding yet struggle with context integration, while shallow layers capture fine spatial details but introduce misalignment issues [28,29]. The effective alignment of these complementary features significantly impacts model performance, particularly in complex UAV imagery [30]. To address these challenges, particularly for drone-view scenarios, we propose a position-aware align context and spatial tuning (PACST) framework. This hierarchical enhancement approach integrates multi-contextual cross fusion (MCCF) for enriching semantic representations and dynamic feature alignment and fusion (DFAF) for orchestrating scale-adaptive fusion with a DRB structure [31], achieving an optimal balance between semantic understanding and spatial precision, which is particularly beneficial for UAV imagery.

In summary, our main contributions are as follows:

1. We design HPFNet, a lightweight backbone network for UAV-based small object detection, featuring innovative MPCBlock and MSPAFBlock designs for efficient multi-scale feature extraction. In shallow layers, the network preserves small object features through partial convolutions and progressive kernel expansion. Meanwhile,

in deeper layers, it enhances global modeling capability with single-head self-attention mechanisms. This hierarchical design significantly improves feature extraction efficiency and accuracy while maintaining the computational efficiency suitable for real-time applications.

2. We propose MultiScaleNet neck architecture, incorporating P2 layer features optimized through SPDConv and CSCT modules. The DomainScaleKernel structure enables the collaborative modeling of spatial and frequency domains, integrating detail enhancement, large-scale receptive field expansion, and frequency domain optimization through frequency-spatial channel attention (FSCA) and feature guidance module (FGM) for cross-domain feature modulation. The innovative semantic compression distillation (SCD) and multi-scale adaptive fusion (MSAF) modules effectively address semantic loss during feature compression, while CSP structure optimization maintains computational efficiency.
3. We develop the PACST framework to address feature representation challenges through hierarchical enhancement. The MCCF module enriches deep features through multi-scale context modeling and attention-based recalibration, while DFAF orchestrates scale-adaptive fusion between shallow and context-enriched deep features, particularly enhancing small object detection capability, while maintaining effective representation across scales.
4. We introduce the AdaptDist-IoU loss function, which combines linear interval mapping for dynamic weight allocation with geometric vertex distance constraints, enhancing detection capabilities for challenging objects. Based on these innovations, we present a hyper scale fusion detection transformer (HSF-DETR), which demonstrates robust performance on the VisDrone2019 dataset and shows significant accuracy improvements over the baseline RT-DETR model while maintaining real-time performance and computational efficiency. Furthermore, to validate the generalization capability of our approach, we conduct extensive experiments on two additional remote sensing datasets, AI-TOD-v2 and DOTA-v1.5, demonstrating the broad applicability of our method across diverse aerial imaging scenarios with varying object scales, densities, and orientations.

The remainder of this paper is organized as follows: Section 2 provides a comprehensive review of related work, Section 3 details our proposed methodology, Section 4 presents experimental results, Section 5 discusses the model proposed in this paper and Section 6 concludes the paper.

2. Related Works

2.1. Transformer-Based Detection Framework

DETR, proposed by Carion et al. [17], revolutionized object detection by reformulating it as a set prediction problem, eliminating the need for hand-crafted components like RPN and NMS. However, DETR struggled with slow convergence and poor performance in small object detection. Subsequent improvements addressed these limitations: Deformable DETR [32] enhanced convergence speed through localized attention mechanisms, while UP-DETR [33] improved accuracy through unsupervised pre-training of transformer parameters. Efficient DETR [34] strengthened decoder queries by selecting top-K positions from encoder predictions, and Meng et al. [35] accelerated training convergence through improved decoder cross-attention. However, these improvements often increased computational complexity due to stacked encoders and decoders, compromising real-time performance. RT-DETR [19] addressed these challenges by leveraging the observation that the final CNN feature layer contains the most global information. By applying the encoder module solely to the final layer and fusing it with shallow multi-scale features, RT-DETR

significantly improved inference speed without sacrificing accuracy. Its single-layer encoder design, combined with IoU-aware query selection strategy, maintains end-to-end training advantages while substantially reducing computational complexity, particularly excelling in multi-scale detection tasks through its hybrid encoder design. These advantages make RT-DETR an ideal baseline model for our work.

2.2. Object Detection in UAV Imagery

The inherent flight characteristics of UAVs introduce complex challenges in captured imagery, including intricate backgrounds, multiple angles, varying platform heights, diverse weather conditions, lighting variations, motion blur, and object occlusion. While these factors significantly impact feature extraction and overall image quality, they also contribute to more diverse image datasets for research and development.

Recent works have addressed these challenges through specialized approaches: TPH-YOLOv5 [36] integrates transformer concepts with YOLOv5 architecture, enhancing detection performance through self-attention mechanisms and CBAM in dense object scenarios. ETAM [37] enriches small object features through magnifying glass and quadruple attention modules. Info-FPN [38] addresses channel information loss and feature misalignment through PSM, FAM, and SEM modules. MFFSODNet [39] introduces hierarchical feature extraction and adaptive fusion for UAV imagery, while DAGN [40] combines lightweight depth-wise separable convolutions with attention mechanisms. HRDNet [41] focuses on high-resolution feature optimization through hierarchical processing and adaptive enhancement strategies.

Despite these significant advancements, critical challenges persist in UAV object detection: computational overhead conflicts with platform resource limitations, multi-scale feature fusion strategies remain insufficient for complex small object detection scenarios, and effective feature alignment and semantic information utilization require further improvement. These persistent limitations motivate our proposed enhancements to the detection framework.

3. Proposed Method

3.1. Overall Framework

Detecting small objects in high-density multi-object and complex background scenarios remains a significant challenge. Traditional convolutional backbones with fixed receptive fields struggle to capture fine-grained local details, while existing feature fusion methods often exhibit redundancy and information loss when processing multi-scale data. Furthermore, current IoU metrics lack the flexibility needed to manage targets of varying shapes and scales. These limitations compromise detection performance and restrict model generalization in high-precision UAV environments. Building upon these observations, we propose hyper scale fusion detection transformer, a systematic improvement strategy that optimizes the detection pipeline through four key enhancements: HPFNet to extract features, with detailed discussions in Section 3.2; MultiScaleNet for cross-domain fusion, with detailed discussions in Section 3.3; PRCST for context-aware alignment in Section 3.4; and AdaptDist-IoU in Section 3.5, as depicted in Figure 1.

Our HSF-DETR framework adopts an end-to-end object detection approach. First, the lightweight HPFNet backbone extracts multi-level feature representations {P2, P3, P4, P5} from input UAV images, where the shallow MPCBlock preserves small object features through partial convolutions and progressive kernel expansion, while the deeper MSPAFBlock enhances global modeling through single-head self-attention. Second, these features are fused through MultiScaleNet, where P2 features are optimized via SPDConv, P3 applies CSCT, and the DomainScaleKernel enables collaborative modeling of spatial and frequency

domains. Subsequently, PACST enriches deep features through MCCF and orchestrates scale-adaptive fusion between shallow and context-enriched features via DFAF. Finally, the model initializes object queries through IoU-aware selection, progressively refining them through an optimized decoder to produce bounding boxes with confidence scores, enhanced by AdaptDist-IoU loss, which improves accuracy through dynamically adjusted penalty allocation. The central innovations of our approach—hierarchical multi-stage feature optimization, domain-collaborative feature fusion, and context-spatial adaptive alignment—significantly enhance detection performance for multi-scale objects in complex UAV imagery while maintaining computational efficiency.

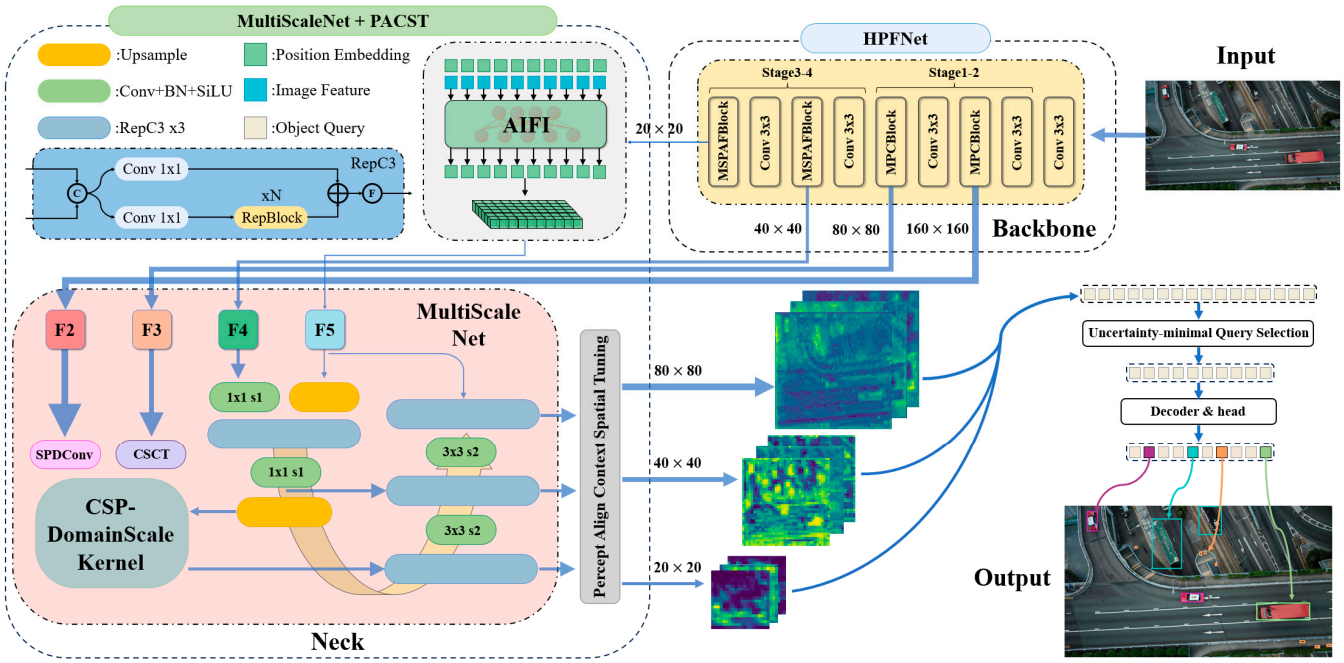


Figure 1. Overview of the proposed HSF-DETR framework. Our main contributions include the following: (1) HPFNet, which integrates MPCBlock and MSPAFBlock to efficiently extract multi-scale local features and model global semantic information using partial convolutions, progressive kernel expansion, and single-head self-attention; (2) MultiScaleNet, which employs SPDConv and CSCT modules to refine small object features and utilizes DomainScaleKernel with a CSP design for adaptive multi-domain and multi-scale feature aggregation; and (3) PACST, which combines MCCF and DFAF components to enhance the contextual recognition of large objects and to dynamically align small object features while suppressing background interference across multiple scales.

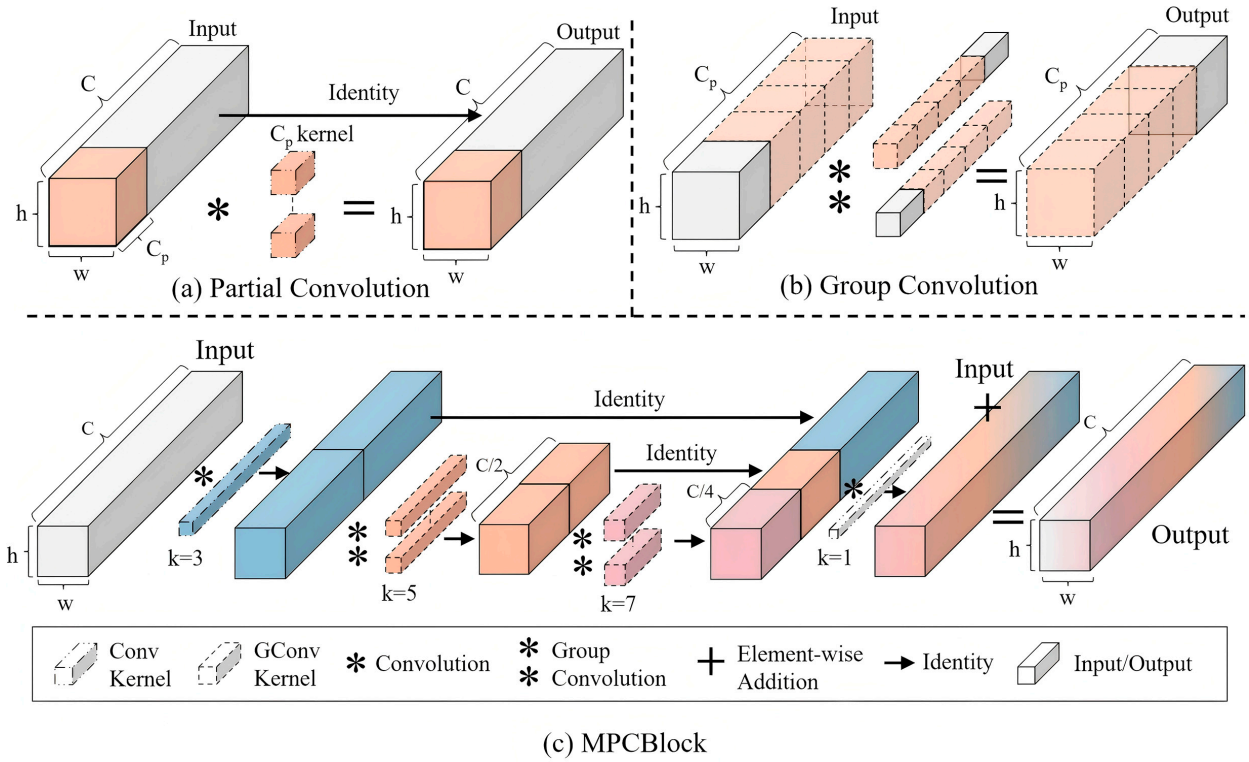
3.2. Hierarchical Multi-Scale Backbone, HPFNet

Detecting small objects in high-density multi-object and complex background scenarios remains challenging. Traditional convolutional backbones like ResNet-18 exhibit limitations in multi-scale object detection due to their fixed receptive fields and high computational cost, leading to suboptimal performance in capturing both global semantic information and fine local details.

3.2.1. MPCBlock

Our investigation of network architecture efficiency led us to examine how computational costs scale with detection performance. While exploring this trade-off, we found particular value in two complementary approaches. Recent studies have introduced two efficient computational paradigms: Partial convolution (PConv) [42] and group convolution (GConv) [43]. PConv reduces computational complexity by performing convolutions on selected channel subsets while maintaining lightweight operations on remaining channels,

though this may compromise feature representation in unused channels (Figure 2a). GConv enhances efficiency by partitioning channels into groups for independent processing, but limited inter-group information exchange may restrict cross-channel feature integration (Figure 2b). These practical constraints motivated our hybrid design approach.



(c) MPCBlock

Figure 2. (**a**) Operational mechanisms of PConv. (**b**) Operational mechanisms of GConv. (**c**) The architectural description of MPCBlock, integrated into stages 1–2 of the HPFNet, combines partial and group convolution mechanisms.

The framework of multi-scale partial convolution block (MPCBlock) is shown in Figure 2c. MPCBlock is a combined partial convolution with group convolution strategies to enhance its multi-scale feature extraction and fusion ability while maintaining its light weight. This novel architecture enables stronger cross-scale semantic interaction and expanded effective receptive fields, capturing rich global and local feature information without significantly increasing computational overhead.

Given an input feature map $X \in \mathbb{R}^{C\times H\times W}$, the feature processing in MPCBlock can be formulated as Equation (1):

$$F_{output} = X + Conv_{1\times1}(Concat[F_{n,1}, \{F_{i,2}\}_{i=1}^{n-1}]) \;\Big|\; (F_{i,1}, F_{i,2}) = Conv_{k_0+2(i-1)}(F_{i-1,1}) \quad (1)$$

where $F_{i-1,1}$ denotes the processed channels from stage $i-1$ and $F_{i,1}$ and $F_{i,2}$ represent the convolution output features at stage i. The kernel size k_i expands progressively across stages, with k_0 as the initial kernel size and n denoting the progression steps. After n times feature operations, multi-scale features $F_{n,1}$ are concatenated with unprocessed channel features ($F_{1,2}$, $F_{2,2}$, etc.) along the channel dimension. The final output integrates channel information through a 1×1 convolution with residual connection.

MPCBlock channel partitioning significantly reduces computational complexity compared with traditional full-channel convolutions. While a standard convolution with kernel size K and channel dimension C requires $K^2 \cdot C \cdot C$ parameters, MPCBlock reduces this to

$K_1^2 \cdot C \cdot C + K_2^2 \cdot \frac{C}{2} \cdot \frac{C}{2} + K_3^2 \cdot \frac{C}{4} \cdot \frac{C}{4}$ (K_i denotes the kernel size employed at the i-th layer), enhancing inference efficiency while maintaining feature diversity.

Despite enhancing the multi-scale modeling capabilities of the backbone, MPCBlock exhibits several inherent limitations. The integration of multi-layer partial convolutions and expanded kernels, while strengthening local feature representations, introduces substantial computational overhead that impacts inference efficiency in deeper architectures. Moreover, the PConv leaves a portion of channels unoptimized, potentially compromising feature utilization and overall representational efficiency.

3.2.2. MSPAFBlock

To enhance global information modeling and cross-channel feature optimization, we propose a multi-scale partial attention fusion block (MSPAFBlock), which integrates single-head self-attention [23] with multi-scale partial convolutions for improved feature representation in complex detection scenarios.

MSPAFBlock achieves efficient feature modeling and multi-scale information fusion through the synergistic integration of depth-wise convolutions, single-head self-attention mechanism, and MPC structures, as illustrated in Figure 3.

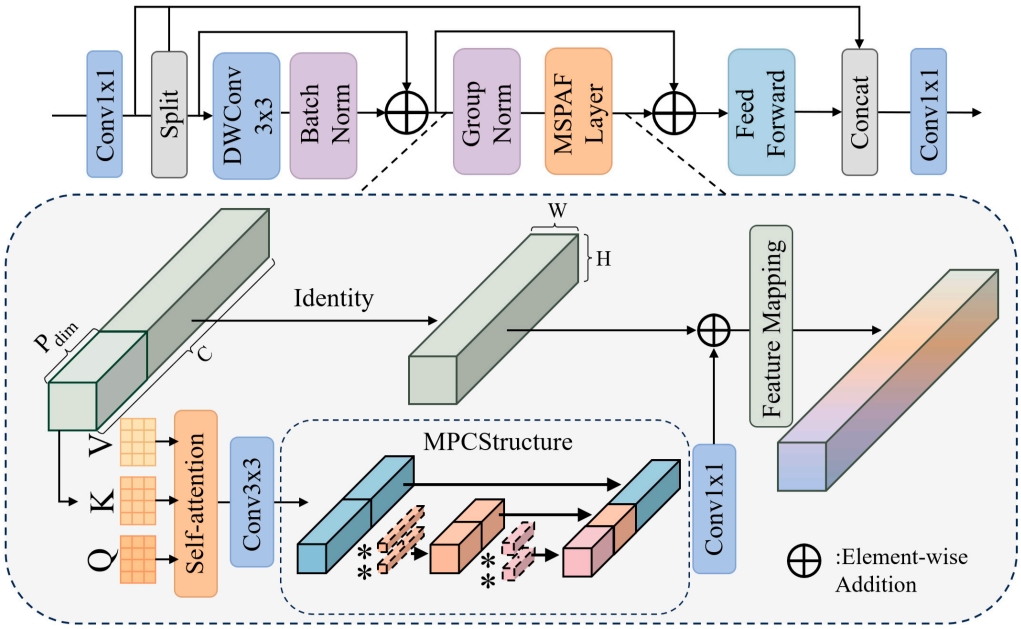


Figure 3. MSPAFBlock, integrated within stages 3–4 of the HPFNet, incorporates single-head self-attention mechanisms to enhance global feature representation capabilities. See text for details. ** (vertically aligned) indicates group convolution.

Given an input feature $X \in \mathbb{R}^{B\times C\times H\times W}$, MSPAFLayer partitions input channels into two branches: $X_1 \in \mathbb{R}^{B\times p_{dim}\times H\times W}$ for global attention modeling and $X_2 \in \mathbb{R}^{B\times (C-p_{dim})C\times H\times W}$ for feature preservation. The branch X_1 undergoes group normalization (GN) followed by convolution operations to generate query, key, and value, where $Q, K \in \mathbb{R}^{B\times qk_{tim}\times HW}$ and $V \in \mathbb{R}^{B\times p_{dim}\times HW}$. Global features are computed through attention weights, as follows:

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{qk_{dim}}}\right)V \tag{2}$$

The attention output X_{attn} guides the subsequent multi-scale feature fusion process through a hierarchical structure similar to MPCBlock. The feature integration pathway employs a series of group convolutions with progressively expanding kernel sizes, where

each layer partitions its output into two channel groups. This progressive feature refinement mechanism, combined with the attention-guided initial features, enables comprehensive multi-scale representation learning. After processing through L convolution layers with each layer applying group convolution (where the group number equals the channel number for computational efficiency), the complete feature transformation of MSPAFBlock can be formulated as per Equation (3):

$$Y = Conv_{1\times1}(SiLU(Concat[X_2, \{X_{i,2}\}_{i=1}^{L}, X_{L+1,1}])) \;\Big|\; X_{i+1} = GConv_{k_i}(X_i) \tag{3}$$

MSPAFBlock enhances multi-scale feature extraction through an organic fusion of MPCBlock and single-head attention mechanisms. The single-head design reduces computational complexity while maintaining global semantic modeling capabilities. The incorporation of residual connections and efficient feature fusion strategies optimizes feature propagation and gradient flow, enhancing model stability and generalization.

HPFNet: Building upon these innovations, we propose a hybrid progressive fusion network as a novel backbone architecture. The network comprises four stages with strategically distributed feature extraction modules: MPCBlock in stages 1–2 for efficient local feature modeling and MSPAFBlock in stages 3–4 for explicit global semantic modeling and cross-channel interaction optimization. This hierarchical design enables comprehensive feature representation from local details to global context while maintaining computational efficiency.

3.3. MultiScaleNet for All Scale Fusions

High-resolution aerial imagery analysis reveals a critical tension between semantic comprehension and detail retention when detecting small objects. While RT-DETR's transformer architecture showed promise, its cross-scale feature fusion (CCFF) component struggled with objects under 32×32 pixels in UAV datasets with extreme scale variations. This limitation stems from the failure of CCFF's pyramidal design to balance semantic understanding with spatial precision, and its single-domain representation inadequately characterizing diverse scales in aerial imagery.

To address these challenges, we propose MultiScaleNet, a novel architecture that achieves precise feature extraction through multi-domain learning. Our framework resolves conflicts between global context modeling and local detail preservation while maintaining computational efficiency through a hierarchical structure that enables adaptive feature learning across scales and domains.

3.3.1. Feature Refinement and Cross-Scale Enhancement

Small object detection in complex scenes requires the preservation of the critical boundary and texture information that is lost during traditional downsampling operations. We integrate high-resolution P2 layer features into our detection pipeline using SPDConv [26], which maps spatial details to channel dimensions while requiring fewer computational resources than standard approaches.

To further enhance feature representation, we introduce the cascaded shuffle convolution tuning (CSCT) module, which integrates dynamic convolution with channel shuffling. This module partitions input features from the P3 layer into static and dynamic components, processing the latter through our dynamic channel reorganization block (DCRB). The enhanced output feature, shown in Equation (4), is our adaptive dual-stream feature enhancement architecture, which enhances small object representation through stable

information flow between high and low-level features without introducing significant computational overhead.

$$X_{out} = Conv^{2}_{1\times1}\Big(Concat\Big(Shuffle\Big(GConv(X_{dyn})\Big), X_{sta}\Big)\Big) + X_{res}. \quad (4)$$

3.3.2. Multi-Domain Feature Fusion

We propose the DomainScaleKernel module (Figure 4a) with a multi-branch architecture that addresses scale-domain heterogeneity through four complementary pathways, as follows: a detail branch using 1×1 depth-wise convolutions [44] for fine-grained information extraction, a wide-field branch employing strip-shaped (1×31, 31×1) [45] and large-scale (31×31) kernels for directional and comprehensive spatial modeling, a bridge branch with a 5×5 convolution for balanced medium-scale representation, and a cross-domain branch that integrates frequency domain information via FFT with spatial features through spectrum-guided attention and adaptive fusion mechanisms.

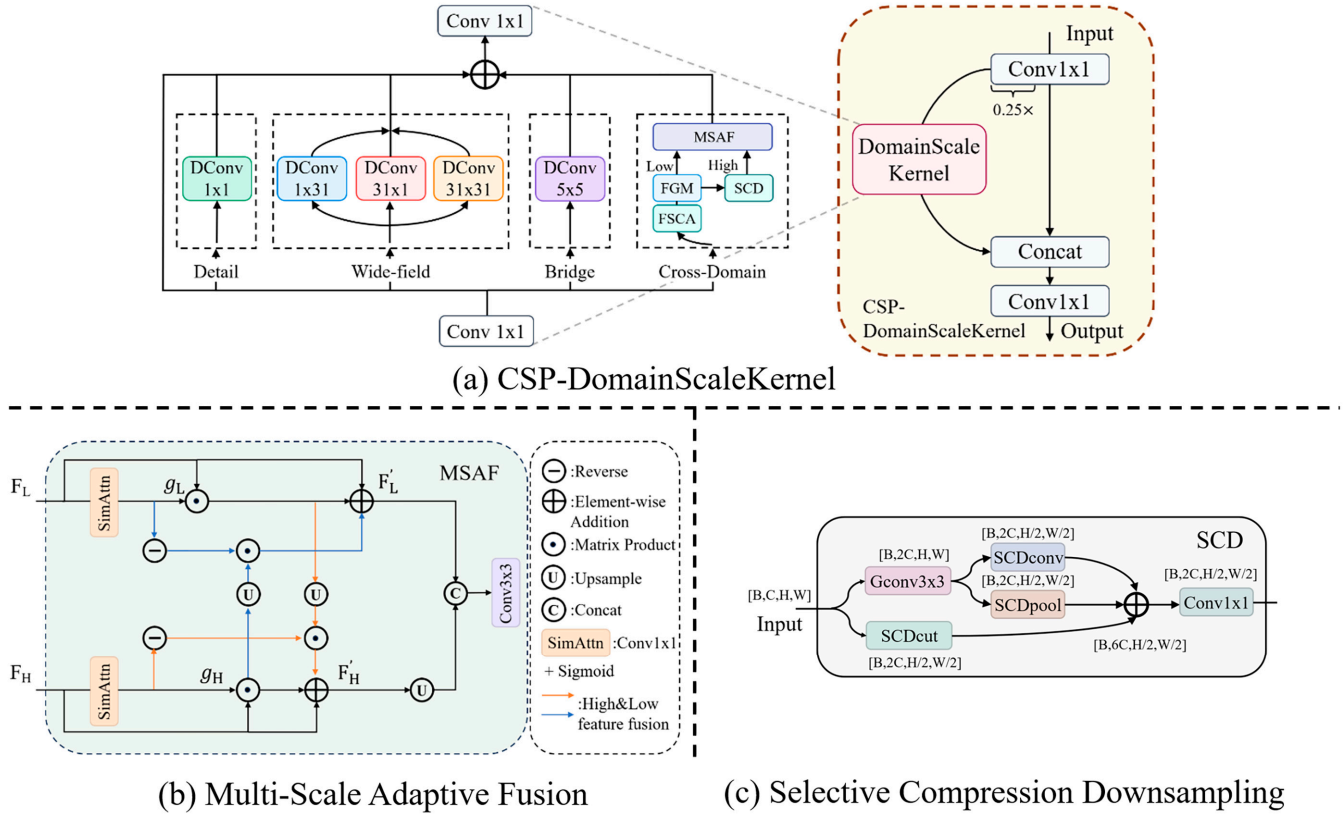


Figure 4. (**a**) Schematic illustration of the DomainScaleKernel architecture. (**b**) Block diagram of the MSAF module, depicting component interactions and symbolic representations. (**c**) Architecture of the proposed SCD module for efficient feature downsampling, transforming input features from [B,C,H,W] to [B,2C,H/2,W/2].

Frequency Extraction Module: We enhance small object detection by leveraging complementary information from both frequency and spatial domains through frequency-spatial channel attention (FSCA) and frequency guided modulation (FGM).

FSCA enhances channel-wise feature representation through frequency domain filtering, transforming features via fast Fourier transform (*FFT*) to capture spectral characteristics missed in purely spatial approaches. FGM then refines this representation through parallel processing paths with learnable parameters (α, β) that adaptively balance original

spatial features and frequency-modulated features. The complete frequency–spatial feature extraction and modulation process can be formulated as per Equation (5):

$$Y = \beta \cdot X + \alpha \cdot IFFT(W_c \odot FFT(X)) \tag{5}$$

where $\odot$ represents the Hadamard product, and α, β are learnable parameters that adaptively balance the contributions of frequency-modulated features and spatial features. This formulation captures the dual-domain feature enhancement process while maintaining computational efficiency, enabling more robust small object detection through complementary feature representation.

Cross-domain Fusion Module: The efficient integration of spatial and frequency domain features is realized through two complementary mechanisms: the selective compression downsampling (SCD) and multi-scale adaptive fusion (MSAF).

SCD (Figure 4c) preserves semantic information during downsampling using a triple-path architecture: depth-wise convolution with GELU activation enhances local features, a slice-based path maintains semantic information through channel expansion, and max-pooling emphasizes salient regions. These complementary features are refined through 1×1 convolution.

The multi-scale adaptive fusion (MSAF) module addresses feature inconsistency through dynamic integration of multi-resolution features. As illustrated in Figure 4b, MSAF operates on dual-scale inputs, as follows: high-resolution features $F_H \in \mathbb{R}^{C \times H \times W}$ and low-resolution features $F_L \in \mathbb{R}^{C \times H/2 \times W/2}$. The module first performs channel alignment through simple attention operations with 1×1 convolutions to get $\hat{F}_H$ and $\hat{F}_L$. The channel-wise feature attention weights g_H, g_L are obtained through sigmoid transformation of $\hat{F}_H$ and $\hat{F}_L$, facilitating adaptive feature modulation.

A bi-directional fusion strategy facilitates comprehensive cross-scale feature interaction through the parallel self-enhancement and complementary information exchange given by Equations (6) and (7), where both high and low-resolution features undergo self-enhancement through the respective attention weights g_H, g_L while incorporating complementary information from the other scale. Final feature integration is achieved through Equation (8).

$$F'_H = \hat{F}_H + g_H \cdot \hat{F}_H + (1 - g_H) \cdot Upsample(g_L \cdot F'_L) \tag{6}$$

$$F'_L = \hat{F}_L + g_L \cdot \hat{F}_L + (1 - g_L) \cdot Upsample(g_H \cdot F'_H) \tag{7}$$

$$F_{out} = Conv_{3\times3}(Concat(Upsample(F'_H), F'_L)) \tag{8}$$

The MSAF module achieves adaptive fusion of shallow detail features and deep semantic features through its interaction mechanism. Compared with conventional feature fusion methods, this module significantly reduces information redundancy and effectively resolves feature conflicts.

3.3.3. CSP-DomainScaleKernel

We enhance our architecture by incorporating cross stage partial (CSP) [27] networks, partitioning input features into two streams after initial 1×1 convolution: a primary path processing 25% of channels through DomainScaleKernel, and a bypass path preserving contextual information in the remaining channels. This dual-path structure optimizes computational resources while strengthening multi-domain representation capabilities, particularly benefiting small object detection in challenging scenarios.

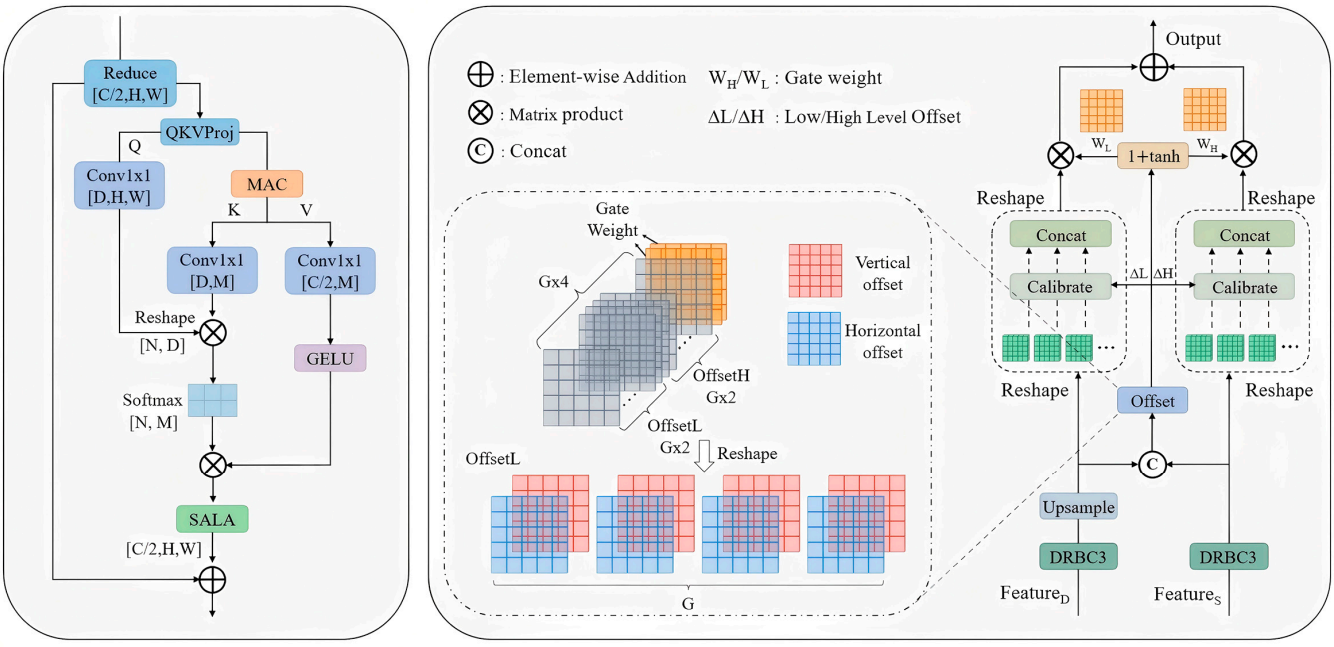
3.4. PACST for Context-Aware Alignment

Object detection architectures face fundamental constraints in feature representation effectiveness. At deeper network levels, extracted features excel in semantic richness yet struggle with context integration and often fail to preserve small object characteristics. Conversely, shallow layer representations capture fine spatial details but introduce misalignment issues and susceptibility to background noise. Such representation deficiencies become particularly pronounced in challenging detection scenarios, where precise feature characterization is crucial.

To address these challenges, particularly for drone-view scenarios, we propose Perceptalign align context and spatial tuning (PACST), a hierarchical feature enhancement framework comprising two key components: Multi-contextual cross fusion (MCCF) for deep feature enrichment and dynamic feature alignment and fusion (DFAF) for scale-adaptive fusion and spatial calibration. While highly effective for most scenarios, PACST shows diminishing returns in extremely dense scenes (>50 objects per image), with stronger improvements observed in sparse environments.

3.4.1. Deep Feature Context Refinement

The MCCF module overcomes inadequate context matching and feature representation challenges through comprehensive context modeling. Utilizing dynamic semantic context modeling, it integrates multi-scale feature enhancement with local attention recalibration, as illustrated in Figure 5a.



(a) Multi-Contextual Cross Fusion (b) Dynamic Feature Alignment and Fusion

Figure 5. Overview of the proposed PACST framework. (**a**) Architecture of the dynamic feature alignment and fusion (DFAF) module for spatial optimization in shallow features. (**b**) Architecture of the multi-contextual cross fusion (MCCF) module for semantic modeling in deep features. The ellipsis (...) indicates multiple similar components of the same type.

Given an input feature map $F \in \mathbb{R}^{C\times H\times W}$, MCCF first applies dimension reduction through a 3×3 convolution layer to generate intermediate features $F_r \in \mathbb{R}^{C_r\times H\times W}$ ($C_r = C/2$), which helps balance computational efficiency and representation capacity. In pursuit of effective context modeling, our multi-scale aggregated context (MAC) module

adaptively captures contextual information at different granularities. MAC generates context matrices through aspect ratio-aware pooling operations across four distinct grid sizes, g_1, g_2, g_3, g_4, corresponding to varying levels of context granularity. For each grid size, an adaptive pooling operation, F^i_{pool}, represents the pooled features at grid size g_i, and the pooling output dimensions are adjusted according to the input aspect ratio $ar = \frac{W}{H}$ to maintain spatial consistency. These four differently pooled representations are then concatenated along the spatial dimension to form a comprehensive multi-scale context representation, as follows:

$$F_{MAC} = Concat[\{F^{g_i}_{pool}(F_r, (g_i, max(1, round(ar \cdot g_i))))\}^4_{i=1}] \quad (9)$$

Building upon these multi-scale features, our attention mechanism establishes semantic correlations through query-key-value interactions. Query features $Q \in \mathbb{R}^{N \times d}$ $(N = H \times W, d = 32)$ emerge from the reduced features F_r, complemented by key $K \in \mathbb{R}^d \times M$ and value $V \in \mathbb{R}^{C_r \times M}$ matrices derived from MAC-processed features. The complete context modeling and feature enhancement process can be expressed as Equation (10):

$$F_{out} = \alpha \cdot F_r + \beta \cdot [(1 + tanh(f_{1\times1}(f_{3\times3}(F_c)))) \odot F_c] \quad (10)$$

where F_c represents the context-aware feature obtained through cross-attention, $\odot$ denotes the Hadamard product, and γ is a learnable scaling factor that dynamically modulates attention intensity. The spatial-aware local attention (SALA) mechanism further refines local representations, while learnable parameters α and β orchestrate the dynamic balance between original features and context-enhanced representations.

3.4.2. Cross-Scale Feature Calibration Fusion

DFAF achieves effective cross-level feature fusion between shallow layers and context-enhanced deep features (P5 processed by MCCF). This design addresses two major challenges: spatial misalignment from different receptive fields and inconsistent feature distributions across semantic levels. The architectural design is depicted in Figure 5b.

Serving as the distribution harmonization component, dilated reparameterized block (DRB) [31] enhances feature representation with an emphasis on small objects while maintaining effective large object detection through complementary convolution branches. Operating beyond conventional fixed receptive fields, this complementary design is formulated as follows:

$$F_{out} = F_{bn}(F_{conv}(x)) + \sum_{k,r} F^{k,r}_{bn}(F^{k,r}_{conv}(x)) \quad (11)$$

where F_{conv} and F_{bn} represent the standard convolution and batch normalization operations in the main branch, respectively, while $F^{k,r}_{conv}$ and $F^{k,r}_{bn}$ respectively denote dilated convolutions with kernel size k and dilation rate r. This formulation enables the module to simultaneously capture fine-grained details and broader contextual information.

To address spatial misalignment, we design an adaptive calibration mechanism that precisely aligns features through learnable spatial transformations. A ConvOffset network processes concatenated features to predict spatial offsets and fusion weights. Group-specific transformation parameters are determined by Equation (12), as follows:

$$(\theta_S, \theta_D, W_S, W_D) = f_{ConvOffset}(Concat(F_S, Upsample(F_D))) \quad (12)$$

where θ_D, θ_S represents the offset corresponding to coarse/fine-grained feature maps and W_S and W_D are adaptive fusion weights generated through learnable modulation. The function $f_{ConvOffset}$ consists of a sequence of convolutional operations to generate the offset fields and attention weights from the concatenated features. Based on these predicted offsets, we construct a normalized sampling coordinate system with a base grid spanning

from -1 to 1 in both dimensions. This coordinate system serves as the reference for our spatial transformation, which is performed through a deformable sampling operation shown by Equation (13):

$$F'_S = G(F_S, grid + \theta_S / N), \quad F'_D = G(F_D, grid + \theta_D / N) \tag{13}$$

where G denotes the bilinear sampling operation, and N is the feature map size used for normalization. This formulation enables precise spatial alignment between features from different network levels.

After spatial calibration, for the obtained calibrated features F'_S and F'_D, we employ a residual design coupled with an adaptive weighting mechanism $1 + tanh(\gamma \cdot W)$ to ensure stable and effective feature integration. The complete fusion process combines the spatially aligned features using learnable weights generated through a tanh-based activation mechanism, where γ is also a learnable parameter. This approach ensures that, during initial training phases, when tanh outputs approach zero, the module defaults to simple feature addition, maintaining baseline performance. As training progresses, the module evolves to adopt more sophisticated feature calibration and fusion strategies, enabling the optimal integration of complementary information from different network levels.

Through this hierarchical design, DFAF effectively bridges the gap between semantically rich deep features and spatially precise shallow features, significantly improving detection performance for small objects while maintaining effective representation of objects across all scales.

3.5. The AdaptDist-IoU Loss Function

In RT-DETR, the loss function integrates L1 loss, VariFocal loss (VFL) [46], and generalized IoU (GIoU) [47], where VFL handles classification while L1 and GIoU govern localization. Traditional localization losses, however, exhibit limitations in scenarios involving small objects, complex backgrounds, and irregular shapes.

Focaler-IoU [48] adopts a linear interval mapping strategy to dynamically adjust IoU weight distribution, enabling focused attention on samples within specific IoU ranges. MPDIoU [49] leverages geometric properties by directly minimizing vertex distances between predicted and ground-truth boxes. It penalizes Euclidean distances of top-left and bottom-right coordinates, which is particularly effective for irregular-shaped objects.

In this work, we leverage a novel *IoU* mechanism, *AdaptDist-IoU*, which integrates dynamic interval sensitivity with geometric vertex optimization. The loss function is defined as Equations (14) and (15):

$$AdaptDist\text{-}IoU = \frac{IoU - d}{u - d} - \frac{d_1}{adhw} - \frac{d_2}{adhw} \tag{14}$$

$$L_{AdaptDist\text{-}IoU} = 1 - AdaptDist\text{-}IoU \tag{15}$$

Unlike the original MPDIoU, which normalizes distance penalties using input image dimensions, we introduce a hyperparameter, adhw, to modulate the penalty weights dynamically. This modification provides flexible control over distance penalties across different detection scenarios, with $adhw = 1$ prioritizing IoU optimization while maintaining balanced positional constraints. The adaptive normalization coefficient enables scenario-specific optimization strategies without over-emphasizing bounding box displacement.

While our combined approach significantly improved detection accuracy for irregular and partially occluded objects (particularly vehicles and small structures in UAV footage), the performance improvements were less pronounced for large, regular-shaped

objects, suggesting that this specialized loss function primarily benefits challenging detection scenarios.

4. Experiments and Results

4.1. Datasets

We validate our model on three UAV weak tiny object datasets: VisDrone2019 [6] is used as our primary benchmark, while AI-TOD-v2 [5] and DOTA-v1.5 [50] are used to evaluate cross-dataset generalization capability, demonstrating our model's robustness across diverse aerial imaging scenarios.

VisDrone2019: The VisDrone2019 dataset is an authoritative resource in the international drone vision community, and features diverse multi-scene and multi-task shooting captured by various drones across 14 cities in China and in various environments (urban and rural), sparse or dense scenes, weather conditions (cloudy and sunny), and lighting conditions (day and night). The dataset contains 10 classes, including pedestrians, people, bicycles, cars, vans, trucks, tricycles, awning tricycles, buses, and motors. Following the official partition protocol of the VisDrone2019 challenge, we utilized the pre-partitioned dataset, consisting of 6471 images for training, 548 for validation, and 1610 for testing.

AI-TOD-V2: AI-TOD-V2 is an extremely challenging dataset designed specifically for the detection of weak, tiny objects. It comprises 28,036 aerial images of 800 $\times$ 800 pixels, including 11,214 images in the train set, 2804 images in the validation set, and 14,018 images in the "test" set. The dataset covers eight categories, including airplane, bridge, and storage tank, with a total of 700,621 object instances. The average object size is only 12.8 pixels, which is much smaller compared with other existing aerial object detection datasets.

DOTA-V1.5: DOTA is a large-scale dataset for evaluating the detection performance of oriented object detection in aerial images. DOTA-V1.5 consists of 2806 images, 402,089 instances, and 16 categories. It contains images of varying orientations, scales, and shapes. The size of these aerial images ranges from 800 $\times$ 800 to 4000 $\times$ 4000. There are 1411 images in the training set, 458 images in the validation set, and 937 images in the testing set. The challenge of DOTA-V1.5 is greater than that of DOTA-V1.0.

For experimental use, these images are cropped into 1024 $\times$ 1024 pixels with a 200-pixel overlap while maintaining their original aspect ratios during the segmentation process. After filtering out images without annotations post-segmentation, the dataset consists of 10,483 training images and 3377 validation images. As DOTA-V1.5 does not provide ground truth annotations for the test set, all comparative experiments were conducted on the validation set.

4.2. Evaluation Metrics

In this section, we employ several key metrics, including precision, recall rate, F1 score, mAP50, mAP50:95, frames per second (FPS), floating-point operations per second (FLOPs), and model parameter quantity, to evaluate the performance of the model on different datasets.

The accuracy of a model in object detection hinges on the correctness of predictions, assessed through IoU and a threshold $\in [0, 1]$ utilizing true positive (TP), false positive (FP), and false negative (FN) metrics. TP signifies a correct prediction when IoU surpasses the specified threshold, while FP denotes an erroneous prediction of a non-existent object or the detection of an object with an IoU lower than the threshold. FN identifies an object in the actual image that the model fails to predict. Notably, true negative (TN) is not utilized in object detection due to the infinite number of prediction boxes that should not be anticipated in each image. The evaluation of object detection models predominantly relies on precision (P) and recall (R) criteria. Furthermore, the primary metric for assessing

the accuracy of object detection models is mean average precision (mAP), calculated based on the mean of average precision (AP).

Precision, outlined in Equation (16), gauges the accurate proportion of all predicted bounding boxes, while recall, defined in Equation (17), signifies the ratio of correctly located and identified objects within the ground truth. F1 score, articulated in Equation (18), serves as the harmonic average of precision and recall, providing a comprehensive evaluation of algorithmic performance.

Additionally, mAP50 measures the average precision of predicting boxes when the intersection over union (IoU) exceeds 0.5, as defined by Equation (19), representing the integral area under the precision–recall curve (P–R curve). The mAP50:95 takes into consideration IoU values between 0.5 and 0.95.

$$Precision = \frac{TP}{TP + FP} \tag{16}$$

$$Recall = \frac{TP}{TP + FN} \tag{17}$$

$$F1 - score = \frac{2PR}{P + R} \tag{18}$$

$$mAP = \frac{1}{n}\sum_{i=1}^{n} AP \tag{19}$$

Finally, the metrics for evaluating model complexity and inference speed are considered:

1. Frames per second (FPS) refers to the number of frames a model can infer per second. Generally, depending on the specific application, real-time object detection models should be capable of processing video frames at a speed of at least 40–50 FPS to align with practical task requirements.
2. FLOPs, or floating-point operations per second, represent the required number of floating-point calculations per second. It is a crucial metric indicating the speed and computational capability of neural network models. Lower FLOPs imply lower computational complexity.
3. The use of megabyte units to denote the size and complexity of the model is one of the key reference metrics for lightweight models. In general, models with fewer parameters can execute tasks more quickly and are suitable for deployment on resource-constrained edge devices.

4.3. Implementation Details

The proposed model in this paper is implemented using a 16 vCPU Intel(R) Xeon(R) Platinum 8481C CPU (Intel Corporation, Santa Clara, CA, USA) and Nvidia GeForce RTX 4090D (24 GB) (NVIDIA Corporation, Santa Clara, CA, USA). The operating system is Ubuntu 22.04. All experiments were conducted using PyTorch 2.1.0, CUDA 12.1, and Python 3.9.0, and all deep learning models were constructed based on the open-source object detection toolbox MMDetection and PyTorch framework.

For fair comparison, we note that our proposed model was trained from scratch without any pre-trained weights, while most baseline models utilized pre-trained weights. Other models trained from scratch will be specifically mentioned where applicable.

The training was performed using AdamW optimizer with an initial learning rate of 1×10^{-4}, momentum of 0.9, and weight decay of 1×10^{-4}. We employed a warmup strategy with 2000 iterations. For data augmentation, Mosaic was applied during training. We set the maximum training epochs to 350, although the models typically converged around 300 epochs. For our proposed model, all experiments were conducted with a batch

size of 4 to ensure fair comparison among different configurations, while the batch sizes for other methods are detailed in their corresponding experimental settings.

4.4. Ablation Study

4.4.1. Ablation Analysis for Component-Wise Contributions

In this subsection, we conducted ablation experiments on the VisDrone2019 test dataset to validate the effectiveness of our proposed method, HSF-DETR. The baseline network was upgraded as follows: the original ResNet-18 was substituted with HPFNet, a hybrid multi-scale backbone, the MultiScaleNet for multi-domain feature fusion, and the PACST module for hierarchical feature refinement across deep context and shallow details. Additionally, AdaptDist-IoU was implemented as the loss function. These experiments were performed by incrementally integrating each improvement module, with the findings detailed in Table 1.

Table 1. Results of ablation experiments.

Methods	HPFNet	MultiScaleNet	PACST	AdaptDist-IoU	P^{Test}	R^{Test}	mAP^{Test}_{50}	$mAP^{Test}_{50:95}$	Parameters	GFLOPs
1. Baseline					0.543	0.395	0.374	0.213	19.88M	57
2	√				0.563	0.413	0.399	0.229	13.54M	47
3		√			0.572	0.417	0.41	0.239	20.72M	67.4
4			√		0.583	0.404	0.4	0.232	20.46M	59.7
5	√	√			0.575	0.421	0.411	0.235	14.65M	60.9
6	√		√		0.583	0.412	0.407	0.238	14.12M	49.7
7		√	√		0.605	0.397	0.414	0.242	21.3M	70.1
8	√	√	√		0.587	0.425	0.42	0.243	15.24M	63.6
9. Ours	√	√	√	√	0.586	0.435	0.428	0.253	15.24M	63.6

The √ symbol indicates an improvement in the structure of the corresponding ordinate.

HPFNet improved mAP50 by 2.5% and mAP50:95 by 1.6%, demonstrating enhanced hierarchical feature learning through joint local-global modeling. MultiScaleNet further increased mAP50 and mAP50:95 by 3.6% and 2.5%, respectively, indicating improved multi-scale feature representation in complex scenarios. The PACST block boosted mAP50 by 2.6%, mAP50:95 by 1.9%, and precision by 4%, effectively bridging semantic–spatial gaps between features. The final HSF-DETR model, incorporating all improvements including AdaptDist-IoU loss, achieved overall gains of 5.3% in mAP50, 3.9% in mAP50:95, 4.3% in precision, and 4% in recall.

4.4.2. Ablation Analysis for MPCBlock and MSPAFBlock

To comprehensively evaluate the effectiveness of our proposed modules, we conduct extensive ablation studies on both MPCBlock and MSPAFBlock, analyzing their performance, computational efficiency, and parameter overhead.

For MPCBlock, Table 2 shows that increasing channel partitioning steps from 2 to 3 improves test mAP50 by 2.1% (0.365 → 0.386) with moderate parameter and computational increases (2.65M → 2.80M and 10.27G → 12.35G FLOPs, respectively). This improvement stems from enhanced cross-scale semantic interaction through progressive kernel expansion. However, further increases to four or five steps show diminishing returns—the four-step variant decreases performance by 0.7% despite having 12.2% more parameters, while the five-step version performs worse despite a 41.9% higher computational cost, indicating excessive partitioning disrupts local–global feature balance.

For MSPAFBlock, Table 3 demonstrates that positioning self-attention before conv3×3 achieves optimal performance (0.399 mAP50) with 12.56G FLOPs and 2.34M parameters. This early integration of global context outperforms later placement options (conv5×5, conv7×7), which show lower performance (0.391/0.396 mAP50) despite reduced parameters (2.24M/2.14M), suggesting that early global semantic modeling more effectively guides multi-scale feature learning while preserving spatial details.

Table 2. Performance comparison of different partitioning steps in MPCBlock. MPC 2x, 3x, 4x, and 5x represent 2, 3, 4, and 5 channel partitioning steps, respectively.

Structure	mAP_{50}^{val}	$mAP_{50:95}^{val}$	mAP_{50}^{Test}	$mAP_{50:95}^{Test}$	Block Parameters	Block GFLOPS
MPC 2x	0.468	0.272	0.365	0.2	2,650,304	10.27G
MPC 3x	0.491	0.297	0.386	0.223	2,802,544	12.35G
MPC 4x	0.493	0.294	0.379	0.218	2,912,774	13.85G
MPC 5x	0.477	0.285	0.368	0.205	3,019,884	15.46G

Table 3. Performance comparison of different self-attention positions in MSPAFBlock.

Structure	mAP_{50}^{val}	$mAP_{50:95}^{val}$	mAP_{50}^{Test}	$mAP_{50:95}^{Test}$	Block Para	Block GFLOPS
SHSA-Conv3×3	0.501	0.308	0.399	0.229	2,337,184	12.56G
SHSA-Conv5×5	0.491	0.294	0.391	0.221	2,243,968	11.54G
SHSA-Conv7×7	0.499	0.307	0.396	0.228	2,139,040	10.53G

4.4.3. Ablation Analysis for Backbone

To validate the effectiveness of our proposed backbone network design, we conduct comprehensive ablation studies by comparing HPFNet with several state-of-the-art backbones. As shown in Table 4, we evaluate different backbone architectures, including CNN-based (ResNet-18, Darknet53 [13], MobileNetV4 [51]), transformer-based (CSwin-Transformer [52]), and hybrid approaches (RepViT [53] and LSKNet [54]).

Table 4. Comparative experiments with other representative backbone architectures.

Backbone	mAP_{50}^{Val}	$mAP_{50:95}^{Val}$	mAP_{50}^{Test}	$mAP_{50:95}^{Test}$	GFLOPs	Parameters	FPS
Resnet-r18 (baseline)	0.472	0.281	0.374	0.213	57	19.88M	68.1
Darknet53	0.489	0.295	0.38	0.218	52.6	15.42M	71.1
EfficientViT	0.425	0.257	0.337	0.193	27.3	10.71M	41.2
MobileNetV4	0.437	0.26	0.337	0.191	39.5	11.32M	72.3
iRMB	0.463	0.28	0.366	0.208	49.2	16.42M	59
CSwinTransformer	0.49	0.301	0.383	0.222	89.9	30.49M	25.5
RepViT	0.467	0.285	0.365	0.208	36.4	13.31M	69.4
LSKNet	0.456	0.278	0.357	0.205	37.6	12.57M	54.9
HPFNet (Ours)	0.501	0.308	0.399	0.229	47	13.54M	70.1

HPFNet achieves the best performance, with 0.501 mAP50 and 0.308 mAP50:95 on the validation set. Compared with the strong CSwinTransformer baseline, it obtains 1.1% improvement while reducing computational cost by 47.7% (from 89.9 to 47 GFLOPs) and parameters by 55.6%. Notably, our approach significantly outperforms CSwinTransformer in inference speed, achieving 70.1 FPS compared with 25.5 FPS, representing a 2.75× increase in speed while maintaining superior accuracy. While lightweight architectures like MobileNetV4 achieve slightly higher speed (72.3 FPS), they significantly compromise on detection accuracy, with substantially lower mAP scores. Moreover, our approach achieves a 7.6% higher mAP50 than EfficientViT, with a moderate computational increase, thus demonstrating a superior efficiency–accuracy trade-off. These comprehensive results validate the observation that HPFNet successfully achieves an optimal balance between accuracy, computational efficiency, and real-time performance.

4.4.4. Ablation Analysis for IoU

We validate the effectiveness of AdaptDist-IoU through extensive ablation studies on the HSF-DETR framework, where we improve upon the original GIoU localization metric used in RT-DETR. Using the VisDrone2019 test set, as shown in Table 5, our experi-

ments demonstrate that AdaptDist-IoU consistently outperforms the baseline GIoU across multiple evaluation metrics.

Table 5. Comparative analysis of models using enhanced loss functions.

IoU	mAP_{50}^{Test}	$mAP_{50:95}^{Test}$	P^{Test}	R^{Test}
GIoU	0.42	0.243	0.584	0.425
Focaler-GIoU	0.406	0.235	0.572	0.416
Wise-GIoU	0.419	0.244	0.581	0.435
MPDIoU	0.42	0.246	0.585	0.428
Wise-MPDIoU	0.424	0.246	0.585	0.439
Focaler-Wise-GIoU	0.423	0.245	0.596	0.43
AdaptDist-IoU (adhw2)	0.427	0.249	0.592	0.433
AdaptDist-IoU (adhw2.5)	0.428	0.253	0.586	0.435
AdaptDist-IoU (adhw1.5)	0.425	0.249	0.591	0.435

Analysis of variant performances reveals that MPDIoU, by incorporating vertex distance constraints, maintains the mAP50 (0.42) while improving mAP50:95 to 0.246. The integration of Focaler strategy with MPDIoU further enhances the model's performance, notably achieving a precision of 0.592 with adhw = 2, indicating substantial improvement in detection accuracy. With adhw = 2.5, our approach achieves mAP50 of 0.428 and mAP50:95 of 0.253, significantly surpassing the baseline GIoU (0.42 and 0.243) while maintaining high precision (0.586) and recall (0.435). These results demonstrate that AdaptDist-IoU effectively enhances localization accuracy while preserving robust detection rates.

4.5. Comparative Result Analysis of Different Detection Models

4.5.1. Enhanced Performance Analysis

To verify the validity of our proposed model, we selected several representative detection methods. For one-stage detectors, we compared with recent YOLO variants (YOLOv8/9/10/11/12-M) and other competitive models like TOOD [55], RetinaNet, EfficientNet, RTMDet-M [56], ATSS [57] and MADet [58]. For two-stage detectors, we selected Faster R-CNN and Cascade R-CNN. For transformer-based approaches, we included vanilla DETR and its variants (deformable-DETR, DAB-DETR [59], conditional-DETR, DINO [60]) as well as the RT-DETR series). Here, R18/R50/R101 denote ResNet-18/50/101 backbones, respectively, -M indicates medium variants, and b3 represents EfficientNet's compound scaling factor. For training, TOOD, Faster R-CNN, Cascade R-CNN, RetinaNet, EfficientNet and DINO follow the MMDetection 1× schedule, while DETR, DAB-DETR, and conditional-DETR need 50 epochs. DAB-DETR converges at 150 epochs and the YOLO variants at 200 epochs. RTMDet, the RT-DETR series, and HSF-DETR each train for 350 epochs. All follow official implementations. Except for the RT-DETR series, HSF-DETR and DINO, the models used pretrained weights.

We evaluate HSF-DETR against state-of-the-art object detection methods on the VisDrone2019 benchmark, as shown in Table 6. Compared with the baseline, our approach achieves substantial improvements in mAP50 by 5.4% (0.374 → 0.428) and mAP50:95 by 4.0% (0.213 → 0.253). These significant enhancements can be attributed to three key factors:

(1) HPFNet multi-scale feature extraction strategy outperforms traditional ResNet backbones, particularly in capturing fine details of small objects. Specifically, MPCBlock preserves spatial details through progressive receptive field expansion and partial convolutions, effectively preventing the loss of critical information during multi-level downsampling. Meanwhile, MSPAFBlock integrates single-head attention mechanisms in deeper layers, enhancing global semantic modeling capabilities, enabling our model to better handle objects in complex backgrounds, thus improving precision from 0.553 to 0.586.

(2) MultiScaleNet cross-domain feature fusion mechanisms address RT-DETR's limitations in consistent representation across scale variations. As evidenced in Table 7, we achieve particularly pronounced improvements in detecting challenging small-scale object categories (e.g., bicycles, tricycles, pedestrians), with the pedestrian category mAP50 increasing from 0.383 to 0.439 (+5.6%). This demonstrates that our P2 layer feature optimization and inter-domain fusion strategies effectively capture discriminative features of these objects.

(3) The PACST module significantly enhances localization accuracy by resolving inconsistencies between shallow spatial features and deep semantic features, improving recall from 0.401 to 0.435. This is particularly crucial for distinguishing visually similar categories (e.g., trucks vs. vans), as demonstrated in Table 7 and the confusion matrices in Figure 6, where our model exhibits stronger discriminative capabilities between these classes.

Table 6. Comparison of the current object detection models on the VisDrone2019 test dataset. Red numbers indicate the best values across all compared models, while blue numbers represent the lowest values. For metrics like mAP, precision, recall and F1 score, higher values (red) are better; for computational costs (parameters, GFLOPs) and inference time, lower values (blue) are preferred.

Model	Input	P^{Test}	R^{Test}	mAP^{Test}_{50}	$mAP^{Test}_{50:95}$	$F1^{Test}$	Parameters	GFLOPs	FPS
One-Stage Object Detector									
TOOD-R50	(768, 1344)	0.474	0.391	0.376	0.231	0.428	32.03M	199G	43.7
RetinaNet-R50	(768, 1344)	0.429	0.339	0.323	0.196	0.378	36.15M	210G	56.7
MADet-R50	(640, 640)	0.482	0.379	0.371	0.233	0.424	33.57M	74.7G	73.1
ATSS-R50	(768, 1344)	0.485	0.39	0.391	0.232	0.432	38.91M	110G	52.4
RTMDet-M	(640, 640)	0.481	0.361	0.381	0.24	0.412	24.66M	**39.1G**	50.9
YOLOV8-M	(640, 640)	0.494	0.365	0.353	0.206	0.419	25.84M	78.7G	135.5
YOLOV9-M	(640, 640)	0.519	0.389	0.385	0.231	0.444	20.07M	77.1G	103.4
YOLOV10-M	(640, 640)	0.491	0.37	0.359	0.21	0.421	16.46M	63.5G	105.4
YOLOV11-M	(640, 640)	0.504	0.383	0.373	0.218	0.435	20.03M	67.7G	117.1
YOLOV12-M	(640, 640)	0.501	0.382	0.368	0.212	0.433	20.11M	67.2G	106.3
Two-Stage Object Detector									
Faster-RCNN-R50	(768, 1344)	0.465	0.384	0.364	0.227	0.420	41.39M	208G	53.3
Cascade-RCNN-R50	(768,1344)	0.471	0.379	0.37	0.229	0.419	69.18M	236G	44.1
Efficietnet-b3	(512, 896)	0.376	0.305	0.283	0.169	0.337	18.52M	62.5G	61.7
DETR-Based Object Detector									
DETR	(750, 1333)	0.343	0.294	0.266	0.13	0.316	41.57M	96.5G	80.7
Deformable-DETR	(750, 1333)	0.383	0.334	0.305	0.166	0.357	40.10M	193G	36.9
DAB-DETR-R50	(750, 1333)	0.438	0.337	0.328	0.17	0.381	43.70M	102G	66.1
Conditional-DETR-R50	(750, 1333)	0.408	0.329	0.316	0.161	0.364	43.45M	101G	73.4
DINO-R50	(750, 1333)	0.52	0.437	**0.431**	0.247	0.475	47.55M	274G	27.5
RT-DETR-R18 (Baseline)	(640, 640)	0.553	0.401	0.374	0.213	0.457	19.88M	57G	78.1
RT-DETR-R50	(640, 640)	0.577	0.403	0.39	0.224	0.474	41.97M	129G	58.1
RT-DETR-R101	(640, 640)	0.584	0.427	0.415	0.244	0.493	74.67M	247G	49.2
Hyper Scale Fusion Detection Transformer									
HSF-DETR	(640, 640)	0.586	0.435	0.428	**0.253**	**0.499**	**15.24M**	63.6G	69.3

4.5.2. Enhanced Efficiency-Accuracy Trade-Off Analysis

From a computational efficiency perspective, HSF-DETR significantly reduces resources compared with models with similar accuracy. When compared with DINO (mAP50 = 0.431), our model maintains comparable accuracy while reducing parameters by 68% (15.24M vs. 47.55M) and computational cost by 77% (63.6G vs. 274G FLOPs), with

2.5× faster inference (69.3 vs. 27.5 FPS). While lightweight models like YOLOv11-M offer higher speed (117.1 FPS), they sacrifice substantial accuracy (−5.5% mAP50).

A detailed component-wise timing analysis comparing HSF-DETR with RT-DETR baseline reveals our strategic redistribution of computational resources (Table 7).

Table 7. Component-wise inference time comparison between HSF-DETR and baseline RT-DETR.

Component	HSF-DETR (ms)	RT-DETR (ms)	Difference
Backbone	57.42	90.10	−36.3%
Feature fusion network	78.58	45.99	+70.8%
Context and alignment	17.45	-	Novel component
Total	153.45	136.09	+12.7%
Accuracy Comparison			
mAP50	0.428	0.374	+14.4%
mAP50:95	0.253	0.213	+18.7%

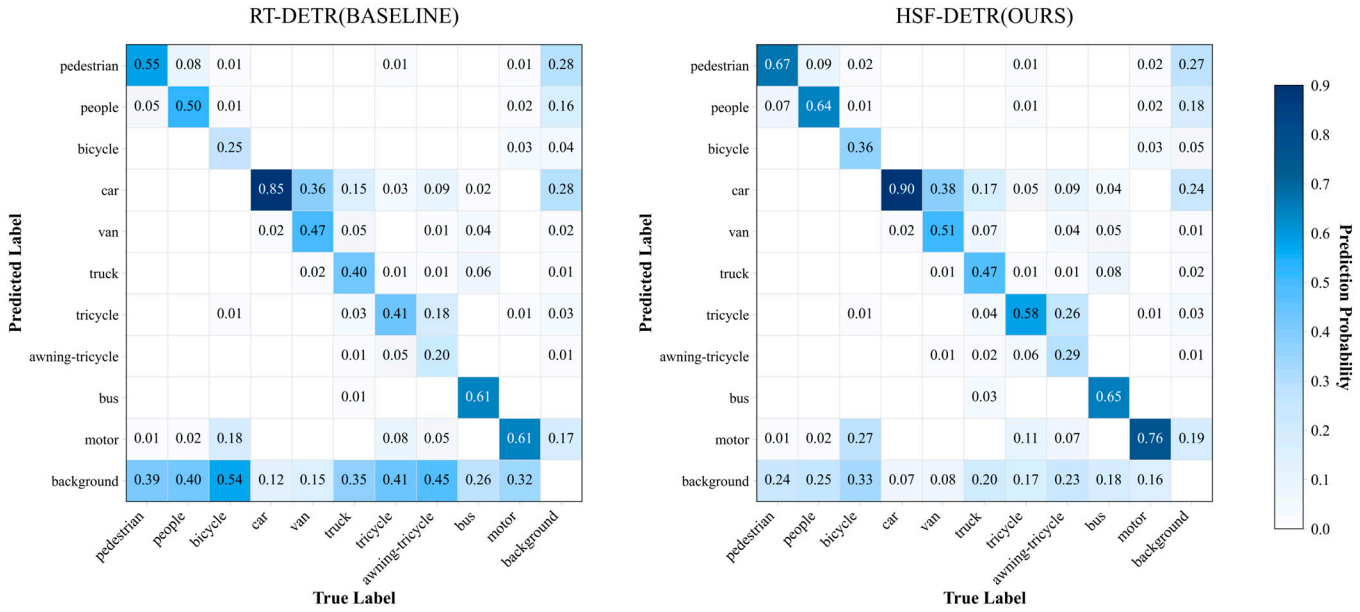


Figure 6. Confusion matrix for the proposed model and RT-DETR-R18 trained on the VisDrone2019 dataset.

Notably, the HPFNet backbone processes significantly faster than the ResNet-18 baseline (−36.3%), demonstrating remarkable efficiency despite enhanced feature extraction capabilities. This speed advantage stems from our novel MPCBlock design with selective channel partitioning and progressive kernel expansion, which reduces computational redundancy while preserving critical feature details. Additional computational resources are strategically allocated toward the feature fusion network and innovative PACST module, components directly responsible for the improved detection of small objects in complex scenes. The DomainScaleKernel within our feature fusion network, while computationally intensive, delivers substantial benefits through its multi-branch architecture and cross-domain processing capabilities, effectively addressing the scale-domain heterogeneity challenges inherent in UAV imagery. Similarly, the PACST module's computational cost is justified by its ability to resolve semantic–spatial misalignments that particularly affect small object detection accuracy in cluttered environments.

Figure 6 depicts the confusion matrices obtained from the VisDrone2019 of our model against RT-DETR-R18, demonstrating robust inter-class discriminative power.

As illustrated in Figure 7, HSF-DETR achieves an optimal balance on the speed–accuracy curve for real-time UAV applications. When considering the mAP50:95 metric (which better reflects localization precision), our model shows clear advantages over all comparable-speed alternatives, ensuring that HSF-DETR remains practical for deployment while delivering the precision critical for aerial surveillance scenarios.

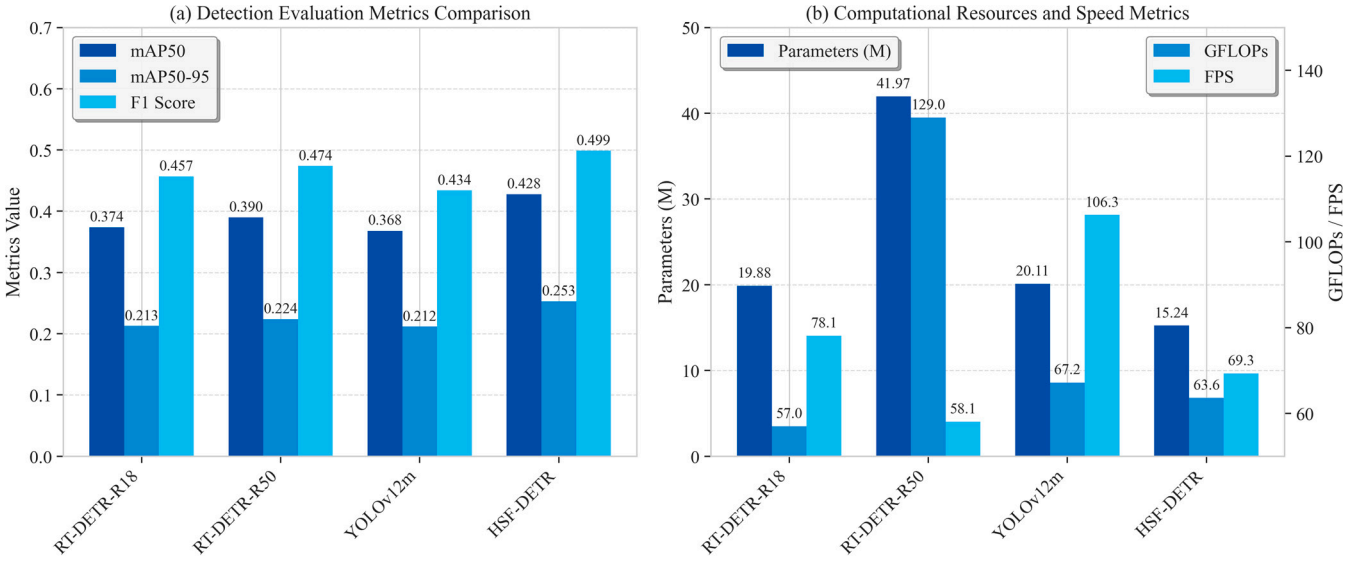


Figure 7. Performance comparison of RT-DETR variants and HSF-DETR.

4.5.3. Enhanced Category-Wise Performance Analysis

Table 8 presents our model's performance across various categories. Particularly noteworthy is HSF-DETR's exceptional performance in detecting challenging categories, as follows:

(1) Small-sized categories: The pedestrian and people categories show significant improvements of 5.6% and 5.9% in mAP50, respectively. This substantial enhancement stems from HPFNet's optimized receptive field design and MultiScaleNet's cross-domain feature learning capability, enabling the model to capture crucial details of these small objects even in dense crowd scenarios (as evident in Figure 8d,f).

(2) Visually similar categories: For challenging-to-distinguish categories such as "truck" vs. "van" and "tricycle" vs. "awning-tricycle," our model demonstrates enhanced discriminative power. For instance, the truck category shows an 8.3% improvement in mAP50 (0.412 $\rightarrow$ 0.495), confirming that the PACST module successfully enhances semantic representation and reduces inter-class confusion, consistent with the confusion matrix results in Figure 7.

(3) Rare categories: For categories with fewer instances in the dataset, our improvements are even more pronounced, with mAP50 increases of 3.7% and 3.6%, respectively, indicating that our approach effectively learns discriminative features even in limited-sample scenarios.

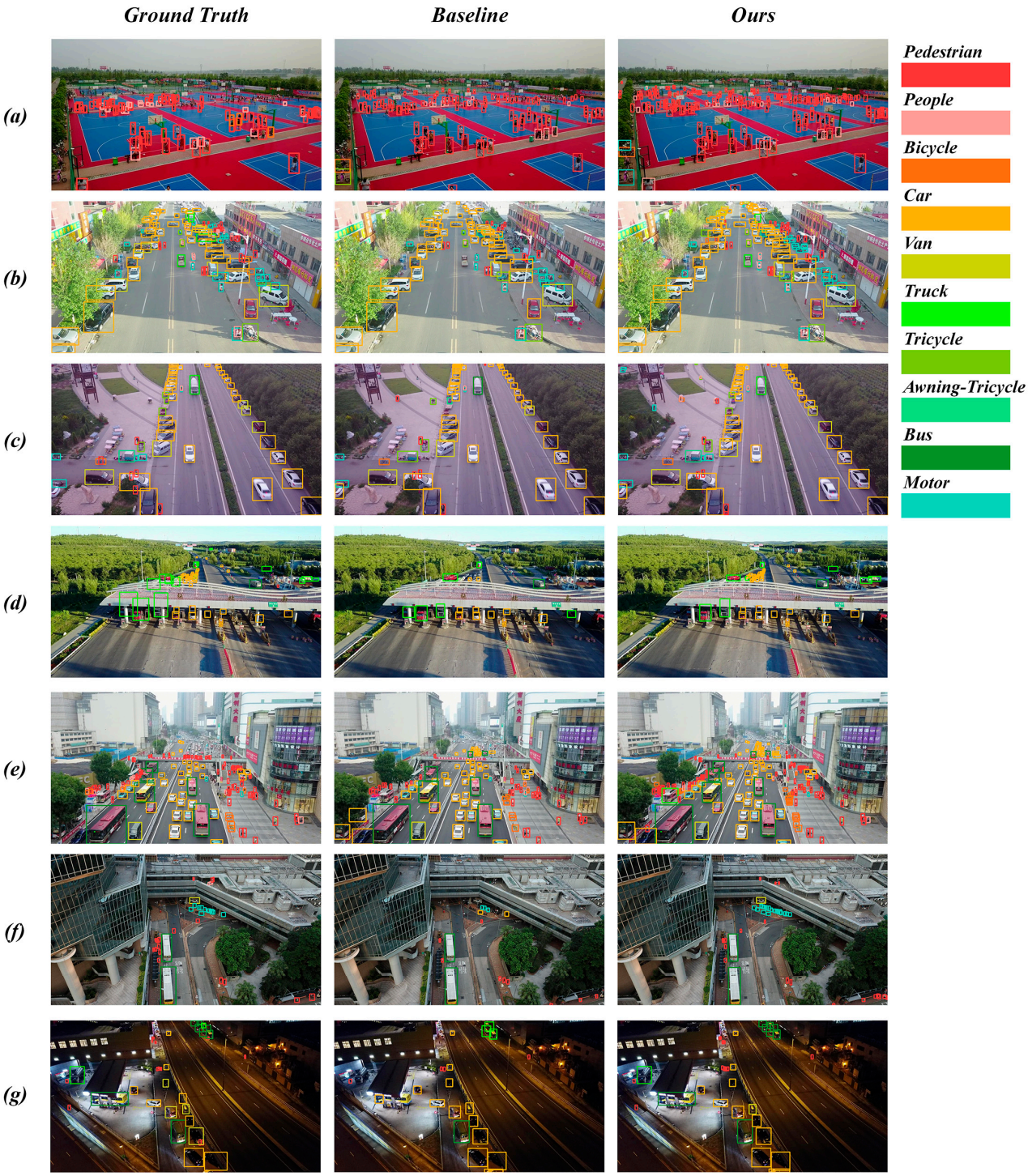


Figure 8. Visualization of detection results using the VisDrone2019 dataset. The bounding box colors indicate different semantic categories of detected objects. (**a**) Small object detection with enhanced identification of tiny targets and vehicles on the left side; (**b**) Dense scenario performance showing accurate detection of bicycle and pedestrian groups in the upper right corner; (**c**) Superior small target detection capability identifying minute objects missed in ground truth annotations; (**d**) Occluded object recognition demonstrating robust detection under partial occlusion conditions; (**e**) Viewpoint adaptation maintaining detection quality across extreme viewing angles and scale variations; (**f**) Dense clustering scenarios with successful detection of motorcycle and pedestrian clusters; (**g**) Boundary precision showing accurate localization of partially occluded bus with improved bounding box accuracy. The bounding box colors indicate different semantic categories of detected objects.

Table 8. Comparative analysis for various categories on the VisDrone2019 test dataset.

Class	Instance	P^{Test}	R^{Test}	mAP^{Test}_{50}	$mAP^{Test}_{50:95}$	P^{Test}	R^{Test}	mAP^{Test}_{50}	$mAP^{Test}_{50:95}$
All	75,102	*Baseline*				*Ours*			
Pedestrian	21,006	0.569	0.376	0.383	0.152	0.634	0.407	0.439	0.183
People	6376	0.572	0.274	0.271	0.0965	0.624	0.298	0.33	0.122
Bicycle	1302	0.388	0.175	0.141	0.0577	0.468	0.2	0.178	0.0775
Car	28,074	0.768	0.784	0.778	0.505	0.778	0.798	0.806	0.529
Van	5771	0.581	0.389	0.372	0.26	0.578	0.439	0.407	0.295
Truck	2659	0.547	0.446	0.412	0.257	0.597	0.514	0.495	0.325
Tricycle	530	0.332	0.372	0.25	0.139	0.376	0.379	0.287	0.171
Awning-Tricycle	599	0.471	0.214	0.184	0.108	0.463	0.227	0.22	0.142
Bus	2940	0.774	0.52	0.554	0.392	0.778	0.565	0.627	0.463
Motor	5845	0.528	0.461	0.408	0.168	0.569	0.519	0.48	0.212

4.6. Visual Analysis

We evaluate HSF-DETR's robustness through systematic tests on the VisDrone2019 dataset across challenging conditions. Figure 8 showcases detection results under varying illumination, flying altitudes, and multi-scale targets, revealing several key findings, as follows:

(1) Enhanced small object detection: Figure 8a,c demonstrate our model's superior capability in detecting minute objects often missed in ground truth annotations. In Figure 8a, HSF-DETR detects significantly more small-pixel targets than the baseline, while correctly identifying motorcycles and bicycles on the left side that were misclassified by the baseline. This stems from the progressive receptive field modulation in HPFNet and improved localization guidance from AdaptDist-IoU for tiny objects.

(2) Superior performance in dense scenarios: Figure 8b,f highlight our model's effectiveness with clustered objects. HSF-DETR accurately identifies bicycle and pedestrian groups in the upper right corner of Figure 8b and successfully detects motorcycle and pedestrian clusters in Figure 8f, all missed by the baseline. This precision in crowded scenes results from cross-domain feature learning in MultiScaleNet combined with vertex distance constraints in AdaptDist-IoU.

(3) Improved boundary precision for challenging objects: Figure 8g demonstrates HSF-DETR's ability to correctly identify and precisely localize the partially occluded bus that the baseline misclassifies. The AdaptDist-IoU loss dynamically adjusts penalty distribution based on object characteristics, particularly benefiting objects with irregular shapes and occlusions.

(4) Robust viewpoint adaptation: Figure 8e highlights our model's consistent performance across extreme viewpoints and scale variations. HSF-DETR maintains detection quality even for distant objects with perspective distortion—a combined effect of our multi-scale architecture and geometric vertex optimization in AdaptDist-IoU for varying viewpoints.

Feature map analysis (Figure 9) demonstrates our backbone superiority through different stages (P2–P5). While the baseline model shows detection limitations, such as missing motor vehicles and making false identifications, HPFNet achieves more accurate results, especially for small objects. This improvement stems from the MPCBlock progressive kernel expansion strategy. Our model generates more concentrated and semantically meaningful attention regions when compared with the baseline's dispersed activation patterns, particularly evident in deeper layers (P5). This validates our hybrid architecture's effective-

ness in combining local–global semantic fusion through group convolutions and channel partitioning strategies.

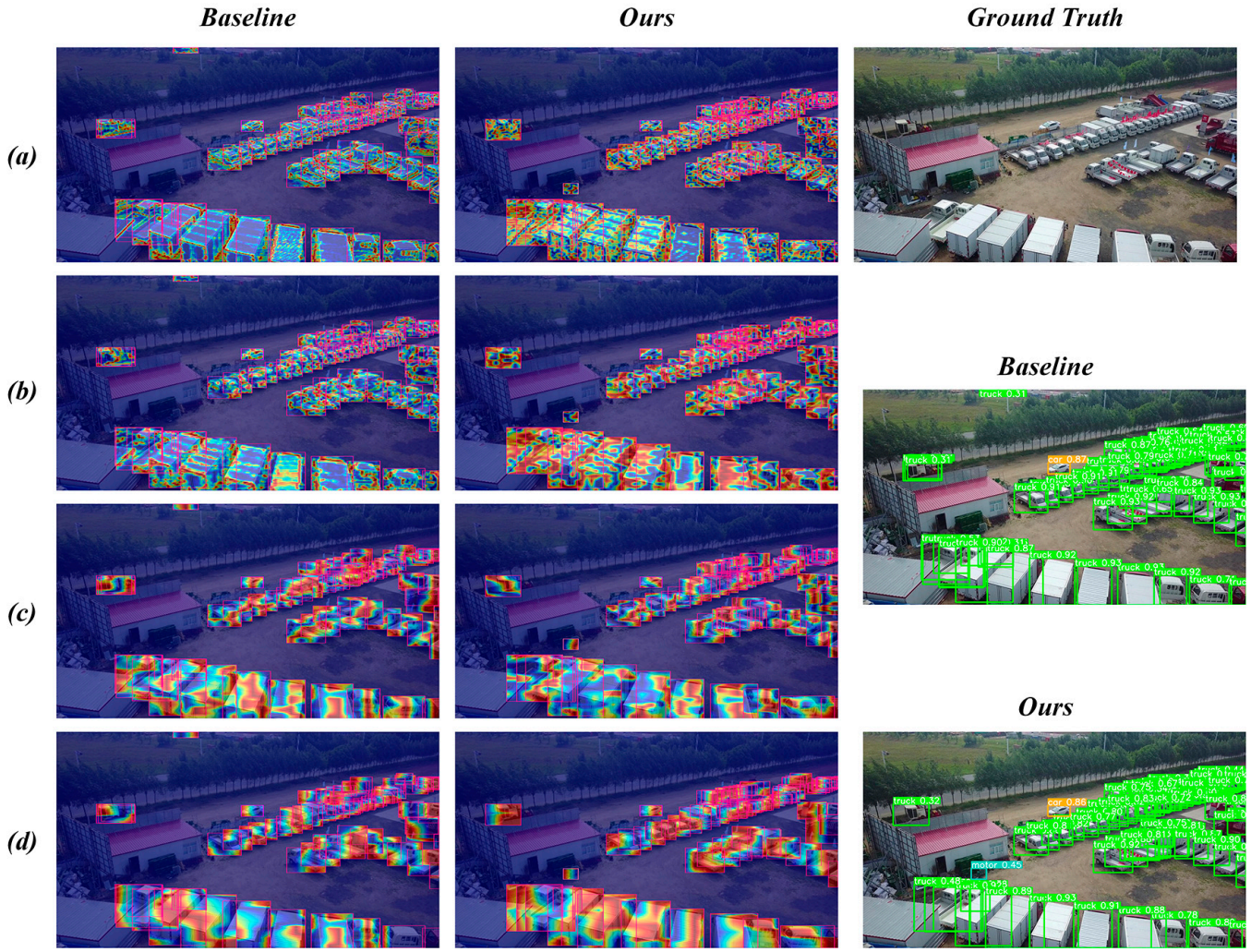


Figure 9. Visualization of detection results using the VisDrone2019 dataset with our backbones and baseline backbone. (**a**) Stage 1 (P2) feature representation showing initial feature extraction capabilities; (**b**) Stage 2 (P3) feature maps demonstrating progressive semantic enhancement; (**c**) Stage 3 (P4) feature activation patterns revealing improved object localization; (**d**) Stage 4 (P5) deep layer features showing concentrated and semantically meaningful attention regions with superior small object detection. The bounding box colors indicate different semantic categories: green for track, orange for car, and blue for motor (consistent with Figure 8 color coding).

4.7. Extend Experiments

To ensure optimal computational efficiency and accommodate the increased model complexity, we conducted these experiments using an NVIDIA RTX 4090 GPU (24 GB) (NVIDIA Corporation, Santa Clara, CA, USA) and a 16 vCPU AMD EPYC 9654 96-core processor (Advanced Micro Devices, Santa Clara, CA, USA). This adjustment was necessitated by the need for higher computational throughput when processing high-resolution remote sensing imagery. Additionally, we refined our training protocol by modifying the number of epochs and batch size settings to fully leverage the computational capabilities.

(1) DOTA-V1.5 dataset

To evaluate model generalization, we conducted additional experiments on the DOTA-V1.5 dataset. Due to the unavailability of official test set labels, all comparisons were performed using validation set results (Table 9). For these experiments, we employed a batch size of 2 and trained the model for 300 epochs, while maintaining all other hyperparameters consistent with our VisDrone2019 experimental configuration.

Table 9. Comparison of the baseline model and our proposed model on the DOTA-v1.5 validation dataset.

Model	P^{Val}	R^{Val}	mAP^{Val}_{50}	$mAP^{Val}_{50:95}$	$F1^{Val}$
RT-DETR-R18 (baseline)	0.689	0.63	0.646	0.409	0.658
HSF-DETR (ours)	0.7	0.676	0.693	0.447	0.688

Our method demonstrates remarkable generalization performance, achieving precision and recall scores of 0.700 and 0.676, respectively, and the F1 score shows a 3.0% improvement. Furthermore, the model exhibits substantial gains in detection accuracy, with respective mAP50 and mAP50:95 values of 0.693 and 0.447, outperforming the baseline scores of 0.646 and 0.409. Particularly noteworthy is the 4.7% enhancement in mAP50 and 3.8% improvement in mAP50:95, validating the model's robust performance in aerial remote sensing scenarios.

(2) AI-TOD-V2 dataset

For the experiments on the AI-TOD-v2 dataset, we maintained hyperparameter settings that were consistent with those used on VisDrone2019, with only necessary adjustments to accommodate different computing resources. Given that AI-TOD-v2 provides comprehensive train, validation, and test splits, along with corresponding annotations, we conducted thorough evaluations on both validation and test sets to ensure robust assessment of our model's generalization capability.

As shown in Table 10, our proposed method demonstrates remarkable generalization capability on the AI-TOD-v2 dataset. The performance gains are consistently maintained on the test set, where our method achieves 0.575 mAP50 and 0.239 mAP50:95, outperforming the baseline by 4.0% and 1.2%, respectively. The precision and recall metrics on the test set show similar improvements, with precision increasing from 0.656 to 0.666 and recall significantly improving from 0.534 to 0.596. The consistent performance improvement across both validation and test sets demonstrates the robustness and reliability of our proposed approach in handling real-world aerial imagery scenarios. Figure 10 shows the visualization of DOTA-V1.5 and AI-TOD-V2 datasets.

Table 10. Comparison of the baseline model and our proposed model on the AI-TOD-v2 dataset.

Model	P^{Val}	R^{Val}	mAP^{Val}_{50}	$mAP^{Val}_{50:95}$	$F1^{Val}$	P^{Test}	R^{Test}	mAP^{Test}_{50}	$mAP^{Test}_{50:95}$	$F1^{Test}$
RT-DETR-R18 (baseline)	0.661	0.548	0.565	0.246	0.599	0.656	0.534	0.535	0.227	0.588
HSF-DETR (ours)	0.676	0.611	0.605	0.259	0.642	0.666	0.596	0.575	0.239	0.629

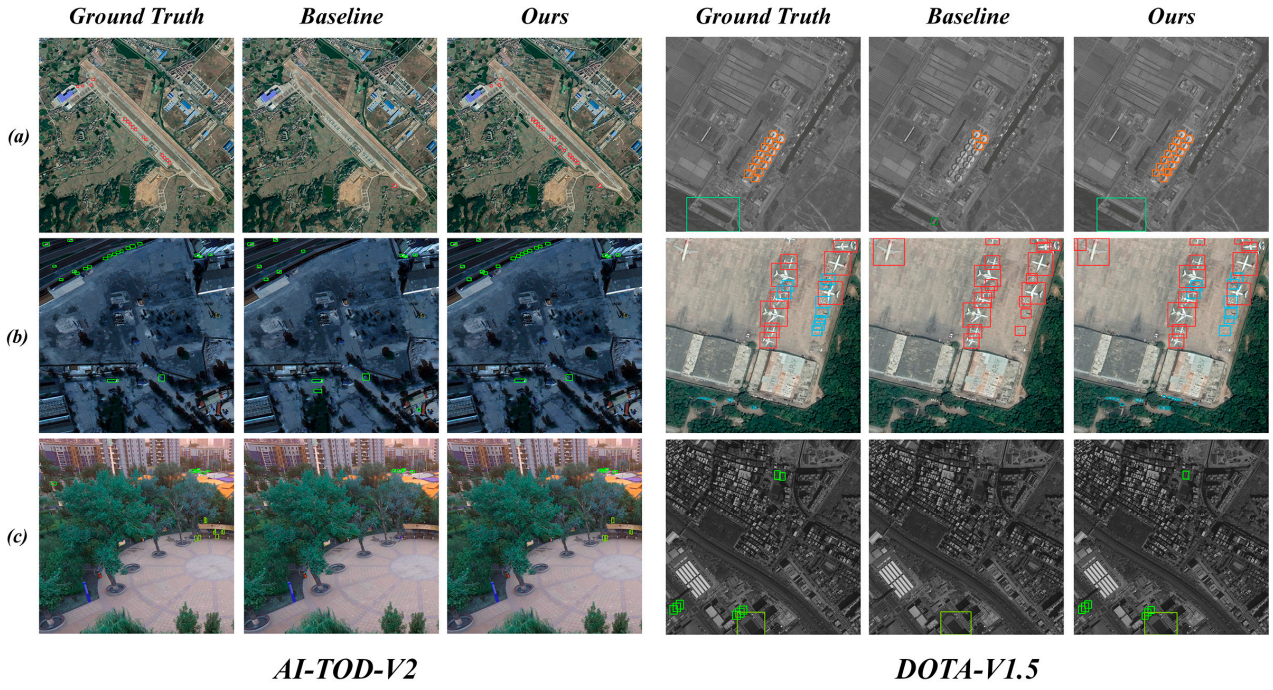


Figure 10. Detection performance comparison on AI-TOD-V2 (left) and DOTA-V1.5 (right) datasets. Each row shows Ground Truth, Baseline, and HSF-DETR results for the same scene. (**a**) Sample 1 detection comparison; (**b**) Sample 2 detection comparison; (**c**) Sample 3 detection comparison.

5. Discussion

The primary contribution of this work is the development of HSF-DETR, a comprehensive and efficient object detection framework that addresses critical challenges in UAV-based detection through four key innovations. HPFNet introduces a hybrid backbone architecture that effectively combines local and global feature modeling. MultiScaleNet advances feature fusion through its DomainScaleKernel with specialized branches for detail enhancement and cross-domain modeling, while PACST addresses feature representation constraints through its dual-module design. Additionally, our AdaptDist-IoU loss significantly improves detection accuracy for irregular-shaped and occluded objects.

Comprehensive experiments demonstrate HSF-DETR's superior performance on the VisDrone2019 test set, with substantial improvements in mAP50 and mAP50:95 (by 5.4% and 4%, respectively) when compared with the RT-DETR-R18 baseline, while reducing the parameter count by 23.3% and maintaining a comparable inference speed. Visualization experiments further validate robust detection capabilities across diverse challenging scenarios.

While our approach demonstrates significant improvements, several important limitations warrant discussion. First, computational complexity in the frequency domain processing within MultiScaleNet, while effective, introduces non-negligible computational overhead. Current FFT implementations on GPUs are not fully optimized for our specific operation patterns, suggesting the potential for specialized implementations or alternative spectral decomposition approaches in future work.

Second, our model shows performance degradation when dealing with severely occluded objects, particularly when distinguishing between visually similar categories. As evident in Figure 8d, while ground truth annotations encompass entire objects, including occluded portions, our model tends to detect only visible regions, producing bounding boxes that do not fully capture the complete object extent. This discrepancy highlights a fundamental limitation in handling occlusion, where insufficient discriminative features

from heavily occluded regions prevent accurate full-object detection, particularly in dense traffic scenarios with multiple overlapping objects.

Third, our architecture exhibits diminishing returns at very high resolutions. While moderate resolution increases improve performance, memory requirements grow quadratically, creating deployment challenges for high-resolution drone footage on memory-constrained devices without architectural optimization.

Finally, despite our improvements, there remains an inherent trade-off between small object detection and computational efficiency. Extremely small objects (under 8×8 pixels) still present challenges even with our enhanced architecture. This limitation is evident in Figure 8e, where pedestrians on the bridge are only partially detected by our model. In this scenario, the combination of small object size and partial occlusion by bridge railings renders many pedestrians nearly indistinguishable, highlighting fundamental limits of current deep learning approaches for ultra-small object detection, even with our specialized architecture.

Future work will focus on addressing these limitations through more efficient spectral processing techniques, occlusion-aware feature learning, adaptive resolution processing, and further optimization for edge deployment, particularly targeting dedicated neural processing hardware for UAV applications.

6. Conclusions

This paper presents HSF-DETR, a novel detector specifically designed to address the challenges of small object detection in drone-view scenarios. Through our comprehensive evaluation, we have demonstrated that the integration of hybrid spatial-frequency feature extraction, multi-scale context modeling, and precise feature alignment significantly enhances detection performance across varying scales and complex environments.

The model's strong generalization ability is evidenced by significant performance gains on both the DOTA-V1.5 validation set (improvements of 4.7% in mAP50 and 3.8% in mAP50:95) and the AI-TOD-V2 test set (increases of 4.0% in mAP50 and 1.2% in mAP50:95). These comprehensive results validate the model's effectiveness in handling complex real-world scenarios while demonstrating substantial potential for practical applications.

Our future work will focus on several important directions to address the current limitations. We plan to optimize the computational efficiency of frequency domain processing through specialized implementations and alternative spectral decomposition approaches that maintain effectiveness while reducing overhead. To improve detection performance for severely occluded objects, we will explore occlusion-aware feature learning mechanisms that better capture complete object representations from partial views, particularly for distinguishing between visually similar categories in dense scenes. Additionally, we aim to develop adaptive resolution processing techniques to overcome memory constraints at high resolutions, enabling more efficient handling of high-resolution drone footage on resource-limited devices. For extremely small objects (under 8×8 pixels), we will investigate enhanced feature extraction methods and specialized attention mechanisms designed specifically for ultra-small object representation. Finally, we plan to explore hardware-aware model optimizations for efficient edge deployment on UAV platforms, balancing computational requirements with detection performance while leveraging dedicated neural processing hardware.

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

Funding: This research was funded by the National Natural Science Foundation of China, grant number 42274051; the Jiangsu Funding Program for Excellent Postdoctoral Talent, grant number 2024ZB335; the Fundamental Research Foundation of National Key Laboratory of Automatic Target Recognition, grant number JKWATR-240101; and the Key Laboratory Program, grant number 6142101230101.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: Author Feng Zhu was employed by the company Nanjing Research Institute of Electronic Engineering. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

UAV	Unmanned aerial vehicle
CNN	Convolutional neural network
DETR	Detection transformer
RT-DETR	Real-time detection transformer
IoU	Intersection over union
NMS	Non-maximum suppression
HPFNet	Hybrid progressive fusion network
MPCBlock	Multi-scale partial convolution block
MSPAFBlock	Multi-scale partial attention fusion block
CCFF	Cross-scale feature fusion
SPDConv	Space-to-depth convolution
CSCT	Cascaded shuffle convolution tuning
CSP	Cross-stage partial
PACST	Perceptalign align context and spatial tuning
MCCF	Multi-contextual cross fusion
DFAF	Dynamic feature alignment and fusion
DRB	Detail refinement block
HSF-DETR	Hyper scale fusion detection transformer
FSCA	Frequency-spatial channel attention
FGM	Feature guidance module
SCD	Semantic compression distillation
MSAF	Multi-scale adaptive fusion
mAP	Mean average precision
FPS	Frames per second
FLOPs	Floating-point operations per second
TP	True positive
FP	False positive
FN	False negative
TN	True negative
P	Precision
R	Recall

References

1. Masuduzzaman, M.; Islam, A.; Sadia, K.; Shin, S.Y. UAV-Based MEC-Assisted Automated Traffic Management Scheme Using Blockchain. *Future Gener. Comput. Syst.* **2022**, *134*, 256–270. [CrossRef]
2. Song, B.D.; Park, H.; Park, K. Toward Flexible and Persistent UAV Service: Multi-Period and Multi-Objective System Design with Task Assignment for Disaster Management. *Expert Syst. Appl.* **2022**, *206*, 117855. [CrossRef]

3. Liu, W.; Zhang, T.; Huang, S.; Li, K. A Hybrid Optimization Framework for UAV Reconnaissance Mission Planning. *Comput. Ind. Eng.* **2022**, *173*, 108653. [CrossRef]
4. Hamzenejadi, M.H.; Mohseni, H. Fine-Tuned YOLOv5 for Real-Time Vehicle Detection in UAV Imagery: Architectural Improvements and Performance Boost. *Expert Syst. Appl.* **2023**, *231*, 120845. [CrossRef]
5. Xu, C.; Wang, J.; Yang, W.; Yu, H.; Yu, L.; Xia, G.-S. Detecting Tiny Objects in Aerial Images: A Normalized Wasserstein Distance and a New Benchmark. *ISPRS J. Photogramm. Remote Sens.* **2022**, *190*, 79–93. [CrossRef]
6. Du, D.; Zhu, P.; Wen, L.; Bian, X.; Lin, H.; Hu, Q.; Peng, T.; Zheng, J.; Wang, X.; Zhang, Y.; et al. VisDrone-DET2019: The Vision Meets Drone Object Detection in Image Challenge Results. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), Seoul, Republic of Korea, 27–28 October 2019; pp. 2778–2788. [CrossRef]
7. Manfreda, S.; McCabe, M.F.; Miller, P.E.; Lucas, R.; Pajuelo Madrigal, V.; Mallinis, G.; Ben Dor, E.; Helman, D.; Estes, L.; Ciraolo, G.; et al. On the Use of Unmanned Aerial Systems for Environmental Monitoring. *Remote Sens.* **2018**, *10*, 641. [CrossRef]
8. Li, X.; Diao, W.; Mao, Y.; Gao, P.; Mao, X.; Li, X.; Sun, X. OGMN: Occlusion-Guided Multi-Task Network for Object Detection in UAV Images. *ISPRS J. Photogramm. Remote Sens.* **2023**, *199*, 242–257. [CrossRef]
9. Chalavadi, V.; Jeripothula, P.; Datla, R.; Ch, S.B.; Mohan C, K. mSODANet: A Network for Multi-Scale Object Detection in Aerial Images Using Hierarchical Dilated Convolutions. *Pattern Recognit.* **2022**, *126*, 108548. [CrossRef]
10. Yuan, Z.; Gong, J.; Guo, B.; Wang, C.; Liao, N.; Song, J.; Wu, Q. Small Object Detection in UAV Remote Sensing Images Based on Intra-Group Multi-Scale Fusion Attention and Adaptive Weighted Feature Fusion Mechanism. *Remote Sens.* **2024**, *16*, 4265. [CrossRef]
11. 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]
12. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollar, P. Focal Loss for Dense Object Detection. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2980–2988. [CrossRef]
13. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**. [CrossRef]
14. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J.; Ding, G. YOLOv10: Real-Time End-to-End Object Detection. *arXiv* **2024**. [CrossRef]
15. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016. [CrossRef]
16. 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]
17. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-End Object Detection with Transformers. *arXiv* **2020**. [CrossRef]
18. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS), Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008. [CrossRef]
19. 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; pp. 16965–16974. [CrossRef]
20. Deng, C.; Wang, M.; Liu, L.; Liu, Y.; Jiang, Y. Extended Feature Pyramid Network for Small Object Detection. *IEEE Trans. Multimed.* **2022**, *24*, 1968–1979. [CrossRef]
21. Paoletti, M.E.; Haut, J.M.; Pereira, N.S.; Plaza, J.; Plaza, A. Ghostnet for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 10378–10393. [CrossRef]
22. Pan, X.; Ge, C.; Lu, R.; Song, S.; Chen, G.; Huang, Z.; Huang, G. On the Integration of Self-Attention and Convolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 18308–18318. [CrossRef]
23. Yun, S.; Ro, Y. SHViT: Single-Head Vision Transformer with Memory Efficient Macro Design. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; pp. 5756–5767. [CrossRef]
24. Pan, X.; Tang, F.; Dong, W.; Gu, Y.; Song, Z.; Meng, Y.; Xu, P.; Deussen, O.; Xu, C. Self-Supervised Feature Augmentation for Large Image Object Detection. *IEEE Trans. Image Process.* **2020**, *29*, 6745–6758. [CrossRef]
25. Zhong, Y.; Li, B.; Tang, L.; Kuang, S.; Wu, S.; Ding, S. Detecting Camouflaged Object in Frequency Domain. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 4494–4503. [CrossRef]
26. Sunkara, R.; Luo, T. No More Strided Convolutions or Pooling: A New CNN Building Block for Low-Resolution Images and Small Objects. *arXiv* **2022**. [CrossRef]

27. Wang, C.-Y.; Mark Liao, H.-Y.; Wu, Y.-H.; Chen, P.-Y.; Hsieh, J.-W.; Yeh, I.-H. CSPNet: A New Backbone That Can Enhance Learning Capability of CNN. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 14–19 June 2020; pp. 1571–1580. [CrossRef]
28. Li, Y.; Yao, T.; Pan, Y.; Mei, T. Contextual Transformer Networks for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 1489–1500. [CrossRef]
29. Han, J.; Ding, J.; Li, J.; Xia, G.-S. Align Deep Features for Oriented Object Detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5602511. [CrossRef]
30. Xie, B.; Li, S.; Li, M.; Liu, C.H.; Huang, G.; Wang, G. SePiCo: Semantic-Guided Pixel Contrast for Domain Adaptive Semantic Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 9004–9021. [CrossRef] [PubMed]
31. Ding, X.; Zhang, Y.; Ge, Y.; Zhao, S.; Song, L.; Yue, X.; Shan, Y. UniRepLKNet: A Universal Perception Large-Kernel ConvNet for Audio Video Point Cloud Time-Series and Image Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; pp. 5504–5514. [CrossRef]
32. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable DETR: Deformable Transformers for End-to-End Object Detection. *arXiv* **2021**. [CrossRef]
33. Dai, Z.; Cai, B.; Lin, Y.; Chen, J. UP-DETR: Unsupervised Pre-Training for Object Detection with Transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 1601–1610. [CrossRef]
34. Yao, Z.; Ai, J.; Li, B.; Zhang, C. Efficient DETR: Improving End-to-End Object Detector with Dense Prior. *arXiv* **2021**. [CrossRef]
35. Meng, D.; Chen, X.; Fan, Z.; Zeng, G.; Li, H.; Yuan, Y.; Sun, L.; Wang, J. Conditional DETR for Fast Training Convergence. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 3651–3660. [CrossRef]
36. 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.
37. Zhang, J.; Xia, K.; Huang, Z.; Wang, S.; Akindele, R.G. ETAM: Ensemble Transformer with Attention Modules for Detection of Small Objects. *Expert Syst. Appl.* **2023**, *224*, 119997. [CrossRef]
38. Chen, S.; Zhao, J.; Zhou, Y.; Wang, H.; Yao, R.; Zhang, L.; Xue, Y. Info-FPN: An Informative Feature Pyramid Network for Object Detection in Remote Sensing Images. *Expert Syst. Appl.* **2023**, *214*, 119132. [CrossRef]
39. Jiang, L.; Yuan, B.; Du, J.; Chen, B.; Xie, H.; Tian, J.; Yuan, Z. MFFSODNet: Multiscale Feature Fusion Small Object Detection Network for UAV Aerial Images. *IEEE Trans. Instrum. Meas.* **2024**, *73*, 1–14. [CrossRef]
40. Zhang, Z.; Liu, Y.; Liu, T.; Lin, Z.; Wang, S. DAGN: A Real-Time UAV Remote Sensing Image Vehicle Detection Framework. *IEEE Geosci. Remote Sens. Lett.* **2020**, *17*, 1884–1888. [CrossRef]
41. Liu, J.; Zhao, D.; Shen, J.; Geng, P.; Zhang, Y.; Yang, J.; Zhang, Z. HRD-Net: High Resolution Segmentation Network with Adaptive Learning Ability of Retinal Vessel Features. *Comput. Biol. Med.* **2024**, *173*, 108295. [CrossRef]
42. Chen, J.; Kao, S.; He, H.; Zhuo, W.; Wen, S.; Lee, C.-H.; Chan, S.-H.G. Run, Don't Walk: Chasing Higher FLOPS for Faster Neural Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 12021–12031. [CrossRef]
43. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
44. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv* **2017**. [CrossRef]
45. 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 (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 4002–4011. [CrossRef]
46. Zhang, H.; Wang, Y.; Dayoub, F.; Sunderhauf, N. VarifocalNet: An IoU-Aware Dense Object Detector. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 8514–8523. [CrossRef]
47. Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; Savarese, S. Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 658–666. [CrossRef]
48. Zhang, H.; Zhang, S. Focaler-IoU: More Focused Intersection over Union Loss. *arXiv* **2024**. [CrossRef]
49. Ma, S.; Xu, Y. MPDIoU: A Loss for Efficient and Accurate Bounding Box Regression. *arXiv* **2023**. [CrossRef]
50. 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/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 3974–3983. [CrossRef]

51. Qin, D.; Leichner, C.; Delakis, M.; Fornoni, M.; Luo, S.; Yang, F.; Wang, W.; Banbury, C.; Ye, C.; Akin, B.; et al. MobileNetV4–Universal Models for the Mobile Ecosystem. *arXiv* **2024**. [CrossRef]
52. Dong, X.; Bao, J.; Chen, D.; Zhang, W.; Yu, N.; Yuan, L.; Chen, D.; Guo, B. CSWin Transformer: A General Vision Transformer Backbone with Cross-Shaped Windows. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 12124–12134. [CrossRef]
53. Wang, A.; Chen, H.; Lin, Z.; Han, J.; Ding, G. Rep ViT: Revisiting Mobile CNN From ViT Perspective. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; pp. 15909–15920. [CrossRef]
54. 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 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; pp. 15464–15475. [CrossRef]
55. Feng, C.; Zhong, Y.; Gao, Y.; Scott, M.R.; Huang, W. TOOD: Task-Aligned One-Stage Object Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 3490–3499. [CrossRef]
56. Lyu, C.; Zhang, W.; Huang, H.; Zhou, Y.; Wang, Y.; Liu, Y.; Zhang, S.; Chen, K. RTMDet: An Empirical Study of Designing Real-Time Object Detectors. *arXiv* **2022**. [CrossRef]
57. Zhang, S.; Chi, C.; Yao, Y.; Lei, Z.; Li, S.Z. Bridging the Gap Between Anchor-Based and Anchor-Free Detection via Adaptive Training Sample Selection. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 9756–9765. [CrossRef]
58. Xie, X.; Lang, C.; Miao, S.; Cheng, G.; Li, K.; Han, J. Mutual-Assistance Learning for Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 15171–15184. [CrossRef]
59. Liu, S.; Li, F.; Zhang, H.; Yang, X.; Qi, X.; Su, H.; Zhu, J.; Zhang, L. DAB-DETR: Dynamic Anchor Boxes Are Better Queries for DETR. *arXiv* **2022**. [CrossRef]
60. Zhang, H.; Li, F.; Liu, S.; Zhang, L.; Su, H.; Zhu, J.; Ni, L.M.; Shum, H.-Y. DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection. *arXiv* **2022**. [CrossRef]