drones

MDPI

Article

ITD-YOLOv8: An Infrared Target Detection Model Based on YOLOv8 for Unmanned Aerial Vehicles

Xiaofeng Zhao, Wenwen Zhang *, Hui Zhang, Chao Zheng, Junyi Ma and Zhili Zhang

Xi'an Research Institute of High-Tech, Xi'an 710025, China; xife_zhao@outlook.com (X.Z.); zh17879839221@outlook.com (H.Z.); zheng00chao@outlook.com (C.Z.); junyi_m_@outlook.com (J.M.);157918018@163.com (Z.Z.)
* Correspondence: ww_z_@outlook.com

Abstract: A UAV infrared target detection model ITD-YOLOv8 based on YOLOv8 is proposed to address the issues of model missed and false detections caused by complex ground background and uneven target scale in UAV aerial infrared image target detection, as well as high computational complexity. Firstly, an improved YOLOv8 backbone feature extraction network is designed based on the lightweight network GhostHGNetV2. It can effectively capture target feature information at different scales, improving target detection accuracy in complex environments while remaining lightweight. Secondly, the VoVGSCSP improves model perceptual abilities by referencing global contextual information and multiscale features to enhance neck structure. At the same time, a lightweight convolutional operation called AXConv is introduced to replace the regular convolutional module. Replacing traditional fixed-size convolution kernels with convolution kernels of different sizes effectively reduces the complexity of the model. Then, to further optimize the model and reduce missed and false detections during object detection, the CoordAtt attention mechanism is introduced in the neck of the model to weight the channel dimensions of the feature map, allowing the network to pay more attention to the important feature information, thereby improving the accuracy and robustness of object detection. Finally, the implementation of XIoU as a loss function for boundary boxes enhances the precision of target localization. The experimental findings demonstrate that ITD-YOLOv8, in comparison to YOLOv8n, effectively reduces the rate of missed and false detections for detecting multi-scale small targets in complex backgrounds. Additionally, it achieves a 41.9% reduction in model parameters and a 25.9% decrease in floating-point operations. Moreover, the mean accuracy (mAP) attains an impressive 93.5%, thereby confirming the model's applicability for infrared target detection on unmanned aerial vehicles (UAVs).

Keywords: infrared target detection; YOLOv8; UAVs; multi-scale small target; lightweight network structure

Citation: Zhao, X.; Zhang, W.; Zhang, H.; Zheng, C.; Ma, J.; Zhang, Z. ITD-YOLOv8: An Infrared Target Detection Model Based on YOLOv8 for Unmanned Aerial Vehicles. *Drones* **2024**, *8*, 161. https://doi.org/10.3390/drones8040161

Academic Editor: Anastasios Dimou

Received: 12 March 2024
Revised: 9 April 2024
Accepted: 12 April 2024
Published: 19 April 2024

1. Introduction

Infrared target detection has the benefits of all-weather, long-range, and strong anti-interference [1], so UAV-based infrared target detection has an important role in military [2], accident search and rescue [3,4], and traffic monitoring [5–7]. However, the aerial images captured by UAVs often contain numerous multi-scale, small targets, which typically have limited features available for extraction [8]. In addition, the actual flying height of unmanned aerial vehicles often changes greatly, and the target proportion in the image changes dramatically, which affects the target detection accuracy [9]. Meanwhile, the aerial environment during UAV flying is typically intricate, with potential occurrences of extensive occlusion among compact targets. Additionally, infrared imagery lacks distinctive attributes such as texture and color [10], thereby intensifying the challenges associated with detecting infrared targets for UAVs. Hence, the investigation of UAVs' infrared target detection technology in intricate scenarios [11,12] holds immense importance and practical implications.

Given that the majority of targets in UAV aerial images are known for their compact size, typically less than 32 × 32 pixels [13], this leads to a limited number of target features and presents difficulties in acquiring sufficient informative data. Consequently, UAV infrared target detection faces a significant challenge. Compared to traditional algorithms, target detection algorithms powered by deep learning provide advantages in terms of improved accuracy in detection and scalability [14]. These advantages contribute to improving the precision of infrared target detection and achieving automated detection, thereby enhancing operational efficiency and cost-effectiveness [15]. At present, there are generally two types of target detection algorithms in deep learning: those that use a two-stage approach and those that use a one-stage approach. The two-stage algorithm is to divide the feature extraction and detection into two stages: first finding out the candidate target region, and then adjusting the candidate region to get the detection result. The two-stage algorithms include R-CNN [16] and faster RCNN [17,18], etc. Several studies suggest that two-step algorithms are suitable for applications requiring accurate detection [19,20]. Single-stage algorithms, on the other hand, directly generate the category probabilities and the location of the bounding box to get the detection results. Single-stage detection algorithms include SDD [21] and YOLO [22–25], etc. Compared to the two-step algorithm, the single-step algorithm offers faster detection speed and is better suited for detecting targets under UAV platforms because it eliminates the need to generate candidate regions. The YOLO algorithms are popular in the domain of UAV infrared target detection due to their exceptional target detection capabilities. Therefore, they are extensively used in this field. Zeng et al. [26] presented an improved YOLOv7-based target detection algorithm for UAV images. This algorithm efficiently captures feature information at various scales and enhances model accuracy by incorporating the DpSPPF module. Zhao et al. [27] suggested a YOLO-ViT grounded infrared target detection method for UAVs. This method improved the YOLOv7 backbone feature extraction network based on a lightweight MobileViT network, enhancing the ability to extract target feature information and enhancing the detection performance of the model. However, detecting targets with UAV infrared technology in complex scenes presents significant challenges due to the features of aerial infrared images trapped by UAVs. These challenges include low resolution, multi-scale imaging, and sensitivity to environmental factors.

In an effort to enhance the precision of detecting multi-scale infrared small targets amidst complex environments encountered by UAVs while simultaneously reducing computational complexity, a novel model named ITD-YOLOv8 is introduced. By implementing the lightweight GhostHGNetV2 network structure to enhance the YOLOv8 backbone network, the model's detection capability is improved while reducing computing resource consumption and accelerating inference speed. The neck structure is improved with the use of the VoVGSCSP module by fusing feature images from different layers to obtain richer feature information about the target. The AKConv module replaces the Conv module to reduce computational and parametric quantities in the model. Meanwhile, the neck structure is further optimized using the attention mechanism, CoordAtt aims to improve network focus on key information and reduce misdetection and omission during model detection. Finally, by incorporating the XIoU loss function, the model gains enhanced capability for acquiring precise details regarding both bounding box location and object class. As a result, this results in an enhancement of the target detection capability. The key findings of this study are outlined below:

1. This paper presents enhancements to the GhostHGNetV2 module for the object detection backbone network. The newly developed network backbone utilizes a combination of the Ghost and HGNetv2 modules. It achieves this by dividing input features into smaller subchannels, performing convolution operations on each subchannel, and then merging the results to produce the final output. And through downsampling, hourglass modules, and upsampling operations, features are extracted and fused at different scales. This enables the model to reduce computation and complexity while enhancing the model's multi-scale target detection performance.

2. By improving the neck structure through the VoVGSCSP module, different stage feature maps are effectively fused to reduce computational complexity and improve inference speed while maintaining the accuracy of the model detection. Replacing the Conv module with the AKConv module provides more efficient and flexible convolution operations in the model by dynamically changing the shape and size of the convolution kernel to adapt to different input data and task requirements. The attention mechanism known as CoordAtt is proposed to enhance the model's ability to identify objectives by learning the spatial relationships between target objects and improving its attention towards various positional features.
3. The utilization of the XIoU loss function enhances the precision of matching between the prediction frame and real frame by addressing their overlap, reaching enhanced detection accuracy for diminutive and compact targets.

2. Related Work

Due to its fast processing speed, high accuracy, and excellent performance in real-world scenarios, the YOLO Series algorithm is widely used for real-time object detection. Consequently, it has been widely used for detecting objects. A deep learning approach based on channel pruning was introduced by Wang et al. [28] to enhance the speed and precision of apple fruit detection in YOLOv5s. ASFF-YOLOv5 is a UAV road detection method introduced by Qiu et al. [29] that utilizes multi-scale feature fusion. The goal is to integrate the ASFF sensor head with the SPPF spatial pyramid pooling structure to improve target detection accuracy by increasing feature scale invariance. Liu et al. [30] proposed CAFFNet, a technique for traffic sign detection. This method employs multi-channel attention and multi-feature fusion in a detection strategy that integrates contextual information at different scales to minimize feature inconsistencies and improve detection accuracy. O. Sahin et al. [31] suggested an enhanced YOLO algorithm, YOLODrone, to address the low detection performance of existing target detection algorithms on UAV aerial images and improve UAV target detection performance; D. Padilla Carrasco et al. [32] presented T-YOLO, a concise vehicle detection model that uses YOLO and multiscale conjunctive neural networks. This model is an enhanced lightweight deep target detection approach based on the YOLOv5 architecture, demonstrating excellent performance in detecting small-sized vehicle targets. Zuo et al. [33] proposed a pyramid network model known as AFFPN, which utilizes an attentional feature fusion mechanism that enhances the shallow and deep positional and semantic information of the model, thereby improving the target detection performance; furthermore, Zhang et al. [34] introduced an innovative technique for detecting infrared targets, known as CHFNet. This method leverages the HLF cross-feature fusion module to enhance the model's expressive capacity, enabling it to better perceive distinct features while minimizing redundancy between them. Consequently, the model's effectiveness is greatly enhanced. To boost the detection capabilities of the model, Dai et al. [35] suggested a novel approach for detecting infrared targets by incorporating asymmetric contextual modulation (ACMM) into their model. The ACMM is designed to improve the extraction of target features. In their study, Liu et al. [36] incorporated GhostNet as an alternative to the conventional convolutional layer. The final backbone layer was enhanced with the SepViT module, and the channel attention mechanism (ECA) was integrated into the YOLOv5 feature extraction network. Guo et al. [37] suggested an LMSD-YOLO model for the one-step detection of the SAS target. This model incorporates a DBA module, an enhanced S-MobileNet module, and a DSASFF module to improve the capability of the network to extract features and decrease its complexity. Wang et al. [38] proposed an improved YOLOv7-tiny method for detecting targets in UAV aerial images. They introduced a bi-directional feature pyramid network (BiFPN) in the neck to enhance the fusion capability of features. Additionally, they incorporated a global attention mechanism (GAM) in the neck to enhance target detection accuracy, specifically for UAV aerial images. Zhong et al. [39] developed a UAV image target detection algorithm utilizing YOLOv7. The algorithm includes a high-level feature-focusing layer (M-FLAM) to im-

prove focus on smaller targets and a low-level feature-focusing layer (M-ELAN) to reduce parameters while maintaining target detection accuracy.

3. Proposed Methodologies, Tools, and Techniques

This section is composed of three parts: an introduction to the overall framework and principles of the ITD-YOLOv8 model, a review of the HiT-UAV dataset [40,41], and an examination of the assessment techniques utilized to verify this paper's methodology.

3.1. ITD-YOLOv8

This research paper presents the ITD-YOLOv8 model for detecting infrared targets using UAVs. This model is visually depicted in Figure 1. The ITD-YOLOv8 model incorporates the GhostHGNetV2 network architecture, which synergistically combines the Ghost module and HGBlock module to accomplish efficient feature extraction and fusion. To enhance our model's detection capabilities, we used the VoVGSCSP module in the neck network to supplement the original C2f module with abundant feature data. Additionally, we replaced the Conv layer in the Neck module with an AXConv layer to enable more efficient and flexible convolutional operations. To optimize network performance, we introduced a CoordAtt module in the neck that assigns weights to feature images, enabling better focus on crucial features. Furthermore, for improved detection accuracy in our model, we incorporated the XIoU loss function into the detection header. This loss function accurately measures the overlap between target frames and prediction frames, thereby enhancing target detection precision.

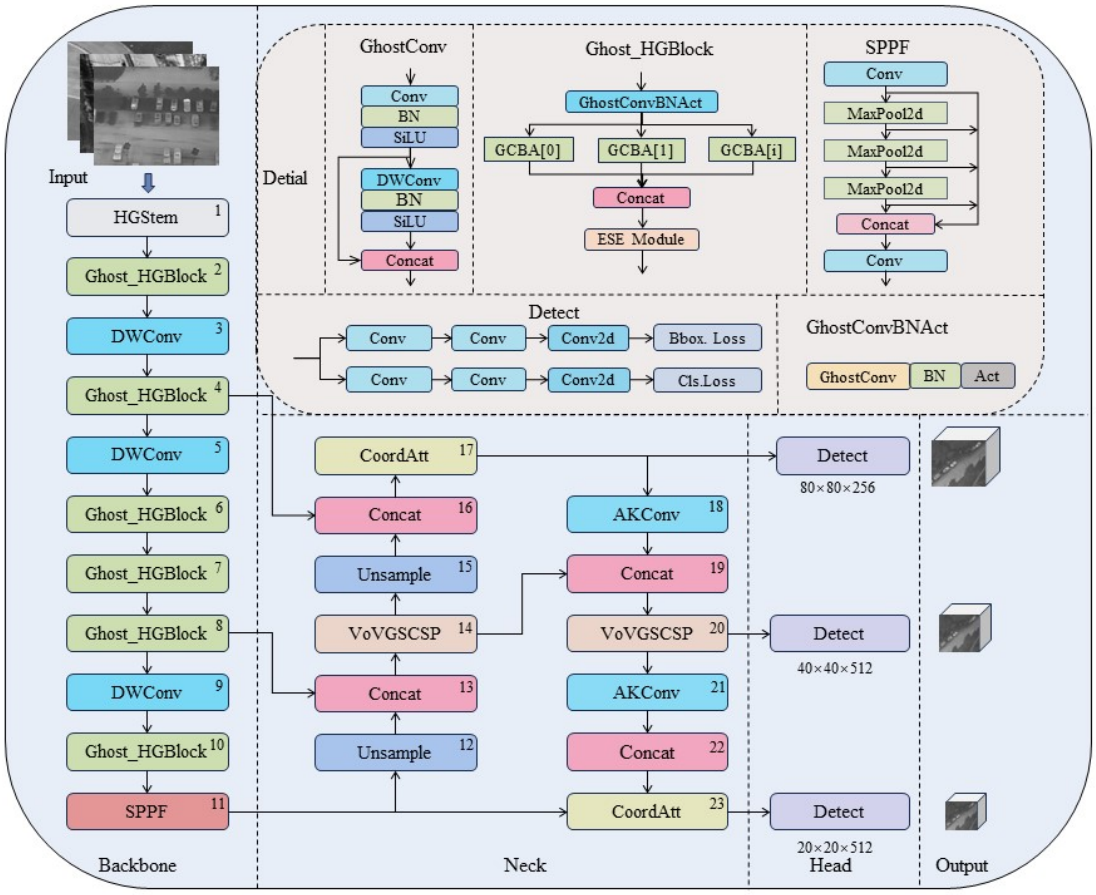

Figure 1. ITD-YOLOv8 network structure diagram.

3.1.1. Enhanced Core Network Utilizing GhostHGNetV2 Architecture

In April 2023, Baidu proposed RT-DETR [42], the first real-time DETR model. As depicted in Figure 2, GhostHGNetV2 is formed by merging the Ghost module and the HGNetv2 module. By optimizing the HGBlock with the GhostConv module, we obtained the Ghost_HGBlock module. This module reduces model parameters and computational requirements while maintaining high detection accuracy. However, the GhostHGNetV2 module consists of multiple Ghost_HGBlock modules to improve the YOLOv8 backbone network. While maintaining good detection performance, this technique reduces the number of model parameters and computations required.

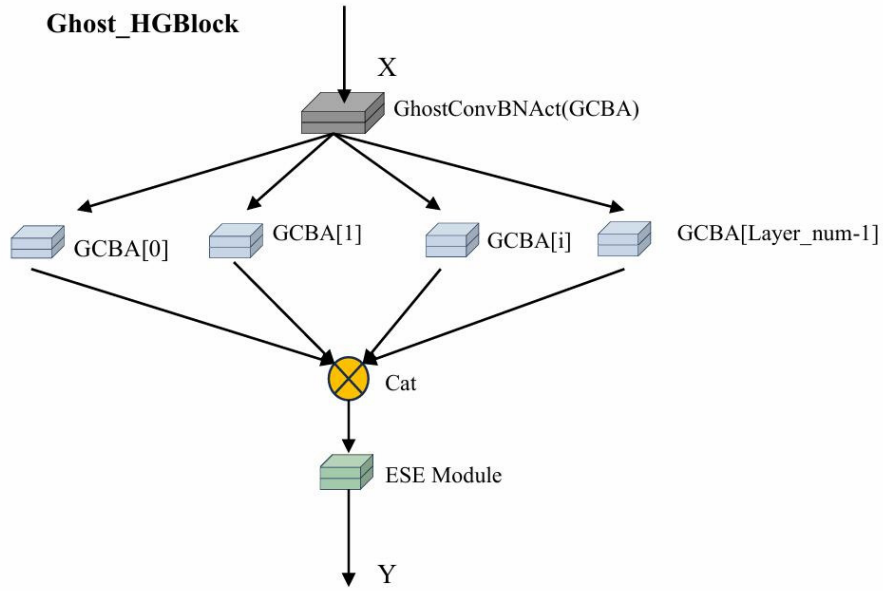

Figure 2. GhostHGNetV2 backbone network diagram.

The accuracy and lightweighting of target detection models are important evaluation indicators for UAV infrared target detection models. Due to the restricted resources of UAV platforms, this model needs to achieve lightweighting to a certain extent. Finding a balance between enhancing model detection accuracy and achieving lightweight models can be challenging. Nevertheless, YOLOv8 models may not be ideal for mobile hardware deployment due to their tendency to produce excessive parameters during training. As a solution to this problem, RT-DETRs [42], a lightweight target detector according to Baidu, effectively reduce the model parameters and computation by combining the Ghost module with HGNetv2 and maintaining good target detection performance.

3.1.2. The VoVGSCSP Module

The VoVGSCSP module [43] is a module used for image processing that improves the learning ability of the model by using GSConv and GSbottleneck and designs cross-level partial networks using a one-time aggregation method. The design of this module aims to reduce computational complexity and network structure while maintaining sufficient accuracy.

As depicted in Figure 3, the incorporation of the Bottleneck module builds upon GSConv to improve the model's capacity for learning. The VoVGSCSP module replaces the Cf2 module in the model neck, while the VoVGSCSP module uses lightweight convolution GSConv to replace standard convolution, thereby reducing computational costs. In addition, GSConv also provides model-learning capabilities comparable to standard convolution. Hence, the utilization of the VoVGSCSP module led to a decrease in model complexity without compromising on accuracy.

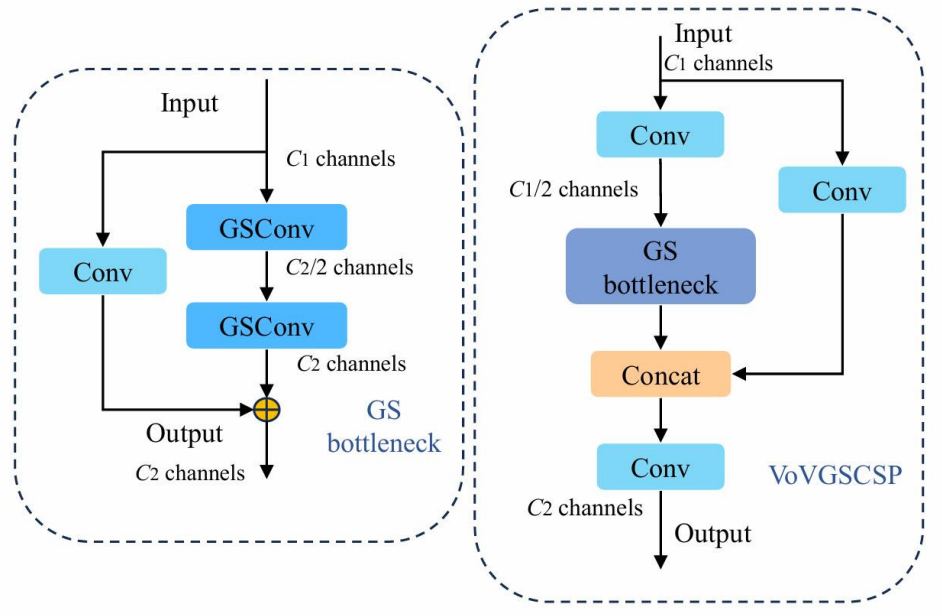

Figure 3. Structure of the VoVGSCSP module.

3.1.3. The Lightweight Convolution Module AXConv

To enhance the efficiency of resource utilization on the UAV platform, we have integrated the AXConv convolution module into our model. This integration significantly reduces both parameter counts and computational demands while maintaining optimal detection performance. The AXConv module replaces traditional fixed-size convolution kernels with convolution kernels of different sizes, thereby achieving model lightweighting.

As shown in Figure 4, the AXConv module [44] uses two convolution kernels of different sizes, namely 1×1 and 3×3. These two convolution kernels are used for dimensionality reduction and dimensionality enhancement operations, respectively. To begin, the input feature map undergoes dimensionality reduction through the use of a 1×1 convolution kernel. The result is a reduction in the number of parameters and calculations. Next, a 3×3 convolution kernel is utilized to extract richer feature information through convolution operations on the reduced feature map. Finally, a 1×1 convolutional kernel is applied to restore the feature map's initial dimensionality and enhance its features.

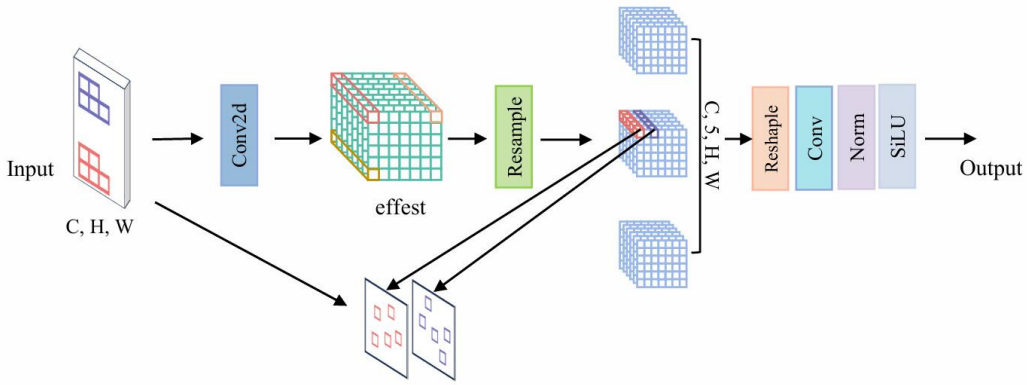

Figure 4. AXConv module structure diagram.

3.1.4. The Attention Mechanism CoordAtt

By incorporating the CoordAttention mechanism within the network's neck, it enhances the network's ability to effectively capture target location information, resulting in improved object detection performance. While traditional attention mechanisms, such as SE attention, only consider information between encoded channels, CoordAttention

considers both inter-channel and positional information. CoordAtt improves object detection accuracy by utilizing positional information to better capture the structure and spatial relationships of objects.

As illustrated in the diagram shown in Figure 5, once the feature map undergoes convolution across multiple layers, a single point encompasses details about a specific region on the original map. To capture extensive long-range information, CoordAtt is devised by performing average pooling both horizontally and vertically. Subsequently, this pooled data is transformed to encode spatial characteristics before being integrated with channel-wise weights for fusion purposes.

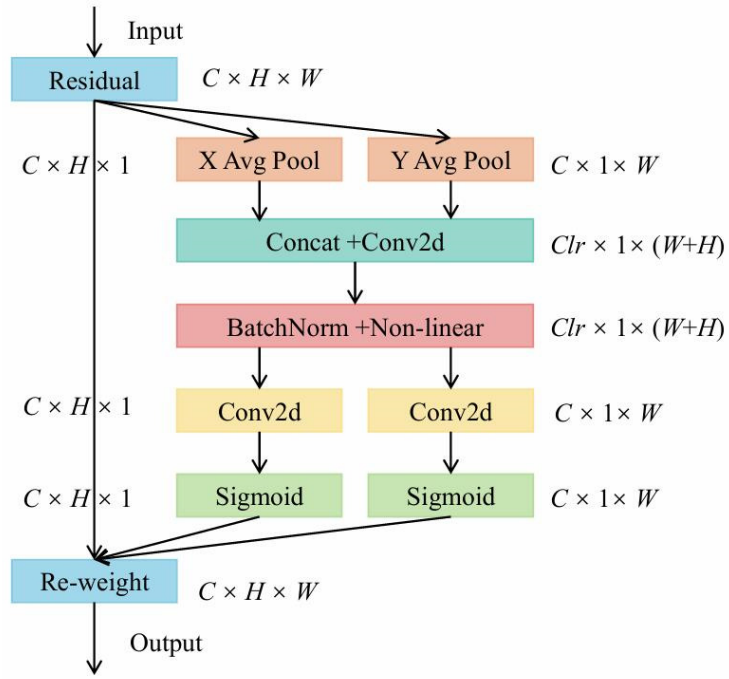

Figure 5. Structure of the attention mechanism CoordAtt.

The CoordAtt [45] module is designed to improve the representation of features in mobile networks. An attention matrix is generated by applying a convolution to the feature map, which assigns weights to the coordinates of the bounding box. This matrix matches the size of the feature map. utilizes accurate positional details to encode relationships between channels and dependencies over a long period of time. This procedure consists of two stages: embedding coordinate information and producing coordinate attention. Although pooling is a common method to encode spatial information using channel attention, it can result in losing positional detail by compressing spatial information into channel descriptors. To overcome this limitation, we propose to transform global pooling into a set of one-dimensional feature encoding operations using the following equation decomposition to accurately capture remote spatial interactions with precise position information:

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} x_c(i, j). \tag{1}$$

The output location related to the cth channel is determined by encoding each channel's input X using two spatial domains (H,1) or (1,W) along the vertical and horizontal directions of the pooled kernel. Therefore, we can express the output of the cth channel at height h as follows:

$$z_c^h(h) = \frac{1}{W} \sum_{0 \le i < W} x_c(h, i). \tag{2}$$

Similarly, the outcome of the c channel having a width of w can be represented as:

$$z_c^w(w) = \frac{1}{H} \sum_{0 \le j < H} x_c(j, w). \tag{3}$$

This section combines the above transformations and applies convolutional transform functions to modify them after the information embedding has been transformed. The $[-,-]$ operation is a sequential process conducted along the spatial dimension and incorporates a non-linear function of the activation. The spatial dimension of tensor f is used to create two different tensors: $f^h \in \mathbb{R}^{C/r \times H}$ and $f^w \in \mathbb{R}^{C/r \times W}$, where f^h represents the feature map in the height direction, and f^w represents the feature map in the width direction. g^h denotes the attention weight in the horizontal direction, and g^w denotes the attention weight in the vertical direction. To introduce spatial information in horizontal and vertical dimensions, as an intermediate step, a feature map called $f \in \mathbb{R}^{C/r \times (h+W)}$ is generated:

$$f = \delta\left(F_1\left(\left[z^h, z^w\right]\right)\right), \tag{4}$$

$$g^h = \sigma\left(F_h\left(\mathrm{f}^h\right)\right), \tag{5}$$

$$g^w = \sigma(F_w(\mathrm{f}^w)). \tag{6}$$

The two 1×1 convolutional operations are employed to convert and into tensors with identical channel dimensions as the input X. g_c^h denotes the position encoding parameter in the horizontal direction, and g_c^w denotes the position encoding parameter in the vertical direction. Σ denotes a sigmoid function, which helps in simplifying the complexity of the overhead model. Finally, the output Y of the attention block can be expressed as follows:

$$y_c(i, j) = x_c(i, j) \times g_c^h(i) \times g_c^w(j). \tag{7}$$

3.1.5. The Loss Function XIoU

The XIoU loss function is a loss function utilized in target detection tasks to assess the level of overlap between the predicted and actual boxes. It represents an enhancement over the IoU (Intersection over Union) loss function, which quantifies the extent of overlap between the predicted and ground-truth boxes through this calculation:

$$\mathrm{IoU} = \frac{A \cap B}{A \cup B} \tag{8}$$

In this given scenario, A symbolizes the anticipated box, whereas B signifies the real frame. The XIoU loss function can be formulated as:

$$\mathrm{XIoU} = \mathrm{IoU} - \frac{(\mathrm{C1 - C2})}{\sigma} \tag{9}$$

where IOU denotes the intersection and concurrency ratio of the two boxes, C1 and C2 denote the distance between the centers of the two boxes, respectively, and σ denotes the sum of the areas of the two boxes. To improve target detection accuracy, the XIoU loss function can better penalize the offset between the predicted and real boxes.

3.2. Datasets

In order to conduct model validation experiments, the HIT-UAV dataset is selected for analysis in this study. The HIT-UAV dataset is specifically designed for UAV infrared target acquisition. As shown in Figure 6, this dataset contains a variety of UAV infrared images covering targets in different scenes and weather conditions. The reduced dataset consists of 2898 infrared images, each of which has a resolution of 640×512 and consists of three different classes: people, bicycles, vehicles, etc. The HIT-UAV dataset consists of infrared images captured by UAVs at high altitude, which are captured from different

heights, viewing angles, and object classes and have different sizes and shapes at different scales. This allows the model to better understand and capture the diversity and complexity of the dataset, which results in more comprehensive information. In addition, the model can be exposed to various scales of data to improve its generalization ability so as to better adapt to different scales of input and enhance its robustness.

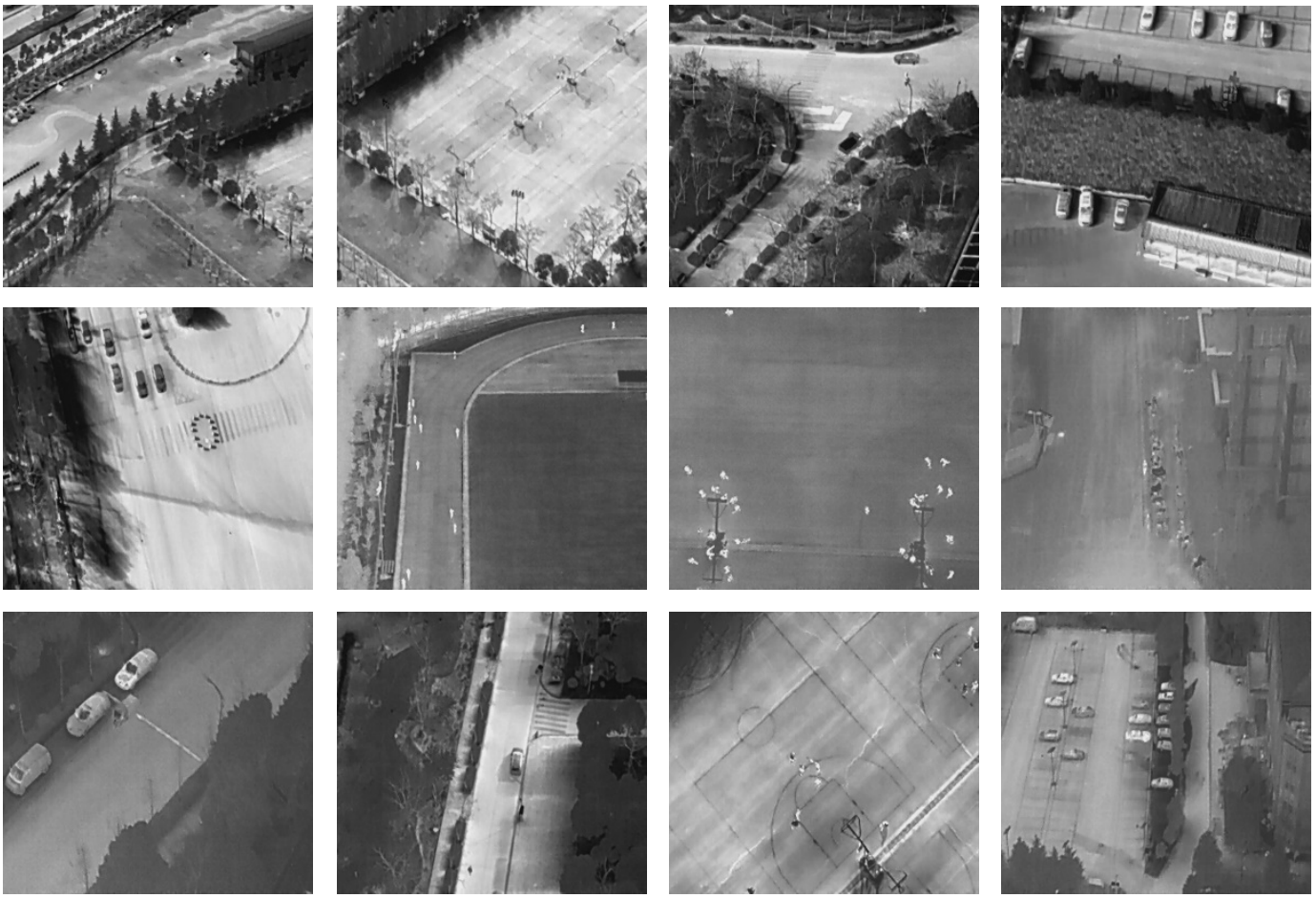

Figure 6. The HIT-UAV dataset.

In this research paper, we have partitioned the HIT-UAV dataset into three segments with a distribution ratio of 7:2:1. The dataset contains 2008 pictures for training, 571 pictures for testing, and 287 pictures for validation. According to Table 1, the HIT-UAV dataset consists of 17,118 small target labels smaller than 32 × 32 pixels, 7249 medium target labels smaller than 96 × 96 pixels, and 384 large target labels. Among them, the smallest target only accounts for 0.01% of the image pixels, and this dataset is important for the research and application of UAV infrared target detection and recognition.

Table 1. HIT-UAV dataset label classification.

	Small (0, 32 × 32)	**Medium (32 × 32, 96 × 96)**	**Large (96 × 96, 640 × 512)**
HIT-UAV	17,118	7249	384
Train set	12,045	5205	268
Test set	3331	1379	70
Validation set	1742	665	46

3.3. Evaluation Indicators

To assess the efficiency of the proposed improved ITD-YOLOv8 model, various evaluation metrics are used, including Precision (P), Recall (R), F1 (F1 Score), AP (Average Precision), mAP (Mean Average Precision), FLOPs, and FPS. F1 scores are combined measures of confidence and recall, while both AP and mAP are final measures of model recognition accuracy. FLOPs is a measure of computer hardware performance and algorithm complexity, while FPS is used to evaluate the real-time performance and efficiency of the target detection algorithm by measuring the number of image frames processed per second. Below is the equation for evaluating these parameters:

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

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

$$\text{F}_1 = \frac{\text{P} * \text{R}}{\text{P} + \text{R}} \tag{12}$$

P stands for precision, R stands for recall, TP indicates true positives, FN indicates false negatives, and FP indicates false positives. The equations for AP and mAP are the following:

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

$$\text{mAP} = \frac{1}{\text{c}} \sum_{j=1}^{c} AP_i \tag{14}$$

4. Experimental Findings

4.1. Platform for Conducting Experiments and Configuring Parameters

This study was carried out using a Windows 10 operating system, and the specific specifications of the platform can be found in Table 2.

Table 2. Configuring the experimental platform.

Name	Related Configurations
Graphics processor	NVIDIA Quadro P6000
Central processor	Intel(R) Core(TM) i9-9900k
Graphic processor memory size	32 G
OS name	Win 10
The computing platform	CUDA10.2
Architecture for deep learning	Pytorch

4.2. Comparison of Experiments

To assess the feasibility and effectiveness of ITD-YOLOv8 in detecting infrared targets in complex UAV scenes, it is compared to the most advanced target detection algorithms currently available under similar conditions. No pre-training weights were used in all the models training processes. The dimensions of the input image for the model were configured as 640 × 640, while a batch size of 16 was used. The training process consisted of 300 epochs. Table 3 shows the results of the experiments comparing the performance of ITD-YOLOv8 and YOLOv8n. Compared to YOLOv8n, the ITD-YOLOv8 parameter set is decreased by 41.9%, the model complexity is decreased by 25.9%, and the vehicle AP is improved by 0.2% with an average accuracy of 93.5%. The ITD-YOLOv8 model effectively maintains target detection accuracy and significantly reduces model complexity. The ITD-YOLOv8 model is implemented on the basis of YOLOv8 by improving the backbone network, neck, and detection head. ITD-YOLOv8 is similar to YOLOv8 in that the ITD-YOLOv8 model adopts a modular design, which can choose different network

structures according to requirements, and hardware acceleration calculations such as GPUs can be used to improve the speed of the algorithm. In addition, the network structure can be optimized by compression and distillation to improve performance. Therefore, ITD-YOLOv8 has good scalability and has a good application prospect for UAV infrared target tasks.

Table 3. Performance comparison between ITD-YOLOv8 and YOLOv8n.

	Parameters	GFLOPs/G	Precision (%)	Recall (%)	F1 (%)	$AP_{Vehicle}$ (%)	mAP50 (%)
YOLOv8n	3.1 M	8.1	91.6	90.3	90.9	98.0	94.6
ITD-YOLOv8	**1.8 M**	6.0	90.3	88.6	89.4	**98.2**	93.5

To evaluate the performance of ITD-YOLOv8, we selected other YOLO models as benchmarks for experimental validation on the HIT-UAV dataset. The dataset consists of images captured at heights ranging from 60 m to 130 m, resulting in diverse and complex backgrounds as well as significant variations in target sizes. Moreover, most targets are small in size, posing a considerable challenge for detection. Table 4 showcases the comparative results of ITD-YOLOv8 against other models. The results show that, compared to the lightweight target detection algorithms within the YOLO series, ITD-YOLOv8 exhibits remarkable reductions in parameters and FLOPs metrics while achieving a vehicle detection accuracy AP of 98.2%. Specifically, it reduces model parameters by 28.0%, 70.5%, and 41.9% compared to YOLOv5n, YOLOv7-tiny, and YOLOv8n, respectively, additionally reducing computational complexity by 15.5%, 38.6%, and 25.9%, respectively, for each model mentioned above. The experimental evaluation results indicate that the designed ITD-YOLOv8 model achieved a significant level of accuracy and precision in infrared vehicle target identification in complex environments while maintaining a lightweight structure.

Table 4. Comparison of performance between ITD-YOLOv8 and other algorithms.

Model	Size	Parameters	F1 (%)	AP_{Person} (%)	$AP_{Vehicle}$ (%)	$AP_{Bicycle}$ (%)	mAP50 (%)	FLOPs /G
YOLOv5s	640	9.1 M	91.1	93.2	98.3	93.1	94.9	23.8
YOLOv5m	640	25.0 M	90.3	92.7	97.9	91.4	94.0	64.0
YOLOv5l	640	53.1 M	91.2	92.5	98.1	91.1	93.9	134.7
YOLOv7	640	36.5 M	86.2	88.2	94.2	88.3	90.2	103.2
YOLOv8s	640	11.2 M	91.1	93.0	98.2	91.4	94.2	28.4
YOLO-ViT	640	17.3 M	90	91.3	98.1	90.6	94.5	33.1
YOLOv5n	640	2.5 M	92.7	92.2	98.0	93.4	94.6	7.1
YOLOv7-tiny	640	6.1 M	89.8	92.5	97.0	91.3	93.6	13.2
YOLOv8n	640	3.1 M	90.9	92.4	98.0	93.4	94.6	8.1
ITD-YOLOv8	640	**1.8 M**	90.3	91.7	**98.2**	90.7	93.5	**6.0**

The visualization comparison experiment results of ITD-YOLOv8 and other YOLO series models in complex scenes are shown in Figure 7. The detection scenes of the model include roads, fields, forests, etc. Among them, the targets in the red detection box represent vehicles, the targets in the blue box represent bicycles, and the targets in the green box represent people. From the third column graph, it can be seen that most models have the problem of missing detection for occluded vehicle targets, while ITD-YOLOv8 and YOLOv7 can accurately detect the occluded vehicle targets. However, from the first column of the graph, it can be seen that YOLOv7 has insufficient ability to detect vehicle targets in the jungle, while ITD-YOLOv8 can accurately detect various occluded vehicle targets.

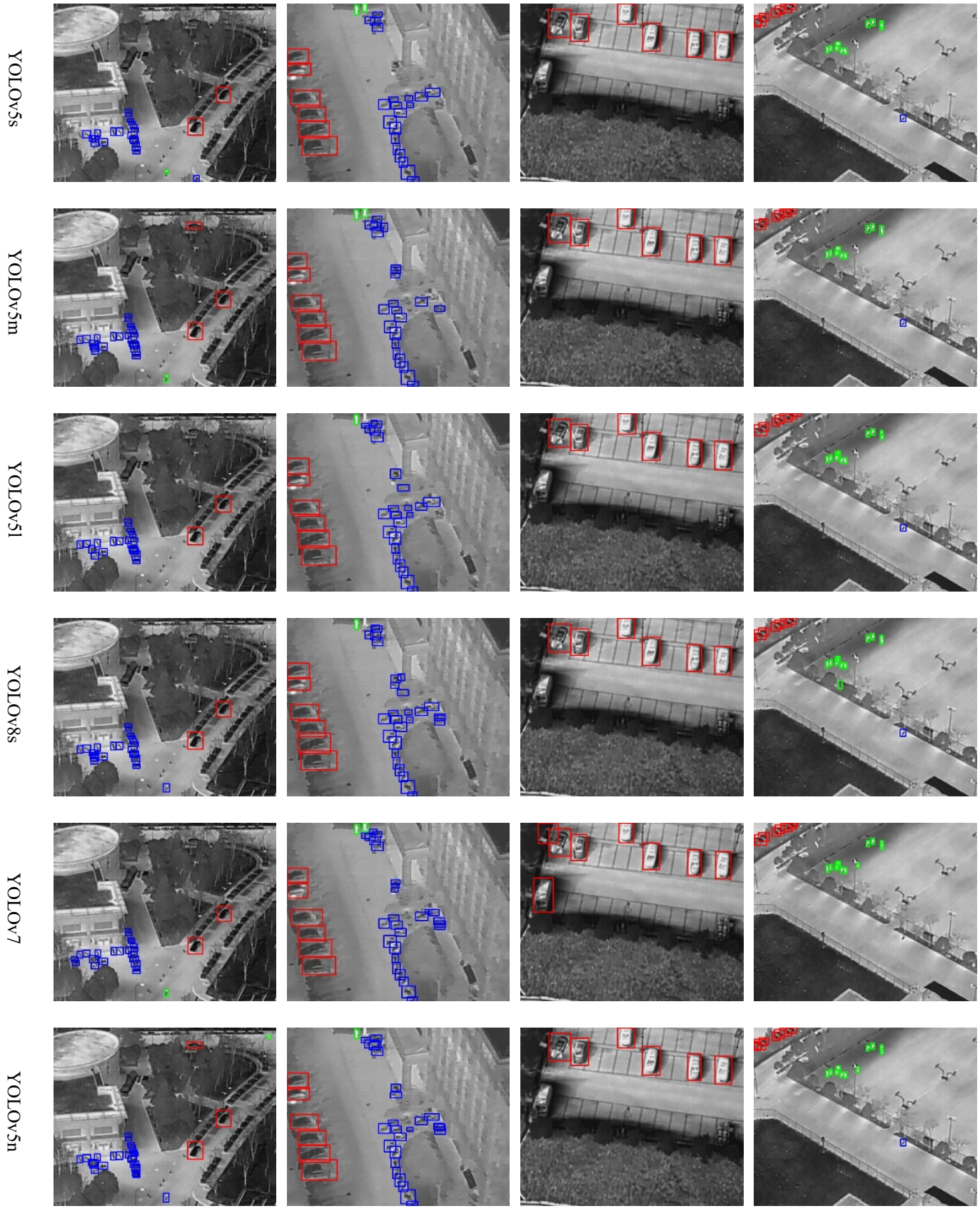

Figure 7. *Cont*.

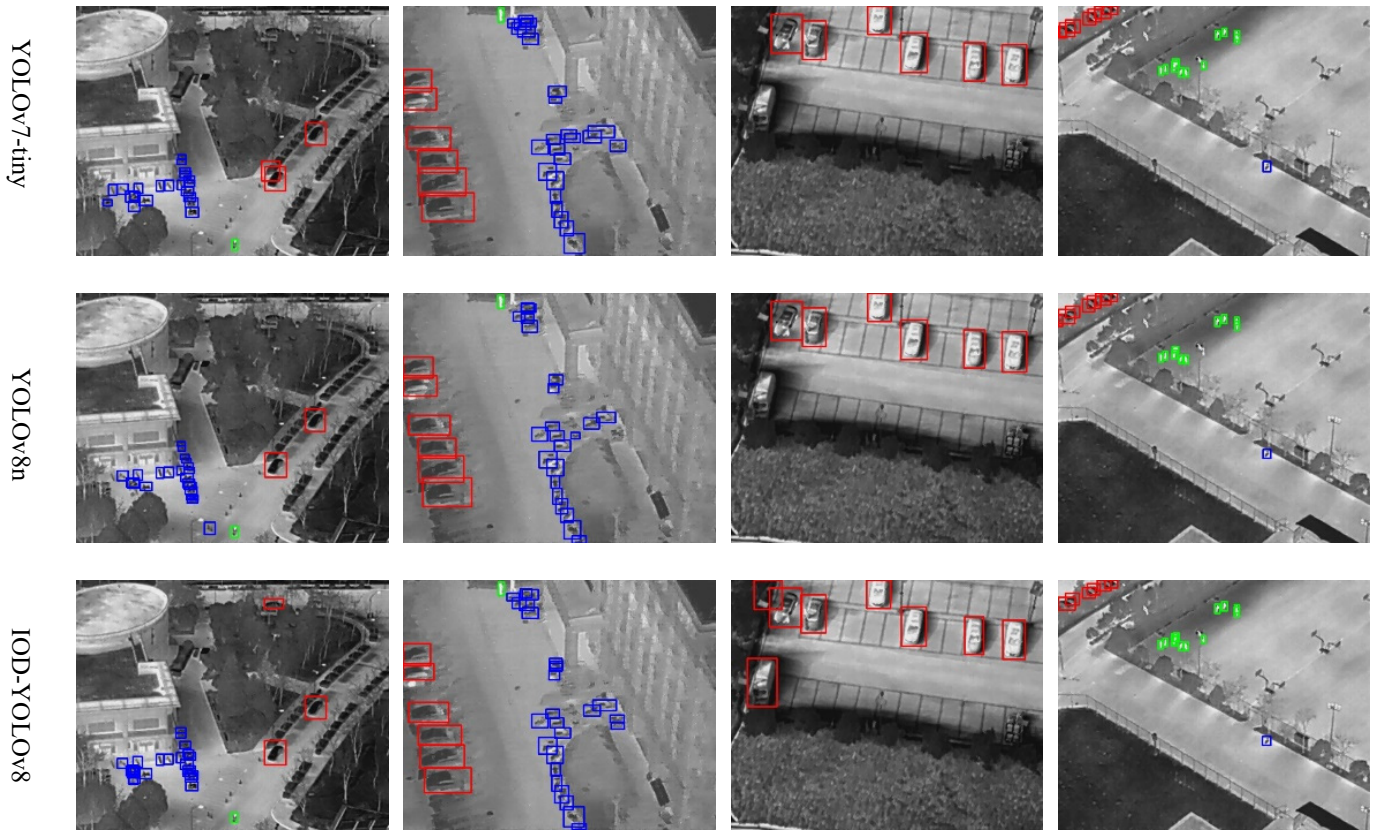

Figure 7. Comparison of the visualization results of different algorithms, where the vehicle target is represented by a red box, and the bicycle and person targets are represented by blue and green boxes, respectively.

Furthermore, the second column of graphs shows that the model improves well on the missed detection of bicycle targets in roadway scenarios. From the fourth column graphs, it can be seen that ITD-YOLOv8 has greatly improved the missed and false detections of people. The model has achieved good detection results in woods, roads, playgrounds, and other environments and can be applied to detection tasks in various environments.

4.3. Ablation Experiments

To assess the individual impact of each component on ITD-YOLOv8, a range of experiments were performed on the HIT-UAV dataset. Ablation experiments have been performed with an input image size of 640 × 640, a batch size of 16, and 300 epochs of training for each network. The trial results are presented in Table 5. It seems that the model backbone network improved based on the GhostHGNetV2 module, which reduced the model parameter count by 22.6% and mAP by 0.9%. After improving the neck with the VoVGSCSP module, the model parameters decreased by 0.1 M, while F1 increased by 0.3% and mAP decreased by 0.3%. Furthermore, by substituting the initial convolutional layer with a lightweight convolutional module, AXConv, in the neck, the parameter count was reduced by 0.2 M, F1 was increased by 0.3%, and mAP was not decreased. The introduction of the attention mechanism module, CoordAtt, in the neck reduced the number of parameters by 0.5 M but only decreased mAP by 0.8%. Finally, the introduction of the XIoU loss function increased the model's mean average precision (mAP) by 0.4% and frames per second (FPS) by 5. The number of ITD-YOLOv8 model parameters is drastically reduced compared to YOLOv8n, with 58.1% of the number of model parameters for YOLOv8n under the same settings. The experiments illustrate that each improved module contributes well to the model.

Table 5. Results from the HIT-UAV vehicle infrared dataset of ablation experiments.

YOLOv8	GhostHGNetV2	VoVGSCSP	AKConv	CoordAtt	XIoU	Parameters	FLOPs/G	F1 (%)	mAP (%)	FPS
√						3.1 M	8.1	90.9	94.6	123
√	√					2.4 M	6.9	90.1	93.7	111
√		√				3.0 M	7.8	91.2	94.3	115
√			√			2.9 M	8.0	91.2	94.6	87
√				√		2.6 M	7.8	91.0	94.5	114
√					√	3.1 M	8.1	90.9	94.6	111
√	√	√				2.3 M	6.6	90.1	93.9	102
√	√	√	√			2.2 M	6.3	90.1	93.6	75
√	√	√	√	√		1.8 M	6.0	89.3	93.1	71
√	√	√	√	√	√	1.8 M	6.0	89.4	93.5	76

Figure 8 demonstrates the enhancement effect of the ITD-YOLOv8 method for each module in the ablation experiment. The individual modules therein effectively reduce the complexity and number of parameters of the original model and retain good detection accuracy. By incrementally incorporating experimental modules into the original model to evaluate its performance under varying conditions, it has been confirmed that ITD-YOLOv8 maintains high detection accuracy, enhances the model's robustness and efficiency, and dramatically reduces the number of parameters and the complexity of the model.

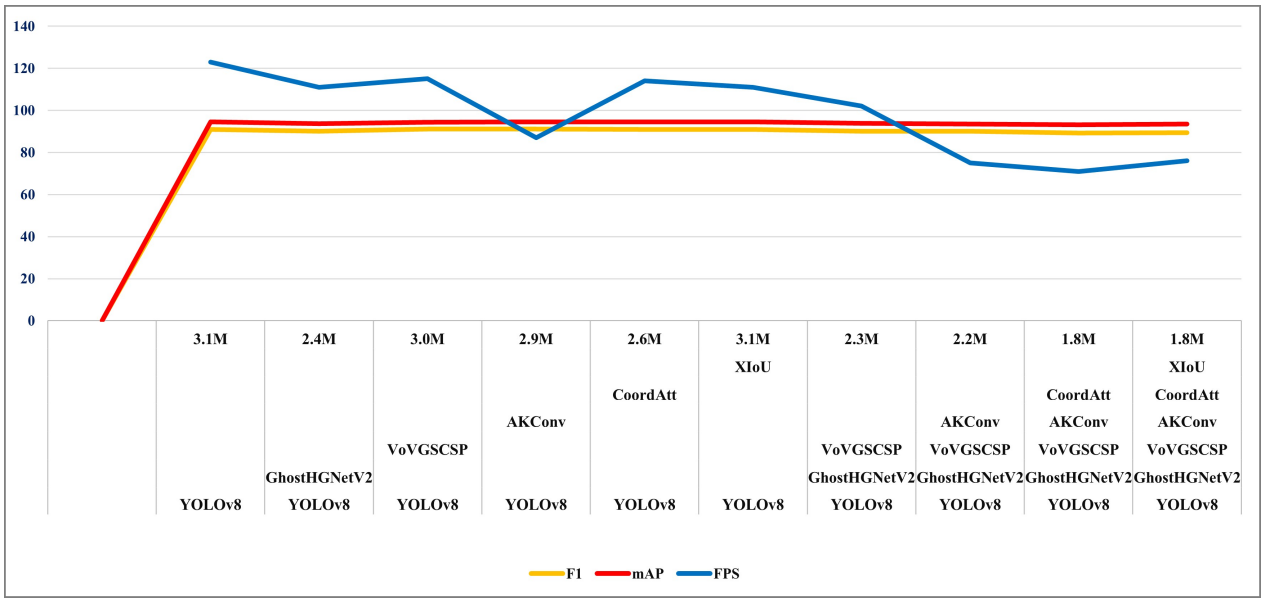

Figure 8. Comparison of ablation module performance.

5. Conclusions

This paper proposes a UAV Infrared Target Detection Model, ITD-YOLOv8, for multi-scale infrared target detection in complex scenes. In the default method, firstly, a backbone feature extraction network is designed based on the GhostHGNetV2 module, which is used as the backbone network module of YOLOv8 to enhance multi-scale feature extraction capability and reduce computational complexity. Secondly, by combining the AKConv module, VoVGSCSP module, and CoordAtt attention mechanism to improve the neck structure, the parameter and computational complexity are effectively reduced while enhancing the model's ability to detect infrared-occluded targets in complex environments.

Finally, to improve the detection accuracy and efficiency of the model, the XIoU loss function is introduced to better measure the degree of matching between detection results and real targets. The experimental results show that the ITD-YOLOv8 model can effectively improve the detection performance of multi-scale small targets in complex environments while significantly reducing model complexity and parameter quantity and reducing the missed detection rate and false detection rate. Compared with YOLOv8n, the number of model parameters decreased by 41.9%, the number of floating-point operations decreased by 25.9%, and the average accuracy (mAP) reached 93.5%. The ITD-YOLOv8 model is a lightweight infrared target detection model that has a good detection effect on multi-scale infrared small targets in complex scenes. It can be deployed on the UAV platform for real-time detection and can implement infrared target detection tasks in scenes such as roads, fields, woods, etc. It has a good application prospect in the multi-scale infrared target detection task of unmanned aerial vehicles in complex scenes.

Author Contributions: Conceptualization, X.Z. and W.Z.; methodology, X.Z. and W.Z.; software, X.Z. and W.Z.; validation, W.Z., H.Z. and C.Z.; formal analysis, W.Z. and J.M.; investigation, X.Z. and W.Z.; Editing, J.M.; Writing—original drafting, X.Z., W.Z. and H.Z.; Writing—review and editing, Z.Z.; Visualization, W.Z. and H.Z.; Oversight, X.Z; Z.Z. was responsible for project administration while X.Z. and Z.Z. were involved in securing funding. All authors have read and agreed to the published version of the manuscript.

Funding: This study received funding from the National Natural Science Foundation of China (grant no. 41404022) and the National Foundation for Enhancing Fundamental Sciences in China (grant no. 2021-JCJQ-JJ-0871).

Data Availability Statement: The corresponding author may provide the data used in this study if appropriate. The corresponding author can also provide the codes used in this study upon request.

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

References

1. Zhang, C.; Li, D.; Qi, J.; Liu, J.; Wang, Y. Infrared Small Target Detection Method with Trajectory Correction Fuze Based on Infrared Image Sensor. *Sensors* **2021**, *21*, 4522. [CrossRef] [PubMed]
2. Cao, S.; Deng, J.; Luo, J.; Li, Z.; Hu, J.; Peng, Z. Local Convergence Index-Based Infrared Small Target Detection against Complex Scenes. *Remote Sens.* **2023**, *15*, 1464. [CrossRef]
3. Fan, X.; Li, H.; Chen, Y.; Dong, D. UAV Swarm Search Path Planning Method Based on Probability of Containment. *Drones* **2024**, *8*, 132. [CrossRef]
4. Oh, D.; Han, J. Smart Search System of Autonomous Flight UAVs for Disaster Rescue. *Sensors* **2021**, *21*, 6810. [CrossRef] [PubMed]
5. Qiu, Z.; Bai, H.; Chen, T. Special Vehicle Detection from UAV Perspective via YOLO-GNS Based Deep Learning Network. *Drones* **2023**, *7*, 117. [CrossRef]
6. Niu, C.; Song, Y.; Zhao, X. SE-Lightweight YOLO: Higher Accuracy in YOLO Detection for Vehicle Inspection. *Appl. Sci.* **2023**, *13*, 13052. [CrossRef]
7. Shokouhifar, M.; Hasanvand, M.; Moharamkhani, E.; Werner, F. Ensemble Heuristic–Metaheuristic Feature Fusion Learning for Heart Disease Diagnosis Using Tabular Data. *Algorithms* **2024**, *17*, 34. [CrossRef]
8. Zhang, Z. Drone-YOLO: An Efficient Neural Network Method for Target Detection in Drone Images. *Drones* **2023**, *7*, 526. [CrossRef]
9. Yang, Z.; Lian, J.; Liu, J. Infrared UAV Target Detection Based on Continuous-Coupled Neural Network. *Micromachines* **2023**, *14*, 2113. [CrossRef]
10. Fan, Y.; Qiu, Q.; Hou, S.; Li, Y.; Xie, J.; Qin, M.; Chu, F. Application of Improved YOLOv5 in Aerial Photographing Infrared Vehicle Detection. *Electronics* **2022**, *11*, 2344. [CrossRef]
11. Tang, G.; Ni, J.; Zhao, Y.; Gu, Y.; Cao, W. A Survey of Object Detection for UAVs Based on Deep Learning. *Remote Sens.* **2024**, *16*, 149. [CrossRef]
12. Wu, X.; Li, W.; Hong, D.; Tao, R.; Du, Q. Deep Learning for Unmanned Aerial Vehicle-Based Object Detection and Tracking: A survey. *IEEE Geosci. Remote Sens. Mag.* **2021**, *10*, 91–124. [CrossRef]
13. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Zitnick, C.L. *Microsoft Coco: Common Objects in Context*; Springer International Publishing: Berlin/Heidelberg, Germany, 2014.
14. 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]

15. Chang, Y.; Li, D.; Gao, Y.; Su, Y.; Jia, X. An Improved YOLO Model for UAV Fuzzy Small Target Image Detection. *Appl. Sci.* **2023**, *13*, 5409. [CrossRef]
16. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. *Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation*; IEEE Computer Society: Piscataway, NJ, USA, 2014.
17. Girshick, R. Fast R-CNN. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
18. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 6. [CrossRef]
19. Liu, M.; Wang, X.; Zhou, A.; Fu, X.; Ma, Y.; Piao, C. UAV-YOLO: Small Object Detection on Unmanned Aerial Vehicle Perspective. *Sensors* **2020**, *20*, 2238. [CrossRef] [PubMed]
20. Wu, X.; Hong, D.; Ghamisi, P.; Li, W.; Tao, R. MsRi-CCF: Multi-Scale and Rotation-Insensitive Convolutional Channel Features for Geospatial Object Detection. *Remote Sens.* **2018**, *10*, 1990. [CrossRef]
21. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. *SSD: Single Shot Multibox Detector*; Springer: Cham, Switzerland, 2016.
22. Redmon, J.; Divvala, S.K.; Girshick, R.B.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
23. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
24. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 7464–7475.
25. Reis, D.; Kupec, J.; Hong, J.; Daoudi, A. Real-Time Flying Object Detection with YOLOv8. *arXiv* **2023**, arXiv:2305.09972.
26. Zeng, Y.; Zhang, T.; He, W.; Zhang, Z. YOLOv7-UAV: An Unmanned Aerial Vehicle Image Object Detection Algorithm Based on Improved YOLOv7. *Electronics* **2023**, *12*, 3141. [CrossRef]
27. Zhao, X.; Xia, Y.; Zhang, W.; Zheng, C.; Zhang, Z. YOLO-ViT-Based Method for Unmanned Aerial Vehicle Infrared Vehicle Target Detection. *Remote Sens.* **2023**, *15*, 3778. [CrossRef]
28. Wang, D.; He, D. Channel pruned YOLO V5s-based deep learning approach for rapid and accurate apple fruitlet detection before fruit thinning. *Biosyst. Eng.* **2021**, *6*, 210. [CrossRef]
29. Qiu, M.; Huang, L.; Tang, B.-H. ASFF-YOLOv5: Multielement Detection Method for Road Traffic in UAV Images Based on Multiscale Feature Fusion. *Remote Sens.* **2022**, *14*, 3498. [CrossRef]
30. Liu, F.; Qian, Y.; Li, H.; Wang, Y.; Zhang, H. Caffnet: Channel attention and feature fusion network for multi-target traffic sign detection. *Int. J. Pattern Recognit. Artif. Intell.* **2021**, *35*, 2152008. [CrossRef]
31. Sahin, O.; Ozer, S. YOLODrone: Improved YOLO Architecture for Object Detection in Drone Images. In Proceedings of the 2021 44th International Conference on Telecommunications and Signal Processing (TSP), Brno, Czech Republic, 26–28 July 2021; pp. 361–365.
32. Carrasco, D.P.; Rashwan, H.A.; García, M.Á.; Puig, D. T-YOLO: Tiny Vehicle Detection Based on YOLO and Multi-Scale Convolutional Neural Networks. *IEEE Access* **2023**, *11*, 22430–22440. [CrossRef]
33. Zuo, Z.; Tong, X.; Wei, J.; Su, S.; Wu, P.; Guo, R.; Sun, B. AFFPN: Attention Fusion Feature Pyramid Network for Small Infrared Target Detection. *Remote Sens.* **2022**, *14*, 3412. [CrossRef]
34. Zhang, M.; Li, B.; Wang, T.; Bai, H.; Yue, K.; Li, Y. CHFNet: Curvature Half-Level Fusion Network for Single-Frame Infrared Small Target Detection. *Remote Sens.* **2023**, *15*, 1573. [CrossRef]
35. Dai, Y.; Wu, Y.; Zhou, F.; Barnard, K. Asymmetric Contextual Modulation for Infrared Small Target Detection. In Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–8 January 2021; pp. 949–958.
36. Li, S.; Yang, X.; Lin, X.; Zhang, Y.; Wu, J. Real-Time Vehicle Detection from UAV Aerial Images Based on Improved YOLOv5. *Sensors* **2023**, *23*, 5634. [CrossRef]
37. Guo, Y.; Chen, S.; Zhan, R.; Wang, W.; Zhang, J. LMSD-YOLO: A Lightweight YOLO Algorithm for Multi-Scale SAR Ship Detection. *Remote Sens.* **2022**, *14*, 4801. [CrossRef]
38. Wang, Z.; Liu, Z.; Xu, G.; Cheng, S. Object Detection in UAV Aerial Images Based on Improved YOLOv7-tiny. In Proceedings of the 2023 4th International Conference on Computer Vision, Image and Deep Learning (CVIDL), Zhuhai, China, 12–14 May 2023; pp. 370–374.
39. Chung, Y.H.; Su, C.Y. Object Detection Algorithm Based on Improved YOLOv7 for UAV Images. In Proceedings of the 2023 IEEE 5th Eurasia Conference on IOT, Communication and Engineering (ECICE), Yunlin, Taiwan, 27–29 October 2023; pp. 18–21.
40. 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]
41. Aibibu, T.; Lan, J.; Zeng, Y.; Lu, W.; Gu, N. An Efficient Rep-Style Gaussian–Wasserstein Network: Improved UAV Infrared Small Object Detection for Urban Road Surveillance and Safety. *Remote Sens.* **2024**, *16*, 25. [CrossRef]
42. Lv, W.; Xu, S.; Zhao, Y.; Wang, G.; Wei, J.; Cui, C.; Du, Y.; Dang, Q.; Liu, Y. DETRs Beat YOLOs on Real-time Object Detection. *arXiv* **2023**, arXiv:2304.08069.
43. 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.

44. 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.
45. Hou, Q.; Zhou, D.; Feng, J. Coordinate Attention for Efficient Mobile Network Design. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 13708–13717.