drones

MDPI

Article

PHSI-RTDETR: A Lightweight Infrared Small Target Detection Algorithm Based on UAV Aerial Photography

Sen Wang [1,2], Huiping Jiang [1,2,*], Zhongjie Li [1,2], Jixiang Yang [1,2], Xuan Ma [1,2], Jiamin Chen [1,2] and Xingqun Tang [1,2]

[1] Key Laboratory of Ethnic Language Intelligent Analysis and Security, Governance of MOE, Beijing 100081, China; 22302044@muc.edu.cn (S.W.); 21301998@muc.edu.cn (Z.L.); 22302018@muc.edu.cn (J.Y.); 22302019@muc.edu.cn (X.M.); 22302049@muc.edu.cn (J.C.); 21301986@muc.edu.cn (X.T.)
[2] School of Information Engineering, Minzu University of China, Beijing 100081, China
* Correspondence: jianghp@muc.edu.cn

Abstract: To address the issues of low model accuracy caused by complex ground environments and uneven target scales and high computational complexity in unmanned aerial vehicle (UAV) aerial infrared image target detection, this study proposes a lightweight UAV aerial infrared small target detection algorithm called PHSI-RTDETR. Initially, an improved backbone feature extraction network is designed using the lightweight RPConv-Block module proposed in this paper, which effectively captures small target features, significantly reducing the model complexity and computational burden while improving accuracy. Subsequently, the HiLo attention mechanism is combined with an intra-scale feature interaction module to form an AIFI-HiLo module, which is integrated into a hybrid encoder to enhance the focus of the model on dense targets, reducing the rates of missed and false detections. Moreover, the slimneck-SSFF architecture is introduced as the cross-scale feature fusion architecture of the model, utilizing GSConv and VoVGSCSP modules to enhance adaptability to infrared targets of various scales, producing more semantic information while reducing network computations. Finally, the original GIoU loss is replaced with the Inner-GIoU loss, which uses a scaling factor to control auxiliary bounding boxes to speed up convergence and improve detection accuracy for small targets. The experimental results show that, compared to RT-DETR, PHSI-RTDETR reduces model parameters by 30.55% and floating-point operations by 17.10%. Moreover, detection precision and speed are increased by 3.81% and 13.39%, respectively, and mAP50, impressively, reaches 82.58%, demonstrating the great potential of this model for drone infrared small target detection.

Keywords: small infrared target; UAV; RT-DETR; lightweight structure; partial convolution; HiLo attention; slimneck; Inner-GIoU

Citation: Wang, S.; Jiang, H.; Li, Z.; Yang, J.; Ma, X.; Chen, J.; Tang, X. PHSI-RTDETR: A Lightweight Infrared Small Target Detection Algorithm Based on UAV Aerial Photography. *Drones* **2024**, *8*, 240. https://doi.org/10.3390/drones8060240

Academic Editor: Seokwon Yeom

Received: 8 May 2024
Revised: 29 May 2024
Accepted: 30 May 2024
Published: 3 June 2024

1. Introduction

With the development of deep learning [1] and edge computing [2], drones can be outfitted with artificial intelligence algorithms as edge computing devices. Infrared target detection using drones is extensively applied in diverse areas such as map drawing [3], traffic monitoring [4], and night-time rescue [5]. This technology is particularly important for target recognition in night-time or visually obstructed conditions because infrared imaging effectively captures the thermal radiation of targets in no-light or low-light environments [6]. However, the aerial targets captured by drones are usually small in size, and objects may obstruct each other, which increases the difficulty of feature extraction [7]. Moreover, different flying heights and shooting angles of drones affect the clarity and accuracy of imaging [8]. The challenges of detecting features in infrared images are heightened by their low contrast and significant noise levels [9]. At the same time, the built-in detection algorithms used in drones face the challenges of limited hardware computing resources and

the need for rapid response [10]. Therefore, creating a lightweight, precise, and efficient drone-based infrared target detection algorithm holds significant social value and practical significance.

Due to the significant distance between the drone and the objects being detected, infrared targets in the images usually occupy only 1 to 10 pixels [11], and objects smaller than 32 × 32 pixels in size are classified as small targets [12]. This challenge of extracting useful feature information from small objects places higher demands on drone infrared detection algorithms. Traditional machine learning techniques used for detecting small infrared targets have certain limitations. For example, filters can usually suppress uniform background clutter but fail to suppress complex background noise, leading to low accuracy and unstable performance [13,14]. Ref. [15] proposed a method for quickly constructing corresponding saliency maps in the spatial domain using spectral residuals, but it was ineffective in suppressing background clutter. Ref. [16] introduced a robust infrared patch-tensor model that can adapt to low-SCR infrared images, but it retained a high error rate in detecting small, occluded targets against complex backgrounds.

Therefore, deep learning has been employed to tackle these challenges while enhancing both the accuracy and speed of detecting targets [17]. Models such as the Faster R-CNN [18], Mask R-CNN [19], SSD [20], and the YOLO [21] series have been extensively used and studied for use in object detection tasks. In 2020, the Facebook AI team developed an end-to-end object Detection Transformer (DETR) model [22]. The model directly predicts the quantity and location of objects, eliminating the need for traditional techniques such as anchor boxes and NMS and yielding notable outcomes within the field. In 2023, Zhao et al. [23] launched the initial real-time end-to-end Detection Transformer (RT-DETR), which efficiently handles multi-scale features through intra-scale interaction and cross-scale fusion. RT-DETR surpasses state-of-the-art YOLO detectors of comparable size in terms of speed and accuracy. However, Transformer-based end-to-end object detection models still face challenges given their high computational demands and insufficient accuracy in detecting small targets. Zhang et al. [24] developed a network structure that runs Transformer and ResNet in parallel to combine their advantages for the more accurate detection of faint, small targets in infrared images. However, this method is computationally intensive and may not be suitable for resource-constrained drones. Li et al. [25] adopted HRNet as the feature extraction framework and proposed an advanced real-time detection method for infrared target detection called ISTD-CenterNet. Although this method improves the accuracy and speed of infrared target detection in complex environments, it still faces difficulties when dealing with extremely small targets and dense target scenes. Chen et al. [26] introduced a lightweight multi-feature fusion network called MFFNet, addressing issues such as indistinct texture features and limited resolution in infrared images to improve the capabilities of drone-mounted smart devices in identifying infrared targets. Nevertheless, the detection accuracy for small and occluded targets still needs improvement. Sun et al. [27] proposed a comprehensive infrared mobile small object detection model named Multi-YOLOv8, which enhances the detection capability by integrating multi-frame data from various sources. However, the complexity and inference latency of this model may hinder its practical application to drones. Li et al. [28] developed an interpretable multi-scale infrared small object detection network, IMD-Net, to enhance the accuracy of detecting and segmenting small objects against complex backgrounds. Meng et al. [29] introduced a locally focused attention-based Swin-transformer technique for thermal infrared moving object detection, termed LAGSwin, which encodes the spatial transformations and directional information of moving objects to enhance interaction and feature integration at varying resolutions. Nevertheless, the computational requirements of this model may exceed the capabilities of drone devices.

Within the domain of automated feature extraction and the handling of high-dimensional data, deep learning has surpassed traditional machine learning techniques, especially in achieving higher precision in detection tasks. However, the limited battery capacity of drones and their more restrictive hardware capabilities compared to terrestrial servers pose

significant challenges. Given these constraints, it is crucial to maximize the practicality of algorithms by significantly reducing the space and computational demands of the model while enhancing detection accuracy and speed. Moreover, robustness is required in order to withstand interference from unstructured external factors such as variable lighting, climatic conditions, and complex geographical environments. To bolster the detection capabilities of drones in challenging environments, particularly for small infrared targets, and to reduce computational complexity, we propose a lightweight infrared small target detection model named PHSI-RTDETR to tackle the challenges mentioned above. The key contributions include the following;

1. "Integration of RPConv with Residual Blocks" by replacing the 2D convolutions in Partial Convolution (PConv) with RepConv to obtain Reparameterized Partial Convolution (RPConv) and utilizing multi-path training combined with merged inference to reduce memory consumption. RPConv is then integrated with BasicBlock to form a lightweight RPConv-Block module, which is used to enhance the backbone architecture. This integration maintains performance while reducing computational load, thereby improving the efficiency of feature extraction;
2. "Introduction of HiLo Attention Mechanism"—HiLo attention is integrated into the intra-scale feature interaction (AIFI) module, replacing the multi-head attention mechanism to form the AIFI-HiLo component. This method processes high-frequency and low-frequency information within the feature map separately, enhancing the ability of the network to capture global dependencies and fine local details of images while reducing complexity and computational costs;
3. "Design of a Lightweight Feature Fusion Architecture"—A novel slimneck-SSFF structure is proposed by integrating the Scale Sequence Feature Fusion (SSFF) framework with the slimneck architecture, which incorporates lightweight GSConv and VoVGSCSP modules. Utilizing this architecture for cross-scale feature fusion not only enhances the detection capabilities related to tiny objects but also reduces computational demands and inference latency;
4. "Loss function optimization"—The Inner-IoU is amalgamated with the GIoU, introducing an auxiliary bounding box within the GIoU controlled by a scale factor ratio to obtain Inner-GIoU. Utilizing the Inner-GIoU loss function, which employs a scaling factor ratio to manage the generation of supplementary bounding boxes at different scales, accelerates convergence and enhances the detection of extremely tiny targets.

2. Materials and Methods

2.1. RT-DETR Algorithm

RT-DETR represents an innovative real-time end-to-end target detection model. Compared to YOLOv8, RT-DETR exhibits higher efficiency and improved balance within identical testing environments, reduces the training duration, and does so without utilizing the mosaic data augmentation strategy while maintaining detection speeds comparable to those of the YOLO series. The RT-DETR model is composed of three main components: the backbone network, a hybrid encoder, and a transformer decoder equipped with an auxiliary prediction head.

The backbone architecture leverages the capabilities of a convolutional network to extract salient features at three distinct scales, with strides of 8, 16, and 32, respectively. The hybrid encoder employs an Attention-based Intrascale Feature Interaction (AIFI) module to process high-level features from the backbone, significantly reducing computational load and enhancing processing speed without compromising performance. It also utilizes a Cross-scale Feature Fusion Module (CCFM) to integrate and interact with multi-scale features. The encoder dynamically adjusts queries based on IoU, choosing a set number of image features from the output sequence that focus on areas most relevant to the detection targets as the initial queries for the decoder. The decoder, equipped with an auxiliary prediction head, continuously refines object queries to produce bounding boxes and confidence scores. Figure 1 depicts the basic architecture of the RT-DETR model.

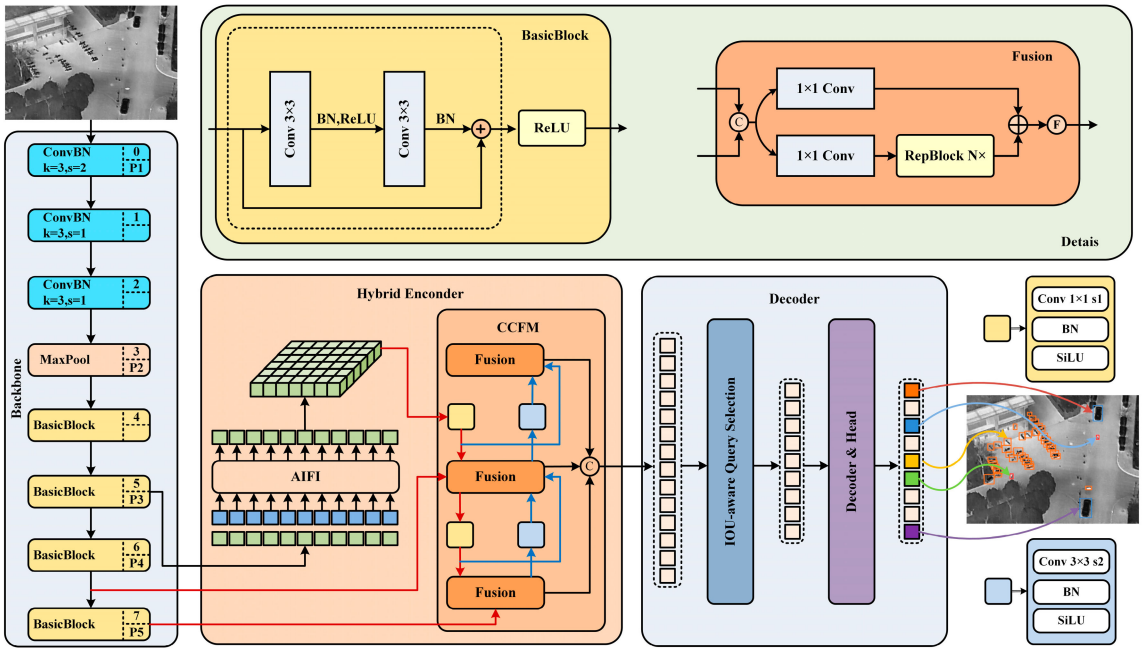


Figure 1. RT-DETR network structure diagram.

2.2. The PHSI-RTDETR Model Architecture

This paper proposes a lightweight PHSI-RTDETR model, as depicted in Figure 2. Addressing the challenges of unmanned aerial vehicle (UAV) infrared image recognition in complex environments, this model not only demands high accuracy and rapid processing capacities, but also faces the challenge of rendering the model lightweight so as to alleviate hardware costs. The model employs RepConv [30] to replace 2D convolution operations in some of the Partial Convolutions (PConv) [31], resulting in the RPConv convolution. By utilizing reparameterization techniques, it reduces computational load and memory consumption. The RPConv convolution is then integrated with residual blocks to form the RPConv-Block module, which optimizes the backbone network to efficiently extract features while significantly reducing the computational burden of the model. Furthermore, HiLo Attention [32], which processes high-frequency and low-frequency information in feature maps separately, is introduced into the Transformer encoder. The resulting AIFI-HiLo module captures both local details and global dependencies in feature maps. Subsequently, the proposed slimneck-SSFF structure for feature fusion, which combines a scaled sequence feature fusion framework [33] with a slim neck design, employs GSConv and VoVGSCSP modules [34] to reduce computational costs and inference latency, enhancing the focus on small target features. Finally, Inner-IoU [35] is combined with GIoU [36] to form Inner-GIoU, using a scale factor ratio to control the auxiliary bounding box size for loss computation in order to accelerate convergence and improve the detection of extremely small targets. This innovative integration of lightweight architecture and advanced feature processing technologies not only optimizes UAV performance in critical missions but also promotes scalability and adaptability in varied operational scenarios, ultimately leading to a more robust and efficient aerial surveillance system.

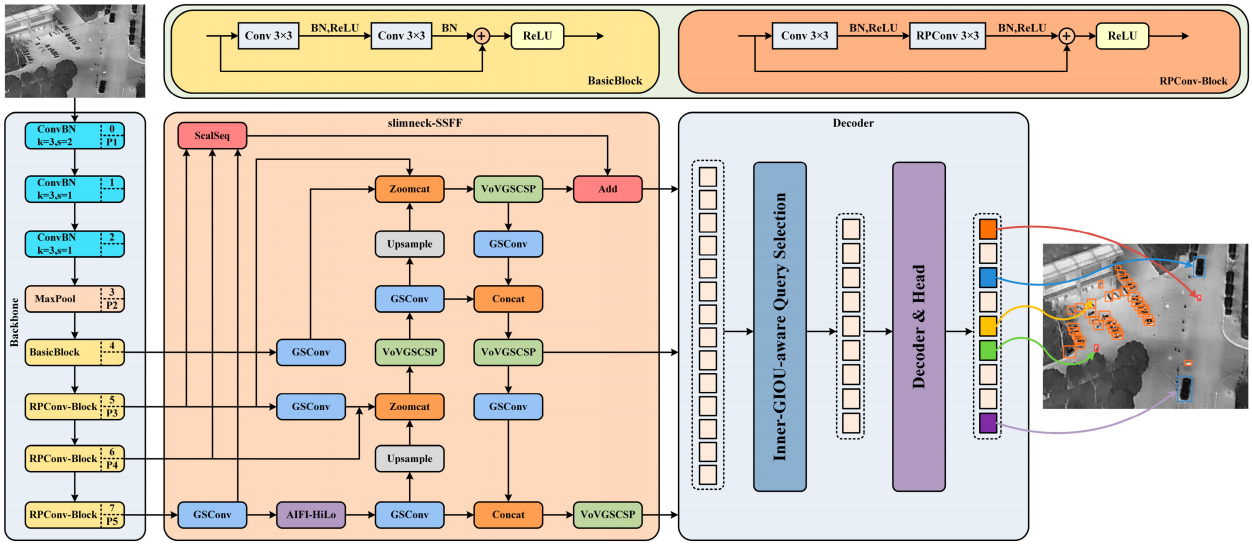


Figure 2. PHSI-RTDETR network structure diagram.

2.2.1. Improvement of Feature Extraction Network

To circumvent the computational redundancies inherent in complex models applied to simple tasks, which consequently diminish detection speeds, this study adopts the relatively lightweight ResNet-18 [37] as the foundational backbone network. Additionally, we innovatively combine RepConv with PConv to develop the reparameterized, lightweight RPConv. This new convolution is then employed to replace conventional convolutions in residual blocks, forming RPConv-Blocks that enhance the backbone. This adaptation not only boosts the capacity for feature extraction but also significantly reduces computational demands and memory consumption, making the model suitable for deployment on UAV mobile hardware. The RPConv tactically utilizes filters only on a selected subset of input channels, preserving the remainder, which ultimately results in lower Floating Point Operations Per Second (FLOPs) compared to standard convolutions. This approach ensures enhanced detection speeds across various UAV models without compromising the precision of recognition. The structural principle of RPConv is depicted in Figure 3.

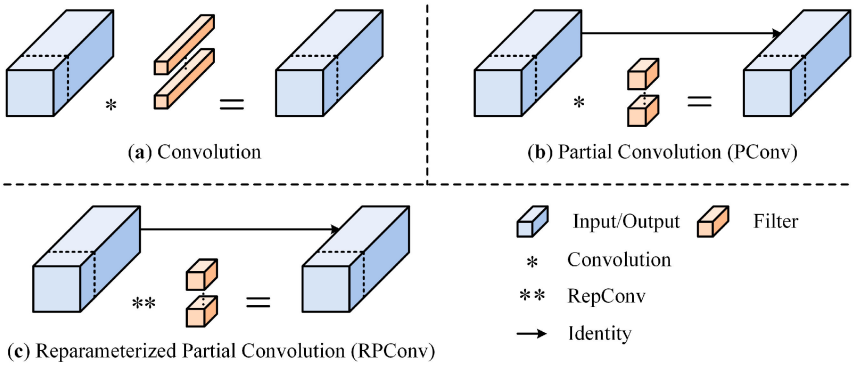


Figure 3. Schematic diagram of the Reparameterized Partial Convolution (RPConv) structure.

The RPConv applies RepConv to a select portion of input channels for extracting spatial features while maintaining other channels as they are. To ensure efficient memory access, the calculation engages the first or last sequence of c_p consecutive channels as a proxy for the computational demand across all feature maps. To maintain methodological consistency, the channel counts for input and output feature maps remain the same. Therefore, the FLOPs of a RPConv are solely

$$h \times w \times k^2 \times c_p^2 \quad (1)$$

In the formula, h and w denote the dimensions of the feature map, k indicates the convolution kernel size, and c_p represents the count of channels used by the RepConv. Furthermore, the smaller memory access of RPConv is

$$h \times w \times 2c_p + k^2 \times c_p^2 \approx h \times w \times 2c_p \tag{2}$$

Memory access for PConv is only a quarter of that of a typical convolution, as the remaining $c \sim c_p$ channels do not participate in the computation. Integrating RPConv into the residual blocks of the feature extraction network substantially reduces the computational and parameter load, thus boosting the inference speed of the model. The backbone structure, incorporating the RPConv-Block modules proposed in this paper, is depicted in Figure 4.

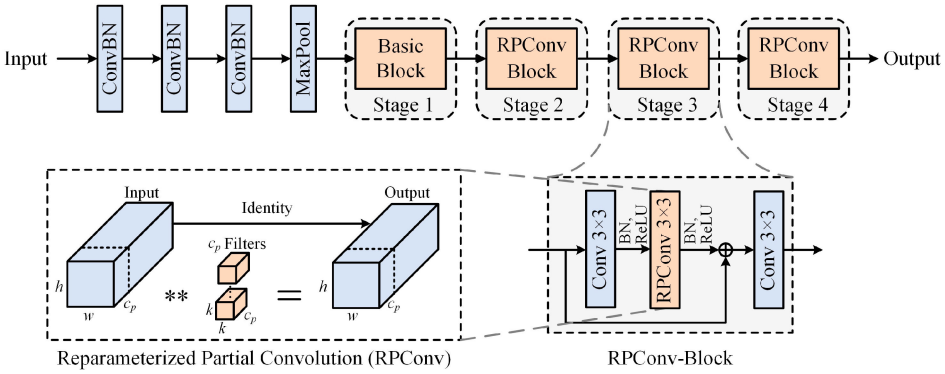


Figure 4. Lightweight feature extraction backbone network structure incorporating RPConv.

2.2.2. Introduce HiLo Attention into the AIFI Module

By incorporating the HiLo attention into the intra-scale feature interaction module, the network is enhanced to capture both the global dependencies and fine local details of images, while also reducing computational costs. The standard Multi-head Self-Attention (MSA) layer applies uniform global attention across all image blocks, ignoring the distinct underlying frequencies in features, which places a huge computational burden on dense, high-resolution images. The HiLo attention, however, partitions the MSA layer into two components: one for encoding high-frequency interactions with local self-attention and high-resolution feature maps, and another for encoding low-frequency interactions through global attention to downsampled features, thereby greatly enhancing efficiency. The structure of HiLo attention is depicted in Figure 5.

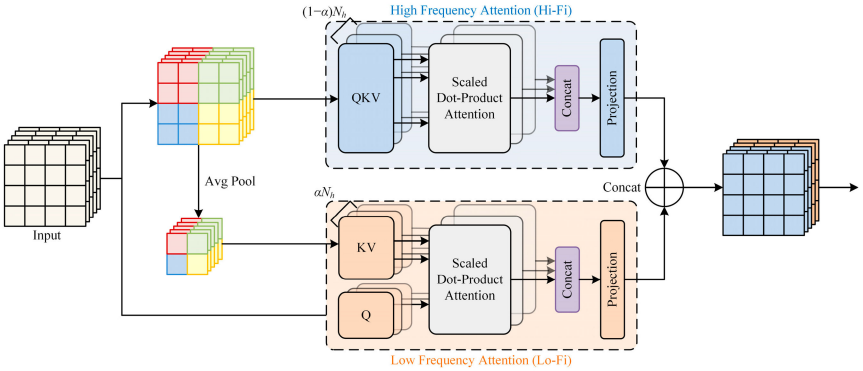


Figure 5. Framework of HiLo attention.

As illustrated above, N_h signifies the total count of self-attention heads in this layer, while α indicates the division ratio for high- or low-frequency heads. The high-frequency attention branch captures the local dependencies of fine details through local attention, which requires a high-resolution feature map. Alternatively, the low-frequency attention

branch captures the global dependencies of the input features using global attention and does not require high-resolution feature maps.

The model uses a separate transformer encoder layer dedicated to processing features from the backbone network. This approach leverages the rich semantic attributes of high-level features, significantly reducing computational demands and enhancing processing speed without compromising performance robustness. This optimized hybrid encoder orchestrates intra-scale feature interaction, transforming multi-scale features into a serialized array of image feature sequences. In this paper, the HiLo attention replaces the standard MSA to separately process high and low frequencies within the feature maps, reducing complexity and ensuring high throughput on GPUs while also enhancing the ability of the model to capture features of dense, high-resolution small targets. The computational process is described as follows:

$$Q = K = V = FlHiLoatten(S_5) \tag{3}$$

$$F_5 = Reshape(HiLoAttn(Q, K, V)) \tag{4}$$

where $HiLoAttn$ represents the HiLo attention, and $Reshape$ represents the restoration of the shape of the feature to match that of S_5, which is the inverse operation of $FlHiLoatten$.

2.2.3. Improvement of Cross-Scale Feature Fusion Network

In the efficient hybrid encoder structure of the model, high-level features are first processed using an attention-based intra-scale feature interaction module, followed by the CCFM for the interaction and fusion of multi-scale features. Compared to YOLO, this hybrid encoder structure has seen an increase in both parameter count and computational load, considering the need for deployment on UAV platforms for high-altitude small target infrared detection. The original model loses small amounts of target information during the convolution and downsampling processes. Therefore, this paper introduces the SSFF module, GSConv, and slimneck technology and proposes a new slimneck-SSFF feature fusion module aimed at making the model more lightweight while improving the precision of small target detection.

As depicted in Figure 6, the GSConv module merges traditional convolution with separable convolution, utilizing the Shuffle method to amalgamate the features produced by both, thereby facilitating inter-channel information exchange and efficiently lowering computational expenses.

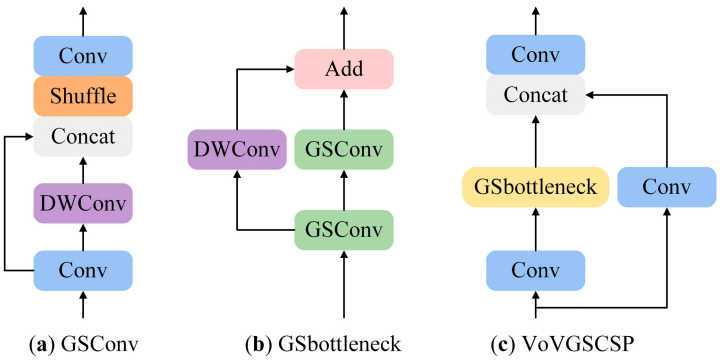


Figure 6. GSConv, GSbottleneck, and VoVGSCSP module structure diagram.

As shown in Figure 6a,b, the GSbottleneck configuration includes two GSConv modules and one DWConv module, with input features fed into each module and their outputs summed. Based on GSbottleneck, VoVGSCSP is constructed using a one-time aggregation strategy, significantly reducing the number of parameters and floating-point operations.

To accurately detect targets of varying scales from aerial images taken at different flying heights by drones, we employ the SSFF module to boost the network's capacity for

capturing features across different scales. This is accomplished by integrating features from multiple layers, providing more nuanced feature representations for objects of various sizes. The architecture of the SSFF module is depicted in Figure 7.

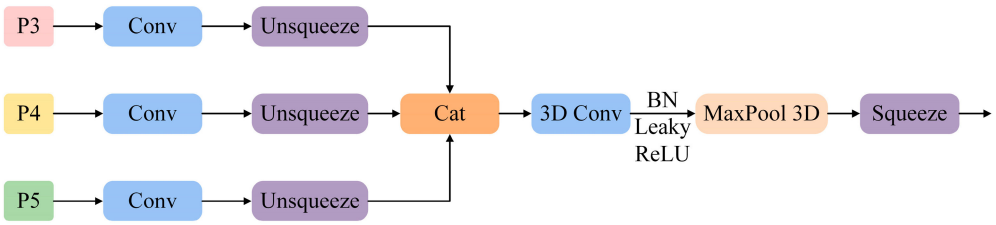


Figure 7. The Scale Sequence Feature Fusion (SSFF) module's structure diagram.

2.2.4. The Loss Function Inner-GIoU

At present, IoU-based bounding box regression concentrates on speeding up convergence through the addition of new loss terms, but it overlooks the intrinsic drawbacks of IoU loss itself. Differentiating regression samples during the process and using auxiliary bounding boxes of various scales to compute loss can effectively speed up the bounding box regression process. In model training, computing loss using smaller auxiliary bounding boxes improves the regression of high IoU samples, whereas the opposite is true for low IoU samples. To boost high-altitude drones' detection of dense small targets, this paper introduces the concept of Inner-IoU loss and proposes Inner-GIoU, which uses a scale factor ratio to manage the generation of auxiliary bounding boxes at various scales for loss calculation, thus enhancing both convergence speed and small target detection accuracy. The definition of Inner-GIoU is as follows:

$$inter = (min\ (b_r^{gt}, b_r) - max\ (b_l^{gt}, b_l)) * (min\ (b_b^{gt}, b_b) - max\ (b_t^{gt}, b_t)) \tag{5}$$

$$union = \left(w^{gt} * h^{gt}\right) * (ratio)^2 + (w * h) * (ratio)^2 - inter \tag{6}$$

$$IoU^{inner} = \frac{inter}{union} \tag{7}$$

$$\mathcal{L}_{GIoU} = 1 - IoU + \frac{A^c - U}{A^c} \tag{8}$$

$$\mathcal{L}_{Inner-GIoU} = \mathcal{L}_{GIoU} + IoU - IoU^{inner} \tag{9}$$

where the ground truth (GT) box and anchor are, respectively, represented as B^{gt} and B. The width and height of the GT box are denoted by w^{gt} and h^{gt}, while the width and height of the anchor box are represented by w and h. The $ratio$ is an auxiliary factor that controls the size of the helper box. In the GIoU loss, IoU represents the Intersection over Union and A^c signifies the area of the smallest enclosing rectangle, while U refers to the union area.

2.3. Datasets

This paper utilizes the infrared HIT-UAV dataset [38], which consists of high-altitude aerial images captured by drones, to validate the improvements made to the model. The HIT-UAV dataset contains 2898 infrared thermal images collected from locations such as schools, highways, and parking lots, as shown in Figure 8. The dataset features images captured at different flight heights ranging from 60 to 130 m, camera angles varying from 30 to 90 degrees, and under various lighting intensities. The HIT-UAV dataset presents a challenging scenario for object detection tasks, as it contains densely packed and small-sized objects from five categories, including people, vehicles, and bicycles.

Figure 8. Some of the HIT-UAV dataset images.

The HIT-UAV dataset is designed to address the challenges in applying UAVs for dark environment applications, such as traffic surveillance and city monitoring at night. By providing infrared thermal images, the dataset expands the application range of UAVs in low-light conditions. Moreover, the dataset records crucial information such as flight altitude, camera perspective, and image date, enabling researchers to investigate the impact of these factors on the precision of object detection. The diversity and complexity of the HIT-UAV dataset, covering a wide range of aspects, enhance its practicality for various tasks and contribute to the development of robust UAV-based object detection systems. Each image in the dataset has a resolution of 640 $\times$ 512, providing sufficient detail for small object detection while maintaining computational efficiency.

To meet the experimental requirements, the dataset was segmented into three sections. Of these, 2029 images were chosen for the training dataset, 290 images for the validation dataset, and 579 images for the test dataset. Detailed annotations of the dataset are shown in Table 1.

Table 1. HIT-UAV dataset labeling information.

Types	Number	Person	Car	Bicycle	OtherVehicle	DontCare
Training	2029	8473	5247	3618	102	110
Validation	290	1152	719	554	12	7
Test	579	2602	1338	792	34	31
Total	2898	12,227	7304	4964	148	148

2.4. Evaluation Indicators

This study evaluates algorithm performance by comparing the differences in image detection by the model before and after improvements, under identical experimental settings. This study employs precision (P), recall (R), mean average precision (mAP), F1 score, GFLOPs, and frames per second (FPS) as evaluation criteria.

Precision is the ratio of true positives to all positive predictions made by the model. Recall measures the portion of true positive samples correctly identified as positive by the model. The equations for precision and recall are presented below:

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

where TP refers to instances correctly identified as positive, FP refers to instances incorrectly identified as positive, and FN refers to instances incorrectly identified as negative. The $F1$ score, serving as the harmonic mean between precision and recall, aims to consolidate these metrics into one indicator.

$$F1 = \left(\frac{2}{Recall^{-1} + Precision^{-1}} \right) = 2 \frac{Precision \cdot Recall}{Precision + Recall} \tag{12}$$

The average precision (AP) is the average highest precision at different recall levels by category. The mean average precision (mAP) calculates the average AP across all categories and provides an overall assessment of the effectiveness of the model. The formula is as follows:

$$AP = \int_0^1 P(r)dr \tag{13}$$

$$\text{mAP} = \frac{\sum_{j=1}^{S} AP(j)}{S} \tag{14}$$

where S denotes the total category count, and the divisor is the aggregate of AP values across all categories. Moreover, GFLOPs measure the computational complexity, and FPS assesses real-time performance. The formula involving t_{avg} represents the FPS calculation.

$$FPS = \frac{1}{t_{avg}} \tag{15}$$

3. Results and Analysis

3.1. Training and Experimental Comparison Platform

The network experimental environment is based on Ubuntu 20.04, Python 3.8.10, and Pytorch 2.0.0, with relevant hardware configurations and model parameters detailed in Table 2. The batch size was set to 4, with a training duration of 150 epochs, and the learning rate was set at 1×10^{-4}. An adaptive image size of 640×640 was selected for the experiments.

Table 2. Hardware configuration and model parameters.

Types	Configuration	Types	Value
GPU	RTX 4090	learning rate	1×10^{-4}
CPU	16 vCPU	momentum	0.9
CUDA	11.8	optimizer	AdamW
CuDNN	8.7.0	batch	4

3.2. Backbone Network Comparative Experiment

Identifying small targets in infrared images captured by drones presents significant challenges in military and civilian contexts. The challenge of this task arises from the requirement for drones to precisely locate small pedestrians or vehicles in complex scenes from high altitudes, while achieving real-time and efficient detection. To address the requirements of the built-in model of the drone, this study enhances the backbone network of the base model. It replaces the original BasicBlock residual blocks with the lightweight RPConv-Block module introduced in this paper. To demonstrate the superiority of using RPConv-Block, several leading convolutional networks were selected for comparative testing, as illustrated in Table 3.

Table 3. Comparative experiment results for different backbone networks.

Model	Precision (%)	Recall (%)	mAP50 (%)	mAP50:95 (%)	Params (M)	GFLOPs (G)
BasicBlock	81.68	76.49	78.77	49.59	19.97	57.3
DualConv-Block	79.40	74.34	76.46	48.79	16.02	50.1
AKConv-Block	82.18	75.36	77.36	47.29	15.55	50.4
DySnakeConv-Block	86.66	71.68	76.65	48.79	27.86	60.8
DCNv2-Block	86.88	74.59	77.77	49.21	20.41	47.4
PConv-Block	76.60	74.21	76.04	48.68	14.17	46.7
RPConv-Block	84.16	76.51	79.52	49.84	14.17	46.7

Table 3 shows that incorporating the lightweight RPConv-Block not only decreases the model parameters and GFLOPs by 29.04% and 18.49%, respectively, but it also enhances the mAP50 by 0.75% compared to the baseline model. Comparative analysis indicates that although improvements to the backbone network with DualConv [39] and AKConv [40] reduced parameter counts and computational loads, they led to reductions of 2.13% and 1.41% in mAP50, respectively, which did not meet the high-precision requirements for drone-based infrared detection. Additionally, this study incorporated DySnakeConv [41] into the original BasicBlock module. While DySnakeConv performed well in terms of the precision metric, with a 4.98% improvement over the original model, the model yielded a significant increase in the parameters by 39.51% and GFLOPs by 6.11%. While the integration of deformable convolution v2 (DCNv2) [42] increased the precision by 5.2%, this enhancement resulted in a 2.2% rise in model parameters. Comparing the performance of several popular operators, we see that the RPConv-Block proposed in this paper demonstrates superior performance.

3.3. Verifying the Role of the AIFI-HiLo Module

To verify and explore the factors contributing to the improved detection of small targets in infrared imagery through the integration of the Intra-scale Feature Interaction module with HiLo attention, this study utilized LayerCAM [43] to generate and compare heatmaps before and after the incorporation of the attention mechanism. To comprehensively evaluate the performance of the HiLo attention mechanism in various scenarios, this study conducted comparative experiments in three representative environments: streets, high-angle parking lots, and low-angle parking lots. The street scene is characterized by dense pedestrians and vehicles, as well as complex and changing backgrounds, posing significant challenges to detection algorithms. In high-angle parking lot environments, the UAV is positioned at a higher altitude, with vehicles appearing relatively small in scale and widely spaced, making them prone to missed detections. In low-angle parking lot environments, the UAV's shooting angle is closer to the ground, leading to frequent occlusions between vehicles, which further complicates the detection task. By analyzing the heatmaps of the RT-DETR model before and after the introduction of the HiLo attention mechanism, we can gain a deeper understanding of the mechanism's effectiveness in improving small target detection performance under complex backgrounds. The comparative evaluation is depicted in Figure 9.

The experimental results demonstrate that the HiLo attention mechanism effectively enhances the RT-DETR model's ability to detect small targets in complex backgrounds. The above figure shows that the RT-DETR model without the introduction of the HiLo attention mechanism ignores certain small targets but focuses on irrelevant backgrounds. After incorporating the HiLo attention mechanism, the focus of the model becomes more concentrated and accurate, enabling it to better focus on tiny pedestrians and bicycles that are obscuring each other. In street scenes, the model can more accurately focus on small targets such as pedestrians and bicycles while effectively suppressing interference from backgrounds like buildings, maintaining high detection accuracy even when targets are mutually occluded. In high-angle parking lot environments, the model's detection

capability for vehicles is significantly improved after using the HiLo attention mechanism, accurately locating each vehicle while reducing the response to background regions. In low-angle parking lot scenes, the model can accurately locate occluded vehicles even when they overlap, with the heatmap clearly reflecting the positions and outlines of these vehicles. These results demonstrate the effectiveness of the HiLo attention mechanism in handling complex occlusion relationships and improving the accuracy of detecting small targets from various UAV perspectives.

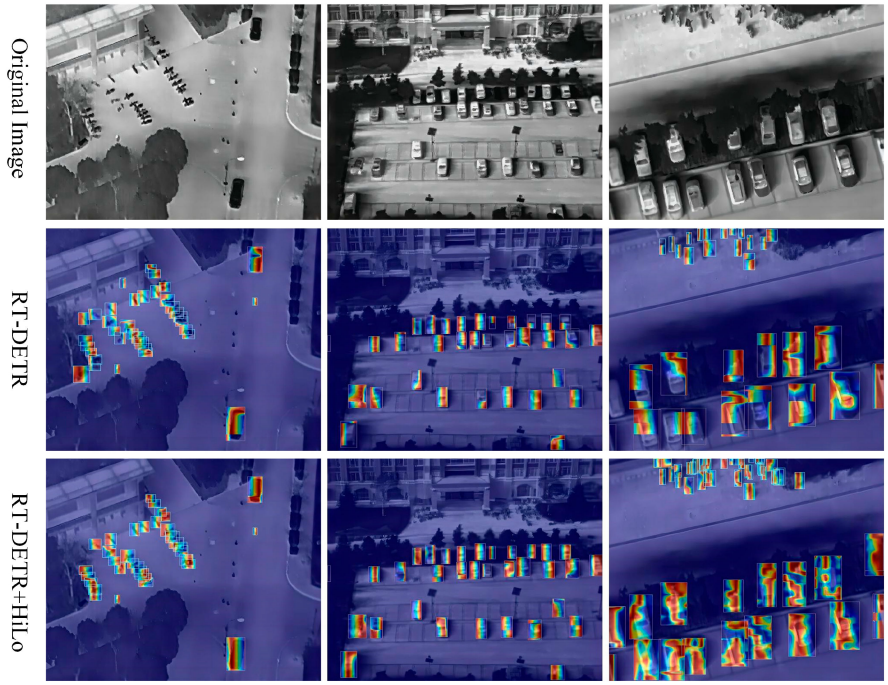


Figure 9. Comparison of feature visualization before and after adding the HiLo attention.

3.4. Verifying the Effectiveness of the Slimneck-SSFF Structure

As shown in the ablation experiments in Table 4, comparing four different model configurations reveals the specific impacts of each feature fusion network improvement strategy on model performance. Model 2, which incorporates the slimneck technology, significantly reduced parameters and GFLOPs by 3.36% and 6.98%, respectively, at the cost of a small amount of accuracy compared to the RT-DETR model. This indicates the significant effect of the slimneck strategy on enhancing the computational efficiency of the model. Model 3, employing the SSFF strategy, significantly increased mAP50 and mAP50:95 by 2.68% and 2.76%, respectively, but slightly increased computational complexity and parameter count, highlighting the additional resources needed for higher detection accuracy. Finally, Model 4, integrating both slimneck and SSFF strategies, achieved a balance, raising precision and recall to 87.09% and 76.96%, enhancing mAP50 by 0.39%, and reducing parameter count by 1.35%, thus achieving the optimization goal of managing computational resource consumption while maintaining high performance.

Table 4. Performance comparison of the model after feature fusion network enhancement.

Model	Precision (%)	Recall (%)	mAP50 (%)	mAP50:95 (%)	Params (M)	GFLOPs (G)
1. RT-DETR	81.68	76.49	78.77	49.59	19.97	57.3
2. RT-DETR + slimneck	81.44	73.32	76.91	48.04	19.30	53.3
3. RT-DETR + SSFF	87.22	74.38	81.45	52.35	20.16	61.5
4. RT-DETR + slimneck + SSFF	87.09	76.96	79.16	49.01	19.70	57.9

3.5. Backbone Network Comparative Experiment

To confirm the effectiveness of the proposed Inner-GIoU loss function, tests were performed using various ratio settings to adjust the dimensions of the auxiliary bounding box. Comparative analyses were performed against established loss functions such as DIoU, CIoU, SIoU, and EIoU [44–46]. The data presented in Table 5 show that the model equipped with the Inner-GIoU loss function at a 1.13 ratio demonstrated enhanced detection accuracy. Compared to the baseline model utilizing the GIoU loss function, there was an increase of 5.09% in the precision indicators and a 1.55% rise in mAP50. This indicates that employing the Inner-GIoU loss function may result in more consistent regression on bounding boxes and improved prediction precision.

Table 5. Performance comparison of models with improved loss functions.

Loss Function	Precision (%)	Recall (%)	mAP50 (%)	mAP50:95 (%)
GIoU	84.21	76.98	81.03	51.99
DIoU	83.49	74.57	79.47	48.90
CIoU	87.04	75.31	78.78	50.62
EIoU	87.23	71.08	79.18	48.59
SIoU	88.01	76.80	80.30	50.11
Inner-GIoU (ratio = 0.70)	87.98	72.56	79.01	51.16
Inner-GIoU (ratio = 0.75)	89.22	72.92	78.09	49.73
Inner-GIoU (ratio = 0.80)	85.53	75.41	79.43	51.01
Inner-GIoU (ratio = 1.10)	87.83	75.79	80.01	51.13
Inner-GIoU (ratio = 1.13)	89.30	76.14	82.58	51.59
Inner-GIoU (ratio = 1.15)	89.32	75.85	81.18	51.48

3.6. Ablation Study

To verify the enhancement effect of the proposed improvement modules on the model, eight groups of ablation experiments were created. On the basis of the RE-DETR network, the following modifications were made: the BasicBlock in the feature extraction network was replaced with the lightweight RPConv-Block module, the Intra-scale Feature Interaction module equipped with a HiLo attention mechanism was added, and the feature fusion network was optimized using slimneck and SSFF. Additionally, the loss function was switched to Inner-GIoU. The experiments were conducted by successively adding each improvement module, and the results are presented in Table 6.

Table 6. Results of ablation experiments.

Methods	RPConv-Block	AIFI-HiLo	Slimneck-SSFF	Inner-GIoU	mAP50 (%)	mAP50:95 (%)	Params (M)	GFLOPs (G)	FPS (f/s)
1. base					78.77	49.59	19.97	57.3	112
2	√				79.52	49.84	14.17	46.7	116
3		√			79.84	51.91	19.94	57.4	143
4			√		79.16	49.01	19.70	57.9	114
5				√	79.35	51.55	19.97	57.3	119
6	√	√			78.11	50.26	14.14	46.8	154
7	√	√	√		81.03	51.99	13.87	47.5	122
8. ours	√	√	√	√	82.58	51.59	13.87	47.5	127

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

The ablation study results in Table 6 indicate that the backbone network, enhanced with the lightweight RPConv-Block module, improved the mAP50, mAP50:95, and FPS metrics by 0.75%, 0.25%, and 3.57%, respectively. These improvements demonstrate more effective feature extraction, faster detection speeds, and a reduction in model weight, with parameter counts and GFLOPs decreasing by 29.04% and 18.49%, respectively. Relative to the baseline model, the integration of the AIFI-HiLo layer into the hybrid encoder led to increases of 1.07% in mAP50 and 27.68% in FPS. As can be seen from Experiment 7, utilizing the slimneck-SSFF architecture as a cross-scale feature fusion network significantly

improved the mAP50 and mAP50:95 by 2.92% and 1.73%, respectively, while maintaining the stability of other metrics. This confirms that the slimneck-SSFF framework effectively enhances the fusion and articulation of features for infrared small target detection. After adopting the Inner-GIoU loss function, the model showed improvements of 0.58%, 1.96%, and 6.25% in mAP50, mAP95, and FPS, respectively. Furthermore, the PHSI-RTDETR model, following the integration of all improvement strategies, exhibited substantial increases in the mAP50, mAP95, and FPS metrics by 3.81%, 2%, and 13.39%, respectively. Meanwhile, there were significant reductions in computational load and parameter count, which decreased by 17.10% and 30.55%, respectively. The experiments demonstrate that each modification made a significant contribution to the performance of the PHSI-RTDETR model.

3.7. Comprehensive Analysis of the Improved Model

Table 7 displays the detection effectiveness of the PHSI-RTDETR model across various categories of infrared targets. Examining the overall performance, we see that the model demonstrated high precision and recall rates, at 89.30% and 76.14%, respectively, achieving an mAP50 of 82.58% and an mAP50:95 of 51.59%, thus confirming its effectiveness in comprehensively assessing small infrared targets from drone aerial imagery. Particularly in detecting pedestrians and cars, the model exhibited exceptionally high precision rates of 92.42% and 93.75%, and also scored highly in mAP50, reaching 94.35% and 97.16%, respectively. Despite the fact that bicycle targets in the dataset are mostly densely distributed and mutually occluded, the model still achieved an mAP50 exceeding 90% for this category, demonstrating its superiority in handling densely packed targets. However, in recognizing OtherVehicle, the mAP50 and mAP50:95 metrics decreased to 65.51% and 46.96%, respectively. This may be due to the less distinct features of these targets, which are easily confused with cars, posing challenges to the detection capabilities of the model. Overall, however, the PHSI-RTDETR model exhibited exemplary performance in detecting infrared targets from drone aerial imagery, particularly excelling in identifying densely packed pedestrians and cars and showcasing significant advantages.

Table 7. Results of the improved model in detecting various categories in the HIT-UAV test set.

Class	Instances	Precision (%)	Recall (%)	mAP50 (%)	mAP50:95 (%)
All	4797	89.30	76.14	82.58	51.59
Person	2602	92.42	87.55	94.35	50.57
Car	1338	93.75	93.10	97.16	69.43
Bicycle	792	92.01	79.29	90.80	53.72
OtherVehicle	34	68.32	70.59	65.51	46.96
DontCare	31	100	50.16	65.07	37.28

To improve the generalization ability and robustness of the model, future work could focus on expanding the range of infrared small target categories that the model can detect. By incorporating a more diverse set of object classes during training, the model can learn to capture a wider variety of features and better distinguish between different types of targets. Additionally, enhancing the ability of the model to differentiate between genuine targets and background clutter or other sources of interference is crucial for reducing false positives and improving overall detection accuracy. This could be achieved through techniques such as data augmentation, where the training dataset is enriched with more challenging examples that include confounding factors, enabling the model to learn more discriminative features. To visually demonstrate the capability of the model to recognize infrared targets in drone aerial imagery, we visualized several evaluation metrics, as shown in Figure 10. The performance curves for various categories of infrared targets clearly illustrate the superiority of the improved model. The graphs in Figure 10 represent the precision, recall, mAP50, and F1 curves of the PHSI-RTDETR model, respectively.

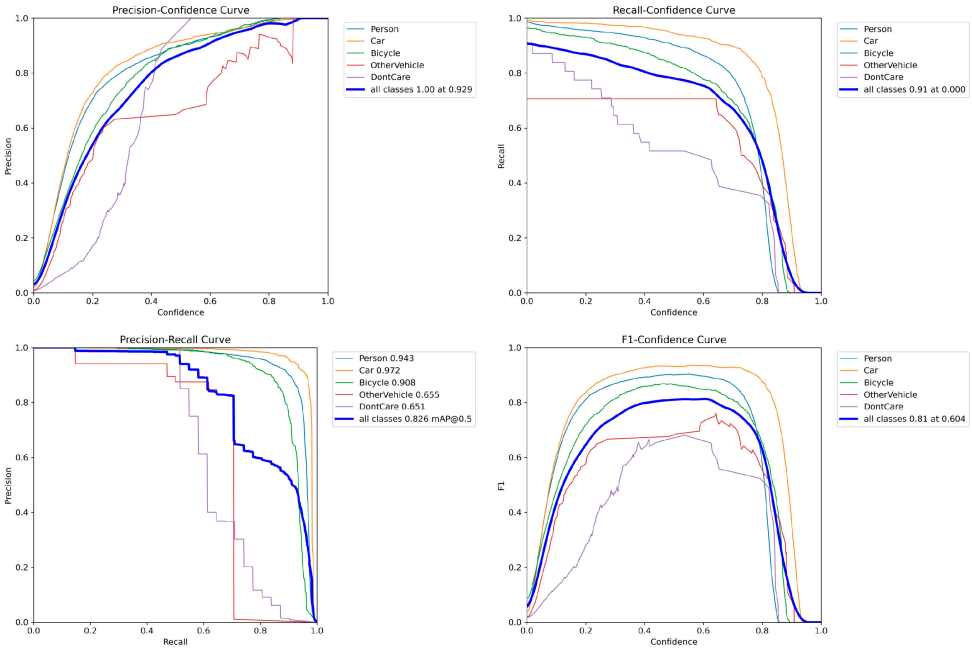


Figure 10. The precision, recall, mAP50, and F1 comparison curves of the PHSI-RTDETR model.

3.8. Comparison of Different Detection Models

Figure 11 displays the confusion matrices for the RT-DETR and PHSI-RTDETR models at their respective optimal performances. From the comparative visualization, it is intuitively clear that the PHSI-RTDETR model surpasses the original model in classification performance across all categories. The new model not only improves the accuracy but also enhances precision and recall rates, leading to fewer false positives and negatives. This comprehensive improvement is particularly notable in challenging categories where previous models struggled. Based on the provided analysis, we see that the PHSI-RTDETR model proposed in this study demonstrated exceptional performance in the task of detecting small infrared targets, thereby confirming its effectiveness and reliability in real-world scenarios. This advancement suggests significant potential for the application of the PHSI-RTDETR model in various surveillance and security contexts where accurate detection of small objects is crucial.

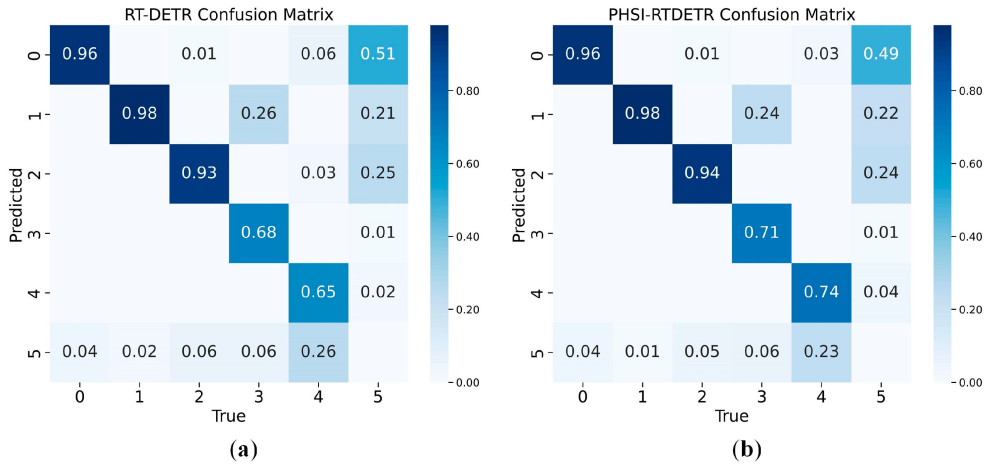


Figure 11. Confusion matrix comparison plot of RT-DETR and PHSI-RTDETR. (**a**) represents the confusion matrix for RT-DETR, (**b**) represents the confusion matrix for PHSI-RTDETR.

The performance of PHSI-RTDETR was compared with those of several other object detection models, including Faster-RCNN, SSD, YOLOv5, YOLOv6, YOLOv8, YOLOv9 [47], YOLOv10 [48], and the RT-DETR series. The specific results are shown in Table 8.

Table 8. Comparison of performance of different models.

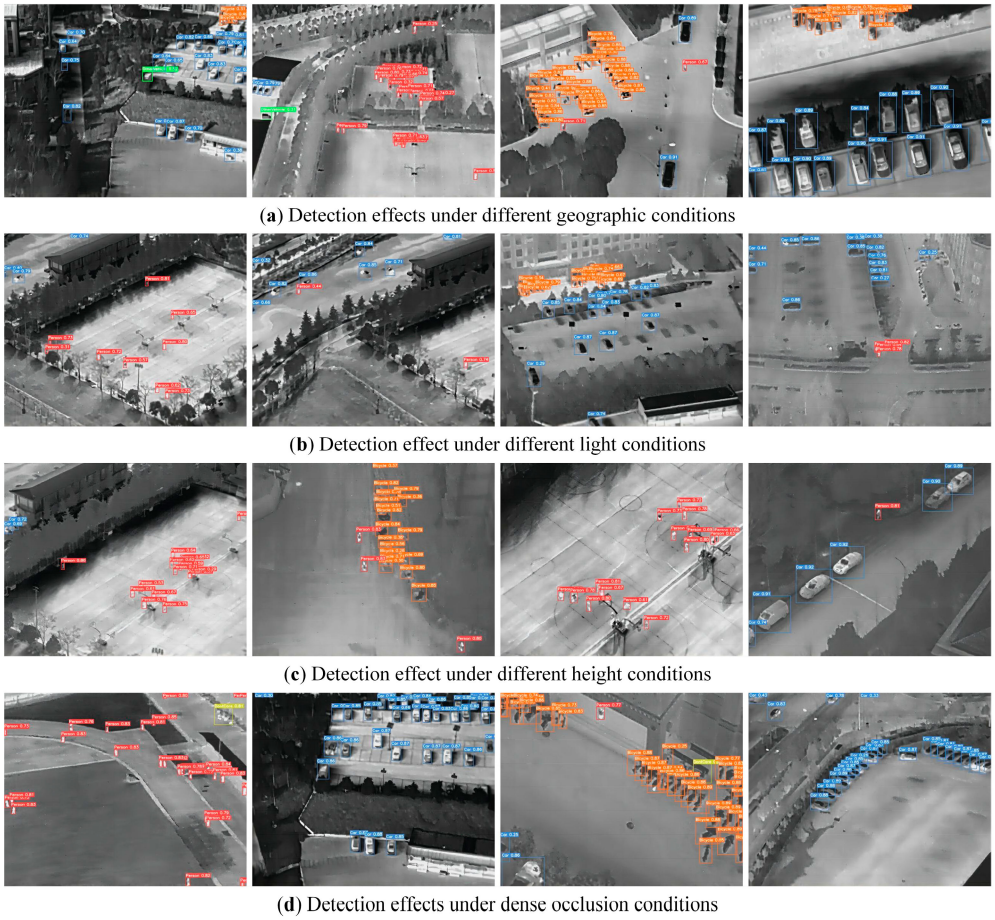
Model	Precision (%)	Recall (%)	mAP50 (%)	mAP50:95 (%)	F1 (%)	Time (ms)
Faster R-CNN	79.71	67.74	70.21	40.19	73	7.7
SSD	82.11	67.83	72.11	41.89	74	9.2
YOLOv5	81.68	71.33	75.47	48.18	76	9.6
YOLOv6	78.41	77.13	78.38	49.73	77	8.0
YOLOv8	82.10	76.58	79.91	51.04	79	7.4
YOLOv9	86.02	75.76	80.89	52.49	80	8.1
YOLOv10	85.81	70.65	77.70	47.39	77	7.5
RT-DETR	81.68	76.49	78.77	49.59	78	8.9
RT-DETR-L	86.28	73.64	79.07	49.65	78	11.6
RT-DETR-R34	85.72	71.78	75.13	46.91	77	8.4
RT-DETR-R50	89.93	76.09	80.33	51.27	81	12.9
PHSI-RTDETR	89.30	76.14	82.58	51.59	81	7.9

As demonstrated in Table 8, the PHSI-RTDETR model exhibits significant improvements in performance metrics compared to various state-of-the-art object detection models. In terms of mAP50, PHSI-RTDETR outperforms Faster-RCNN, SSD, YOLOv5, YOLOv6, YOLOv8, YOLOv9, and YOLOv10 by 12.37%, 10.47%, 7.11%, 4.2%, 2.67%, 1.69%, and 4.88%, respectively. These substantial gains in mAP50 highlight the exceptional detection accuracy of PHSI-RTDETR, particularly for small infrared targets in drone aerial imagery. Moreover, PHSI-RTDETR achieves a remarkable detection speed of 7.9 milliseconds per image, surpassing the real-time detection criteria. This swift inference time positions PHSI-RTDETR as an efficient model for practical applications demanding real-time performance.

Furthermore, compared to the RT-DETR-L, RT-DETR-R34, and RT-DETR-R50 models, mAP50 was improved by 3.51%, 7.45%, and 2.25%, respectively. These results underscore the effectiveness of the architectural optimizations and design choices incorporated in PHSI-RTDETR, contributing to its superior detection accuracy. Furthermore, compared to the original RT-DETR model, PHSI-RTDETR showcases notable advancements, with mAP50 and mAP50:95 metrics increasing by 3.81% and 2%, respectively, indicating improved performance across different IoU thresholds. Simultaneously, the F1 score and inference speed are improved by 3.85% and 11.24%, respectively, emphasizing the model's balanced approach to accuracy and efficiency.

3.9. Visual Analysis

In complex environments, infrared targets are affected by various factors, including differing geographical locations, uneven lighting, varying altitudes of drone aerial photography, and obstructions caused by overlapping objects. To address these challenges, we evaluated the performance of the enhanced model across various scenarios, as depicted in Figure 12. These detection outcomes show that the model accurately identifies small infrared targets in complex settings from drone aerial images, demonstrating robustness.



(**a**) Detection effects under different geographic conditions

(**b**) Detection effect under different light conditions

(**c**) Detection effect under different height conditions

(**d**) Detection effects under dense occlusion conditions

Figure 12. Examples of infrared small target detection results under different conditions.

Figure 12 presents a comparison of the effects of small target detection using infrared imaging technology under different conditions. The algorithm demonstrates good adaptability in various geographic environments, effectively detecting tiny targets in scenes such as streets and parking lots, with high accuracy and stability in the detection results. The comparison of detection effects under different light intensities indicates that light changes have a certain impact on the detection results, but overall, the algorithm exhibits strong robustness to light variations. The analysis of the influence of target height on the detection effect shows that as the target height increases, the occlusion relationships become more complex, making small target detection more challenging. Despite this, the algorithm maintains a high detection accuracy. The demonstration of the detection effect under dense occlusion conditions reveals that the algorithm can detect most of the occluded targets, but a small portion of targets are not correctly identified, indicating that there is room for further optimization in handling severe occlusion problems. Overall, the infrared small target detection algorithm exhibits good performance in complex environments, but still requires targeted improvements in specific scenarios to further enhance its practical value.

Figure 13 displays the results of the comparison of small infrared target detection using drone aerial imagery in various environmental conditions between PHSI-RTDETR and other RT-DETR series models.

These images clearly demonstrate that PHSI-RTDETR surpasses the original model in detection accuracy. Comparative detections were performed in various settings, including schools, playgrounds, roads, and parking lots. From the first column of images, it is evident that PHSI-RTDETR achieves higher detection accuracy compared to other models. In the second column, we see that the original model failed to detect some pedestrians and cars, whereas PHSI-RTDETR accurately identified all small infrared targets. In the third column, we see that PHSI-RTDETR significantly reduced the miss and false detection rates

of pedestrians. In the fourth column, we see that, while RT-DETR incorrectly detected cars and RT-DETR-R34 missed pedestrians, PHSI-RTDETR exhibited superior detection accuracy over the other models. Overall, PHSI-RTDETR demonstrated exceptional detection performance on small infrared targets across various complex aerial environments, making it suitable for high-altitude drone detection tasks.

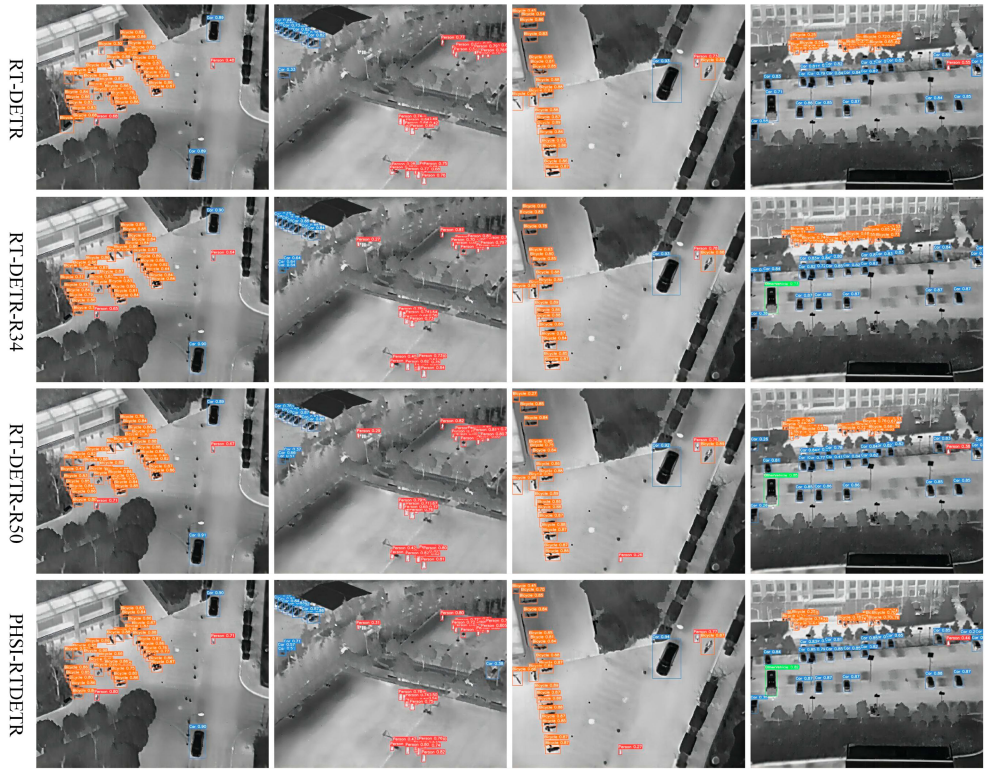


Figure 13. Comparative visualization of detection results for different algorithms.

4. Discussion

The experimental results demonstrate that the proposed PHSI-RTDETR model achieves significant improvements in both accuracy and efficiency for infrared small target detection in UAV aerial photography. The ablation studies confirm the effectiveness of each proposed improvement module, including the lightweight RPConv-Block, the AIFI-HiLo module, the slimneck-SSFF structure, and the Inner-GIoU loss function.

The introduction of the RPConv-Block module in the backbone network not only reduces the computational cost and parameter count but also improves the detection accuracy, which can be attributed to the ability of reparameterized partial convolution to extract more discriminative features by focusing on informative regions while suppressing irrelevant background information. The AIFI-HiLo module further enhances the feature representation by separately processing high-frequency and low-frequency information in the feature maps, capturing both local details and global dependencies, as revealed by the visualization of attention maps, which demonstrates that the AIFI-HiLo module enables the model to focus more on the target objects while suppressing the interference from background clutter. In the neck network, the slimneck-SSFF structure effectively balances the trade-off between accuracy and efficiency, with the slimneck strategy reducing the computational complexity and parameter count, while the SSFF strategy improves the feature fusion and representation capability across different scales, as demonstrated by the ablation experiments, which show that the combination of slimneck and SSFF achieves the best performance, indicating the complementary nature of these two strategies. Furthermore, the Inner-GIoU loss function promotes more precise bounding box regression by introducing an auxiliary bounding box controlled by a scale factor ratio to accelerate

convergence and improve the detection of extremely small targets, and the comparison with other state-of-the-art loss functions validates the superiority of Inner-GIoU in terms of convergence speed and detection accuracy.

A comprehensive analysis of the improved model reveals its strong capability in detecting small infrared targets under various complex backgrounds and dense object distributions. The high precision and recall rates demonstrate that the model can capture the unique features associated with small targets, even in challenging scenarios. Comparisons with other object detection models, including Faster R-CNN, SSD, YOLOv9, YOLOv10, and the original RT-DETR, highlight the superior performance of PHSI-RTDETR in terms of accuracy and speed. Visual analysis under various environmental conditions proves the model's robustness in handling challenges such as low contrast, significant noise levels, and object occlusion. When applied to UAV-based infrared target detection tasks in the future, this approach has the potential to enhance the performance of applications such as search and rescue, wildlife monitoring, and military reconnaissance by improving the detection accuracy and efficiency of small targets in complex environments.

5. Conclusions

In this study, a UAV aerial photography infrared small target detection model, PHSI-RTDETR, is proposed, which can provide more accurate decision support for traffic supervision and night rescue. Firstly, the backbone feature extraction architecture was crafted using the lightweight RPConv-Block module introduced in this paper, which simplified the model by reducing its complexity and computational demands. Then, an AIFI-HiLo module was obtained by combining the HiLo attention mechanism with the in-scale feature interaction module, which was incorporated into the hybrid encoder to enhance the attention of the model to small or dense targets and to reduce leakage and false detection rates. In addition, the slimneck-SSFF architecture was proposed as a cross-scale feature fusion architecture for the model, which utilized the GSConv and VoVGSCSP modules to enhance the adaptability of the model to infrared targets at different scales, generating more detailed semantic information while reducing the network computation. Finally, the original GIoU loss was substituted by the Inner-GIoU loss, which utilized the scale factor control assistance box to accelerate the convergence speed and improve the accuracy of judging small targets.

Comprehensive experiments were conducted to validate the performance of PHSI-RTDETR. The comparative results indicate that PHSI-RTDETR improved precision, mAP50, and mAP50:95 by 7.62%, 3.81%, and 2%, respectively, compared to the original model. Concurrently, the number of parameters and GFLOPs were reduced by 30.55% and 17.10%, respectively, with the detection time per image being only 7.9 milliseconds. Moreover, the experiments confirmed the robustness of PHSI-RTDETR, demonstrating its capability to effectively detect small infrared targets in various complex scenarios involving different geographical locations, lighting, and drone flight altitudes. To further verify the versatility and robustness of the proposed method, future work will focus on conducting experiments on more infrared small target detection datasets. By testing the algorithm on different datasets, we aim to gain deeper insights into its strengths and limitations under various scenarios, enabling the improved model to better adapt to diverse UAV-based infrared detection applications. Furthermore, future research will explore the potential of multi-drone collaborative detection of small targets, leveraging the information shared among multiple drones to enhance detection performance, especially for partially occluded or low-visibility targets in large-scale scenarios.

Author Contributions: Conceptualization, S.W. and Z.L.; methodology, S.W.; software, S.W.; validation, S.W., J.Y. and X.T.; formal analysis, X.M.; investigation, X.M.; resources, H.J.; data curation, J.C.; writing—original draft preparation, S.W., Z.L. and X.T.; writing—review and editing, S.W. and H.J.; visualization, J.C.; supervision, H.J.; project administration, S.W. and J.Y.; funding acquisition, H.J. 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 No. 61773416 and supported by the Graduate Research and Practice Projects of Minzu University of China, grant No. SJCX2024021.

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

Acknowledgments: Our sincere thanks are given to the Key Laboratory of Ethnic Language Intelligent Analysis and Security Governance of MOE for their generous support.

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

References

1. Pouyanfar, S.; Sadiq, S.; Yan, Y.; Tian, H.; Tao, Y.; Reyes, M.P.; Shyu, M.-L.; Chen, S.-C.; Iyengar, S.S. A survey on deep learning: Algorithms, techniques, and applications. *ACM Comput. Surv. (CSUR)* **2018**, *51*, 92. [CrossRef]
2. Shi, W.; Cao, J.; Zhang, Q.; Li, Y.; Xu, L. Edge computing: Vision and challenges. *IEEE Internet Things J.* **2016**, *3*, 637–646. [CrossRef]
3. Samad, A.M.; Kamarulzaman, N.; Hamdani, M.A.; Mastor, T.A.; Hashim, K.A. The potential of Unmanned Aerial Vehicle (UAV) for civilian and mapping application. In Proceedings of the 2013 IEEE 3rd International Conference on System Engineering and Technology, Shah Alam, Malaysia, 19–20 August 2013; pp. 313–318.
4. Heintz, F.; Rudol, P.; Doherty, P. From images to traffic behavior-a uav tracking and monitoring application. In Proceedings of the 2007 10th International Conference on Information Fusion, Quebec, QC, Canada, 9–12 July 2007; pp. 1–8.
5. Bravo, R.Z.B.; Leiras, A.; Cyrino Oliveira, F.L. The use of UAVs in humanitarian relief: An application of POMDP-based methodology for finding victims. *Prod. Oper. Manag.* **2019**, *28*, 421–440. [CrossRef]
6. Ma, J.; Ma, Y.; Li, C. Infrared and visible image fusion methods and applications: A survey. *Inf. Fusion* **2019**, *45*, 153–178. [CrossRef]
7. Zhang, Z. Drone-YOLO: An efficient neural network method for target detection in drone images. *Drones* **2023**, *7*, 526. [CrossRef]
8. Yang, Z.; Lian, J.; Liu, J. Infrared UAV Target Detection Based on Continuous-Coupled Neural Network. *Micromachines* **2023**, *14*, 2113. [CrossRef] [PubMed]
9. Bai, X.; Bi, Y. Derivative entropy-based contrast measure for infrared small-target detection. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 2452–2466. [CrossRef]
10. Mohsan, S.A.H.; Othman, N.Q.H.; Li, Y.; Alsharif, M.H.; Khan, M.A. Unmanned aerial vehicles (UAVs): Practical aspects, applications, open challenges, security issues, and future trends. *Intell. Serv. Robot.* **2023**, *16*, 109–137. [CrossRef] [PubMed]
11. Zhang, M.; Zhang, R.; Yang, Y.; Bai, H.; Zhang, J.; Guo, J. ISNet: Shape matters for infrared small target detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 877–886.
12. 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 *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September, 2014, Proceedings, Part V 13*; Springer Internation Publishing: Cham, Switzerland, 2014; pp. 740–755.
13. Bai, X.; Zhou, F. Analysis of new top-hat transformation and the application for infrared dim small target detection. *Pattern Recognit.* **2010**, *43*, 2145–2156. [CrossRef]
14. Deshpande, S.D.; Er, M.H.; Venkateswarlu, R.; Chan, P. Max-mean and max-median filters for detection of small targets. In Proceedings of the Signal and Data Processing of Small Targets 1999, Denver, CO, USA, 19–23 July 1999; pp. 74–83.
15. Hou, X.; Zhang, L. Saliency detection: A spectral residual approach. In Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 17–22 June 2007; pp. 1–8.
16. Zhang, L.; Peng, Z. Infrared small target detection based on partial sum of the tensor nuclear norm. *Remote Sens.* **2019**, *11*, 382. [CrossRef]
17. Wang, Y.; Tian, Y.; Liu, J.; Xu, Y. Multi-Stage Multi-Scale Local Feature Fusion for Infrared Small Target Detection. *Remote Sens.* **2023**, *15*, 4506. [CrossRef]
18. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
19. 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.
20. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. Ssd: Single shot multibox detector. In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016, Proceedings, Part I 14*; Springer Internation Publishing: Cham, Switzerland, 2016; pp. 21–37.
21. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 779–788.
22. 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.
23. Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; Chen, J. Detrs beat yolos on real-time object detection. *arXiv* **2023**, arXiv:2304.08069.

24. Zhang, M.; Bai, H.; Zhang, J.; Zhang, R.; Wang, C.; Guo, J.; Gao, X. Rkformer: Runge-kutta transformer with random-connection attention for infrared small target detection. In Proceedings of the 30th ACM International Conference on Multimedia, Lisboa, Portugal, 10–14 October 2022; pp. 1730–1738.
25. Li, N.; Huang, S.; Wei, D. Infrared Small Target Detection Algorithm Based on ISTD-CenterNet. *Comput. Mater. Contin.* **2023**, *77*, 3511–3531. [CrossRef]
26. Chen, Y.; Liu, Z.; Zhang, L.; Wu, Y.; Zhang, Q.; Zheng, X. MFFNet: A lightweight multi-feature fusion network for UAV infrared object detection. *Egypt. J. Remote Sens. Space Sci.* **2024**, *27*, 268–276.
27. Sun, S.; Mo, B.; Xu, J.; Li, D.; Zhao, J.; Han, S. Multi-YOLOv8: An Infrared Moving Small Object Detection Model Based on YOLOv8 for Air Vehicle. *Neurocomputing* **2024**, *588*, 127685. [CrossRef]
28. Li, D.; Lin, S.; Lu, X.; Zhang, X.; Cui, C.; Yang, B. IMD-Net: Interpretable multi-scale detection network for infrared dim and small objects. *Math. Biosci. Eng.* **2024**, *21*, 1712–1737. [CrossRef] [PubMed]
29. Meng, H.; Si, S.; Mao, B.; Zhao, J.; Wu, L. LAGSwin: Local attention guided Swin-transformer for thermal infrared sports object detection. *PLoS ONE* **2024**, *19*, e0297068. [CrossRef] [PubMed]
30. 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, 20–25 June 2021; pp. 13733–13742.
31. 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.
32. Pan, Z.; Cai, J.; Zhuang, B. Fast vision transformers with hilo attention. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 14541–14554.
33. Kang, M.; Ting, C.-M.; Ting, F.F.; Phan, R.C.-W. ASF-YOLO: A Novel YOLO Model with Attentional Scale Sequence Fusion for Cell Instance Segmentation. *arXiv* **2023**, arXiv:2312.06458. [CrossRef]
34. Li, H.; Li, J.; Wei, H.; Liu, Z.; Zhan, Z.; Ren, Q. Slim-neck by GSConv: A better design paradigm of detector architectures for autonomous vehicles. *arXiv* **2022**, arXiv:2206.02424.
35. Zhang, H.; Xu, C.; Zhang, S. Inner-iou: More effective intersection over union loss with auxiliary bounding box. *arXiv* **2023**, arXiv:2311.02877.
36. 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, 15–20 June 2019; pp. 658–666.
37. 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.
38. 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] [PubMed]
39. Zhong, J.; Chen, J.; Mian, A. DualConv: Dual convolutional kernels for lightweight deep neural networks. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, *34*, 9528–9535. [CrossRef] [PubMed]
40. Zhang, X.; Song, Y.; Song, T.; Yang, D.; Ye, Y.; Zhou, J.; Zhang, L. AKConv: Convolutional Kernel with Arbitrary Sampled Shapes and Arbitrary Number of Parameters. *arXiv* **2023**, arXiv:2311.11587.
41. Qi, Y.; He, Y.; Qi, X.; Zhang, Y.; Yang, G. Dynamic snake convolution based on topological geometric constraints for tubular structure segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 6070–6079.
42. Zhu, X.; Hu, H.; Lin, S.; Dai, J. Deformable convnets v2: More deformable, better results. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 9308–9316.
43. Jiang, P.-T.; Zhang, C.-B.; Hou, Q.; Cheng, M.-M.; Wei, Y. Layercam: Exploring hierarchical class activation maps for localization. *IEEE Trans. Image Process.* **2021**, *30*, 5875–5888. [CrossRef] [PubMed]
44. 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; pp. 12993–13000.
45. Gevorgyan, Z. SIoU loss: More powerful learning for bounding box regression. *arXiv* **2022**, arXiv:2205.12740.
46. Zhang, Y.-F.; Ren, W.; Zhang, Z.; Jia, Z.; Wang, L.; Tan, T. Focal and efficient IOU loss for accurate bounding box regression. *Neurocomputing* **2022**, *506*, 146–157. [CrossRef]
47. 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.
48. 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.