

## Article

# Hybrid-DETR: A Differentiated Module-Based Model for Object Detection in Remote Sensing Images

Mingji Yang <sup>1,\*</sup>, Rongyu Xu <sup>1</sup>, Chunyu Yang <sup>2,3</sup>, Haibin Wu <sup>1</sup>  and Aili Wang <sup>1</sup> 

<sup>1</sup> Heilongjiang Province Key Laboratory of Laser Spectroscopy Technology and Application, Harbin University of Science and Technology, Harbin 150080, China; 2220610137@stu.hrbust.edu.cn (R.X.); woo@hrbust.edu.cn (H.W.); aili925@hrbust.edu.cn (A.W.)

<sup>2</sup> School of Astronautics, Harbin Institute of Technology, Harbin 150001, China; sparkyyy@163.com

<sup>3</sup> The 54th Research Institute of China Electronics Technology Group Corporation, Shijiazhuang 050081, China

\* Correspondence: yangmingji@hrbust.edu.cn

**Abstract:** Currently, embedded unmanned aerial vehicle (UAV) systems face significant challenges in balancing detection accuracy and computational efficiency when processing remote sensing images with complex backgrounds, small objects, and occlusions. This paper proposes the Hybrid-DETR model based on a real-time end-to-end Detection Transformer (RT-DETR), featuring a novel HybridNet backbone network that implements a differentiated hybrid structure through lightweight RepConv Cross-stage Partial Efficient Layer Aggregation Network (RCSPPELAN) modules and the Heat-Transfer Cross-stage Fusion (HTCF) modules, effectively balancing feature extraction efficiency and global perception capabilities. Additionally, we introduce a Small-Object Detection Module (SODM) and an EIFI module to enhance the detection capability of small objects in complex scenarios, while employing the Focaler-Shape-IoU loss function to optimize bounding box regression. Experimental results on the VisDrone2019 dataset demonstrate that Hybrid-DETR achieves mAP50 and mAP50:95 scores of 52.2% and 33.3%, respectively, representing improvements of 5.2% and 4.3% compared to RT-DETR-R18, while reducing model parameters by 29.33%. The effectiveness and robustness of our improved method are further validated on multiple challenging datasets, including AI-TOD and HIT-UAV.



**Citation:** Yang, M.; Xu, R.; Yang, C.; Wu, H.; Wang, A. Hybrid-DETR: A Differentiated Module-Based Model for Object Detection in Remote Sensing Images. *Electronics* **2024**, *13*, 5014. <https://doi.org/10.3390/electronics13245014>

Academic Editor: Toshifumi Moriyama

Received: 5 November 2024

Revised: 17 December 2024

Accepted: 19 December 2024

Published: 20 December 2024



**Copyright:** © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

**Keywords:** UAV remote sensing images; RT-DETR; small-object detection; feature fusion; Focaler-Shape-IoU

## 1. Introduction

With the increasing volume and quality of remote sensing images, object detection has been widely applied in fields such as traffic monitoring [1], military reconnaissance [2], and rescue operations [3]. The methods based on optical images can automate the processing of large-scale aerial data by combining advanced deep learning algorithms, achieving accurate target identification and tracking in complex, dynamic scenes. However, as application requirements continue to increase, particularly in the domain of UAV remote sensing, there is a growing demand for higher accuracy in target detection. UAV target detection still faces significant challenges, including the following:

1. Small-sized targets and complex backgrounds: targets in remote sensing images are often small, making it difficult for detection models to extract sufficient information from these small-object images [4], thereby reducing detection accuracy. Additionally, remote sensing images typically feature complex backgrounds, including buildings, vegetation, vehicles, and other elements, whose complexity makes it challenging for object detection models to effectively distinguish the targets from the background;
2. Varying scales and environmental conditions: the significant variation in shooting angles and flying altitudes of UAVs leads to a wide range of target scales, creating

inconsistencies in object appearances across images. Furthermore, the changes in lighting and weather conditions also impact detection performance;

3. Long-tailed data distribution: the common targets are represented in larger quantities in UAV aerial image target detection datasets, while small or rare targets are less frequently seen, resulting in a noticeable long-tail distribution. This imbalanced data distribution limits the generalization ability of deep learning models, leading to good performance for common objects but poor detection of rare or small targets, which severely affects the overall accuracy of small-object detection.

In recent years, YOLO (You Only Look Once) series methods [5–10] have achieved remarkable progress in object detection, gaining widespread adoption in industrial applications due to their superior performance and practicality. The iterative version YOLOv8 [8] has achieved an impressive balance between detection speed and accuracy, demonstrating exceptional overall performance. Meanwhile, the emergence of RT-DETR [11], a transformer-based end-to-end real-time detector, has pioneered new research directions. Compared to YOLO models of similar scale, RT-DETR demonstrates advantages with relatively fewer parameters and lower computational costs, while achieving superior speed and accuracy performance. However, in specific scenarios such as UAV remote sensing imagery, RT-DETR still faces technical challenges that require further investigation.

RT-DETR employs ResNet [12] as its backbone network for feature extraction, utilizing a transformer encoder for global correlation at the final layer of the feature extraction network. In the neck network, a CCFM structure similar to PAFPN [13] is established, where fusion modules merge feature maps through top-down and bottom-up pathways. After selecting top-k queries, the dense predictions from fusion modules are fed into the decoder for final predictions. However, in UAV aerial imagery applications, several limitations persist: first, the multi-level downsampling modules result in substantial information loss regarding small objects, hindering the accurate capture of fine-grained features [14]. Second, the high correlation between feature maps leads to redundant computations in standard convolutions [15]. Furthermore, the neck network's traditional pyramid structure fails to effectively utilize the correlations between different feature maps, resulting in suboptimal fusion performance. Finally, the model's Generalized Intersection over Union (GIoU) loss degenerates to IoU when predicted boxes overlap with ground truth boxes, leading to increased vertical errors and slower convergence rates [16].

In summary, we propose a target detection model called Hybrid-DETR to address the aforementioned challenges. The main improvements and innovations of this article are summarized as follows:

1. SODM integrated into CCFM for enhanced small object detection: enhances the detection capability of small objects in UAV remote sensing imagery;
2. Differentiated module design strategy: this study presents HybridNet, a backbone network based on CSPDarknet [17], with tailored module designs for different depths. In the shallow layers, the lightweight RGCSPELAN module ensures efficient feature extraction while minimizing computational cost. In the deeper layers, the HTCF module, incorporating the Heat Conduction Operator (HCO) [18] into the CSP module, enhances global information capture and maintains local feature extraction, addressing the inefficiencies of Vision Transformers (ViTs) [19]. This design achieves a balance between accuracy and efficiency;
3. Efficient additive attention-based intrascale feature interaction (EIFI) module: incorporates efficient additive attention (EAA) [20] into the attention-based intrascale feature interaction (AIFI) module to form EIFI, achieving improved small-object detection under equivalent computational resources;
4. Loss function optimization: develops Focaler-Shape-IoU loss function by combining Shape-IoU [21] with Focaler-IoU [22], improving bounding box regression accuracy through adaptive focus on samples of varying difficulty;

The remainder of this paper is organized as follows: Section 2 briefly introduces relevant prior research. Section 3 focuses on HybridNet, while also discussing the integra-

tion of SODM into CCFM for enhanced small-object detection, as well as the RCSPELAN, HTCF, EIFI modules, and Focaler-Shape-IoU. Section 4 presents the experimental setup, including the environmental parameters, datasets, backbone network comparison, loss function comparison, model comparison, generalization experiments, and visualization analysis. Section 5 provides an in-depth analysis of the ablation experiments, evaluating the effectiveness of the proposed improvements. Finally, Section 7 summarizes the findings and discusses potential directions for future work.

## 2. Related Work

In recent years, Convolutional Neural Networks (CNNs) have demonstrated remarkable performance in computer vision, driving extensive research in deep learning. Object detection methods based on CNNs can be broadly categorized into two-stage and single-stage approaches. As deep-learning-based object detection technologies continue to evolve, two-stage detection algorithms, exemplified by the R-CNN series [23–27], enhance detection accuracy by first selecting candidate regions followed by refined object classification and localization. Conversely, single-stage detection algorithms, such as SSD [28] and YOLO series [5–10], reduce computational steps to achieve efficient real-time detection. Although both approaches have their distinct characteristics, they require threshold filtering and non-maximum suppression (NMS) [29], which to some extent compromises model robustness and detection speed. Moreover, these CNN-based methods still face challenges in complex scenarios, particularly in UAV aerial image object detection tasks.

UAV aerial images typically contain small objects, multi-scale overlapping objects, and complex backgrounds, posing significant challenges to traditional CNN-based methods. While CNNs excel at extracting local features due to their inherent nature, they have inherent limitations in capturing global image information and long-range relationships. Although increasing network depth and expanding receptive fields can partially mitigate this issue, such strategies often lead to dramatic increases in model complexity and optimization difficulties, potentially causing problems such as gradient vanishing.

Transformers [30] possess natural advantages over traditional CNNs in processing sequential data, capturing global dependencies through multi-head self-attention mechanisms, which enables better understanding of complex relationships between objects [31], thus compensating for CNNs' limitations in global information extraction. The Vision Transformer (ViT) model [19] pioneered the use of a pure transformer architecture for image classification, outperforming CNN networks when trained on large-scale datasets. In 2020, Facebook AI Research introduced DETR [32], the first end-to-end baseline network applying transformers to object detection. DETR reformulates object detection as a set prediction problem, streamlining the detection process by eliminating dependence on manually designed components such as anchor generation and non-maximum suppression. However, DETR's high computational requirements limit its practicality, especially in scenarios demanding real-time performance.

Consequently, numerous DETR variants have emerged to address the challenges of high parameter count and computational cost. Wang et al. proposed the Deformable DETR [33] in 2020, improving performance by focusing attention on key sampling points around reference points, reducing training time by 90% compared to DETR. Li et al. introduced DN-DETR [34] in 2022, utilizing query denoising mechanisms to eliminate convergence issues caused by Hungarian matching ambiguity. Chen et al. proposed Group DETR [35] in 2022, introducing multiple object queries to maintain DETR's end-to-end inference advantages while leveraging single-to-multiple training benefits to enhance performance and accelerate convergence. Zhang et al. presented DINO [36] in 2022, integrating and improving previous work by significantly reducing model size through enhanced encoder and decoder designs. While these works have addressed transformers' speed limitations, their inherent high computational complexity still presents significant limitations for real-time UAV remote sensing image object detection tasks.

To address DETR's real-time performance issues, Lv et al. proposed the Real-time Detection Transformer (RT-DETR) [11] in 2023. They designed an efficient hybrid encoder that effectively processes multi-scale features by decoupling intrascale interactions and cross-scale fusion, enabling RT-DETR to outperform state-of-the-art YOLO detectors of similar scale in both speed and accuracy. However, significant limitations persist in detecting small objects within complex backgrounds [37].

### 3. Materials and Methods

#### 3.1. RT-DETR Model

RT-DETR introduces an innovative real-time, end-to-end object detection model. The model is composed of three main components: a backbone network, a hybrid encoder, and a transformer decoder with an auxiliary prediction head. The backbone network leverages convolutional architecture to extract prominent features at three distinct scales, with strides of 8, 16, and 32, respectively. The hybrid encoder incorporates an attention-based intrascale feature interaction (AIFI) module, which processes high-level features from the backbone, reducing the computational burden and enhancing processing speed without compromising performance. Furthermore, the encoder utilizes a Cross-scale Feature Fusion Module (CCFM) to integrate and interact with multi-scale features. To optimize the detection process, the encoder dynamically adjusts queries based on Intersection over Union (IoU), selecting a set of image features most relevant to the detection targets, which are then used as the initial queries for the decoder. The decoder, aided by the auxiliary prediction head, iteratively refines the object queries to produce bounding boxes and confidence scores. Figure 1 illustrates the core architecture of the RT-DETR model.

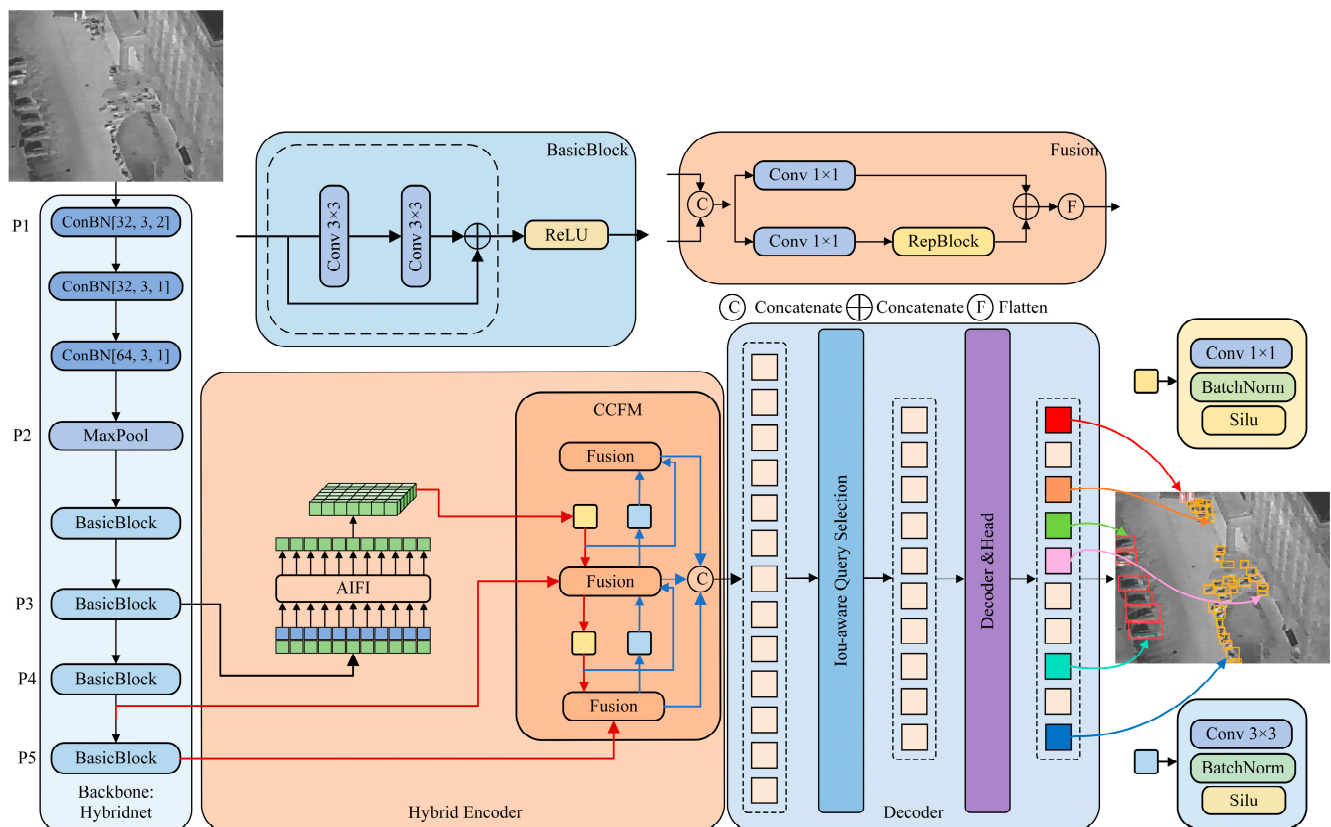


**Figure 1.** RT-DETR network structure diagram.

#### 3.2. Hybrid-DETR Model

In this study, we propose a novel network called Hybrid-DETR based on RT-DETR. As shown in Figure 2, the backbone network draws inspiration from the CSPDarknet

(Cross-stage Partial Darknet) framework [17], which maximizes gradient flow differences by partitioning gradient streams, effectively reducing redundant gradient learning. Within the backbone, we introduce an innovative differential network module called Hybrid-Net, aimed at optimizing feature extraction efficiency at different levels of the network. This strategy employs a different module design at shallow and deep network layers to balance fine-grained feature extraction with global perception capability. The module effectively establishes long-range dependencies, enhancing the capture of global semantic information, which in turn improves large target detection performance. Moreover, the neck introduces a specialized layer for small-object detection, which is fused with the P2 layer of the RCSPELAN module. This improves performance while introducing only a tolerable increase in parameters, thus effectively balancing performance and efficiency. We further incorporate an efficient additive attention mechanism into the transformer encoder, significantly reducing computational costs while maintaining or enhancing performance. Lastly, the original GIoU loss was substituted by the Focaler-Shape-IoU loss, which dynamically adjusts to balance difficult and easy samples, particularly excelling in detecting small and irregularly shaped UAV aerial targets. This multi-level, multi-angle optimization strategy aims to balance model size and detection accuracy, while enhancing scalability and adaptability in various application scenarios, forming a more powerful and efficient aerial surveillance system.

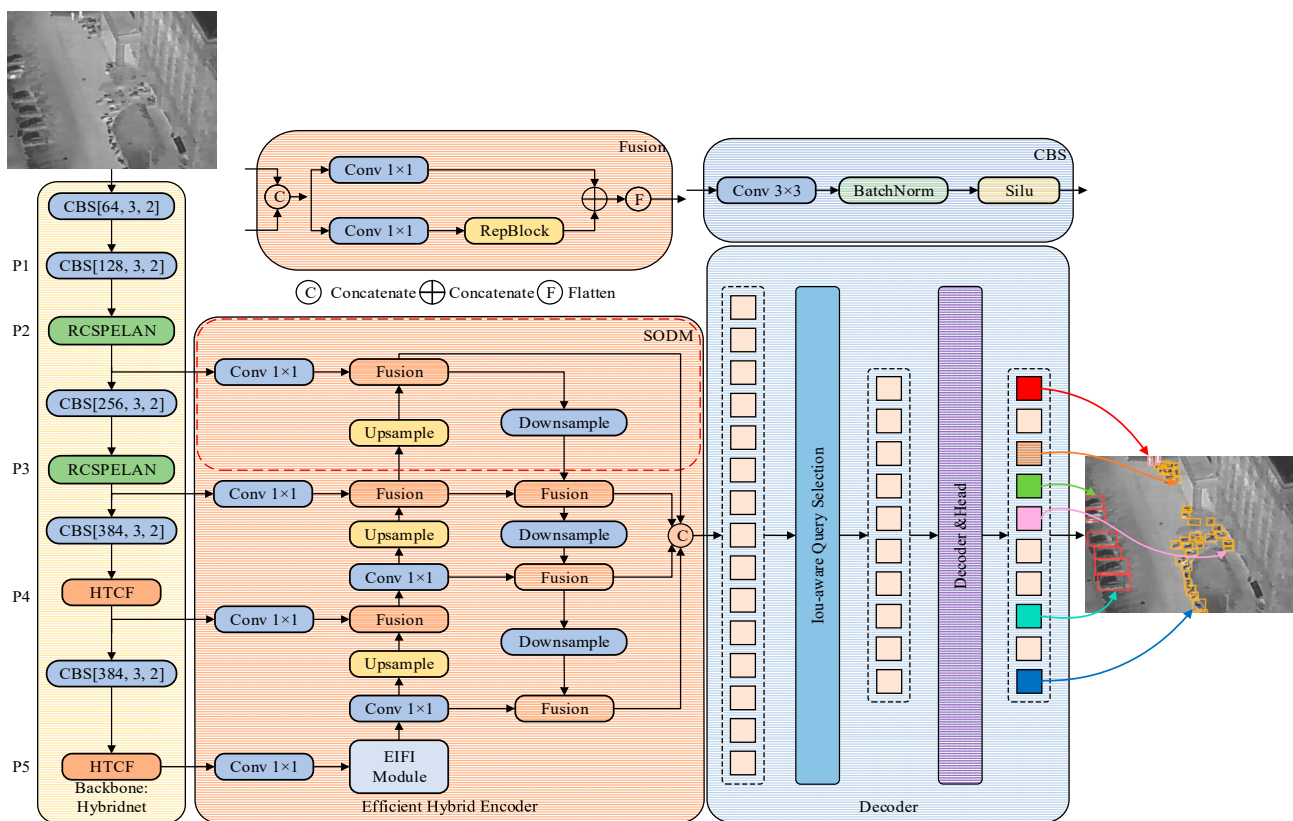


**Figure 2.** Hybrid-RTDETR network structure diagram.

### 3.2.1. SODM Integrated into CCFM for Enhanced Small-Object Detection

In UAV aerial image object detection tasks, factors such as imaging distance, high-speed flight, and low-altitude operations frequently result in motion blur, causing target objects to become smaller and more difficult to detect, thereby affecting network training efficiency. Although the original RT-DETR model employs detection heads at three scales ( $80 \times 80$ ,  $40 \times 40$ , and  $20 \times 20$ ), its high downsampling rate renders small objects indistinguishable, often merging with the background. To address this challenge, we propose

the Small-Object Detection Module (SODM). While the stacked convolutional layers in the backbone network expand the receptive field to capture comprehensive image details, excessive downsampling may lead to the loss of fine-grained information related to small objects. By integrating SODM into the neck network, we provide high-quality feature maps for both top-down and bottom-up branches of PAFPN [13]. Within the model's efficient hybrid encoder structure, SODM works in conjunction with attention-based intrascale feature interaction modules and CCFM to fuse rich small-object semantic information from shallow features with global information from deep semantic layers at the  $160 \times 160$  scale. This innovative architecture effectively overcomes the information loss typically associated with traditional convolution and downsampling processes, significantly enhancing both the model's resolution and small-object detection capabilities, thereby substantially improving detection performance in UAV aerial images. Figure 3 shows the P2 enhancement layer and PAFPN structure in Hybrid-DETR.

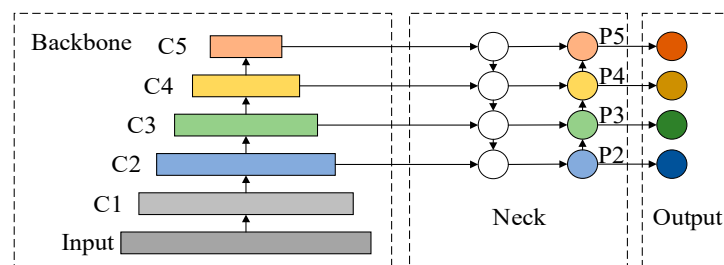


**Figure 3.** The P2 enhancement layer and PAFPN structure in Hybrid-DETR.

### 3.2.2. Lightweight Module Design: RCSPELAN

Introducing a specialized Small-Object Detection Module (SODM) in the neck component undoubtedly increases the model's complexity, leading to a rise in the number of parameters and GFLOPs. Given that UAV aerial image target detection applications often need to be deployed on resource-constrained edge devices, enhancing the model's lightweight nature while maintaining detection accuracy becomes particularly critical. Therefore, improving the model's lightweight characteristics without sacrificing detection accuracy is the primary issue to be addressed next. In this paper, we design a lightweight detection capability with high-speed inference in resource-constrained environments, especially focusing on small-object detection. We propose a module named RCSPELAN. The RCSPELAN module inherits the advantages of the CSPNet [38] structure in balancing computational cost and feature representation capability while optimizing gradient flow. Drawing on the ideas of ELAN, it enhances the network's scalability and flexibility by optimizing gradient length and hierarchical aggregation. Inspired by GhostNet [39], this module generates rich feature maps through inexpensive operations via RepConv [40], effectively extracting low-level visual features while significantly reducing computational complexity and optimizing the inference process. This integrated design not only greatly reduces computational redundancy and lowers the number of floating-point operations but also avoids the increased memory access issues that may arise from solely using methods like depthwise convolution (DWConv) [41]. This innovative design is particularly suitable for resource-constrained scenarios, such as UAV platforms, capable of providing efficient object detection performance while maintaining low latency and low computational complexity.

The structure of the module is detailed below. RCSPELAN integrates two neural network architectures, CSPNet [38] and ELAN [42], both employing gradient path optimization designs. It is a general and efficient layer aggregation network that balances lightweight design, fast inference speed, and high precision, as shown in Figure 4.

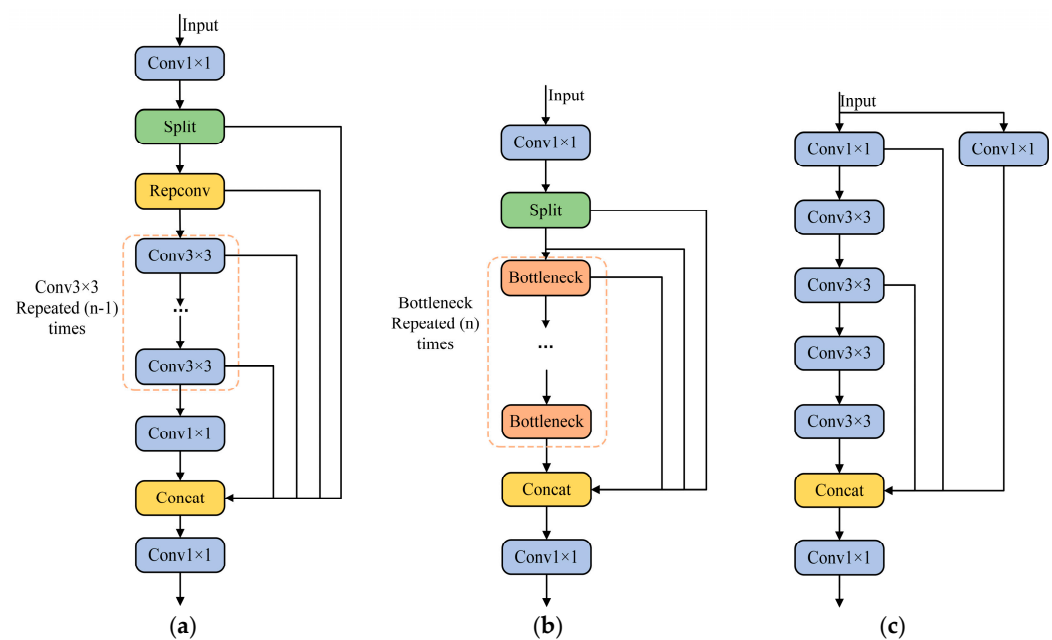


**Figure 4.** The structure of RGCSPELAN, CSPNet, and ELAN. (a) RGCSPELAN; (b) CSPNet; (c) ELAN.

CSPNet enhances gradient combination by splitting the feature maps of the base layer into two parts. One part undergoes computational blocks (such as Res blocks, ResX blocks, or Dense blocks), while the other part is directly combined with the processed feature maps in the next stage. This cross-stage local operation increases the diversity of learned features and reduces the overall computational load and memory usage. The gradient flow truncation operation introduces transition layers at the ends of the two paths, which helps increase the richness of gradient combinations and improve the network's learning ability. Therefore, CSPNet not only enhances gradient combination but also improves inference speed without sacrificing accuracy, making it suitable for real-time applications.

On the other hand, ELAN optimizes gradient length by effectively stacking computational blocks. This reduces the shortest gradient path and ensures effective gradient propagation. ELAN combines the hierarchical aggregation of VoVNet with the gradient optimization of CSPNet, improving the network's overall efficiency. By analyzing and optimizing gradient paths, ELAN ensures better convergence, especially for deep networks, and promotes effective gradient flow. This makes the network scalable and flexible, adaptable to various tasks and datasets without degrading performance.

In the RGCSPELAN module, RepConv (re-parameterizable convolution) serves as a key feature extraction and fusion module, cleverly balancing model performance and inference efficiency. During training, RepConv employs a multi-branch structure that includes  $3 \times 3$  convolutions,  $1 \times 1$  convolutions, and identity mappings. This design enhances feature extraction capability, improves gradient flow, and effectively compensates for the potential performance loss from discarding traditional residual blocks. In the inference phase, RepConv fuses the multi-branch structure into a single  $3 \times 3$  convolution through re-parameterization techniques, significantly improving computational efficiency and reducing memory access. This dual characteristic of complex training and simple inference makes RepConv particularly suitable for resource-constrained UAV platforms. It not only provides rich feature representation and good gradient propagation but also enhances model performance without increasing inference complexity. This innovative design offers a new solution for achieving efficient object detection in resource-constrained environments.

### 3.2.3. Efficient Long-Range Feature Extraction Module Design: HTCF

To address the limitations of CNNs in handling high-resolution images and complex scenes, and to establish long-range feature dependencies in the image space, we have

innovatively combined the Heat Conduction Operator (HCO) with CSP (cross-stage partial feature fusion module) to develop the Heat-Transfer–Cross-stage Fusion (HTCF) module. This fusion not only retains the computational efficiency advantages of CSPNet but also significantly enhances the network’s understanding of complex scenes through the global perception capability of HCO, performing exceptionally well in handling long-tail distribution data.

Moreover, the physical principles underlying HCO improve the model’s interpretability. As illustrated in Figure 5, the HTCF module retains the extended feature extraction capabilities of CSPNet and replaces the original bottleneck network structure with a Heat Conduction Operator neck (HCOneck).

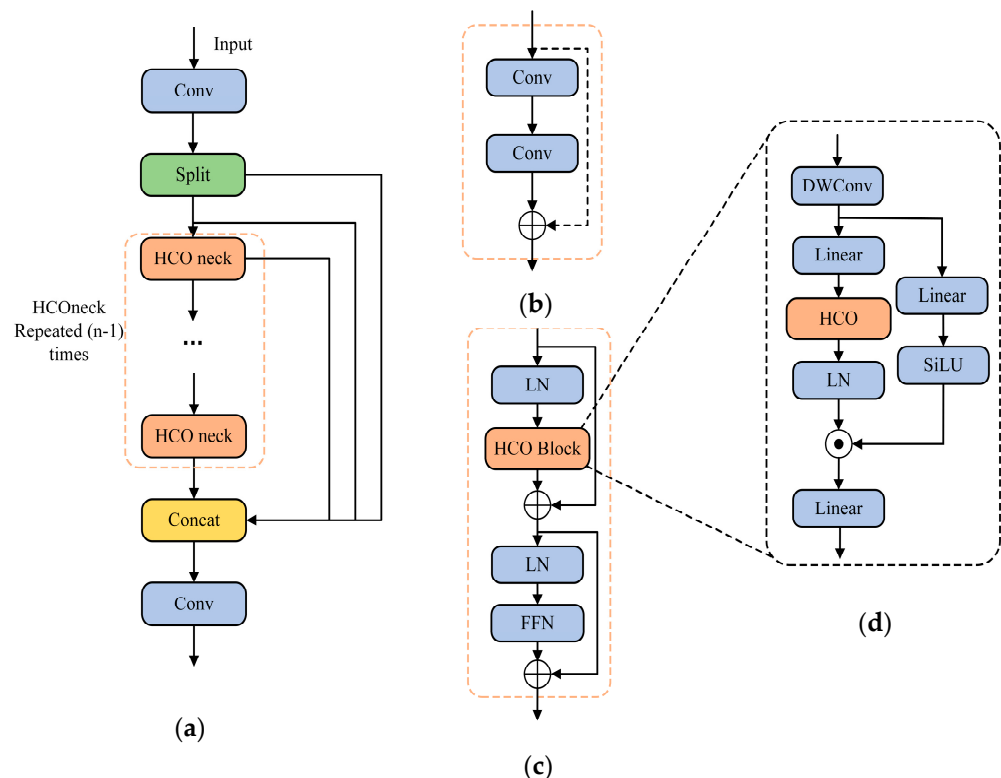


**Figure 5.** The structure of the HTCF, bottleneck, HCOneck, and HCOneck. (a) HTCF; (b) bottleneck; (c) HCOneck; (d) HCOblock.

This design introduces HCO, an attention mechanism that leverages the highly parallel characteristics of two-dimensional Discrete Cosine Transform (DCT) and Inverse Discrete Cosine Transform (IDCT). It not only significantly improves training and inference efficiency but also effectively expands the global receptive field. Simultaneously, it allows the module to capture global contextual information with a complexity of  $O(N^{1.5})$ , which is significantly better than the  $O(N^2)$  complexity of traditional self-attention mechanisms. Within the HCO Block (as shown in Figure 5d), depthwise separable convolution (DW-Conv), linear transformations, layer normalization (LN), and the SiLU activation function are seamlessly integrated to form a multi-branch structure. The SiLU activation function is specifically chosen for its smooth, non-monotonic nature and self-gating mechanism, which enables adaptive regulation of information flow while maintaining stable gradient propagation. Combining the beneficial properties of both ReLU and Sigmoid, SiLU is smooth and differentiable everywhere, facilitating enhanced gradient flow and potentially accelerating convergence during training. Compared to ReLU, SiLU better mitigates the issue of vanishing gradients in deeper networks. In the context of the HTCF module, SiLU plays a pivotal role in efficiently extracting both local and global features while ensuring computational stability, thus contributing to the overall effectiveness of the HCO Block.

This setup retains the ability for local feature extraction while enhancing the propagation of global information. Through this design, HCOneck greatly improves the network's capability to model multi-scale features and capture long-range dependencies while maintaining computational efficiency. This structure is not only applicable to various visual tasks but also provides a more efficient solution for processing high-resolution images.

### 3.2.4. Improvement of Feature Extraction Network: The HybridNet Architecture

This study proposes an innovative multi-scale feature extraction optimization strategy, drawing on the CSPDarknet framework as the backbone network. As shown in Figure 6, by implementing a differentiated module design across various depth levels of the network, the strategy not only achieves a comprehensive enhancement in feature extraction efficiency but is also specifically optimized to address the challenges posed by long-tailed datasets such as VisDrone. The core of this strategy lies in identifying and adapting to the unique requirements of each network layer, thereby achieving an optimal balance between performance and efficiency within the overall architecture.

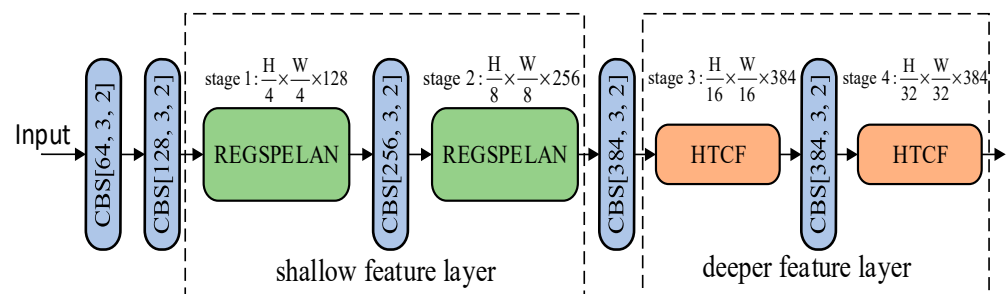


**Figure 6.** The structure of HybridNet.

The introduction of the Small-Object Detection Module (SODM) increases computational complexity; however, the lightweight RGCSPELAN module ensures high-quality feature extraction with low computational overhead, making it particularly suitable for resource-constrained edge computing devices. To balance performance and efficiency, the RGCSPELAN module is incorporated into the shallow layers of the network (P2 and P3), where high-resolution feature maps are prevalent. Introducing the computationally intensive HTCf module at these layers could marginally improve performance but would significantly increase computational load, potentially affecting the system's response speed and energy consumption in real-time applications such as unmanned aerial vehicles (UAVs). By contrast, the HTCf module excels in establishing long-range dependencies and integrating global semantic information, which is more effective when applied to the deeper layers (P4 and P5) that process high-level semantic features. This differentiation strategy not only optimizes performance in multi-scale object detection but also achieves a balance between computational efficiency and model expressiveness.

The design of HybridNet adopts a four-stage architecture inspired by the CSPDarknet framework, employing differentiated feature extraction modules at varying depth levels. In the shallow layers (P2 and P3), the RGCSPELAN module is applied to control the model's complexity while maintaining efficient feature extraction, avoiding the significant increase in parameters and GFLOPs associated with directly using the HTCf module on high-resolution feature maps. Conversely, in the deeper layers (P4 and P5), the HTCf module enhances detection accuracy with an acceptable trade-off in computational cost, leveraging the lower resolution of feature maps. This targeted modular design, applied across hierarchical levels, not only optimizes HybridNet's performance on UAV datasets with substantial variations in object scales and frequencies but also significantly improves its capability to handle long-tail distribution data. By integrating these design principles, HybridNet achieves an optimal balance between detection accuracy and computational complexity, forming a robust and efficient architecture for multi-scale object detection.

### 3.2.5. EIFI: An Enhanced Version of AIFI

We propose the EIFI module, which integrates efficient additive attention (EAA) [29] into the intrascale feature interaction module to enhance the computational efficiency and performance of the model, while maintaining its ability to detect complex scenes and small objects. The traditional multi-head self-attention (MHSA) mechanism computes the similarity between query, key, and value across multiple attention heads, followed by a weighted summation of these values. The specific formula is given as follows:

$$\text{MHSA}(\mathbf{X}) = \text{Softmax}\left(\frac{\mathbf{Q} \cdot \mathbf{K}^T}{\sqrt{d}}\right) \cdot \mathbf{V} \quad (1)$$

Specifically, self-attention takes an input  $\mathbf{X} \in \mathbb{R}^{n \times d}$ , where  $n$  denotes the number of tokens, and  $d$  represents the embedding dimension of each token. The input  $\mathbf{X}$  is projected into query ( $\mathbf{Q}$ ), key ( $\mathbf{K}$ ), and value ( $\mathbf{V}$ ) spaces through three learnable weight matrices  $\mathbf{W}_Q$ ,  $\mathbf{W}_K$ , and  $\mathbf{W}_V$ , each with dimensions  $\mathbb{R}^{d \times d}$ . The attention scores between each pair of tokens in  $\mathbf{Q}$  and  $\mathbf{K}$  are computed using the dot-product, followed by normalization and weighting via the Softmax function to quantify the interactions between tokens. The weighted interactions are then multiplied by  $\mathbf{V}$  to produce the final weighted output. Each self-attention layer comprises  $h$  heads, allowing the model to attend to different views of the input. Overall, the computational complexity of the self-attention mechanism is  $O(n^2 \cdot d)$ . As the number of tokens increases, the computational and memory demands for  $\mathbf{Q} \cdot \mathbf{K}^T$  grow quadratically, leading to slower inference speeds and high memory usage, making real-time processing of long sequences or high-resolution images infeasible.

In contrast, the proposed efficient additive attention employs linear transformations to replace matrix multiplication, thereby reducing computational complexity. The core formulation is given as follows:

$$\alpha = \frac{\mathbf{Q} \cdot \mathbf{w}_a}{\sqrt{d}} \quad (2)$$

$$\mathbf{q} = \sum_{i=1}^n \alpha_i * \mathbf{Q}_i \quad (3)$$

$$\text{EAA}(\mathbf{X}) = \hat{\mathbf{Q}} + \mathbf{T}(\mathbf{K} * \mathbf{q}) \quad (4)$$

where  $\mathbf{w}_a \in \mathbb{R}^d$  is a learnable weight vector,  $\hat{\mathbf{Q}}$  denotes the normalized query matrix,  $*$  denotes the element-wise multiplication operation, and  $\mathbf{T}$  is the linear transformation layer. Through this design, the computational complexity of EAA is reduced from  $O(n^2 \cdot d)$  to  $O(n \cdot d)$ , significantly improving computational efficiency while maintaining the ability to capture global context. Firstly, UAV-captured images often cover large areas, with small objects that are typically numerous. EAA, by establishing a more efficient global context, can accurately detect and recognize these small targets under the same computational resources. Additionally, due to the varying altitudes and angles of UAV flights, backgrounds in the captured images may undergo significant changes. EAA can swiftly adjust and adapt to such complex background variations by reducing unnecessary computations and focusing only on the most relevant features. Furthermore, UAVs often need to process large amounts of imagery continuously. EAA, with its linear complexity, can handle these long sequences more effectively, ensuring the model's practicality and high efficiency. Therefore, integrating EAA into frameworks like RT-DETR facilitates optimal speed–accuracy trade-offs in resource-constrained environments, making it particularly suitable for real-time target detection in UAV applications.

### 3.2.6. The Loss Function Focaler-Shape-IoU

In the field of object detection, the accuracy of bounding box regression has a decisive impact on the overall performance of the model. However, the existing bounding box regression loss functions, such as GIoU [43], CIoU [44] and Shape-IoU [21], have limitations in handling the difficulty differences among samples and in considering the shape

characteristics of bounding boxes. To address these issues, this study proposes a novel loss function Focaler-Shape-IoU. By integrating the strengths of Focaler-IoU and Shape-IoU, this loss function not only improves the accuracy of bounding box regression but also achieves adaptive focusing based on sample difficulty, thereby significantly enhancing the overall detection performance of the model. Focaler-IoU introduces a linear interval mapping mechanism to reconstruct the traditional IoU loss function. This improvement enables the loss function to adaptively adjust its focus according to the difficulty of the samples in different detection tasks, thereby more effectively handling challenging samples and improving detection performance. Meanwhile, Shape-IoU emphasizes the shape and size characteristics of bounding boxes, further enhancing regression accuracy. The definition diagram of Focaler-Shape-IoU is shown in Figure 7.

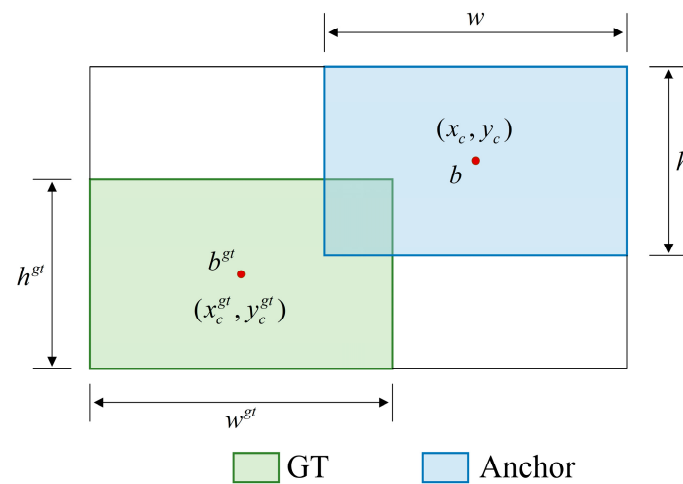


**Figure 7.** Focaler-Shape-IoU definition diagram.

The Focaler-Shape-IoU formula is written as follows:

$$\text{IOU}^{\text{focaler}} = \begin{cases} 0, & \text{IOU} < d \\ \frac{\text{IOU}-d}{u-d}, & d \leq \text{IOU} \leq u \\ 1, & \text{IOU} > u \end{cases} \quad (5)$$

$$w_0 = \frac{2 \times (w^{\text{gt}})^{\text{scale}}}{(w^{\text{gt}})^{\text{scale}} + (h^{\text{gt}})^{\text{scale}}} \quad (6)$$

$$h_0 = \frac{2 \times (h^{\text{gt}})^{\text{scale}}}{(w^{\text{gt}})^{\text{scale}} + (h^{\text{gt}})^{\text{scale}}} \quad (7)$$

$$d^{\text{shape}} = h_0 \times \frac{(x_c - x_c^{\text{gt}})^2}{c^2} + w_0 \times \frac{(y_c - y_c^{\text{gt}})^2}{c^2} \quad (8)$$

$$\Omega^{\text{shape}} = \sum_{t=w,h} (1 - e^{-w_t})^{\theta}, \theta = 4 \quad (9)$$

$$w_w = w_0 \times \frac{|w - w^{\text{gt}}|}{\max(w, w^{\text{gt}})} \quad (10)$$

$$h_h = h_0 \times \frac{|h - h^{\text{gt}}|}{\max(h, h^{\text{gt}})} \quad (11)$$

where  $u$  and  $d$  are tunable coefficients used to determine the attention between different IOU regions. Specifically, the width and height of the ground truth box are denoted as  $w^{\text{gt}}$  and  $h^{\text{gt}}$ , while  $\text{scale}$  represents a scaling factor associated with the size of the objects in the

dataset. The weighting coefficients  $W_0$  and  $H_0$  enable the model to adapt dynamically to the actual size of the object, adjusting the loss function's sensitivity to shape discrepancies. The center coordinates of the predicted box are denoted by  $x_c$  and  $y_c$ , and those of the ground truth box are represented by  $x_c^{gt}$  and  $y_c^{gt}$ , with  $c$  representing the diagonal length of the smallest enclosing box.

The shape discrepancy term  $d^{shape}$  is computed using the relative positions of the center points of the predicted and ground truth boxes through  $W_0$  and  $H_0$ . Furthermore, the shape loss weighting term  $\Omega^{shape}$  is balanced by the relative differences in width and height between  $w_w$  and  $w_h$ . The parameter  $\theta$ , which typically takes a value of 4, is used to optimize robustness. The final Focaler-Shape-IoU loss function leverages the advantages of both Focaler-IoU and Shape-IoU to achieve more precise boundary localization defined as follows:

$$\text{Loss} = 1 - \text{IOU}^{\text{focaler}} + d^{\text{shape}} + 0.5 \times \Omega^{\text{shape}} \quad (12)$$

## 4. Results

### 4.1. Experimental Environment Settings and Dataset Description

#### 4.1.1. Experimental Environment Settings

During the training process, the input image size is set to  $640 \times 640$ , the batch size is set to 4, and the number of epoch is set to 300. In this paper, the hyperparameters are based on RT-DETR, and the first 300 coded features are selected to initialize the object query module of the decoder. Our detector is trained using the AdamW optimizer,  $\text{base\_learning\_rate} = 0.0001$ ,  $\text{weight\_decay} = 0.0001$ ,  $\text{Weight decay} = 0.0001$ ,  $\text{linear\_warmup\_steps} = 2000$ ,  $\text{patience} = 30$ , and  $\text{Momentum} = 0.9$ . To ensure the fairness and comparability of model performance, this study adopts hyperparameters based on RT-DETR. No pre-trained weights were used in any of the training processes for all models involved in ablation and comparative experiments. The detailed experimental setup parameters are shown in Table 1.

**Table 1.** Experimental setup parameters.

| Parameter            | Value        |
|----------------------|--------------|
| Operating System     | Windows 11   |
| Programming Language | Python3.8.19 |
| GPU                  | 4080S        |
| GPU Memory           | 16 G         |
| CPU                  | Intel-13700K |

#### 4.1.2. Evaluation Indicators

In this study, we evaluate the algorithm's performance by comparing the differences between the pre-improved and post-improved models under identical experimental settings for image detection. Precision (P), recall (R), mean Average Precision (mAP), GFLOPs, and frames per second (FPS) as the evaluation metrics for model performance. The details are listed below.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (13)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (14)$$

$$\text{AP} = \int_0^1 P(R) dR \quad (15)$$

$$\text{mAP} = \frac{1}{N} \sum_{n=1}^N \text{AP}(n) \quad (16)$$

Specifically, TP, FP, and FN represent True Positive, False Positive, and False Negative detections, respectively. Recall refers to the recall rate, and Precision refers to accuracy. P(R) represents the precision–recall curve. N represents the number of classes, and mAP is the mean of the AP values for all classes. To evaluate the network’s efficiency, model parameters Params(M) and computational complexity GFLOPs(G) are used to assess the effectiveness of the network in UAV application scenarios.

#### 4.1.3. Experimental Datasets

To evaluate the performance of our proposed method, we conducted experiments on three UAV image datasets: VisDrone2019 [45], AI-TOD [46], and HIT-UAV [47]. Each dataset has distinct characteristics:

1. VisDrone2019 emphasizes small-object detection and object occlusion issues. This dataset is released by the Machine Learning and Data Mining Laboratory at Tianjin University. It includes 10 categories of aerial detection targets, among which small objects (less than  $32 \times 32$  pixels) account for up to 44.7%, and nearly 50% of the targets are occluded;
2. AI-TOD contains annotations for eight target categories, primarily focusing on tiny object detection problems. The average target size in this dataset is only 12.8 pixels, significantly smaller than those in other datasets;
3. HIT-UAV is the first publicly available high-altitude UAV-based infrared thermal image dataset, concentrating on object detection under night-time and low-light conditions. It includes dense and small-sized objects from five categories (including humans, vehicles, and bicycles), captured at varying flight altitudes from 60 to 130 m and camera angles from 30 to 90 degrees. This dataset expands the application scope of UAVs under low-light conditions.

The number of instances for each category in the three datasets can be organized as shown in Table 2. These datasets cover a wide range of target types, from tiny objects to large ground facilities. This offers a comprehensive testing foundation for evaluating the accuracy and robustness of object detection algorithms.

**Table 2.** Class distribution of VisDrone2019, AI-TOD, and HIT-UAV datasets.

| Dataset | VisDrone2019    |           | AI-TOD        |           | HIT-UAV      |           |
|---------|-----------------|-----------|---------------|-----------|--------------|-----------|
|         | Class           | Instances | Class         | Instances | Class        | Instances |
| 1       | pedestrian      | 109,187   | airplane      | 793       | Car          | 7311      |
| 2       | people          | 38,966    | bridge        | 652       | Person       | 12,312    |
| 3       | bicycle         | 13,069    | storage-tank  | 7757      | OtherVehicle | 148       |
| 4       | car             | 187,005   | ship          | 17,332    | Bicycle      | 4980      |
| 5       | van             | 32,702    | swimming-pool | 327       | DontCare     | 148       |
| 6       | truck           | 16,284    | vehicle       | 307,992   |              |           |
| 7       | tricycle        | 6387      | person        | 17,968    |              |           |
| 8       | awning-tricycle | 4377      | wind-mill     | 243       |              |           |
| 9       | bus             | 9117      |               |           |              |           |
| 10      | motor           | 40,378    |               |           |              |           |

#### 4.2. Performance Comparison of Feature Network Fusion Models

In both military and civilian applications, recognizing small objects in images captured by drones remains a significant challenge. These challenges stem from the need for drones to accurately locate targets of different scales within complex environments, which often involve severe occlusion. To address these challenges, this study enhances the backbone network of the base model, introducing a proposed HybridNet as the backbone. Considering that incorporating a Small-Object Detection Module (SODM) significantly increases computational overhead, this study compares several representative backbone networks combined with SODM to demonstrate the superiority of using the HybridNet module in conjunction with small-object detection layers.

Given that the HTCF module excels in establishing long-range dependencies and integrating global semantic information, these characteristics are particularly effective when processing high-level semantic features. Therefore, the HTCF module is applied only in the deeper layers of the network (P4 and P5) and is omitted in the P2 and P3 layers. Experimental results for different fusion strategies are presented in Table 3.

**Table 3.** Comparison of experimental results of different feature network fusion models.

| Hybrid Net1 | Hybrid Net2 | Hybrid Net3 | Hybrid Net4 | P     | R     | mAP50 | mAP50:95 | Params (M) | GFLOPs (G) |
|-------------|-------------|-------------|-------------|-------|-------|-------|----------|------------|------------|
| -           | -           | -           | -           | 0.634 | 0.486 | 0.504 | 0.315    | 18.60      | 78.2       |
| ✓           |             |             |             | 0.627 | 0.496 | 0.514 | 0.325    | 12.90      | 72.9       |
|             | ✓           |             |             | 0.631 | 0.502 | 0.518 | 0.328    | 14.37      | 73.4       |
|             |             | ✓           |             | 0.635 | 0.496 | 0.514 | 0.324    | 12.59      | 66.5       |
|             |             |             | ✓           | 0.638 | 0.494 | 0.518 | 0.324    | 14.05      | 67.0       |

Note: HybridNet1 represents replacing the CSP module in the P4 and P5 layers of the backbone with RCSPPELAN and fusing with SODM. HybridNet2 represents replacing the CSP module in the P4 and P5 layers of the backbone with HTCF and fusing with SODM. HybridNet3 represents replacing the CSP module in the P2–P5 layers of the backbone with RCSPPELAN and fusing with SODM. HybridNet4 represents replacing the CSP module in the P2 and P3 layers of the backbone with RCSPPELAN, replacing the CSP module in the P4 and P5 layers with HTCF, and finally fusing with SODM.

Table 3 shows that compared to the baseline model, HybridNet4 demonstrated significant improvements in both efficiency and performance: model parameters were reduced by 24.46%, GFLOPs decreased by 14.32%, and mAP50 increased by 1.4%. This result strongly supports the superiority of the proposed hybrid architecture in balancing model lightweighting with detection accuracy. A comparative analysis reveals that while HybridNet2 achieves the same mAP50 as HybridNet4 and slightly higher mAP50:95, it comes at the cost of increased parameters and computational complexity. HybridNet3 which replaces the CSP modules in the P2–P5 layers entirely with RCSPPELAN has a slight advantage over HybridNet4 in terms of parameter count and computational load, but its mAP50 decreases by 0.4%. This loss in accuracy is insufficient for high-demand applications such as UAV infrared detection, underscoring the more balanced modular design strategy of HybridNet4.

Further research indicates that the application of the HTCF module in the deeper layers of the network (P4 and P5) can significantly improve detection accuracy, albeit at the cost of a slight increase in parameters and computational overhead. In contrast, the use of the RGCSPPELAN module in the shallower layers (P2 and P3) not only enhances accuracy but also substantially reduces the parameter count and computational complexity. Therefore, in the backbone network, this study adopts RGCSPPELAN for the P2 and P3 layers, while employing HTCF in the P4 and P5 layers, thereby establishing an efficient differentiated configuration strategy named the HybridNet architecture. Through this innovative module configuration strategy, the HybridNet architecture effectively addresses the increased computational complexity introduced by the Small-Object Detection Module (SODM), while significantly enhancing overall detection accuracy. This novel architecture is particularly suitable for application scenarios that require the deployment of high-performance object detection systems in environments with limited computational resources.

#### 4.3. Ablation Study on Loss Functions

To evaluate the performance of Focaler-shape-IOU, a quantitative comparative experiment was conducted between FocalerShape and GIoU (used by RTDETR), Shape-IOU, and other commonly used IOUs (GIoU, CIoU, and Shape-IOU). The comparison results are presented in Table 4. In this table, Focaler-Shape-IOU achieved the highest mAP50, validating the superiority of the improved model's loss function.

**Table 4.** Performance comparison of different loss functions on the VisDrone validation set.

| IoU     | GIoU  | CIoU  | Shape-IoU | Focaler-Shape-IoU |
|---------|-------|-------|-----------|-------------------|
| mAP50/% | 0.470 | 0.468 | 0.473     | 0.480             |

In the Focaler-Shape-IoU loss function, the parameter  $d = 0.0$  is used as the lower bound threshold, ensuring that IoU values below this threshold are directly set to 0, which helps in eliminating the impact of poorly overlapping boxes. The parameter  $u = 0.95$  serves as the upper bound, emphasizing high-quality overlaps and enhancing shape similarity optimization, particularly improving bounding box alignment in complex scenarios.

#### 4.4. Comparative Experimental Results and Analysis with Other SOTA Algorithms

To further validate the advantages of the improved algorithm, we conducted comparative experiments with other SOTA algorithms on the VisDrone2019 dataset, considering both the model size and detection performance of lightweight models. Eight advanced algorithms were selected for comparison: SSD, Faster R-CNN, YOLOv5m, YOLOv7, YOLOv8m, YOLOv9m, YOLOv10m, Deformable DETR, and RT-DETR-R18. The experimental parameters were set according to Table 1. The results of the comparative experiments are shown in Table 5.

**Table 5.** Performance comparison of different models on the VisDrone validation set.

| Models          | mAP50/% | mAP50-95/% | Params (M) | GFLOPS/G |
|-----------------|---------|------------|------------|----------|
| SSD             | 24.0    | 10.3       | 26.2       | 62.7     |
| Faster R-CNN    | 33.0    | 16.5       | 137.1      | 370.2    |
| YOLOv5m         | 37.2    | 19.5       | 21.2       | 49.0     |
| YOLOv7          | 45.8    | 23.7       | 37.2       | 103.3    |
| YOLOv8m         | 43.1    | 26.5       | 25.9       | 78.7     |
| YOLOv9m         | 44.2    | 27.1       | 20.3       | 77.0     |
| YOLOv10m        | 43.8    | 26.5       | 15.4       | 58.9     |
| Deformable DETR | 42.8    | 26.8       | 40.0       | 196.0    |
| RT-DETR-R18     | 47.0    | 28.7       | 19.9       | 57.0     |
| Ours            | 52.2    | 33.3       | 14.1       | 67.2     |

Our proposed Hybrid-DETR addresses these challenges by leveraging the global modeling strengths of transformers while introducing architectural optimizations tailored for UAV aerial object detection. Hybrid-DETR enhances the detection of small objects in complex backgrounds while maintaining a low computational cost and parameter count. Compared to SSD, Faster R-CNN, and the YOLO series, Hybrid-DETR exhibits superior performance in handling small-scale objects and multi-scale complexities. Moreover, it outperforms YOLOv8m and RT-DETR-R18 by further optimizing detection efficiency and accuracy through its innovative design.

In terms of mAP50, the Hybrid-DETR model outperformed SSD, Faster R-CNN, YOLOv5m, YOLOv7, YOLOv8m, YOLOv9m, and YOLOv10m by 28.2%, 19.2%, 15%, 6.4%, 9.1%, 8%, and 8.4%, respectively. These substantial improvements in mAP50 highlight the superior detection accuracy of the Hybrid-DETR model, especially in the detection of small objects in UAV-based Remote Sensing Images.

Compared to the Deformable DETR and RT-DETR-R18 models, the results of mAP50 are improved by 9.4% and 5.2%, respectively. These results underscore the effectiveness of the architectural optimizations and design choices in Hybrid-DETR, contributing to its enhanced detection accuracy.

#### 4.5. Generalization Experiments

To verify the applicability and generalization ability of the small-object detection by Hybrid-DETR algorithm, we conducted comparative experiments on the VisDrone2019, HIT-UAV, and AI-TOD datasets. The experimental results are shown in Table 6.

**Table 6.** Comparative experimental results on different datasets.

| Dataset      | Methods     | P     | R     | mAP50 | mAP50:95 |
|--------------|-------------|-------|-------|-------|----------|
| VisDrone2019 | RT-DETR-R18 | 0.619 | 0.455 | 0.47  | 0.287    |
| VisDrone2019 | Ours        | 0.639 | 0.505 | 0.522 | 0.33     |
| HIT_UAV      | RT-DETR-R18 | 0.883 | 0.753 | 0.819 | 0.527    |
| HIT_UAV      | Ours        | 0.914 | 0.733 | 0.826 | 0.537    |
| AITOD        | RT-DETR-R18 | 0.582 | 0.476 | 0.475 | 0.206    |
| AITOD        | Ours        | 0.595 | 0.508 | 0.491 | 0.225    |

The results of Table 6 indicate that Hybrid-DETR achieved substantial improvements in P, R, and mAP on the AI-TOD and HIT-UAV datasets. The model demonstrates significant advantages in detecting small objects and targets in low-light environments, particularly excelling in tasks involving dense and small-sized objects, such as people and vehicles, where it shows enhanced generalization capabilities. Although recall slightly decreased on the HIT-UAV dataset, the model's significant improvement in precision ensures high detection performance while reducing false positives. On the AI-TOD dataset, mAP50 increased by 1.6% and mAP50:95 improved by 1.9%, indicating the great potential and advantages of the Hybrid-DETR model in micro-object detection tasks.

#### 4.6. Visual Analysis

In the field of UAV-based remote sensing image object detection, we face numerous challenges, particularly in processing high-resolution images with large numbers of small objects, complex backgrounds, and occlusions. To address these challenges, we evaluated the performance of the improved model in various complex scenarios as shown in Figure 8. The experimental results demonstrate that the model can accurately identify small objects in complex scenes from various UAV aerial images, exhibiting strong robustness. Whether in the high-occlusion scenarios of VisDrone2019, the extremely small object detection tasks of AI-TOD, or the low-light conditions of HIT-UAV, the improved model demonstrated excellent detection capabilities. This not only proves the model's adaptability in tackling the specific challenges of different datasets but also highlights its potential in practical applications, providing reliable technical support for UAV object detection tasks in complex environments.

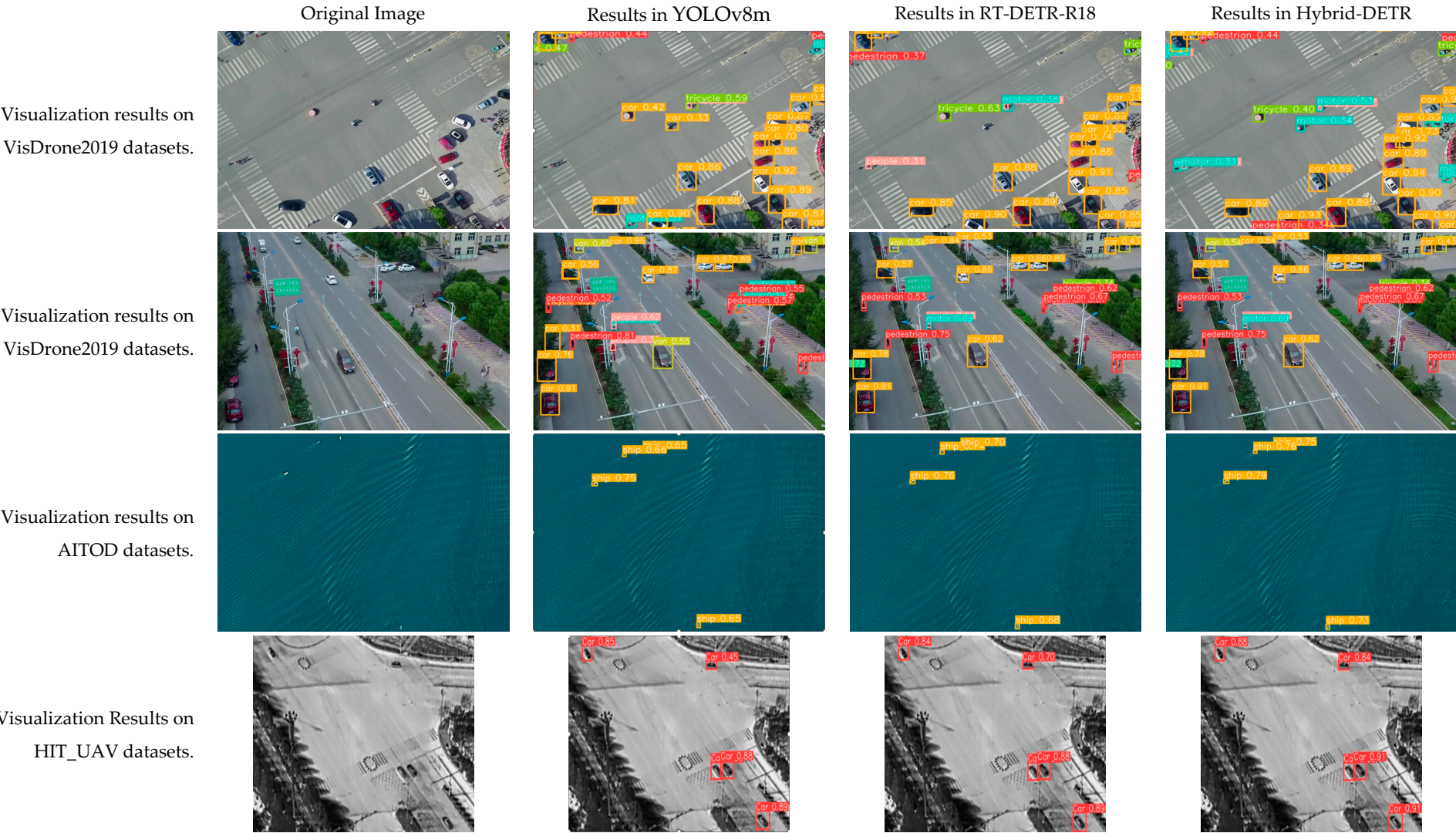


Figure 8. Visualization of detection results with different datasets.

The figures show a comparison of small-object detection between YOLOv8m, Hybrid-DETR, and RT-DETR-R18 on UAV aerial images across the three datasets. These images clearly illustrate that Hybrid-DETR surpasses the original model in terms of detection accuracy. In the first row of images, it is evident that the original model failed to detect some motorcycles and pedestrians, whereas Hybrid-DETR accurately identified all small objects, reducing both missed detections and false positives while achieving higher detection accuracy. In the second row, although Hybrid-DETR made an incorrect detection of a person, its detection accuracy was still noticeably superior to the original model. In the third row, we observe that Hybrid-DETR excels in detecting very small objects. In the fourth row, when handling infrared imaging of small objects, Hybrid-DETR also shows better detection performance compared to the original model. Overall, Hybrid-DETR demonstrates outstanding performance in processing high-precision images with large numbers of small objects, complex backgrounds, and occlusions.

#### 4.7. Comparison of Evaluation Metrics

Throughout the training process, we recorded the performance of each model on the training set in terms of precision, recall, mAP50, and mAP50:95, and plotted the corresponding curves to track training progress, as shown in Figure 9.

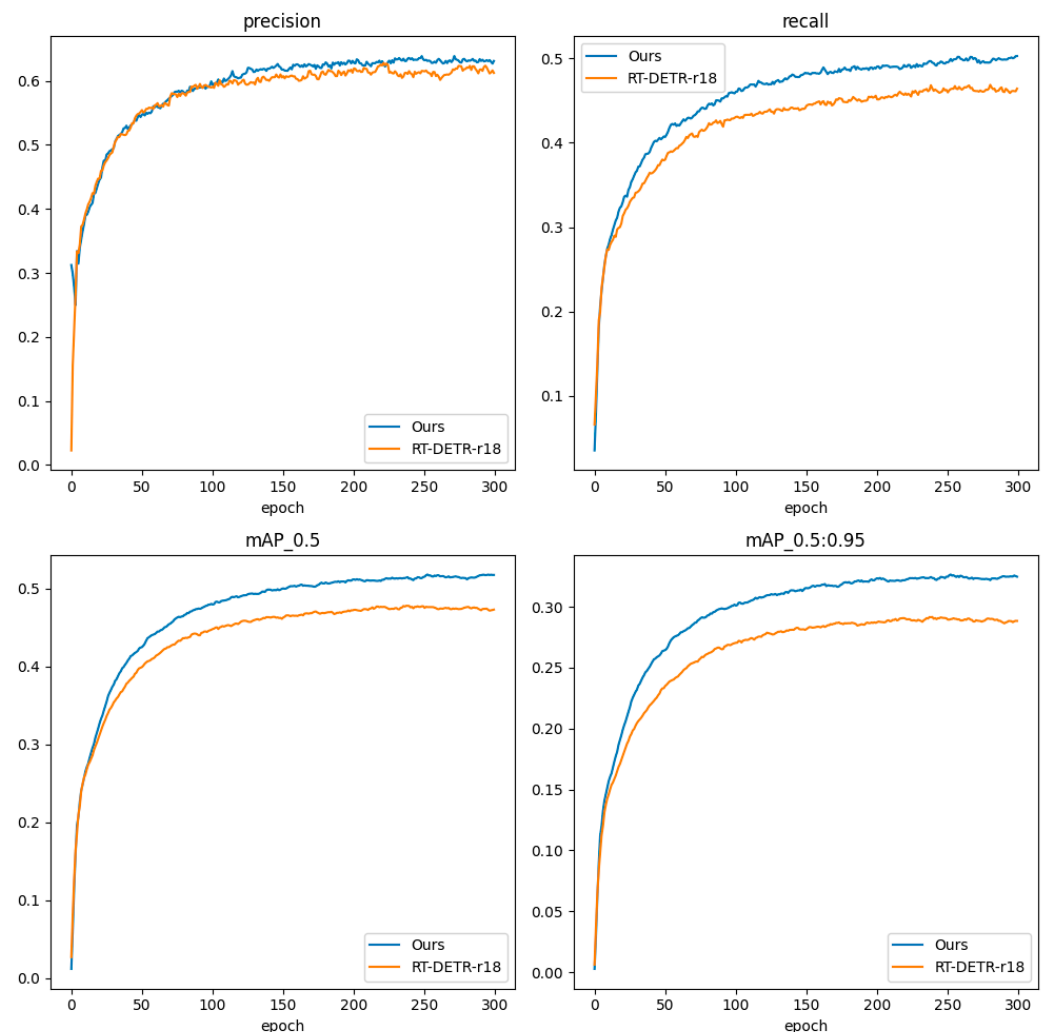


**Figure 9.** Comparison of evaluation metrics between Hybrid-DETR and RT-DETR-R18.

As shown in Figure 8, the results demonstrate that our model significantly outperforms the original RT-DETR-R18 in both precision and recall. The precision curve shows that our

model maintains high stability during training, quickly converging to a high precision level. In contrast, RT-DETR-R18 exhibits precision similar to ours in the early stages of training but experiences considerable fluctuations in later stages, eventually stabilizing at a slightly lower precision than our model.

In terms of recall, our model also performs excellently, with the curve showing rapid growth in the early stages of training and stable improvement thereafter, whereas RT-DETR-R18's recall increases more slowly. Additionally, in the object detection evaluations of mAP50 and mAP50:95, our model consistently outperforms RT-DETR-R18 across various confidence thresholds. Particularly at higher confidence thresholds, our model demonstrates greater stability and average precision, while RT-DETR-R18 performs noticeably worse on the mAP50:95 curve. Overall, the experimental results clearly show that our model not only surpasses RT-DETR-R18 in precision and recall but also offers significant advantages in detection robustness and overall performance.

## 5. Discussion

Based on the RT-DETR network architecture, we implemented several modifications to enhance its performance. First, we incorporated our proposed lightweight HybridNet as the feature extraction network. Additionally, we introduced a Small-Object Detection Module (SODM) to enrich the semantic representation of small objects, thereby optimizing the feature fusion network. Subsequently, we integrated the efficient additive attention (EAA) into the attention-based intrascale feature interaction (AIFI) module. Finally, we introduced our Focaler-Shape-IoU loss. To evaluate the effectiveness of each improvement, we conducted a series of ablation studies by sequentially adding these enhanced modules. The experimental results are presented in Table 7, demonstrating the incremental performance gains achieved through each modification.

**Table 7.** Ablation study results on the VisDrone validation set.

| Methods | HybridNet | SODM | EIFI | Focaler-Shape-IoU | P     | R     | mAP50 | mAP50:95 | Params (M) | GFLOPs (G) | FPS  |
|---------|-----------|------|------|-------------------|-------|-------|-------|----------|------------|------------|------|
| 1.base  |           |      |      |                   | 0.619 | 0.455 | 0.470 | 0.287    | 19.88      | 57.0       | 74.5 |
| 2       | ✓         |      |      |                   | 0.611 | 0.461 | 0.479 | 0.291    | 15.32      | 45.0       | 67.6 |
| 3       |           | ✓    |      |                   | 0.634 | 0.486 | 0.504 | 0.315    | 18.60      | 78.2       | 64.1 |
| 4       |           |      | ✓    |                   | 0.611 | 0.458 | 0.472 | 0.290    | 19.88      | 57.2       | 78.6 |
| 5       |           |      |      | ✓                 | 0.608 | 0.468 | 0.480 | 0.296    | 19.88      | 57.0       | 79.1 |
| 6       | ✓         | ✓    |      |                   | 0.639 | 0.490 | 0.515 | 0.326    | 14.05      | 67.0       | 55.7 |
| 7       | ✓         | ✓    | ✓    |                   | 0.630 | 0.502 | 0.518 | 0.326    | 14.05      | 67.2       | 58.1 |
| 8.ours  | ✓         | ✓    | ✓    | ✓                 | 0.639 | 0.505 | 0.522 | 0.330    | 14.05      | 67.2       | 62.3 |

The ablation study results presented in Table 7 demonstrate that employing the efficient HybridNet backbone architecture increased mAP50 by 0.9%, while reducing the parameter count and GFLOPs by 22.94% and 21.05%, respectively. These improvements indicate more effective feature extraction and lighter model weights. Experiment 6 shows that HybridNet successfully mitigates the substantial increase in parameters and computations introduced by SODM, resulting in 1.1% improvements in both mAP50 and mAP50:95, along with a 17.84% reduction in parameter count. This highlights the effectiveness of incorporating RCSPELAN in the shallow layers and modifying PAFPN to enhance the fusion and integration of features for small-object detection.

Compared to the baseline model, Experiment 7 reveals that integrating the EIFI layer into the hybrid encoder yielded a 0.3% increase in mAP50, underscoring the synergy among EIFI, HybridNet, and SODM. Further, adopting the Focaler-Shape-IoU loss function led to 0.4% improvements in both mAP50 and mAP95, confirming that this IoU formulation enhances model precision in complex object detection tasks. After incorporating all optimization strategies, the PHSI-RTDETR model achieved significant improvements, with mAP50 and mAP95 increased by 5.2% and 4.3%, respectively, while reducing the parameter count by 29.33%. Although the FPS is slightly lower than that of the original RT-DETR algorithm, the improved model demonstrates a substantial enhancement in both accuracy

and parameter efficiency, meeting the requirements for real-time aerial detection. These experiments collectively indicate that each modification contributes significantly to the performance improvement of the HSAF-DETR model.

The experimental results aptly highlight the commendable performance of Hybrid-DETR. To further investigate the specific performance of our model in the field of small-object detection, we present the precision–recall (PR) curves of RT-DETR-R18 and Hybrid-DETR for each category on the VisDrone validation set, as illustrated in Figure 10.

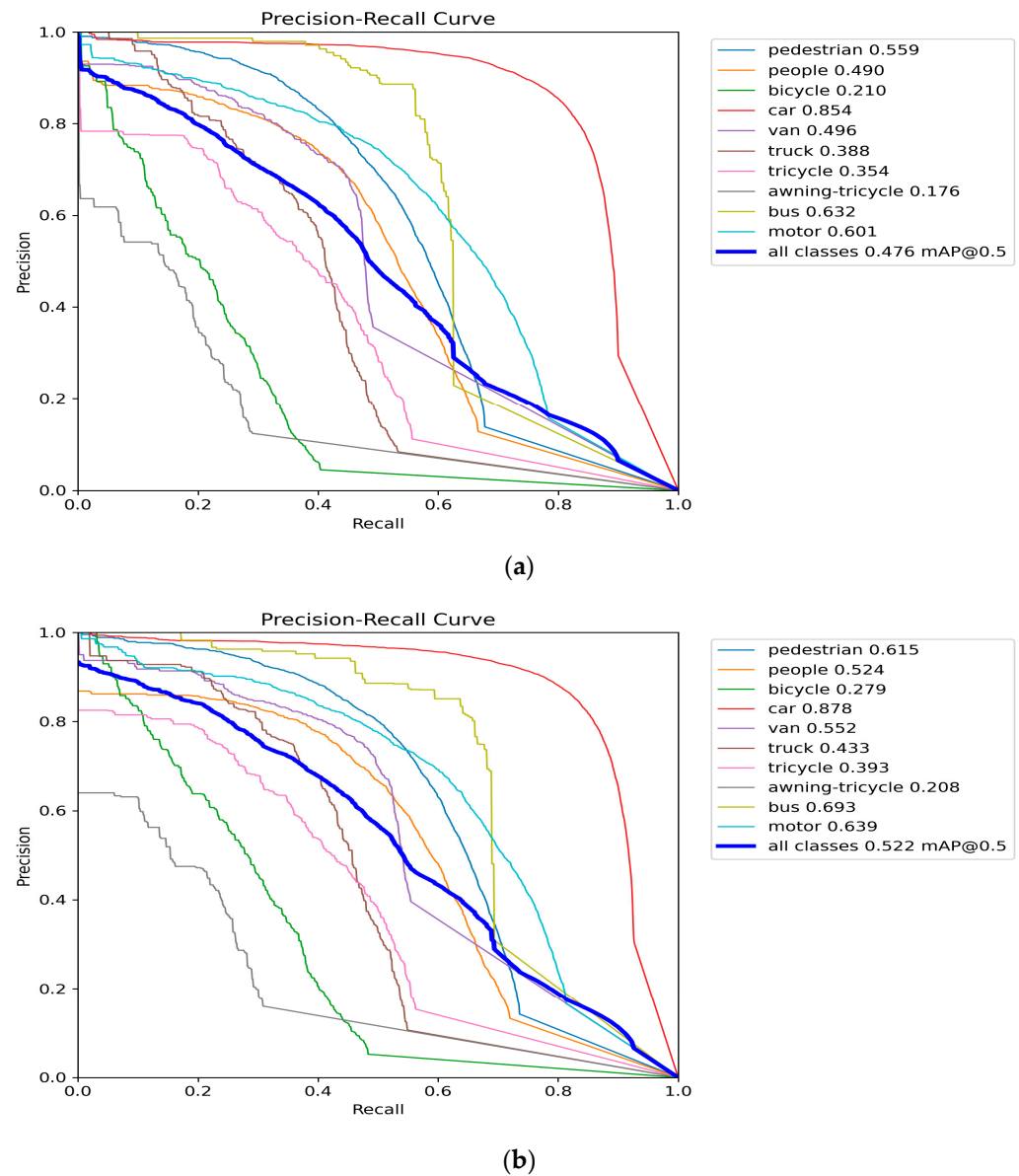


**Figure 10.** Precision–recall curve on the VisDrone validation set: (a) result of RT-DETR-R18; (b) result of Hybrid-DETR.

These curves were generated by setting the Intersection over Union (IoU) threshold to 0.5 and subsequently calculating the mean Average Precision (mAP) from the area under each curve. Notably, the “Car” category stands out with the highest precision, which draws significant attention, particularly due to its considerable representation in the dataset. In contrast, the “Motor”, “Tricycle” and “Awning-tricycle” categories exhibit more challenging characteristics, such as variable poses, high confusion, and small object sizes. Among these, the “Awning-tricycle” category, being the least represented, experienced the most challenging detection scenarios. Despite the complexities within these categories, our

model demonstrated superior performance compared to RT-DETR-R18, showcasing its outstanding capability in capturing intricate details and overcoming occlusion-related challenges. Furthermore, it is worth noting that the improved model also achieved significant performance gains in detecting larger objects, such as the “Van”, “Truck”, and “Bus” categories. The results of these experiments clearly reaffirm the effectiveness of our approach in comprehensively enhancing network performance.

## 6. Limitations and Future Research Directions

The proposed Hybrid-DETR model has achieved significant advancements in object detection within UAV (unmanned aerial vehicle) remote sensing imagery; however, certain limitations remain that warrant further investigation. A critical challenge lies in the model’s performance under extreme environmental conditions, such as heavy fog, rain, or low-light scenarios, which are prevalent in UAV applications. In these situations, detection accuracy tends to decrease, limiting the model’s effectiveness in adverse weather conditions—key for scenarios like search and rescue or disaster response.

Moreover, the proposed method has been primarily evaluated on UAV aerial imagery datasets. Its generalization to other imaging domains, including satellite imagery, terrestrial camera feeds, or multi-angle drone captures, remains unexplored. Finally, although the model demonstrates competitive computational efficiency, scenarios with ultra-low latency requirements, such as real-time surveillance or fast-moving object tracking, could still challenge its real-time applicability.

Future research will focus on overcoming these limitations to enhance the robustness and scalability of Hybrid-DETR. Integrating complementary multi-sensor data, such as LiDAR or thermal imaging, could significantly improve detection accuracy in extreme weather and lighting conditions. Additionally, employing advanced data augmentation techniques and domain adaptation strategies could extend the model’s applicability to diverse imaging domains. Developing lightweight, hardware-optimized variants of Hybrid-DETR will further ensure its viability for real-time, resource-constrained UAV operations. These advancements aim to meet the growing demand for versatile and adaptive detection systems in UAV-based remote sensing applications.

## 7. Conclusions

The Hybrid-DETR model introduced in this study offers an innovative and efficient solution to the challenging task of small-object detection in UAV remote sensing imagery. By incorporating novel modules and differentiated design strategies, the model effectively addresses key technical challenges, including detecting small objects in high-resolution images, managing complex and dynamic backgrounds, and operating effectively under challenging conditions such as low light and high occlusion. Experimental results on the VisDrone2019 dataset underscore the model’s effectiveness, achieving mAP50 and mAP50:95 scores of 52.2% and 33.3%, respectively, reflecting improvements of 5.2% and 4.3% over RT-DETR-R18, alongside a significant 29.33% reduction in parameter count. Its strong performance on additional datasets, such as AI-TOD and HIT-UAV, further highlights its generalization capabilities across diverse UAV application scenarios.

Beyond its technical contributions, Hybrid-DETR showcases substantial potential for real-world applications in fields such as military reconnaissance, disaster management, environmental monitoring, and urban planning. The model strikes an effective balance between accuracy, efficiency, and adaptability, making it a robust and scalable solution for UAV-based object detection in practical scenarios.

Looking ahead, this study provides a solid foundation for future advancements in UAV-based small-object detection. By addressing the identified limitations and pursuing the outlined future directions, we aim to further enhance the robustness, precision, and scalability of Hybrid-DETR. These improvements will ensure its applicability in increasingly complex and dynamic real-world environments, enabling transformative progress

in both the technical and interdisciplinary applications of UAV-based remote sensing and object detection.

**Author Contributions:** Conceptualization, M.Y., R.X., C.Y., H.W. and A.W.; methodology, M.Y., R.X. and C.Y.; software R.X.; validation, M.Y., R.X. and C.Y.; writing—review and editing, M.Y., R.X., C.Y., H.W. and A.W. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was funded by the Key Research and Development Plan Project of Heilongjiang (JD2023SJ19), the Natural Science Foundation of Heilongjiang Province (LH2023F034), and the Science and Technology Project of Heilongjiang Provincial Department of Transportation (HJK2024B002).

**Data Availability Statement:** VisDrone2019: <https://github.com/VisDrone/VisDrone-Dataset>, accessed on 14 April 2018; HIT-UAV: <https://github.com/suojiaashun/HIT-UAV-Infrared-Thermal-Dataset>, accessed on 7 April 2022.

**Conflicts of Interest:** Author Chunyu Yang was employed by the company The 54th Research Institute of China Electronics Technology Group Corporation. 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.

## References

- Kaleem, Z.; Rehmani, M.H. Amateur drone monitoring: State-of-the-art architectures key enabling technologies and future research directions. *IEEE Wirel. Commun.* **2018**, *25*, 150–159. [CrossRef]
- Rossiter, A. Military technology and revolutions in warfare: Priming the drone debate. *Def. Secur. Anal.* **2023**, *39*, 253–255. [CrossRef]
- Emimi, M.; Khaleel, M.; Alkrash, A. The current opportunities and challenges in drone technology. *Int. J. Electr. Eng. Sustain.* **2023**, *1*, 74–89.
- Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
- Jocher, G.; Chaurasia, A.; Stoken, A.; Borovec, J.; Kwon, Y.; Michael, K.; Fang, J.; Wong, C.; Yifu, Z.; Montes, D. Ultralytics/yolov5:v6. 2-yolov5 classification models, apple m1, reproducibility, clearml and deci. ai integrations. *Zenodo* **2022**. [CrossRef]
- Li, C.; Li, L.; Jiang, H.; Weng, K.; Geng, Y.; Li, L.; Ke, Z.; Li, Q.; Cheng, M.; Nie, W.; et al. YOLOv6: A single-stage object detection framework for industrial applications. *arXiv* **2022**, arXiv:2209.02976.
- Wang, C.-Y.; Bochkovskiy, A.; Liao, H.-Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 7464–7475.
- Reis, D.; Kupec, J.; Hong, J.; Daoudi, A. Real-time flying object detection with YOLOv8. *arXiv* **2023**, arXiv:2305.09972.
- Wang, C.-Y.; Yeh, I.-H.; Liao, H.-Y.M. YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information. *arXiv* **2024**, arXiv:2402.13616.
- Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J.; Ding, G. YOLOv10: Real-Time End-to-End Object Detection. *arXiv* **2024**, arXiv:2405.14458.
- 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, Seattle, WA, USA, 17–22 June 2024; pp. 16965–16974.
- He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
- Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8759–8768.
- Deng, C.; Wang, M.; Liu, L.; Liu, Y.; Jiang, Y. Extended feature pyramid network for small object detection. *IEEE Trans. Multimed.* **2021**, *24*, 1968–1979. [CrossRef]
- Chen, J.; Kao, S.-H.; 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, Vancouver, BC, Canada, 17–24 June 2023; pp. 12021–12031.
- Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. Distance-IoU loss: Faster and better learning for bounding box regression. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 12993–13000.
- Zhang, H.; Zhuang, X.; Zheng, J. Optimizing human abnormal behavior detection method of YOLO network. *Comput. Eng. Appl.* **2023**, *59*, 242–249.
- Wang, Z.; Liu, Y.; Liu, Y.; Yu, H.; Wang, Y.; Ye, Q.; Tian, Y. vHeat: Building Vision Models upon Heat Conduction. *arXiv* **2024**, arXiv:2405.16555.

19. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
20. Shaker, A.; Maaz, M.; Rasheed, H.; Khan, S.; Yang, M.-H.; Shahbaz Khan, F. SwiftFormer: Efficient Additive Attention for Transformer-based Real-time Mobile Vision Applications. *arXiv* **2023**, arXiv:2303.15446.
21. Zhang, H.; Zhang, S. Shape-iou: More accurate metric considering bounding box shape and scale. *arXiv* **2023**, arXiv:2401.10525.
22. Zhang, H.; Zhang, S. Focaler-IoU: More Focused Intersection over Union Loss. *arXiv* **2024**, arXiv:2401.10525.
23. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
24. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
25. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
26. Xu, Y.; Yu, G.; Wang, Y.; Wu, X.; Ma, Y. Car detection from low-altitude UAV imagery with the faster R-CNN. *J. Adv. Transp.* **2017**, *2017*, 2823617. [[CrossRef](#)]
27. Avola, D.; Cinque, L.; Diko, A.; Fagioli, A.; Foresti, G.L.; Mecca, A.; Pannone, D.; Piciarelli, C. MS-Faster R-CNN: Multi-stream backbone for improved Faster R-CNN object detection and aerial tracking from UAV images. *Remote Sens.* **2021**, *13*, 1670. [[CrossRef](#)]
28. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Part I 14; pp. 21–37.
29. Xu, H.; Tang, S.; Wang, J.; Dong, B.; Wang, X.; Zhao, K.; Zhu, Y.; Geng, J. Rock fracture identification algorithm based on the confidence score and non-maximum suppression. *Bull. Eng. Geol. Environ.* **2024**, *83*, 1–20. [[CrossRef](#)]
30. 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 Advances in Neural Information Processing Systems 30: 31st Annual Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017.
31. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 10012–10022.
32. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 213–229.
33. Wang, D.; Li, Z.; Du, X.; Ma, Z.; Liu, X. Farmland obstacle detection from the perspective of uavs based on non-local deformable detr. *Agriculture* **2022**, *12*, 1983. [[CrossRef](#)]
34. Li, F.; Zhang, H.; Liu, S.; Guo, J.; Ni, L.M.; Zhang, L. Dn-detr: Accelerate detr training by introducing query denoising. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 13619–13627.
35. Chen, Q.; Chen, X.; Zeng, G.; Wang, J. Group detr: Fast training convergence with decoupled one-to-many label assignment. *arXiv* **2022**, arXiv:2207.13085.
36. Li, F.; Zhang, H.; Xu, H.; Liu, S.; Zhang, L.; Ni, L.M.; Shum, H.-Y. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 3041–3050.
37. Zhang, H.; Hao, C.; Song, W.; Jiang, B.; Li, B. Adaptive slicing-aided hyper inference for small object detection in high-resolution remote sensing images. *Remote Sens.* **2023**, *15*, 1249. [[CrossRef](#)]
38. Wang, C.-Y.; Liao, H.-Y.M.; 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, Seattle, WA, USA, 14–19 June 2020; pp. 390–391.
39. Han, K.; Wang, Y.; Tian, Q.; Guo, J.; Xu, C.; Xu, C. Ghostnet: More features from cheap operations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 1580–1589.
40. Ding, X.; Zhang, X.; Ma, N.; Han, J.; Ding, G.; Sun, J. Repvgg: Making vgg-style convnets great again. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 13733–13742.
41. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258.
42. Wang, C.Y.; Liao, H.Y.M.; Yeh, I.H. Designing network design strategies through gradient path analysis. *J. Inf. Sci. Eng.* **2023**, *39*, 975–995.
43. 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, Long Beach, CA, USA, 16–20 June 2019; pp. 658–666.
44. Zheng, Z.; Wang, P.; Ren, D.; Liu, W.; Ye, R.; Hu, Q.; Zuo, W. Enhancing geometric factors in model learning and inference for object detection and instance segmentation. *IEEE Trans. Cybern.* **2021**, *52*, 8574–8586. [[CrossRef](#)] [[PubMed](#)]

45. Du, D.; Zhu, P.; Wen, L.; Bian, X.; Ling, H.; Hu, Q.; Zheng, J.; Peng, T.; Wang, X.; Zhang, Y. VisDrone-SOT2019: The vision meets drone single object tracking challenge results. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Republic of Korea, 27–28 October 2019; pp. 1569–1578.
46. 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](#)]
47. Suo, J.; Wang, T.; Zhang, X.; Chen, H.; Zhou, W.; Shi, W. HIT-UAV: A high-altitude infrared thermal dataset for Unmanned Aerial Vehicle-based object detection. *Sci. Data* **2023**, *10*, 227. [[CrossRef](#)]

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