

Article

Improved YOLOv7-Tiny for Object Detection Based on UAV Aerial Images

Zitong Zhang ^{1,*}, Xiaolan Xie ^{1,*} , Qiang Guo ^{2,*} and Jinfan Xu ¹

¹ College of Computer Science and Engineering, Guilin University of Technology, Guilin 541006, China; 1020221059@glut.edu.cn (Z.Z.); 2120221139@glut.edu.cn (J.X.)

² Guangxi Key Laboratory of Embedded Technology and Intelligent System, Guilin University of Technology, Guilin 541006, China

* Correspondence: xxl@glut.edu.cn (X.X.); 2005030@glut.edu.cn (Q.G.)

Abstract: The core task of target detection is to accurately identify and localize the object of interest from a multitude of interfering factors. This task is particularly difficult in UAV aerial images, where targets are often small and the background can be extremely complex. In response to these challenges, this study introduces an enhanced target detection algorithm for UAV aerial images based on the YOLOv7-tiny network. In order to enhance the convolution module in the backbone of the network, the Receptive Field Coordinate Attention Convolution (RFCACConv) in place of traditional convolution enhances feature extraction within critical image regions. Furthermore, the tiny target detection capability is effectively enhanced by incorporating a tiny object detection layer. Moreover, the newly introduced BSAM attention mechanism dynamically adjusts attention distribution, enabling precise target–background differentiation, particularly in cases of target similarity. Finally, the innovative inner-MPDIoU loss function replaces the CIoU, which enhances the sensitivity of the model to changes in aspect ratio and greatly improves the detection accuracy. Experimental results on the VisDrone2019 dataset reveal that relative to the YOLOv7-tiny model, the improved YOLOv7-tiny model improves precision (P), recall (R), and mean average precision (mAP) by 4.1%, 5.5%, and 6.5%, respectively, thus confirming the algorithm’s superiority over existing mainstream methods.

Keywords: UAV image; object detection; YOLOv7-tiny; BSAM attention mechanism



Citation: Zhang, Z.; Xie, X.; Guo, Q.; Xu, J. Improved YOLOv7-Tiny for Object Detection Based on UAV Aerial Images. *Electronics* **2024**, *13*, 2969. <https://doi.org/10.3390/electronics13152969>

Academic Editor: Felipe Jiménez

Received: 29 June 2024

Revised: 24 July 2024

Accepted: 26 July 2024

Published: 27 July 2024



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

1. Introduction

As drone technology continues to evolve, it is increasingly revealing its potential applications across diverse fields. Drones have transcended their initial status as high-tech toys to become versatile and indispensable tools for contemporary applications [1]. Accompanied by technological advances and lower drone costs, drones have transitioned from the laboratory to the commercial sector and daily life. The application domains of drones have expanded beyond military reconnaissance to include agriculture inspection, logistics, environmental monitoring, and disaster response, among others, and their role is becoming increasingly prominent in society [2,3]. In comparison to traditional methods such as ground observation, manual inspection, and aerial and satellite remote sensing, drone technology offers numerous substantial advantages. For instance, in the scenario of forest fire monitoring, drones can facilitate rapid deployment and real-time data transmission [4], thereby enhancing emergency response efficiency, minimizing dependence on human resources, and reducing long-term operational costs. However, despite the wide range of UAV uses and increasing adoption rates, the complexity of operational environments and the diversity of user needs still pose challenges to current UAV target recognition techniques.

Object detection through UAV platforms is confronted with numerous significant challenges [5–7], which primarily include the following aspects. Firstly, the frequent alterations in lighting conditions, including transitions from direct sunlight to shadowed areas, can

induce significant variations in the brightness range of images. In such high dynamic range environments, the contrast of target edges may be greatly enhanced, whereas details within shaded regions may disappear, significantly complicating the detection process. When UAVs navigate through diverse complex terrains and backgrounds, these intricate settings complicate the separation of the target from its surroundings, potentially leading to the target being obscured by other objects or terrain features [8]. Furthermore, UAV devices typically lack robust computational capabilities. Consequently, reconciling the need for efficient target detection algorithms with the constraint of limited computational resources presents a pivotal challenge for researchers today.

Consequently, to address the previously outlined challenges, we propose an enhanced network model derived from the YOLOv7-tiny architecture. We tested the enhanced algorithm on the VisDrone2019 dataset, demonstrating its excellent detection performance. The main contributions of this paper can be summarized as follows:

- In the backbone of the ELAN-S, the partial convolution has been substituted by a Receptive Field Coordinate Attention Convolution (RFCACConv). It not only addresses the issues related to parameter sharing in the convolution kernel but also enhances the network's ability to pinpoint key image regions with greater precision through the coordinate attention mechanism, thereby allowing it to focus on key features. This improvement substantially improves the detection performance with only a slight increase in parameters.
- An additional layer has been incorporated specifically for tiny object detection, the sampling scale has been enhanced, and multi-scale feature fusion has been implemented at the neck of the model. This approach effectively solves the problem of missing or misdetecting tiny objects, which often occurs in traditional object detection models, and greatly enhances the generalization ability of the algorithm.
- The BSAM attention mechanism, enhanced by CBAM and BiFormer attention, has been integrated into the feature integration phase of the framework. It dynamically adjusts the distribution of attention so that the network can more accurately distinguish targets from the background. This improvement greatly enhances the model's ability to discriminate between tiny targets.
- The enhanced loss function inner-MPDIoU is used to replace the CIoU in the initial model, which solves the problem whereby the CIoU is insensitive to the similarity of the aspect ratio between the predicted frames and the real frames and thus greatly improves the detection capability of the model.

2. Related Work

Traditional object detection methods are also known as artificial feature-based object detection methods. The core principle involves utilizing artificially designed features, such as Haar, LBPs (Local Binary Patterns), and HOGs (Histogram of Oriented Gradients) features, to describe target regions within images. Subsequently, machine learning techniques have been employed to enhance object detection capabilities. However, there are some obvious limitations to this approach, namely that feature selection and extraction often requires specialized knowledge and experience, which makes the whole process difficult to initiate and operate. In addition, because the features of the model are manually extracted only for a specific dataset, the generalization ability of the model is limited, often making it difficult to obtain satisfactory results.

As advancements in neural networks and computing power continue, object detection has evolved to an advanced deep learning phase. Compared to traditional object detection approaches, deep learning methods substantially reduce the learning challenges associated with object detection by leveraging neural networks to autonomously learn features. At first, there were mainly two-stage object detection algorithms, such as RCNN [9], fast-RCNN [10], and faster R-CNN [11]. The core idea was to separate the work of object identification into two phases: object categorization and candidate region extraction. Although these methods offer high detection precision, they necessitate performing convolutions on each candidate

region, thereby imposing a significant computational burden. Consequently, they are not well suited for real-time detection applications. In this case, single-stage object detection algorithms emerged, exemplified by the SSD [12] and YOLO series [13–18], which extract features only once without moving through complex regional division and classification stages and then directly predict the classification scores and coordinates. This end-to-end object detection approach significantly boosted the processing speed and effectiveness of the models.

Compared to traditional target detection, the background of UAV aerial images is more complex and the possibility of occlusion between targets is higher, which greatly increases the complexity of the detection process. Additionally, given that drones are equipped with chips of limited computing power, considerations for real-time processing and computational power consumption are imperative. To address these challenges, scholars have proposed some effective strategies so far.

In 2020, Liu et al. [19] enhanced the YOLOv3 architecture by integrating two ResNet units of the same dimension in the Dark network backbone, which significantly extends the perceptual range of the network. Additionally, the anchor frames were clustered using data augmentation and the k-means algorithm to refine the accuracy of model detection. This strategy improves its ability to handle the various sizes and types of targets encountered in UAV images. In 2021, Tan et al. [20] incorporated the RFB module into the character extraction phase of YOLOv4, enhancing the model's capability to extract features from sampling and convolution. During the feature pyramid stage, the ULSAM (Ultra Lightweight Subspace Attention Mechanism) was introduced to generate a different attention feature map for each feature map subspace. It enabled the achievement of multi-scale feature representation, boosting the ability to detect tiny targets against complicated backdrops. Ultimately, the Soft-NMS method was used to assess the score and overlap degree of detection boxes, dynamically adjusting the score of overlapping boxes to minimize target detection omissions caused by occlusions. In 2022, Luo et al. [21] added an improved efficient channel attention (IECA) module to the YOLOv4 algorithm, which enabled the network to capture inter-channel information interactions more efficiently through maximization and global average pooling methods. Furthermore, an adaptive spatial feature fusion module was developed to optimize the integration of feature maps at different scales, thereby augmenting the model's ability to detect objects across multiple scales. In 2023, Zhao et al. [22] introduced a Swin Transformer unit, complemented by a multi-scale detecting head and a CBAM attention module, to effectively capture global information. They also implemented a novel SPPFS pyramid pooling module to facilitate enhanced interaction among feature information and integrated both the SoftNMS and Mish activation functions. This significantly improved the performance of the model in detecting small and dense targets in UAV aerial images. In 2023, Zhai et al. [23] implemented SPD-Conv to replace traditional convolution layers, enhancing the retention of small target features and improving the model's ability to detect small targets. Moreover, the incorporation of the GAM attention mechanism in the neck section of the model significantly reduced the probability of erroneous detections. In 2024, Bai et al. [24] introduced the dynamic snake convolution technology to cope with changes in the target shape. It can flexibly modify the convolution kernel's shape to effectively capture slender pipe structures such as roads or rivers, thus adapting well to the characteristics of UAV datasets. In 2024, Zeng et al. [25] developed a hybrid attention module based on YOLOv5, integrating spatial and coordinate attention (SCA) mechanisms to enhance feature extraction capabilities for small objects. They constructed a multilayer feature fusion structure using channel splicing technology to integrate shallow and deep feature maps, thus enriching the semantic content of shallow features and improving the performance of the model in recognizing small objects.

3. Methods and Improvements

3.1. YOLOv7-Tiny Network Structure

YOLOv7 stands out as one of the leading currently available one-stage target detection algorithms. YOLOv7-tiny represents a compact variant of YOLOv7 specifically designed for edge devices. Compared with other versions of YOLOv7, YOLOv7-tiny boasts a faster detection speed and fewer parameters. Although the detection effect of the YOLOv7-tiny model is slightly worse than that of other versions, considering the limited computational resources of UAV platforms and the high demand for real-time detection in aerial photography applications, the YOLOv7-tiny algorithm was selected as the base method in this work. Figure 1 illustrates the YOLOv7-tiny network architecture, which primarily consists of the backbone, neck, and head components.

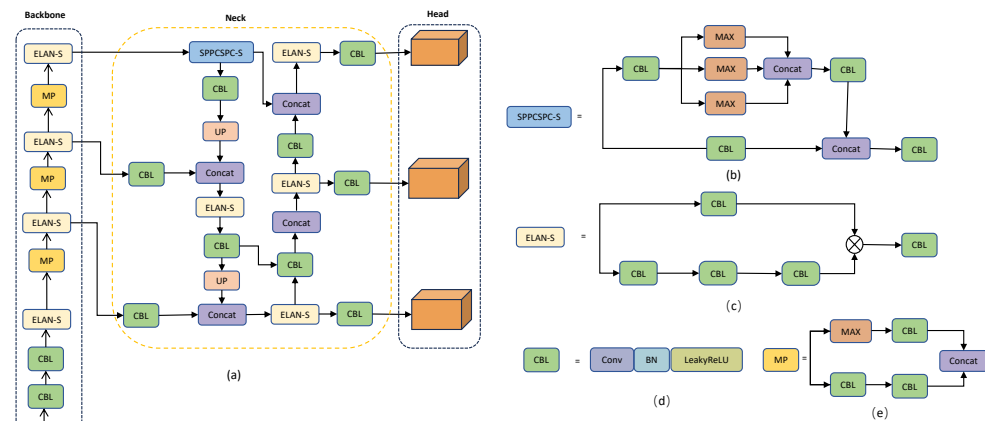


Figure 1. Network structure diagram of each part of YOLOv7-tiny; (a) the overall network structure of YOLOv7-tiny; (b) the network structure of SPPCSPC-S; (c) the network structure of ELAN-S; (d) the network structure of CBL; and (e) the network structure of MP.

The backbone network consists of the CBL layer, the Lightweight Efficient Remote Aggregation Network (ELAN-S) layer, and the MPConv layer. The fundamental aim of this part is to extract the features of the detected objects. The ELAN-S architecture significantly accelerates feature extraction and facilitates easier model convergence.

The neck component of the YOLOv7-tiny incorporates the PAFPN architecture, which ingeniously merges the top-down robust semantic information from the Feature Pyramid Network (FPN) [26] with the bottom-up precise localization capabilities of the Path Aggregation Network (PANet) [27]. This module focuses on the multi-scale feature fusion of feature maps after 8-fold, 16-fold, and 32-fold downsampling, which greatly enhances the model's ability to detect targets of different sizes, realizing multi-scale learning targets and improving the flexibility and accuracy of detection.

In the head module, unlike YOLOv7, YOLOv7-tiny only uses standard convolution and does not use RepConv to process the feature map for feature fusion. This strategy somewhat diminishes the model's detection efficiency. However, the adoption of standard convolution circumvents issues associated with a large parameter count and the vanishing gradient problem typically induced by RepConv. The feature maps after three standard convolutions are fed into three different sizes of detection heads to predict targets of different dimensions and determine the location and confidence scores of the targets.

3.2. Improved YOLOv7-Tiny UAV Network Model

3.2.1. Receptive Field Coordinate Attention Convolution

Receptive Field Attention Convolutional Operations (RFACConv) [28] effectively solve the difficulties associated with sharing convolutional kernel parameters by focusing on the spatial features of the receptive field. Within RFACConv, the enhancement of the network performance is achieved through the mutual learning of attention maps derived from

receptive field features. However, this methodology may entail additional computational expenses. To mitigate this, average pooling is employed to aggregate global information from each receptive field feature. Furthermore, the softmax function is utilized to enhance the correlation of each component within the receptive field. The computation of the RFA can be represented as follows:

$$F = \text{Softmax}(g^{1 \times 1}(\text{AvgPool}(X))) \times \text{ReLU}(\text{Norm}(g^{k \times k}(X))) \quad (1)$$

where the symbol $g^{i \times i}$ is defined as a grouped convolution operation of size $i \times i$. k represents the size of the convolution kernel used, Norm is used to refer to the normalization process, and X represents the input feature map. F represents the field space features, which are obtained by performing multiplication operations on the attention map and the appropriately transformed perceptual map.

Coordinate attention (CA) [29] not only focuses on the channel itself but also considers its spatial location relationship, which realizes the effective combination of channel dominant and spatial attention.

The design of the Receptive Field Coordinate Attention Convolution (RFCACnv) [28] not only addresses the issue of shared convolution parameters but also markedly enhances the network's comprehension of the spatial relevance of the input features through the integration of the coordinate attention mechanism. This enhancement enables the network to more precisely identify key regions within the image, thereby allowing for a more focused analysis of critical features. RFCACnv employs a methodology akin to the self-attention mechanism, enabling it to capture long-range dependencies within the information. Compared to the conventional self-attention mechanism, RFCACnv significantly reduces computational demands and parameter counts while simultaneously enhancing the efficacy of the convolution process.

Figure 2 illustrates the contrast between the enhanced RFA module and the original CA module.

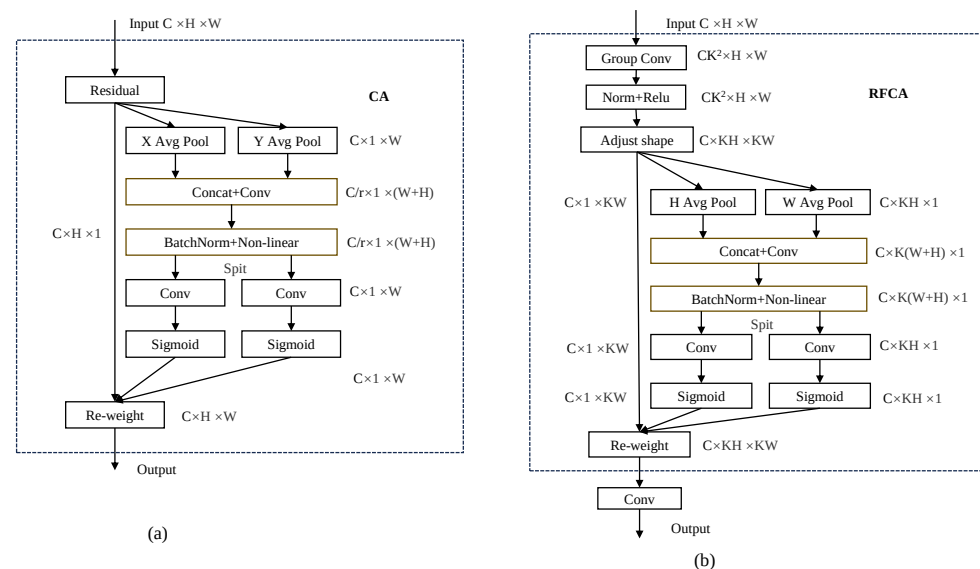


Figure 2. Detailed structure comparison diagram of CA and RFCACnv: (a) the structure of CA; and (b) the structure of RFCACnv.

3.2.2. Multi-Scale Sampling Feature Fusion

- Adding a tiny object detection layer

In UAV target detection tasks, particularly under conditions of long-distance imaging, high-speed motion, and low-altitude flights, accurately detecting small targets presents a significant challenge. The main reason for this challenge is the small number of features available for small targets coupled with the reduction of feature information due to the

pooling layer and convolutional kernel operations. This situation significantly increases the difficulty of recognizing small targets and increases the likelihood of detection errors or misses. In addition, the Feature Pyramid Network (FPN) structure does not fully utilize the feature map output from the backbone during the fusion phase, which may lead to partial information loss and interference during the upsampling process, thus adversely affecting the detection of tiny targets. To address this problem, we add a new detection layer, the P2 layer, to the neck module of the YOLOv7-tiny model, which fuses deep global information with shallow features, thus preserving more image details and significantly enhancing the model's ability to detect and recognize smaller objects. The structure of the additional tiny object detection layer is shown in Figure 3.

Figure 3 displays the updated YOLOv7-tiny network structure.

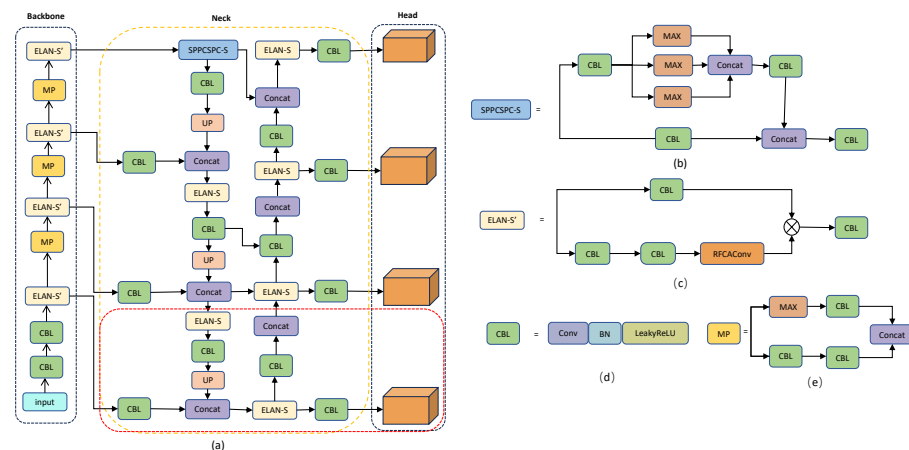


Figure 3. Network structure diagram of each part of the improved YOLOv7-tiny: (a) the overall network structure of the improved YOLOv7-tiny; (b) the network structure of SPPCSPC-S; (c) the network structure of ELAN-S'; (d) the network structure of CBL; and (e) the network structure of MP.

3.2.3. Bilevel Spatial Attention Module

The integration of attention mechanisms has been proven to be a pivotal technological advancement in target detection research, significantly enhancing detection accuracy. The basic principle of this mechanism is to assign different weights to different regions of the feature map in order to highlight the features of interest. Especially in the process of analyzing remote sensing images, the target information is usually smaller than the complex background information, and the traditional convolutional neural network is easily affected by the non-target region in the feature extraction stage, thus affecting the detection performance.

To address this challenge, researchers have devised numerous attention mechanism strategies to optimize the network's capability to extract target features. Typically, these strategies encompass two principal dimensions: channel attention and spatial attention. The channel attention mechanism concentrates on identifying "what" the target is by enhancing the representation of target features through differential weighting of feature channels, while the spatial attention mechanism is tasked with pinpointing "where" the target is located.

This study employs the Bilevel Routing Attention Module from BiFormer [30] and the Spatial Attention Module (SAM) from CBAM [31] to enhance the recognition of small objects in images while effectively filtering out irrelevant information. By integrating the two modules in parallel, the Bilevel Spatial Attention Module (BSAM) is constructed, enhancing the model's overall attentional capabilities. The BSAM module, which integrates BiFormer's dynamic sparse attention with spatial attention mechanisms, exhibits superior feature extraction capabilities when compared to the traditional CBAM attention module, particularly in terms of location sensitivity. It can flexibly modify the allocation of attention depending on the content, focusing on a small number of key tokens rather than dispersing

attention to other unrelated tokens, thereby improving the computational efficiency while maintaining high feature extraction accuracy. The operational principle of the BSAM module begins with a detailed segmentation of the input feature map into numerous smaller regions. Building on this, the module computes three vectors—query (Q), key (K), and value (V)—and employs an adjacency matrix to pinpoint regions with substantial semantic correlations. Subsequently, the module constructs an indexed routing matrix to facilitate the token-to-token attention mechanism (K^g, V^g) among these semantically linked regions.

$$K^g = \text{gather}(K, I^r) \quad (2)$$

$$V^g = \text{gather}(V, I^r) \quad (3)$$

where K^g and V^g denote the gather key and the value tensor, respectively, while I^r represents an index containing the most relevant or important regions or elements representing the attention mechanism.

Following global maximum pooling and global average pooling, the feature maps processed via the Bilevel Attention mechanism are transformed into two-dimensional representations. Subsequently, these maps are concatenated along the channel dimension and compacted into a single channel via a convolutional layer, streamlining the feature integration process. A sigmoid activation function is then employed to generate the spatial attention feature map, refining the focus of the module on relevant features. Finally, the spatial attention feature map is multiplied with the input feature map in elemental order to generate the final output. The architecture of the Bilevel Spatial Attention Module is illustrated in Figure 4.

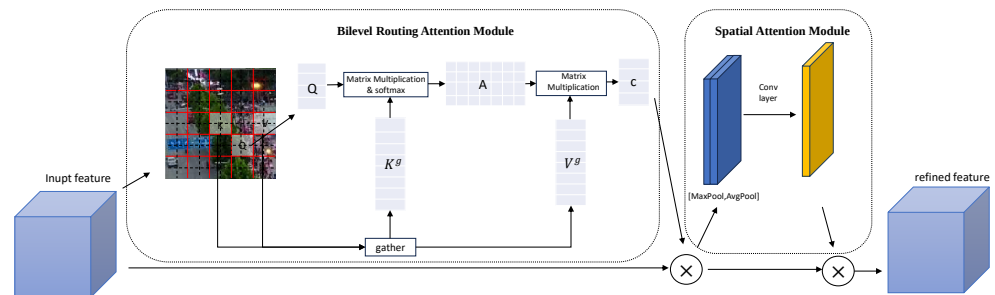


Figure 4. Detailed structure diagram of the Bilevel Spatial Attention Module.

3.2.4. Optimization of Loss Function

YOLOv7-tiny employs the CIoU [32] loss function for bounding box prediction, which extends the traditional IoU by incorporating a penalty term that accounts for the Euclidean distance between the centers of bounding boxes and their aspect ratios. However, this penalty mechanism may not perform effectively when the detected objects possess aspect ratios similar to those of the predicted boxes.

Conversely, the inner-IoU [33] introduces a novel method for calculating loss functions across datasets of varying scales, enhancing adaptability and efficiency. Auxiliary frames of varying scales were utilized with their sizes regulated by the scale factor ratio to optimize the efficiency of loss function computation. The inner-IoU calculation formula is as follows:

$$b_l = x_c - \frac{w * ratio}{2}, b_r = x_c + \frac{w * ratio}{2} \quad (4)$$

$$b_t = y_c - \frac{h * ratio}{2}, b_b = y_c + \frac{h * ratio}{2} \quad (5)$$

$$inter = (\min(b_r^{gt}, b_r) - \max(b_l^{gt}, b_l)) * \min(b_b^{gt}, b_b) - \max(b_t^{gt}, b_t) \quad (6)$$

$$union = (w^{gt} * h^{gt}) * (ratio)^2 + (w * h) \quad (7)$$

$$IoU^{inner} = \frac{inter}{union} \quad (8)$$

By integrating an auxiliary bounding box, the inner-IoU leverages the IoU between this auxiliary and the target bounding box as a component of the loss calculation. For prediction frames exhibiting a high intersection ratio with the actual frame, Inner-IoU employs a smaller auxiliary bounding box to calculate the loss, thereby facilitating faster convergence. In instances where prediction frames have a low intersection ratio with the actual frame, a larger auxiliary bounding box is utilized to expand the effective regression range, aiding the regression of low IoUs and improving the generalization capabilities of the IoU metric.

The MPDIoU loss function improves the accuracy of the overlap measurement by taking into account the minimum distance between the vertical edges of the predicted and actual frames.

$$MPDIoU = IoU - \frac{\rho^2(P_1^{pred}, P_1^{gt})}{w^2 + h^2} - \frac{\rho^2(P_2^{pred}, P_2^{gt})}{w^2 + h^2} \quad (9)$$

where P_1^{pred} , P_2^{pred} , P_1^{gt} , and P_2^{gt} refer to the points in the upper left corner and lower right corner of the prediction box and the real box, respectively, and $\rho^2(P_1^{pred}, P_1^{gt})$ calculates the separation of the respective points.

The inner-MPDIoU improves MPDIoU by utilizing the inner-IoU to optimize the evaluation of overlaps between bounding boxes, making the method more effective when dealing with multiple objects in complex or crowded scenes. This approach improves the response speed of the model to changes in the position of the bounding box and enhances the model's ability to adapt to complex visual environments. The formula of the inner-MPDIoU is as follows:

$$MPDIoU^{inner} = IoU^{inner} - \frac{\rho^2(P_1^{pred}, P_1^{gt})}{w^2 + h^2} - \frac{\rho^2(P_2^{pred}, P_2^{gt})}{w^2 + h^2} \quad (10)$$

4. Analysis of Experimental Results

4.1. Dataset

The VisDrone2019 [34] dataset, provided by Tianjin University, is tailored for UAV vision tasks and contains 10,209 static images designated for target detection. Of these images, 6471 were assigned to the training set, 548 to the validation set, and the remaining 3190 to the testing set. The dataset captures a variety of real-world environments, including city streets, rural areas, and parks under different weather conditions and times of day, and is equipped with annotations for 10 target categories to facilitate the evaluation and enhancement of the target detection model. Before target detection using the YOLOv7-tiny model, a series of pre-processing steps were performed on the images in the VisDrone2019 dataset. First, all the images were resized to 640×640 pixels, and then the training efficiency of the network was accelerated by normalizing the images such that the scaled pixel values were within the range of $[0, 1]$. In addition, a random rotation technique was applied for data enhancement to improve the robustness of the model to complex backgrounds. These pre-processing steps ensured the consistency and quality of the input data, thus helping the YOLOv7-tiny model more accurately recognize and localize targets in images.

4.2. Experimental Environment

For this study's experimental configuration, we used a computing platform equipped with Intel(R) Xeon(R) Bronze 3106 CPU @ 1.70 GHz, 256 GB of RAM, and running a CentOS 7.6 operating system. The deep learning experiments were conducted under the PyTorch framework, utilizing NVIDIA Tesla V100 GPUs and a CUDA 11.8 environment. Furthermore, to ensure the consistency of the experimental conditions, consistent hyperparameter settings were shared across all training programs. The weights obtained on the

COCO dataset were used for model pre-training. Table 1 displays the settings and most crucial hyperparameter settings used during training.

Table 1. Experimental environment configuration.

Device	Configuration
CPU	Intel(R) Xeon(R) Bronze 3106 CPU @ 1.70 GHz
GPU	NVIDIA Tesla V100
System	CentOS 7.6
Python	3.7.0
CUDA	11.8
Pytorch	1.13.1 + cu117
Batch Size	16
Optimizer	SGD
Momentum	0.937
Epochs	300

4.3. Assessment Metrics

In order to objectively assess the improvement in model performance and prediction accuracy, we used several key metrics: precision (P), recall (R), and mean average precision (mAP). In particular, mAP@0.5 indicates that a prediction is considered correct when there is at least 50% overlap between the predicted frame and the true frame. It is an important metric used to evaluate the prediction effectiveness of the model.

$$P = \frac{TP}{(TP + FP)} \quad (11)$$

$$R = \frac{TP}{(TP + FN)} \quad (12)$$

$$AP = \int_0^1 p(r) dr \quad (13)$$

$$mAP = \frac{1}{k} \sum_{i=1}^k AP_i \quad (14)$$

where TP indicates that the model accurately predicts instances of positive classes as positive, FP indicates that the model misclassifies instances of negative classes as positive, TN indicates that the model correctly identifies instances of negative classes as negative, and FN indicates that the model incorrectly determines instances of positive classes as negative.

Furthermore, $p(r)$ is the precision at a given recall, and AP is the average precision, which is the integral of the precision at different recall levels. It provides a comprehensive evaluation of performance by integrating the precision at different recall points.

Inference time (also known as response time), measured in milliseconds (ms), evaluates the time required by the model to complete the target detection. To thoroughly assess the model's performance, we additionally took into account the model's parameters and GFLOPS to assess the computing power.

4.4. Ablation Experiment

Experiments using a single upgraded module or several improved combined modules were conducted on the VisDrone2019 dataset in order to validate the efficacy of the enhanced modules on the model. The outcomes of the experiment are displayed in Table 2.

Table 2. Results of the ablation experiment on the dataset VisDrone2019.

Method	RFCA	TODL	BSAM	Inner-MPDIoU	Parameters	GFLOPS	P%	R%	mAP@0.5%
YOLOv7-tiny					6.03 M	13.3	47.1	36.6	35.0
A	✓				6.09 M	13.7	47.8	38.8	37.8
B		✓			6.14 M	15.8	49.0	41.2	38.5
C			✓		6.22 M	17.9	48.9	39.5	36.5
D				✓	6.03 M	13.3	47.1	39.0	35.8
E	✓	✓			6.20 M	16.2	50.1	40.6	39.7
F	✓		✓		6.28 M	18.4	49.3	39.7	39.3
G	✓	✓	✓		6.39 M	21.0	50.5	41.3	40.7
H	✓	✓	✓	✓	6.39 M	21.0	51.2	42.1	41.5

As shown in Table 2, the initial group of experiments used YOLOv7-tiny as the baseline model for experiments on the VisDrone2019 dataset, and the results showed that the mAP@0.5 was 35.0%. The subsequent experiments, labeled A, B, C, and D, were based on the YOLOv7-tiny model and incorporated the RFCA, TODL, BSAM, and inner-MPDIoU modules, respectively, to investigate the effects of each module on the functionality of the framework.

Experiment A solved the problem of convolution kernel sharing through replacing part of the convolutions of the ELAN-S structure with RFCACnv and achieved an mAP@0.5 improvement of 2.8% relative to the baseline model. In experiment B, the introduction of the tiny object detection layer increased the model parameters, but the improvement effect in the model was the most obvious. Compared to the baseline model, the P, R, and mAP@0.5 increased by 1.9%, 4.6%, and 3.5%, respectively. In experiment C, the BSAM attention mechanism formed by combining the BRA attention mechanism in BiFormer and the spatial attention mechanism in CBAM was integrated into the neck part of the model, which significantly enhanced the detection ability of small targets with a 1.7% improvement in mAP@0.5 compared to the baseline model. Experiment D replaces the loss function with inner-MPDIoU and improves mAP@0.5 by 0.8% compared to the baseline model with essentially no change in the number of parameters of the original model. Experiments E, F, and G further optimized the performance based on the existing improvements by incorporating different improvement strategies into the model in a combinatorial form. In Experiment H, integrating four distinct improvement strategies into the model resulted in a modest parameter increase of just 0.36 M compared to the previous one, but it significantly boosted P, R, and mAP@0.5 by 4.1%, 5.5%, and 6.5%, respectively.

Ablation experiments were conducted using varying ratios for the inner-MPDIoU loss function to test the impact of different ratio values on the experimental outcomes. The experimental outcomes are displayed in Table 3.

Table 3. Effect of different ratio values on the experimental results.

Ratio	P%	R%	mAP@0.5%	mAP@0.5:0.95%
0.5	50.1	40.9	40.3	23.1
0.8	50.4	40.3	40.1	22.8
1	51.6	41.3	40.7	23.3
1.1	50.7	41.9	41.3	23.4
1.2	50.8	41.6	40.9	23.5
1.3	51.2	42.1	41.5	23.6
1.4	50.7	41.7	41.2	23.4

When the ratio = 1, the inner-MPDIoU loss function essentially becomes the MPDIoU loss function. The experimental results show that since there are mostly small targets in UAV aerial images when the ratio is larger than 1, the auxiliary frame is larger than the actual frame, which is favorable for low IoU regression. The best experimental results are obtained when the ratio = 1.3. However, when the ratio = 1.2 or 1.4, the experimental

effect decreases. Therefore, the exact ratio value must be varied according to the different characteristics of the experimental dataset. In general, the ratio should be less than 1 when the targets in the dataset are large and greater than 1 when the targets in the dataset are small.

4.5. Comparative Experiments

To ascertain the effects of incorporating an attention mechanism into the neck component of the model on its performance, this study conducted four sets of comparative experiments, contrasting the proposed BSAM module with the extensively employed SENet, CBAM, and BiFormer attention modules. Table 4 shows a comparison of the detection performance in YOLOv7-tiny before and after the addition of various attention mechanism modules. The experimental results demonstrate that the inclusion of diverse attention mechanism modules enhances the overall accuracy of the network model. However, the inclusion of the BSAM attention mechanism in the YOLOv7-tiny model resulted in better detection performance than other attention mechanisms.

Table 4. Comparison of various attention mechanism algorithms' detection performance.

Method	Parameters	mAP@0.5	Inference Time (ms)	GFLOPS
YOLOv7-tiny	6.02 M	35.0	2.9	13.3
+SE	6.05 M	35.2	3.1	13.3
+CBAM	6.04 M	35.3	3.3	13.3
+BiFormer	6.06 M	35.8	3.7	13.9
+BSAM	6.22 M	36.5	4.4	17.9

In order to enhance the comparative understanding of the BSAM, CBAM, and BiFormer attention mechanisms, we used gradient-weighted class activation mapping [35] for visualization. The specific outcomes of these visualizations are depicted in Figure 5. As shown in Figure 5, the model with the integration of the CBAM attention mechanism and the BiFormer attention mechanism shows higher attention to targets such as people and vehicles in the image compared to the model without the integration of the attention mechanism. Figure 5d shows the effect of combining the baseline model with the BSAM attentional mechanism, and the results indicate that the BSAM attentional mechanism not only focuses on a wider target area but also effectively distinguishes between the target and the background information compared with CBAM and BiFormer. This suggests that the BSAM attention mechanism module has significant advantages over CBAM and BiFormer.

In order to comprehensively assess the performance enhancement of the improved model, this study compares and analyzes it with several current mainstream lightweight target detection algorithms. The specific experimental results are displayed in Table 5. The analysis of Table 5 reveals that two-stage target detection algorithms, such as Faster-RCNN [10] and Cascade R-CNN [36], generally have low mAP values, while the YOLO family of algorithms can usually achieve higher mAP values. In particular, the algorithm proposed in this study improves the mAP value by 6.5% compared to YOLOv7-tiny, demonstrating significant superiority over RetinaNet [37], MSA-YOLO [38], and other algorithms.

In addition, the experimental data show that the improved model also significantly outperforms the original model in terms of average precision (AP) for each category. Specifically, the AP values of the 10 categories were improved by 6.4%, 6.2%, 3.5%, 7.6%, 7.3%, 9.7%, 5%, 9.8%, 4.6%, and 4.5%, respectively, with the trucks and awning-tricycles categories showing the most significant improvement. Based on these test results, it can be concluded that the algorithm proposed in this paper performs better in UAV target detection, further confirming the effectiveness of the method and its applicability in real-time UAV detection scenarios.

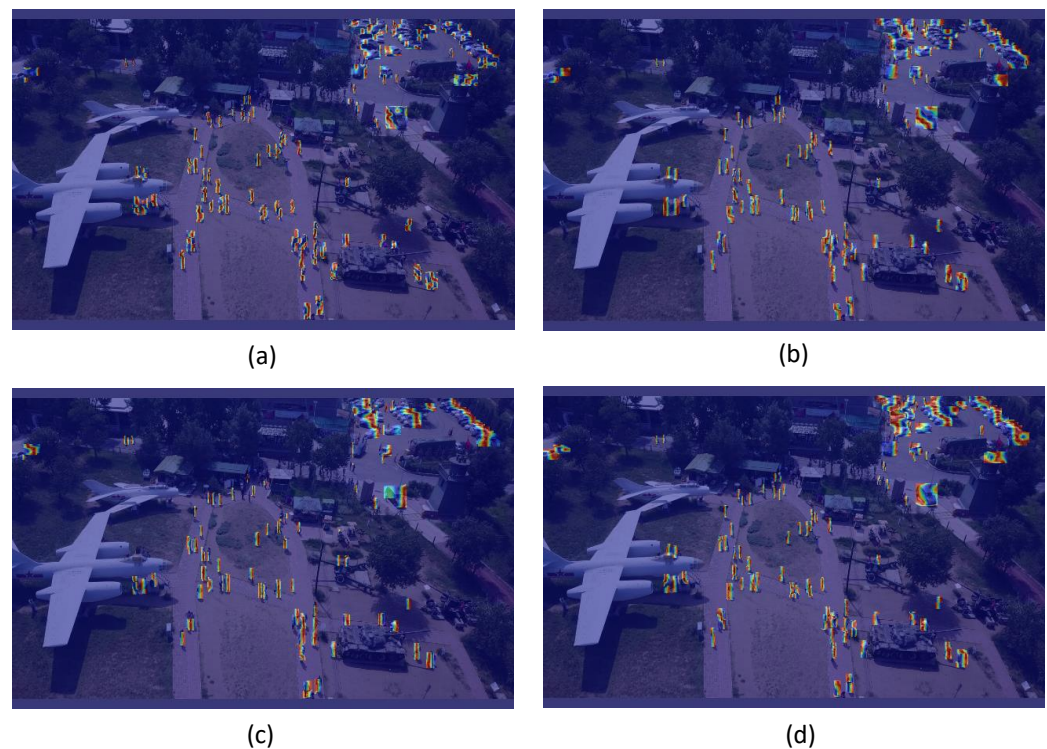


Figure 5. The heat map comparison of different attention mechanisms is added to the original network structure. (a) The heat map without an attention mechanism; (b) the heat map with CBAM; (c) the heat map with Biformer; (d) the heat map with BSAM.

Table 5. Comparison of experimental results on the VisDrone2019 dataset.

Method	Pedestrian	Person	Bicycle	Car	Van	Truck	Tricycle	A-T	Bus	Motor	mAP@0.5
Faster R-CNN	21.4	15.6	6.7	51.7	29.5	19.0	13.1	7.7	31.4	20.7	21.7
Cascade R-CNN	22.2	14.8	7.6	54.6	31.5	21.6	14.8	8.6	34.9	21.4	23.2
DMNet [39]	28.5	20.4	15.9	56.8	37.9	30.1	22.6	14.0	47.1	29.2	30.3
YOLOv3-LITE [40]	34.5	23.4	7.9	70.8	31.3	21.9	15.3	6.2	40.9	32.7	28.5
CDNet [41]	35.6	19.2	13.8	55.8	42.1	38.2	33.0	25.4	49.5	29.3	34.2
MSC-CenterNet	33.7	15.2	12.1	55.2	40.5	34.1	29.2	21.6	42.2	27.5	31.1
DBAI-Det	36.7	12.8	14.7	47.4	38.0	41.4	23.4	16.9	31.9	16.6	28.0
YOLOv5s	35.8	30.5	10.1	65.0	31.5	29.5	20.6	11.1	41.1	35.4	31.1
MSA-YOLO	33.4	17.3	11.2	76.8	41.5	41.4	14.8	18.4	60.9	31.0	34.7
YOLOv7-tiny	38.9	35.5	9.6	76.2	38.0	28.8	20.6	10.9	48.7	43.2	35.0
Ours	45.3	41.7	13.1	83.8	45.3	38.5	25.6	20.7	53.3	47.7	41.5

4.6. Detection Effect Analysis

In order to visualize the superiority of the algorithm proposed in this paper for UAV small target detection, we compare the detection effect of the algorithm proposed in this paper with some typical algorithms in different scenes.

Figure 6 illustrates that compared with other algorithms, the improved YOLOv7-tiny algorithm is able to recognize more small targets, such as distant pedestrians and trucks, during the daytime, and it significantly improves the confidence level of target recognition.

Figure 7 demonstrates the detection performance of the model proposed in this study in a nighttime environment. The results show that the model is able to effectively detect more pedestrians in dimly lit conditions, demonstrating its ability to adapt to low-light environments. This phenomenon suggests that the improved algorithm exhibits lower sensitivity in the face of changing light conditions, resulting in stronger anti-interference ability and higher network robustness.



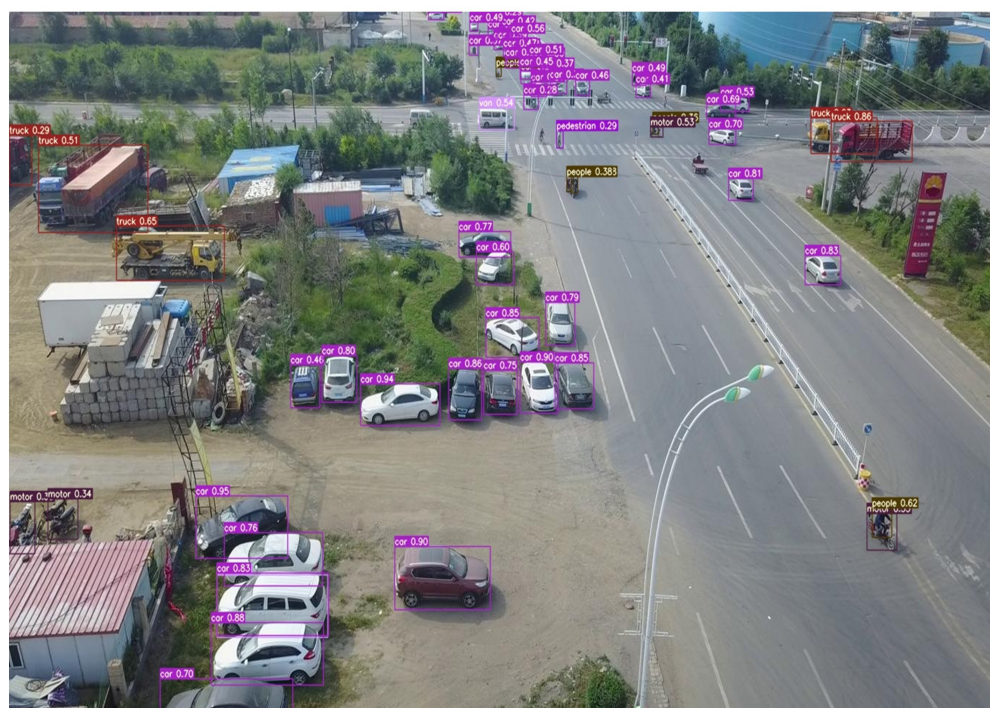
(a)



(b)



(c)



(d)

Figure 6. (a–d) represent the detection results of YOLOv5s, MSA-YOLO, YOLOv7-tiny and improved YOLOv7-tiny algorithms during daytime, respectively.



(a)



(b)



(c)

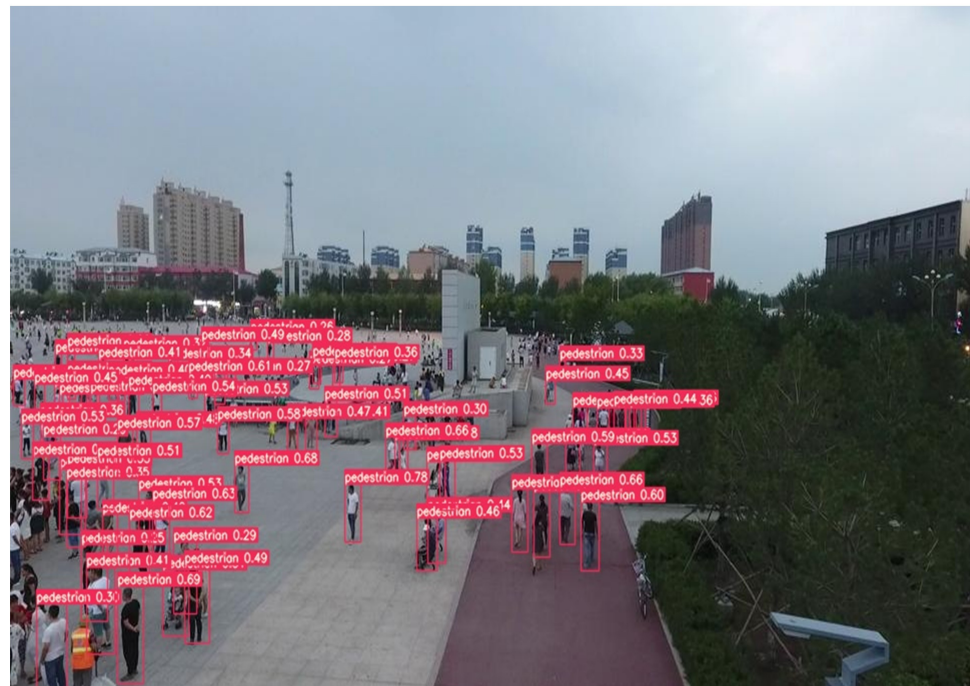


(d)

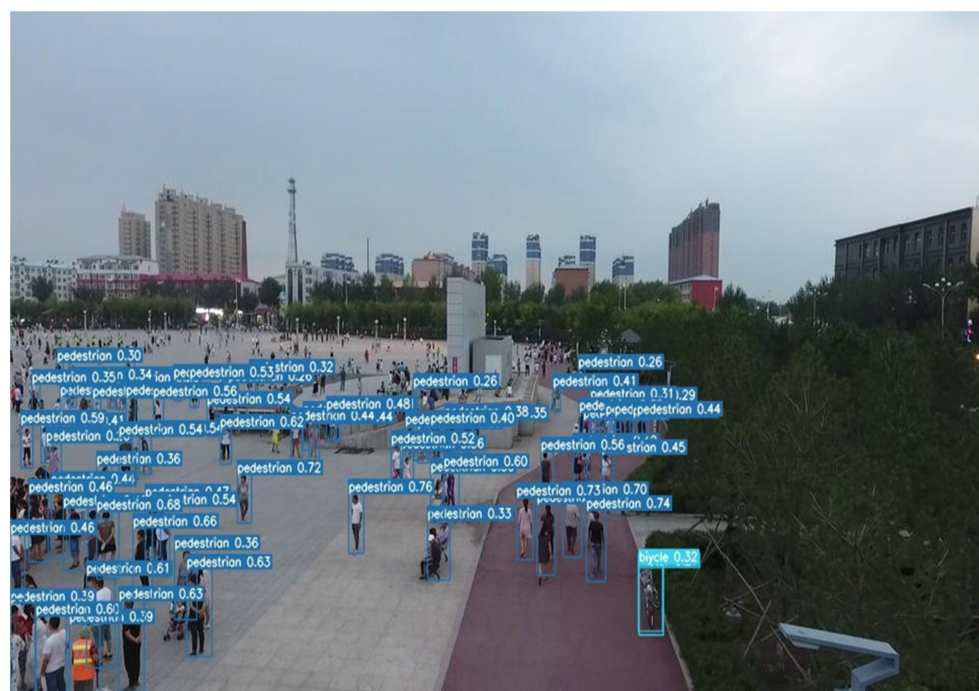
Figure 7. (a–d) represent the detection results of YOLOv5s, MSA-YOLO, YOLOv7-tiny and improved YOLOv7-tiny algorithms in dark light scenes, respectively.

Figures 8 and 9 show the comparative analysis of the target detection effectiveness of the improved YOLOv7-tiny algorithm in dense target environments and sparse target environments, respectively. In the target-dense scenario, the presence of a large number of small targets and their mutual occlusion significantly increases the complexity of target de-

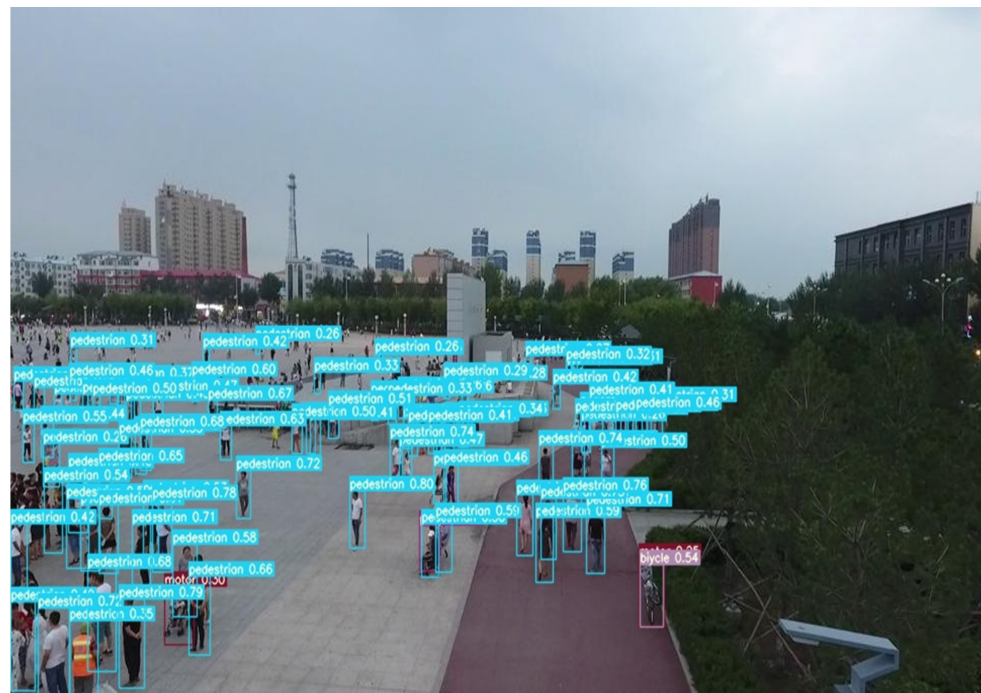
tection in images captured by UAVs. Nevertheless, the optimized YOLOv7-tiny algorithm successfully improves the detection of pedestrians and accurately identifies motorcycles, showing its effectiveness in dealing with high-density target environments. In sparse target scenarios, the model continues to show significant advantages in recognizing cars, trucks, and motorcycles, proving its good generalization. Overall, compared with other algorithms, the model has obvious advantages in reducing leakage and false detection rates. The experimental results fully validate the improvements made to the YOLOv7-tiny algorithm, which effectively enhance its target detection capability in complex environments.



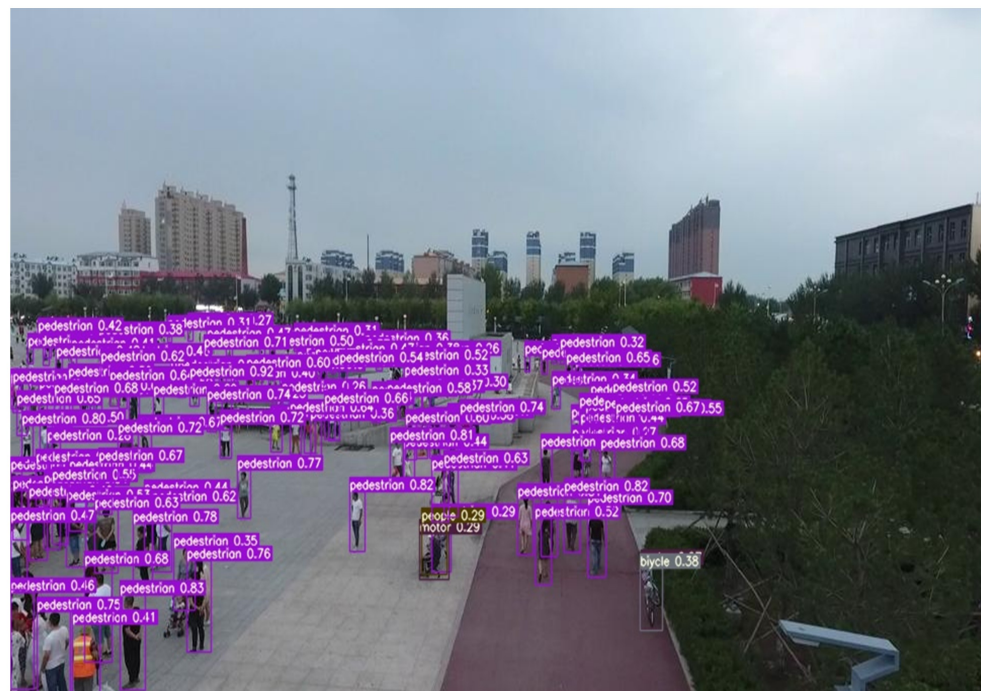
(a)



(b)

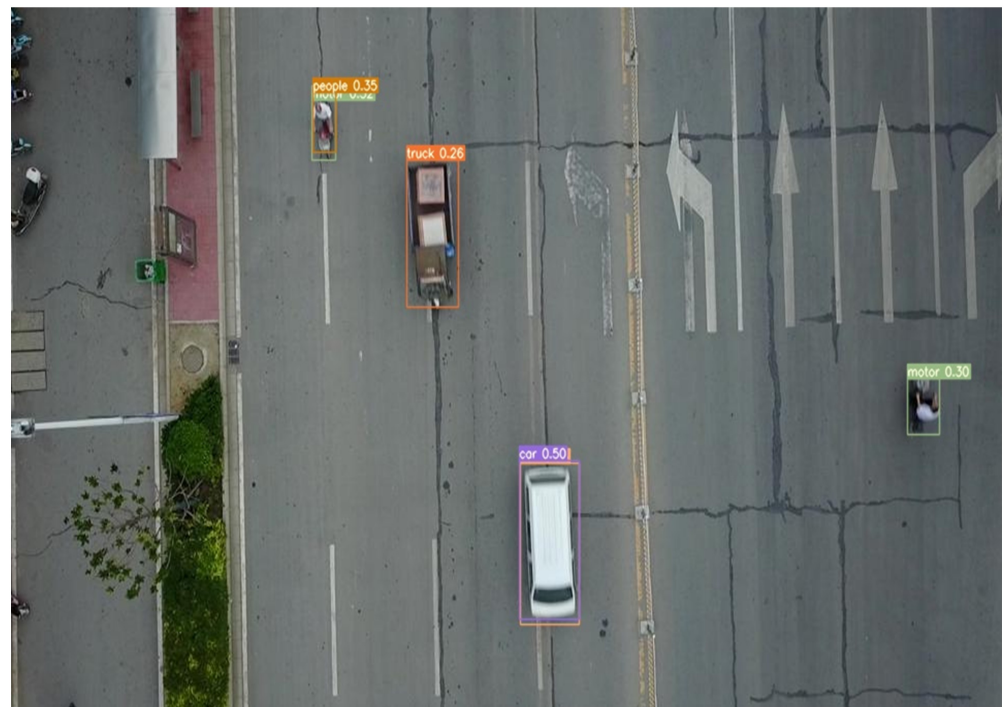


(c)

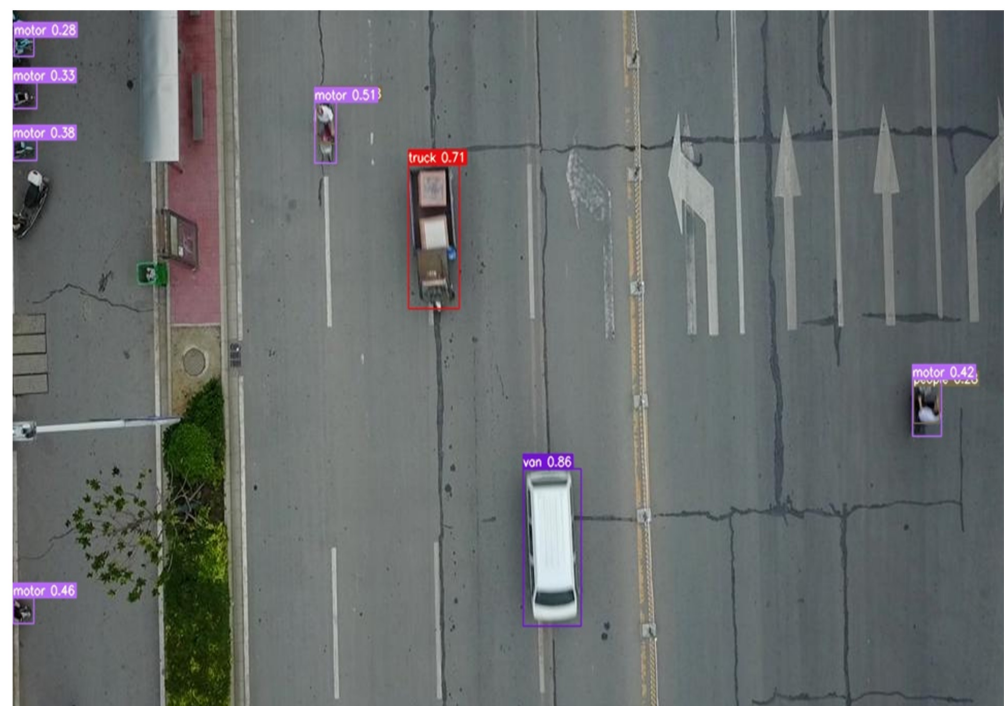


(d)

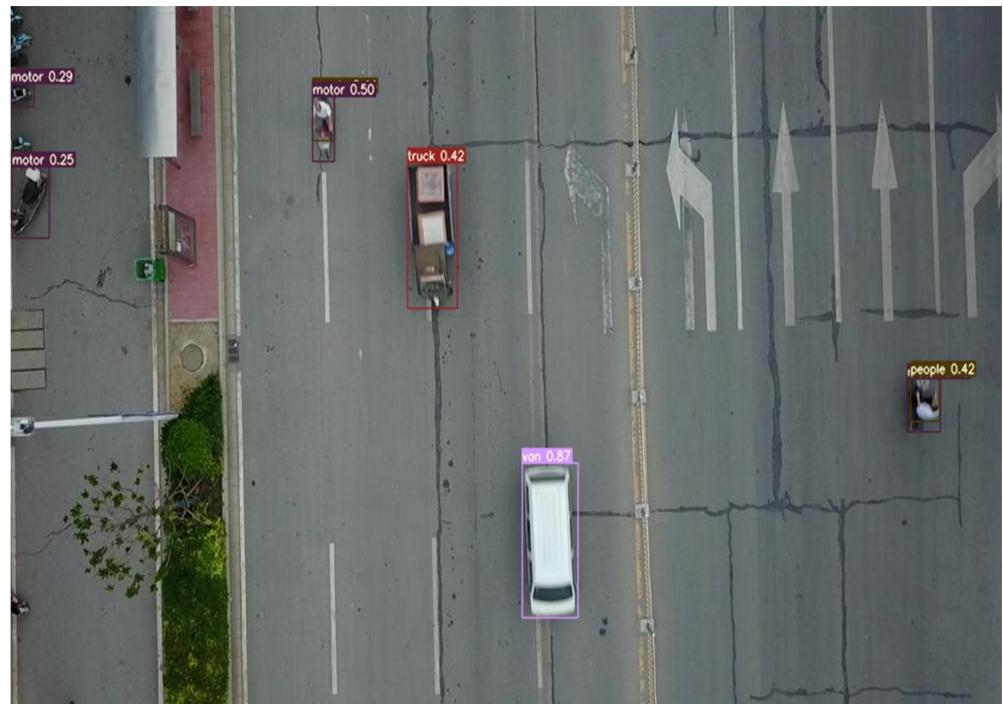
Figure 8. (a–d) represent the detection results of YOLOv5s, MSA-YOLO, YOLOv7-tiny and improved YOLOv7-tiny algorithms in dense scenes, respectively.



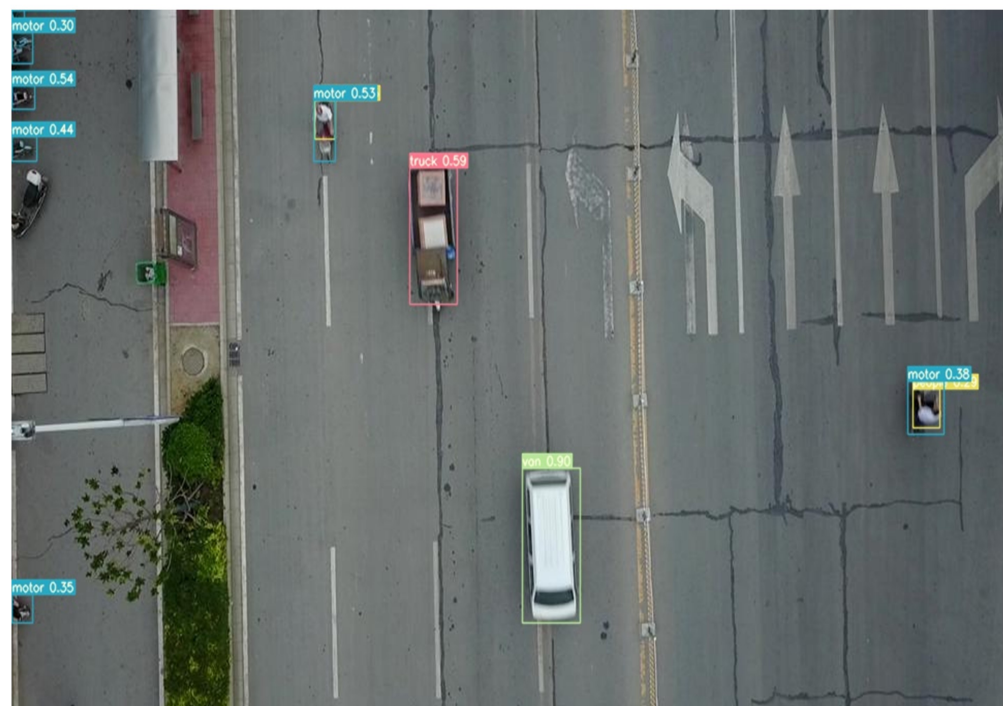
(a)



(b)



(c)



(d)

Figure 9. (a–d) represent the detection results of YOLOv5s, MSA-YOLO, YOLOv7-tiny and improved YOLOv7-tiny algorithms in sparse scenarios, respectively.

5. Conclusions and Outlook

Faced with the challenge of accurately recognizing objects in UAV aerial photographs, which are often subject to errors and omissions, this study enhances the ELAN-S convolution in the backbone of the model by integrating the receptive field coordinate attention convolution, thus improving the feature-capturing capability of the network. In terms of network structure, a tiny target detection layer is added to the neck to scale up the sampling and enrich the feature map information used for subsequent upsampling and downsampling. The attention to the small target detection region is strengthened by adding a BSAM attention mechanism. Finally, by introducing the inner-MPDIoU loss function, the model is made to exhibit enhanced learning capability in detecting small and challenging samples. Based on the test results on the VisDrone2019 dataset, the model outperforms other popular models in the UAV target detection task.

In existing research on UAV target detection algorithms, various methods have been used to optimize algorithm performance. In addition to the improvements presented in this study, future research could explore improving detection results by improving the downsampling method so that more detailed information is retained during the downsampling of the dataset. In addition, pruning of the algorithm module could be considered to optimize the size of the model while maintaining good detection accuracy.

Author Contributions: Conceptualization, Z.Z. and X.X.; methodology, Z.Z. and J.X. software, Z.Z.; validation, Z.Z., J.X. and X.X.; formal analysis, Z.Z. and Q.G.; investigation, Q.G.; resources, X.X. and Q.G.; data curation, Z.Z.; writing—original draft preparation, Z.Z.; writing—review and editing, Z.Z.; visualization, Z.Z.; supervision, X.X. and Q.G.; project administration, X.X.; funding acquisition, X.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant number No. 62262011, and the Guangxi Key Research and Development Program, grant number Guike AB23049001.

Data Availability Statement: Data set: <https://github.com/VisDrone> (accessed on 1 December 2023).

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

References

- Maghazei, O.; Lewis, M.A.; Netland, T.H. Emerging technologies and the use case: A multi-year study of drone adoption. *J. Oper. Manag.* **2022**, *68*, 560–591. [\[CrossRef\]](#)
- Rao, B.; Gopi, A.G.; Maione, R. The societal impact of commercial drones. *Technol. Soc.* **2016**, *45*, 83–90. [\[CrossRef\]](#)
- Aydin, B. Public acceptance of drones: Knowledge, attitudes, and practice. *Technol. Soc.* **2019**, *59*, 101180. [\[CrossRef\]](#)
- Mahmudnia, D.; Arashpour, M.; Bai, Y.; Feng, H. Drones and blockchain integration to manage forest fires in remote regions. *Drones* **2022**, *6*, 331. [\[CrossRef\]](#)
- Cao, Y.; He, Z.; Wang, L.; Wang, W.; Yuan, Y.; Zhang, D.; Zhang, J.; Zhu, P.; Van Gool, L.; Han, J.; et al. VisDrone-DET2021: The vision meets drone object detection challenge results. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 2847–2854.
- Mittal, P.; Singh, R.; Sharma, A. Deep learning-based object detection in low-altitude UAV datasets: A survey. *Image Vis. Comput.* **2020**, *104*, 104046. [\[CrossRef\]](#)
- Du, D.; Qi, Y.; Yu, H.; Yang, Y.; Duan, K.; Li, G.; Zhang, W.; Huang, Q.; Tian, Q. The unmanned aerial vehicle benchmark: Object detection and tracking. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 370–386.
- López, Y.Á.; García-Fernández, M.; Álvarez-Narciandi, G.; Andrés, F.L.H. Unmanned aerial vehicle-based ground-penetrating radar systems: A review. *IEEE Geosci. Remote Sens. Mag.* **2022**, *10*, 66–86. [\[CrossRef\]](#)
- Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
- Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
- Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In Proceedings of the Advances in Neural Information Processing Systems 28, Montreal, QC, Canada, 7–12 December 2015.

12. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part I 14; Springer: Cham, Switzerland, 2016; pp. 21–37.
13. 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, 27–30 June 2016; pp. 779–788.
14. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.
15. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
16. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
17. Li, C.; Li, L.; Jiang, H.; Weng, K.; Geng, Y.; Li, L.; Ke, Z.; Li, Q.; Cheng, M.; Nie, W.; et al. YOLOv6: A single-stage object detection framework for industrial applications. *arXiv* **2022**, arXiv:2209.02976.
18. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 7464–7475.
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](#)]
20. Tan, L.; Lv, X.; Lian, X.; Wang, G. YOLOv4_Drone: UAV image target detection based on an improved YOLOv4 algorithm. *Comput. Electr. Eng.* **2021**, *93*, 107261. [[CrossRef](#)]
21. Luo, X.; Wu, Y.; Zhao, L. YOLOD: A target detection method for UAV aerial imagery. *Remote Sens.* **2022**, *14*, 3240. [[CrossRef](#)]
22. Zhao, L.; Zhu, M. MS-YOLOv7: YOLOv7 based on multi-scale for object detection on UAV aerial photography. *Drones* **2023**, *7*, 188. [[CrossRef](#)]
23. Zhai, X.; Huang, Z.; Li, T.; Liu, H.; Wang, S. YOLO-Drone: An Optimized YOLOv8 Network for Tiny UAV Object Detection. *Electronics* **2023**, *12*, 3664. [[CrossRef](#)]
24. Bai, Z.; Pei, X.; Qiao, Z.; Wu, G.; Bai, Y. Improved YOLOv7 Target Detection Algorithm Based on UAV Aerial Photography. *Drones* **2024**, *8*, 104. [[CrossRef](#)]
25. Zeng, S.; Yang, W.; Jiao, Y.; Geng, L.; Chen, X. SCA-YOLO: A new small object detection model for UAV images. *Vis. Comput.* **2024**, *40*, 1787–1803. [[CrossRef](#)]
26. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
27. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8759–8768.
28. Zhang, X.; Liu, C.; Yang, D.; Song, T.; Ye, Y.; Li, K.; Song, Y. Rfaconv: Innovating spatital attention and standard convolutional operation. *arXiv* **2023**, arXiv:2304.03198.
29. Hou, Q.; Zhou, D.; Feng, J. Coordinate attention for efficient mobile network design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 13713–13722.
30. Zhu, L.; Wang, X.; Ke, Z.; Zhang, W.; Lau, R.W. Biformer: Vision transformer with bi-level routing attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 10323–10333.
31. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
32. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. Distance-IoU loss: Faster and better learning for bounding box regression. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 12993–13000.
33. Zhang, H.; Xu, C.; Zhang, S. Inner-iou: More effective intersection over union loss with auxiliary bounding box. *arXiv* **2023**, arXiv:2311.02877.
34. Du, D.; Zhu, P.; Wen, L.; Bian, X.; Ling, H.; Hu, Q.; Zheng, J.; Peng, T.; Wang, X.; Zhang, Y.; et al. VisDrone-SOT2019: The vision meets drone single object tracking challenge results. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Republic of Korea, 27–28 October 2019.
35. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 618–626.
36. Cai, Z.; Vasconcelos, N. Cascade r-cnn: Delving into high quality object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6154–6162.
37. Ale, L.; Zhang, N.; Li, L. Road damage detection using RetinaNet. In Proceedings of the 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 10–13 December 2018; IEEE: New York, NY, USA, 2018; pp. 5197–5200.
38. Su, Z.; Yu, J.; Tan, H.; Wan, X.; Qi, K. Msa-yolo: A remote sensing object detection model based on multi-scale strip attention. *Sensors* **2023**, *23*, 6811. [[CrossRef](#)]
39. Li, W.; Zhang, X.; Peng, Y.; Dong, M. DMNet: A network architecture using dilated convolution and multiscale mechanisms for spatiotemporal fusion of remote sensing images. *IEEE Sens. J.* **2020**, *20*, 12190–12202. [[CrossRef](#)]

40. Li, Y.; Han, Z.; Xu, H.; Liu, L.; Li, X.; Zhang, K. YOLOv3-lite: A lightweight crack detection network for aircraft structure based on depthwise separable convolutions. *Appl. Sci.* **2019**, *9*, 3781. [[CrossRef](#)]
41. Wang, Y.; Jodoin, P.M.; Porikli, F.; Konrad, J.; Benezeth, Y.; Ishwar, P. CDnet 2014: An expanded change detection benchmark dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Columbus, OH, USA, 23–28 June 2014; pp. 387–394.

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