electronics

MDPI

Article

Small Target Detection Algorithm for UAV Aerial Photography Based on Improved YOLOv5s

Jingcheng Shang [1], Jinsong Wang [1], Shenbo Liu [2], Chen Wang [1] and Bin Zheng [1,*]

[1] School of Computer and Communications Engineering, Changsha University of Science and Technology, Changsha 410114, China
[2] School of Physics and Electronic Science, Changsha University of Science and Technology, Changsha 410114, China
* Correspondence: zhengbin@csust.edu.cn

Abstract: At present, UAV aerial photography has a good prospect in agricultural production, disaster response, and other aspects. The application of UAVs can greatly improve work efficiency and decision-making accuracy. However, owing to inherent features such as a wide field of view and large differences in the target scale in UAV aerial photography images, this can lead to existing target detection algorithms missing small targets or causing incorrect detections. To solve these problems, this paper proposes a small target detection algorithm for UAV aerial photography based on improved YOLOv5s. Firstly, a small target detection layer is applied in the algorithm to improve the detection performance of small targets in aerial images. Secondly, the enhanced weighted bidirectional characteristic pyramid Mul-BiFPN is adopted to replace the PANet network to improve the speed and accuracy of target detection. Then, CIoU was replaced by Focal EIoU to accelerate network convergence and improve regression accuracy. Finally, a non-parametric attention mechanism called the M-SimAM module is added to enhance the feature extraction capability. The proposed algorithm was evaluated on the VisDrone-2019 dataset. Compared with the YOLOV5s, the algorithm improved by 7.30%, 4.60%, 5.60%, and 6.10%, respectively, in mAP@50, mAP@0.5:0.95, the accuracy rate (P), and the recall rate (R). The experiments show that the proposed algorithm has greatly improved performance on small targets compared to YOLOv5s.

Keywords: small target detection; Mul-BiFPN; M-SimAM; Focal EIoU

Citation: Shang, J.; Wang, J.; Liu, S.; Wang, C.; Zheng, B. Small Target Detection Algorithm for UAV Aerial Photography Based on Improved YOLOv5s. *Electronics* **2023**, *12*, 2434. https://doi.org/10.3390/electronics12112434

Academic Editor: Yeong Jun Koh

Received: 5 May 2023
Revised: 20 May 2023
Accepted: 25 May 2023
Published: 27 May 2023

1. Introduction

As UAV technology improves by leaps and bounds, UAV aerial photography has become widely used in various fields, including natural disaster detection, traffic safety monitoring, search and rescue, and agricultural and forestry management. This technology has significantly reduced labor costs and improved monitoring efficiency, leading to better management and service. However, compared to the application scenarios of traditional target detection algorithms, images captured by UAVs present several challenges, including a wide range of target scales, diverse angle changes, and complex backgrounds [1]. These factors significantly impact the accuracy rate and recall rate of target detection results.

At present, object detection algorithms can be divided into two categories: two-stage detection algorithms and one-stage detection algorithms. The two-stage detection algorithm, including convolutional neural networks [2,3], CNN [4] and R-CNN [5], generates candidate boxes that contain potential targets and then uses a region classifier to predict them. The one-stage detection algorithm, such as SSD [6] and the YOLO [7–13] series, directly classifies and predicts the target at each position on the feature map, resulting in a faster detection speed and increased practicality. Academically, the defini-

tion of small goals can be divided into two categories: relative scale-based, and absolute scale-based. The former defines the small target by its proportion to the whole image. Chen et al. [14] defined small targets as follows: when the ratio of the boundary box area to the image area falls between 0.08% and 0.58%, it can be considered a small target. The latter defines small targets from the perspective of the absolute pixel size of targets, defining small targets as those with a resolution of fewer than 32 pixels on each side.

Based on the above definition, most targets in UAV aerial photography images can be defined as small targets. However, small target detection is a hugely challenging task. Lim et al. [15] proposed FA-SSD, which aims to improve the detection of small targets by fusing feature information from F-SSD and A-SSD network structures. However, the two-stage detection algorithm used in FA-SSD results in a slow detection speed. Liu et al. [16] improved the algorithm accuracy and generalization ability by introducing a shallow feature extraction network in the P1 layer and fusing shallow features in the FPN and PAN layer. Yang et al. [17] aimed to improve the detection of small targets by adding the scSE attention mechanism module and a small target detection layer to enhance feature information. However, there were still issues with missing and the false detections of small targets. A novel object detection network called DCLANet was proposed by Zhang et al. [18], which utilizes intensive clipping and local attention techniques to enhance the feature representation of small targets. They further incorporated the bottleneck attention mechanism (BAM) into the network, leading to a substantial improvement in the detection accuracy. Jin et al. [19] proposed a scale sensing network that accurately determines the scale of predefined anchor points. This network could effectively narrow the scale search range, reduce the risk of overfitting, and improve the detection speed and accuracy of aerial images. Liu et al. [20] constructed the SPPCSPG module and introduced the shuffle attention (SA) mechanism into YOLOv5s to achieve a new lightweight network, which greatly improved the detection efficacy. The algorithm models mentioned above have significantly improved the performance of small target detection. However, there is still a lot of space for improvement in terms of detection efficacy.

In this paper, the YOLOv5s target detection algorithm is improved and tested on the public dataset VisDrone-2019 [21], which indicates that the proposed algorithm improves the detection efficacy. Compared with YOLOv5s, the model has been improved as follows:

- We add a small target detection layer to the YOLOv5s by introducing a 160×160 scale feature map for detecting small and medium targets in aerial images. This modification significantly improves the detection efficacy of small targets in the dataset.
- We add the enhanced, weighted bidirectional feature pyramid Mul-BiFPN to replace the PANet [22] network in YOLOv5s, aiming to balance the information transfer between feature maps of different scales in the network. This improves the ability to detect targets of different scales and sizes.
- To balance samples with good and poor regression quality, accelerate network convergence, and enhance the regression accuracy, we introduced the Focal EIoU [23] loss function. Compared with CIoU [24], the EIoU loss function is faster to calculate and better suited to deal with small target boxes and overlapping occluded target boxes, making it more conducive to small target detection. Additionally, the utilization of the Focal loss devotes a more concentrated emphasis on high-quality anchor boxes during the regression process.
- Finally, the M-SimAM module based on the non-parametric attention mechanism SimAM [25] is proposed and added to the backbone network. The module is designed to emphasize the more critical information of the current task, reduce, or filter out attention to other irrelevant information, and thereby enhance the detection efficacy.

The rest of the article is structured as follows: In Section 2, we describe the related work. In Section 3, the improved method is introduced elaborately. In Section 4, the relevant experiments are conducted, and the corresponding results are presented to demonstrate the superior performance of the new model. In Section 5, the paper is summarized.

2. Related Work

YOLO (You Only Look Once) is a typical one-stage detection algorithm, meaning that the network only needs to look at the picture once to output the results. Joseph Redmon et al. proposed YOLOv1 [7] in 2015, which uses a convolutional neural network (CNN) to take the entire image as the input and output the boundary box and category of the target in a forward transmission. Compared with the two-stage detection algorithm, this algorithm has significant advantages in detection speed, but the detection accuracy and generalization ability are relatively poor. In 2017, YOLOv2 [8] was proposed. Compared with YOLOv1, YOLOv2 used anchor boxes to predict target location and size, adopted a convolution layer instead of a full connection layer, etc. These improvements greatly improved YOLOv2 in accuracy and speed. Then, YOLOv3 [9], YOLOv4 [10], YOLOv5 [11], YOLOv7 [13], YOLOv8, and other models were proposed successively.

YOLOv5 is a one-stage target detection algorithm. This model has high computational efficiency and a simple structure. YOLOv5 uses a deep convolutional neural network (CNN) to detect objects. Its implementation method mainly involves dividing the entire image into a series of grids and predicting the presence of objects in each grid, as well as their location, size, category, and other information. The structure of the YOLOv5 model includes the following parts:

Backbone: The backbone of YOLOv5 consists of CBS, C3, and SPPF. It converts the original input images into multi-layer feature maps and extracts features for subsequent target detection. As a basic module commonly used in convolutional neural networks, the CBS module functions in the feature extraction of input images and is composed of Conv [26], BatchNorm [27], and SiLU [28]. Conv is the most basic layer in convolutional neural networks, mainly used to extract local spatial information in input features. The BatchNorm layers are appended after the convolutional layers, which help to normalize the distribution of eigenvalues in the network. The SiLU activation function is a nonlinear function that introduces a nonlinear transformation capability to the neural network. The C3 module can significantly enhance the computational efficiency of the network and enhance the speed and efficiency of target detection while maintaining high accuracy. The SPP module is a spatial pyramid pool layer that can process input images with different resolutions and realize adaptive size output, which can increase the receptive field and extract the overall feature of the detection target. The latest version of YOLOv5 uses SPPF instead of SPP, reducing the amount of computation by half with the same effect.

Neck: YOLOv5 adopts a combination of Feature Pyramid Networks (FPN) [29] and Path Aggregation Network PANet [22] in the neck. Compared with FPN, PANet uses a combination of bottom-up network and top-down network structures to fuse feature maps of different scales. The feature maps of different scales complement each other to enhance the detection efficacy.

Head: The head of YOLOv5 uses three detection heads to detect target objects and predict the category and location of the target. It can detect the feature maps of 80×80, 40×40, and 20×20 at different scales, respectively.

According to the different depth and width of the network, YOLOv5 has five versions: YOLOv5n, YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x. This paper focuses on improving YOLOv5s to effectively enhance the detection of small and medium targets in aerial images. The structure of YOLOv5s is shown in Figure 1.

The attention mechanism is a popular technology used in deep learning models. Its function is to focus on the most important information for the current task and filter out

irrelevant information. This can improve the model's performance. Currently, many attention modules have been introduced into the field of target detection, among which SENet [30], CBAM [31], and CA [32] are relatively classical. The SE-Net network introduces squeeze and excitation blocks to learn the relationship between feature map channels, which improves the model's performance. The CBAM network introduces channel attention mechanisms and spatial attention mechanisms to weight feature maps and extract more effective information. The coordinate attention CA introduces spatial position information and constantly adjusts the channel and weight of each position in the feature map to obtain more important feature information.

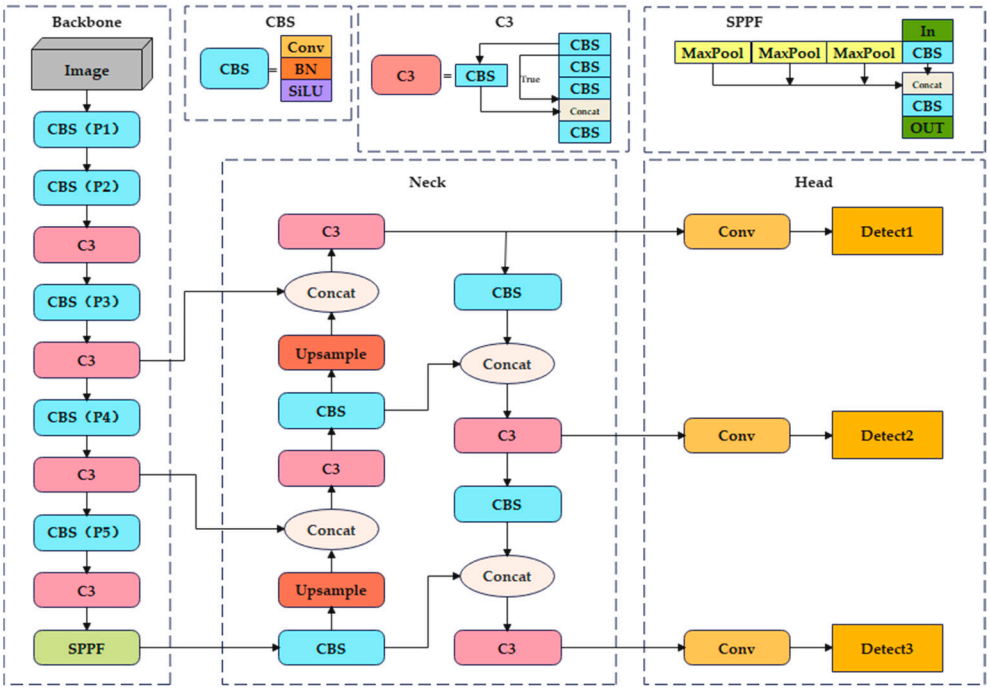


Figure 1. Block diagram of YOLOv5s.

3. Methods

3.1. Add a Small Target Detection Layer

Although YOLOv5 demonstrates exceptional performance across various application scenarios, it suffers from inadequate detection outcomes when dealing with small targets in aerial images. The underlying reason lies in the convolution-based feature extraction mechanism employed by YOLOv5. With the increasing depth of the network, it continuously reduces the size of objects due to the pooling layer and convolution kernel operations. Consequently, this phenomenon can cause missed or false detections.

In comparison to the YOLOv5s model, we added a new Detect-P2 small target detection layer to enhance the detection performance in aerial images. Due to the loss of shallow, small target information caused by multiple convolution and pooling operations on the deep feature map, the new layer fuses shallow and deep features to process small targets. Specifically, the shallow feature contains low-level semantic features, and the deep feature contains high-level semantic features. Additionally, the new detection layer detects the shallow feature layer, which can effectively detect the feature information of small targets and improve the detection efficacy of small targets in aerial images. Our model includes four detection heads that detect feature maps of different scales, denoted as 160 × 160, 80 × 80, 40 × 40, and 20 × 20, respectively.

The network structure after adding the new Detect-P2 detection layer is shown in Figure 2.

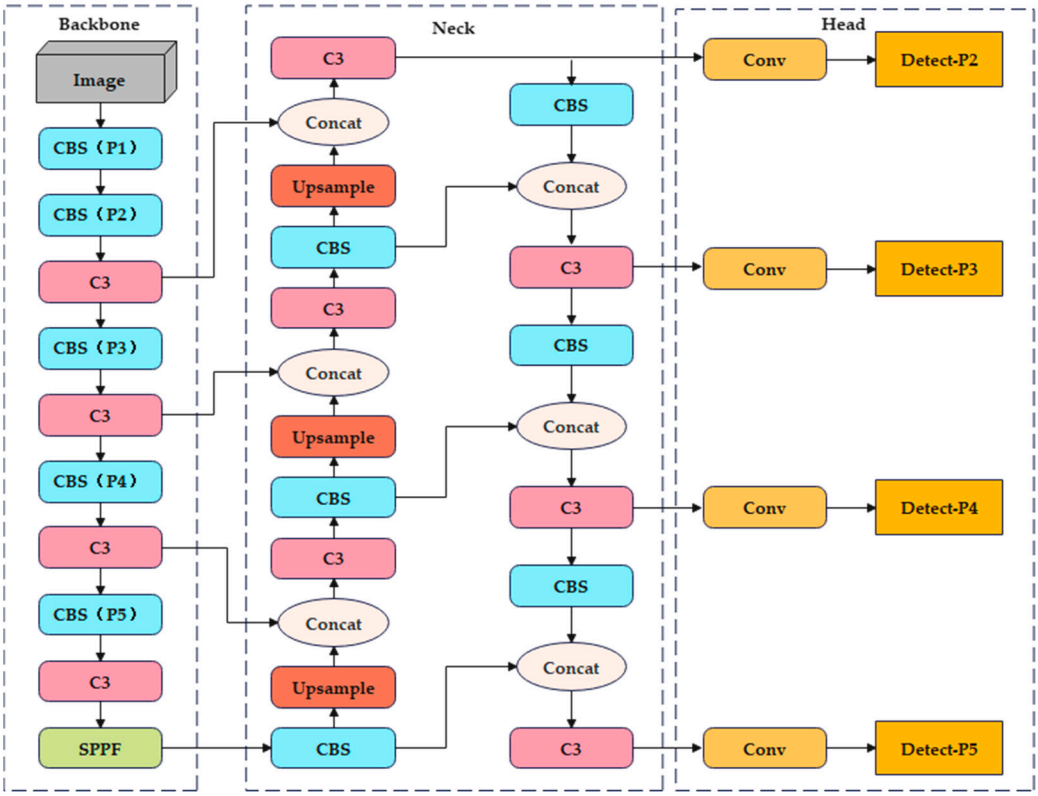


Figure 2. YOLOv5s with small target detection layer added.

3.2. Mul-BiFPN Network Structure

The YOLOv5s model utilizes the PANet as a neck for feature fusion, which is a pyramid-like structure that fuses feature map information from the bottom to the top. Although this network can somewhat address the problem of large mesoscale differences in detection images, it suffers from unidirectional information flow and poor scalability. To address these limitations, we introduce the Weighted Bidirectional Feature Pyramid Network (BiFPN) [33] in this paper. BiFPN incorporates a bidirectional connection mechanism to enhance the feature transfer capability and enable better handling of targets of different scales. Additionally, it uses a weighted feature fusion mechanism to balance information transfer between feature maps of different scales, which enhances the detection performance of both small and large targets.

The BiFPN network exhibits excellent performance in various complex scenarios. However, it sometimes misses extracting characteristic information from small targets. This is due to its main focus on extracting features from deeper layers, while being relatively weaker in extracting shallow-level feature information. To solve this problem, this paper proposes an enhanced weighted bidirectional feature pyramid network called Mul-BiFPN, which is based on the BiFPN network.

Due to the presence of more shallow feature information in the C3 layer, the Mul-BiFPN network performs additional weighted feature fusion operations in this layer to strengthen the fusion of small target features. As a result, the feature layer detected by the P3 detection head contains richer, shallow, small target feature information, which can enhance the detection efficacy of small targets during the prediction process. The Mul-BiFPN network carries out multiple information exchanges and fusions between feature maps at different levels. Through multi-layer feature fusion, the network enhances features between different layers, thereby improving the receptive field and semantic expression ability of the target detection model. This leads to a significant improvement in the robustness and accuracy of target detection.

The improvement in the neck network is shown in Figure 3.

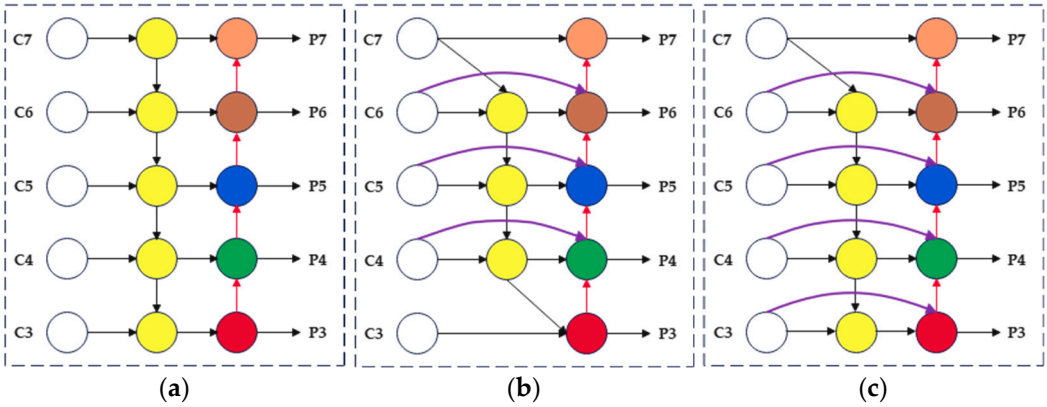


Figure 3. Network structure improvement. (**a**) PANet network structure. (**b**) BiFPN network structure. (**c**) Mul-BiFPN network structure.

3.3. Focal EIoU Loss

YOLOv5s uses the CIoU loss function, which employs a scientific and precise method to calculate the distance between target boxes. This enables accurate calculation of the actual distance between target boxes and improves training accuracy. However, the CIoU loss function places more weight on large targets, making it more sensitive to large targets and slightly insufficient for detecting small targets. To address this problem, this paper introduces a more efficient loss function called Focal EIoU [23], which replaces the CIoU loss function to improve small target detection performance. The EIOU loss function considers the aspect ratio of the target box's length and width when calculating the distance, enabling better handling of target boxes of different sizes and a more balanced weighting of different target sizes in the loss function. Additionally, EIOU uses a more efficient calculation method for the distance between target boxes, making the network model more efficient in training for small targets. Focal loss gives more weight to rare categories of targets, making it more sensitive to small target detection. Furthermore, by introducing a balance factor, it can alleviate category imbalance problems to some extent. Based on EIOU, Focal EIoU combines Focal loss to focus on better anchor boxes, thereby significantly improving the accuracy of detecting small targets. The Focal EIoU formula is shown below.

$$L_{EIOU} = L_{IOU} + L_{dis} + L_{asp} = 1 - IOU + \frac{\rho^2(b, b^{gt})}{c^2} + \frac{\rho^2(w, w^{gt})}{C_w^2} + \frac{\rho^2(h, h^{gt})}{C_h^2} \quad (1)$$

$$L_{Focal-EIOU} = IOU^{\gamma} L_{EIOU} \quad (2)$$

In Formula (1), L_{IOU} stands for IOU loss, L_{dis} stands for distance loss; L_{asp} stands for side length loss; $\rho^2(b, b^{gt})$ stands for Euclidean Distance between b (center coordinates of the prediction box) and b^{gt}(center coordinates of the real box); $\rho^2(w, w^{gt})$ stands for the square of the difference between w (width of the prediction box) and w^{gt}(width of the real box); $\rho^2(h, h^{gt})$ stands for the square of the difference between h (height of the prediction box) and h^{gt} (height of the real box); C_w and C_h, respectively, represent the width and height of the minimum enclosing rectangle of the predicted box and the target box; and c represents the diagonal length of the minimum enclosing rectangle. In Formula (2), γ is a hyperparameter used to control the curve.

3.4. M-SimAM Attention Mechanism

Aiming at the problem that most existing attention mechanisms can only focus on features in one dimension of channel or space, lacking flexibility in the simultaneous processing of space and channel. This paper introduces the non-parametric attention mechanism module called SimAM [25], and proposes the M-SimAM module based on it.

This module can help the network to extract the feature and enhance the detection performance of small targets without increasing other parameters.

SimAM is a 3D attention module whose core idea is to focus attention on the most relevant parts using a similarity-based weighting method. SimAM calculates the attention weight by designing an energy function. Most of its operations are energy function-based solutions that can flexibly and effectively enhance the extraction of features in neural networks. The structure of the SimAM module is shown in Figure 4.

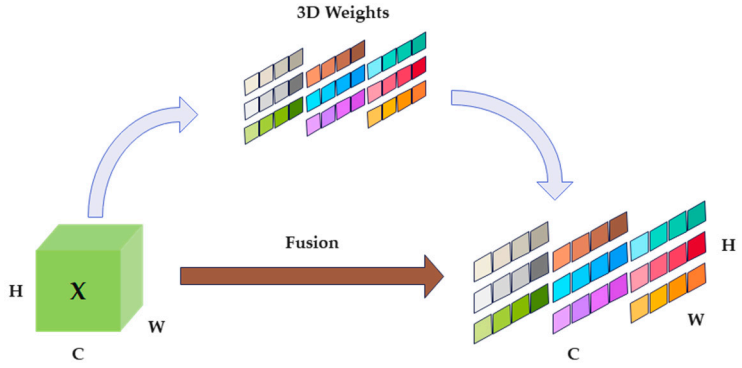


Figure 4. The structure of SimAM module.

In neuroscience, neurons that carry more information exhibit different firing patterns than other peripheral neurons and inhibit adjacent neurons. The simplest way to identify important neurons is to measure the linear separability between them. The energy function is defined as follows:

$$e_t(w_t, b_t, y, x_i) = (y_t - \hat{t})^2 + \frac{1}{M-1}\sum_{i=1}^{M-1}(y_0 - \hat{x}_i)^2 \tag{3}$$

$$\hat{t} = w_t t + b_t \tag{4}$$

$$\hat{x}_i = w_t x_i + b_t \tag{5}$$

In the above formula, t is the target neuron in the single channel of the input feature, x_i is other neurons, $\hat{t}$ and $\hat{x}_i$ are the linear change in t and x_i, respectively, w_t and b_t are linear changes in weight and deviation, respectively, y is the variable, and M is the number of neurons on the channel. By minimizing Formula (3), it can be seen as training the linear separability between the target neuron t and other neurons in the identical channel. Finally, the formula is streamlined by utilizing binary labels and incorporating regular terms, as shown below:

$$e_t(w_t, b_t, y, x_i) = \frac{1}{M-1}\sum_{i=1}^{M-1}(-1 - (w_t x_i + b_t))^2 + (1 - (w_t t + b_t))^2 + \lambda w_t^2 \tag{6}$$

In Formula (6), λ is the coefficient. The solution of the energy formula is as follows:

$$\begin{cases} w_t = -\dfrac{2(t-\mu_t)}{(t-\mu_t)^2 + 2\sigma_t^2 + 2\lambda} \\ b_t = -\dfrac{1}{2}(t+\mu_t)w_t \\ \mu_t = \dfrac{1}{M-1}\displaystyle\sum_{i=1}^{M-1} x_i \\ \sigma_t^2 = \dfrac{1}{M-1}\displaystyle\sum_{i=1}^{M-1}(x_i - \mu_t)^2 \end{cases} \tag{7}$$

In Formula (7), σ_t^2 and μ_t are the variance and mean of all neurons except t. Therefore, the minimum energy can be calculated by the following formula:

$$e_t^* = \frac{4(\hat{\sigma}^2 + \lambda)}{(t - \hat{\mu})^2 + 2\hat{\sigma}^2 + 2\lambda} \tag{8}$$

From Formula (8), it is evident that lower energy corresponds to a greater disparity between the neuron and its neighboring neuron, indicating the significance of the target neuron. Therefore, $1/e_t^*$ represents the weight of neurons, and the formula for incorporating the SimAM module into the feature map can be calculated using the following formula:

$$\tilde{X} = sigmoid(\frac{1}{E}) \odot X \tag{9}$$

In Formula (9), E represents the aggregation of e_t^* across channel and spatial dimensions. X represents the raw feature values input into the SimAM module. $\tilde{X}$ represents the feature output values after being enhanced by the SimAM module. The $sigmoid$ activation function is used to constrain excessively large values.

In this paper, we introduce the $mish$ activation function to replace the $sigmoid$ activation function in the SimAM. The new module is named M-SimAM. Due to its smoothness and better gradient performance, the $mish$ activation function can enhance the model's ability to detect small targets and avoid problems such as disappearing gradients or exploding gradients, thus improving the model's accuracy and stability. Moreover, the $mish$ activation function can enhance the model's generalization ability, making it perform better in more challenging scenarios. The formula for the feature map incorporating the M-SimAM attention mechanism is shown below:

$$\tilde{X} = mish(\frac{1}{E}) \odot X \tag{10}$$

In Formula (10), the $mish$ activation function is introduced to replace the $sigmoid$ activation function. X represents the raw feature values that are input into the M-SimAM module. $\tilde{X}$ represents the feature output values after being enhanced by the M-SimAM module.

Based on the aforementioned improvements, the network structure of the algorithm proposed in this paper is illustrated in Figure 5.

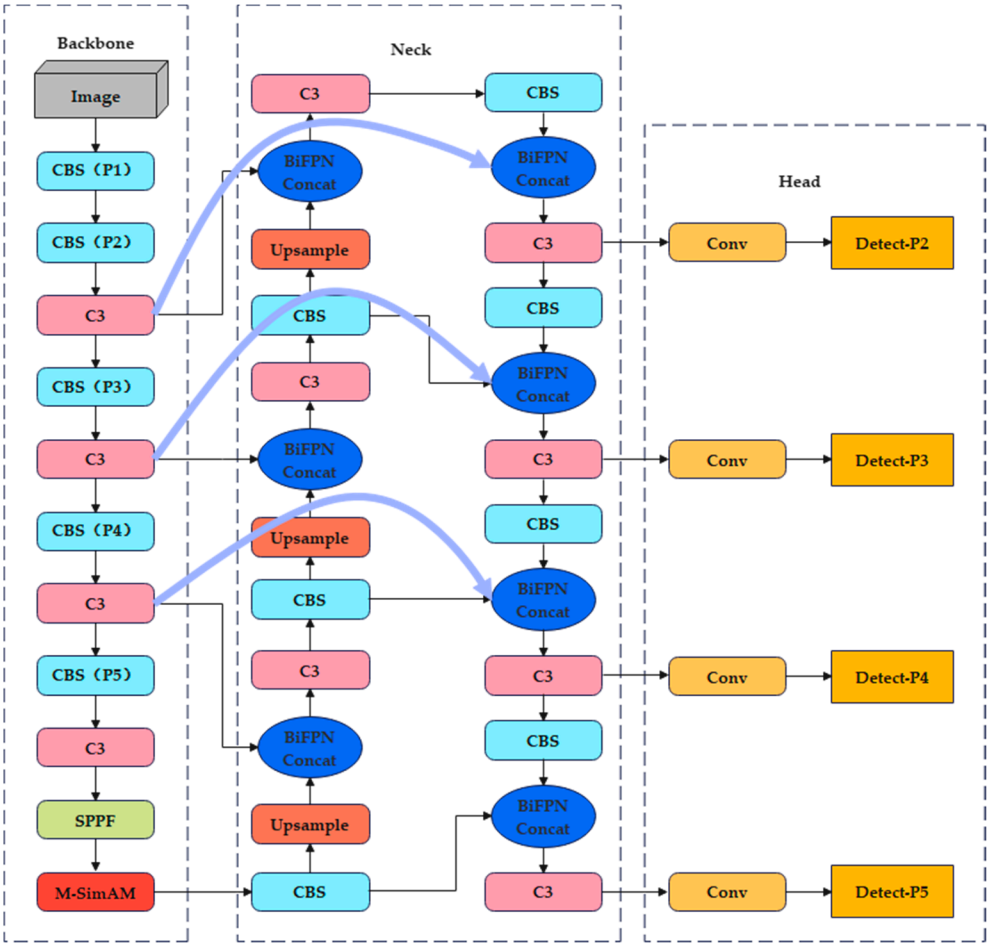


Figure 5. Algorithm network structure in this paper.

4. Experiments

4.1. Dataset

The experiment in this paper utilized the VisDrone-2019 dataset, which was publicly released by the AISKYEYE team at Tianjin University. The dataset was collected by various types of drones in diverse weather, scenarios, and lighting conditions, encompassing 288 video clips, 261,908 frames, and 10,209 still images. The images are annotated with ten categories of labels, including awning-tricycle, bicycle, bus, motor, people, pedestrian, car, truck, tricycle, and van. The category distribution of examples in this dataset is shown in Figure 6.

The VisDrone-2019 dataset contains 6471, 548, and 1610 images for training, validation, and testing, respectively. Figure 7 depicts the distribution of all category label sizes in the training dataset. The horizontal axis of the figure stands for the width of the object label frame, while the vertical axis stands for the height of the object label frame. Most of the data points are situated in the lower-left corner of the figure, indicating that the VisDrone-2019 dataset has a greater proportion of small target objects. This observation aligns with the problem explored in this paper.

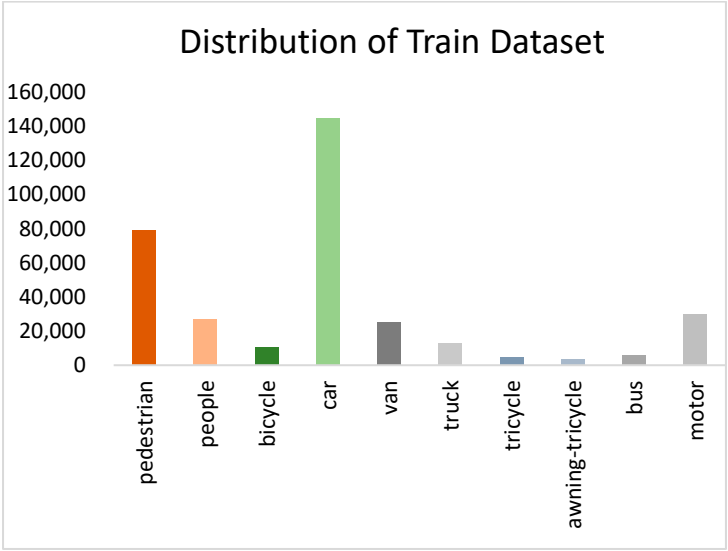


Figure 6. VisDrone-2019 data label distribution.

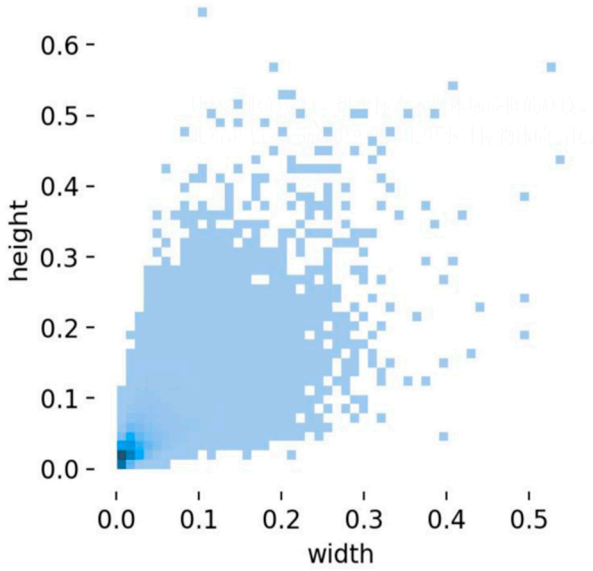


Figure 7. The distribution of all category label sizes in the training dataset.

4.2. Experimental Platform

The experiment employed an Ubuntu 18.04 system and utilized Python 3.9.16, Pytorch 2.0.0, and CUDA 11.8. The experimental model training platform consisted of RTX 3090 GPUs, and training, validation, and testing were conducted using identical hyperparameters. Specifically, the training epoch was set to 300, the batch size was set to 16, and the image resolution was 640×640. The pre-training model employed was YOLOv5s.pt, which was provided by the official source.

4.3. Evaluation Criteria

The experimental results were assessed using the cross-validation method. Following training and validation with the respective datasets, a final performance evaluation of the model was performed using the test dataset. In the experiment, the network's performance was evaluated based on three performance metrics: the accuracy rate (P), recall rate (R), and mean average precision (mAP).

The accuracy rate (P) stands for the percentage of targets that are correctly predicted in all detected targets. The accuracy rate can be calculated by the following formula: TP stands for the correct prediction of the model and FP stands for the wrong prediction of the model.

$$P = \frac{TP}{TP + FP} \tag{11}$$

The recall rate (R) stands for the proportion of targets that are correctly predicted in all targets. The recall rate can be calculated by the following formula: FN stands for the targets that need to be predicted but are incorrectly detected by the model.

$$R = \frac{TP}{TP + FN} \tag{12}$$

The average accuracy (AP) stands for the area enclosed by the axis of the curve formed by the precision rate and recall rate, and the average accuracy (mAP) is the mean of the average accuracy of all samples, which can be calculated by the following formula:

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

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

4.4. Experimental Results

In the experiment, the performance of the new model was evaluated on the VisDrone-2019 dataset. Comparing the experimental results of the new model and the YOLOv5s model, it can be concluded that the proposed algorithm achieves better performance in detecting small targets than YOLOv5s.

Tables 1 and 2 present a comparative performance evaluation of YOLOv5s and the proposed algorithm on the VisDrone-2019 validation and test dataset. The results show that the proposed algorithm outperforms the YOLOv5s model. In the validation dataset, the proposed algorithm achieves a 5.60% increase in P and a 5.90% increase in R, as well as a 6.60% increase in mAP@0.5 and a 4.60% increase in mAP@0.5:0.95, compared to the original model. In the test dataset, the proposed algorithm achieves a 5.60% increase in P and a 6.10% increase in R, as well as a 7.30% increase in mAP@0.5 and a 4.60% increase in mAP@0.5:0.95. The considerable improvement in performance of the new model on the VisDrone-2019 dataset suggests its effectiveness in small target detection.

Table 1. Comparison of the detection effect between YOLOv5s and our model on the validation dataset.

Models	**P(%)**	**R(%)**	**mAP@0.5 (%)**	**mAP@0.5:0.95 (%)**
YOLOv5s	45.90	31.60	31.80	17.60
Ours	51.50	37.50	38.40	22.20

Table 2. Comparison of the detection effect between YOLOv5s and our model on the test dataset.

Models	**P(%)**	**R(%)**	**mAP@0.5 (%)**	**mAP@0.5:0.95 (%)**
YOLOv5s	41.30	31.90	29.10	15.50
Ours	46.90	38.00	36.40	20.10

Figure 8 illustrates the comparison of detection effects between the YOLOv5s model and the proposed model on the VisDrone-2019 test dataset. We can directly observe from the figure that, compared to Figure 8(a), more small targets have been successfully

detected in Figure 8(b). The experimental results explain how the algorithm in this paper enhances the detection ability of small target objects, significantly solving the problem of small targets being easily missed and incorrectly detected in the existing object detection literature. At the same time, the detection confidence of almost all objects is improved, and the accuracy of detection is significantly enhanced.

(**a**) (**b**)

Figure 8. Comparison of detection effect between the YOLOv5s model and the proposed model. (**a**) The detection effect of the YOLOv5s model. (**b**) The detection effect of the new model.

4.5. Ablation Experiment

To further verify the effectiveness of the proposed algorithm, an ablation experiment was conducted on the VisDrone-2019 dataset. Using YOLOv5s as the baseline model, multiple improved methods mentioned in this paper were added one-by-one or in combination to verify the improvement in the target detection performance of each method.

Table 3 shows the results of the ablation experiment conducted on the VisDrone-2019 validation dataset. The experiment involved adding various improvements to the basic YOLOv5s model, including a P2 layer small target detection head (shown in table +P2), BiFPN (shown in table +BF), Mul-BiFPN network structures (shown in table +MBF), Focal EIoU (shown in table +FE), M-SimAM attention module (shown in table +MSimA), and combinations of these improvements. The results in Table 3 show that each improvement significantly enhances P, R, mAP@0.5, and mAP@0.5:0.95 on the VisDrone-2019 validation dataset compared to the YOLOv5s model. For instance, the final model proposed in this paper achieved a 5.60% increase in P and a 5.90% increase in R, along with a 6.60% increase in mAP@0.5 and a 4.60% increase in mAP@0.5:0.95, compared to the original model on this validation dataset.

Table 3. Ablation experiments of VisDrone-2019 validation dataset.

Models	P(%)	R(%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)
YOLOv5s	45.90	31.60	31.80	17.60
+P2	49.80	35.30	36.60	21.10
+BF	45.50	32.80	32.70	17.90
+MBF	48.00	34.60	35.30	19.70
+FE	46.90	32.80	32.90	18.10
+MSimA	45.10	32.40	32.00	17.50
+P2+MBF	49.30	35.80	36.50	21.20
+P2+MBF+FE	51.50	37.50	38.70	22.50
+P2+MBF+FE+MSimA	51.50	37.50	38.40	22.20

Table 4 shows the ablative experiment results of the VisDrone-2019 test dataset, which is utilized to reflect the detection effect of the new model in unknown and complex scenarios. Through our analysis, we discovered that adding the Detect-P2 detection layer can significantly enhance the detection performance of smaller targets, resulting in an increase of 3.40% in P, 4.00% in R, and 5.00% in mAP@50. Furthermore, the Mul-BiFPN and BiFPN proposed in this paper were individually added to YOLOv5s for comparative testing, and it was concluded that Mul-BiFPN has better detection performance than BiFPN. With the addition of Focal EIoU and M-SimAM Attention, the detection efficacy of small targets in aerial images was significantly enhanced. Compared to YOLOv5s, the proposed algorithm achieved a 5.60% increase in P and a 6.10% increase in R on the VisDrone-2019 test dataset, as well as a 7.30% increase in mAP@50 and a 4.60% increase in mAP@0.5:0.95, which validates the effectiveness of the proposed algorithm improvements.

Table 4. Ablation experiment of VisDrone-2019 test dataset.

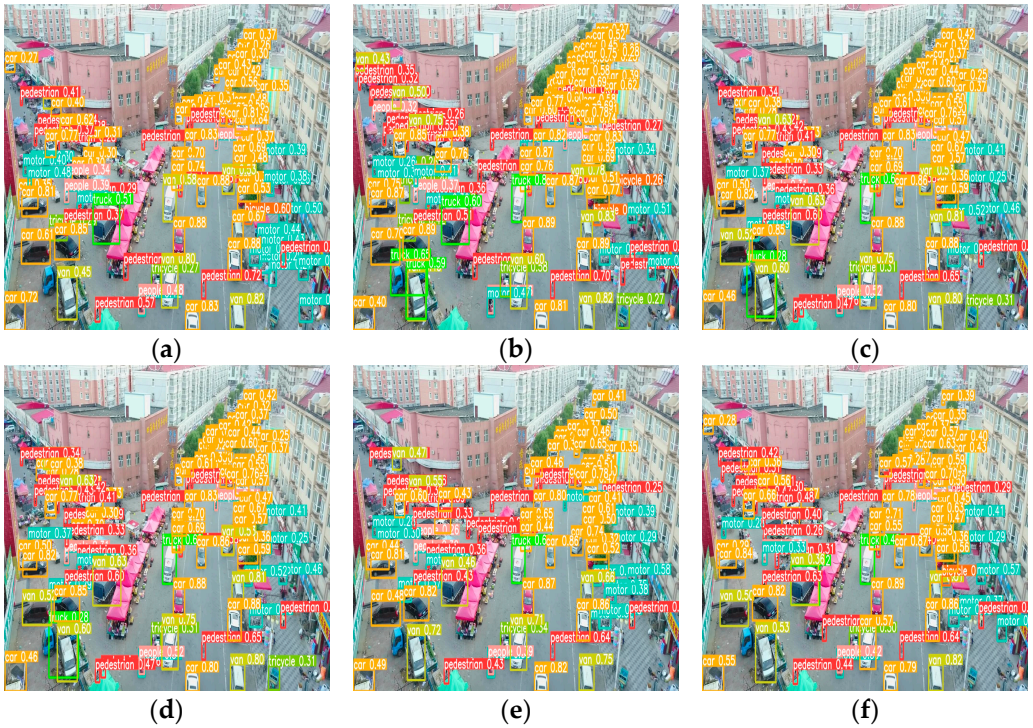
Models	P(%)	R(%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)
YOLOv5s	41.30	31.90	29.10	15.50
+P2	44.70	35.90	34.10	18.70
+BF	41.40	32.40	29.80	15.90
+MBF	43.30	35.10	31.90	17.20
+FE	42.20	33.10	30.70	16.40
+MSimA	40.60	32.10	29.50	15.70
+P2+MBF	46.10	37.20	35.50	19.60
+P2+MBF+FE	46.30	38.00	36.00	19.80
+P2+MBF+FE+MSimA	46.90	38.00	36.40	20.10

Table 5 presents the accuracy of various target detections in the VisDrone-2019 test dataset. According to our data analysis, the accuracy of various target detections in unknown scenarios has been significantly improved in the proposed algorithm. The detection accuracy of bicycles, truck, and motor showed the greatest improvement, increasing by 12.90%, 8.10%, and 6.20%, respectively. Although the accuracy of some types of target detections may decrease to some extent with the addition of each module, the overall detection accuracy is still improved.

Table 5. Precision of various objects in VisDrone-2019 test dataset.

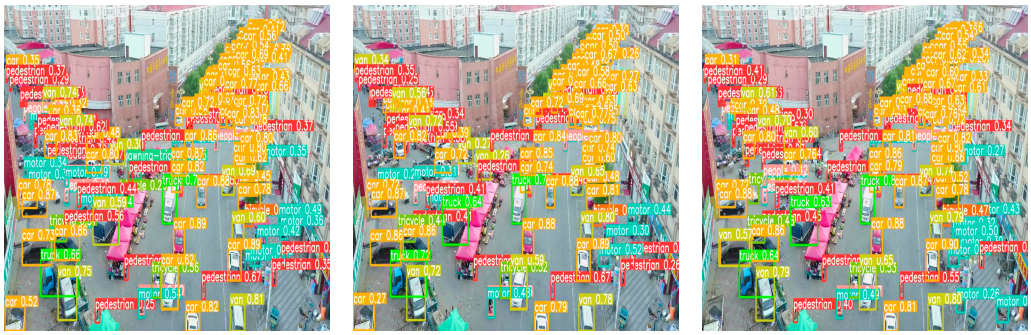
	YOLOv5s (%)	+P2 (%)	+P2+MBF (%)	+P2+MBF+FE (%)	+P2+MBF+FE+MSimA (%)
pedestrian	43.90	44.40	46.50	52.30	48.60
people	41.70	42.60	44.90	44.00	44.00
bicycle	26.30	35.50	35.60	34.80	39.20
car	59.50	62.40	63.50	66.90	65.40
van	37.30	41.40	42.90	41.20	43.20
truck	39.00	44.70	46.90	45.90	47.10
tricycle	25.00	25.00	24.90	26.70	28.10
awning-tricycle	34.30	38.80	42.90	37.30	37.20
bus	64.70	69.40	68.30	68.70	69.00
motor	41.00	43.30	45.10	45.80	47.20
all	41.30	44.70	46.30	46.30	46.90

The ablation experiment under the same scenario of different models is shown among (a)~(i) in Figure 9. As can be seen clearly from the picture comparison, when a module is added or improved to the original model (a), the image detection effect will be improved compared with the (a) model. When there are target groups with a large density and coverage in the images, the (a) model has a high probability of missing and false detection, but the (i) model can deal with this situation well, greatly improving the recall rate and accuracy rate and improving the detection confidence of each target.



(**a**) (**b**) (**c**)

(**d**) (**e**) (**f**)



(**g**) (**h**) (**i**)

Figure 9. Comparison of ablation experiments. (**a**) YOLOv5s model. (**b**) YOLOv5s+P2 model. (**c**) YOLOv5s+BF model. (**d**) YOLOv5s+MBF model. (**e**) YOLOv5s+FE model. (**f**) YOLOv5s+MSimA model. (**g**) YOLOv5s+P2+MBF model. (**h**) YOLOv5s+P2+MBF+FE model. (**i**) Algorithm model of this paper.

To verify the proposed algorithm's significant improvement in detecting small targets, a comparative evaluation experiment was conducted between the proposed algorithm and other models, including YOLOv3, YOLOv5s, YOLOv5l, YOLOv7 [13], and YOLOv8s. The comparison of different models is presented in Table 6.

Table 6. Comparison of different models in VisDrone-2019 test dataset.

Models	P(%)	R(%)	mAP@0.5 (%)	Pedestrian	Car	Bus	Model Size (MB)	Inference (ms)
YOLOv3	44.80	34.80	32.70	0.48	0.63	0.66	123.50	13.5
YOLOv5s	41.30	31.90	29.10	0.44	0.60	0.65	14.40	9.40
YOLOv5l	47.60	37.30	35.90	0.37	0.57	0.63	92.90	12.10
YOLOv7	39.90	36.60	31.50	0.34	0.72	0.33	12.30	4.80
YOLOv8s	45.00	36.30	34.00	0.49	0.68	0.62	22.50	7.10
Ours	46.90	38.00	36.40	0.49	0.65	0.69	15.30	11.90

As shown in Table 6, our model outperforms other models in terms of performance evaluation on the VisDrone-2019 test dataset. Compared to the YOLOv5s model, the new model achieves an improvement of 5.60% in detection accuracy, surpassing YOLOv5l and YOLOv8s as well. Furthermore, compared to other mainstream algorithms, the proposed algorithm exhibits better performance in detection accuracy (P), recall rate (R), and mAP@0.5. Notably, it achieves higher detection accuracy for pedestrians and buses than other models. In summary, the proposed algorithm shows superiority in detecting small targets, which validates the feasibility of the algorithm.

5. Conclusions

In this paper, we propose a new model based on YOLOv5s that significantly enhances the small target detection efficacy in UAV aerial photography images with large scale differences and high overlap rates. We achieve this by adding a new Detect-P2 detection layer to the original YOLOv5s model to enhance the detection of shallow, small target information. Furthermore, we replace the PANet network with the Mul-BiFPN network to enhance the fusion of information from different feature layers. To improve regression accuracy, we introduce the Focal EIoU loss function, which also accelerates network convergence. Lastly, we add the proposed M-SimAM module to the last layer of the backbone to enhance effective information extraction and improve detection accuracy.

The results of the experiment demonstrate that our model outperforms YOLOv5s in detecting objects in the VisDrone-2019 dataset. It significantly reduces false detection

and missing detection of small targets and greatly improves the detection accuracy of various objects. Furthermore, the new model shows superior detection performance compared to other models.

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

Funding: This research received no external funding.

Data Availability Statement: The code for the proposed model in this paper is openly available at https://github.com/sjchaha/1_yolov5s_improve.git (accessed on 20 May 2023)

Acknowledgments: The authors want to thank the editor and anonymous reviewers for their valuable suggestions for improving this paper.

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

References

1. 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.
2. Ahmed, S.; Kamal, U.; Hasan, K. DFR-TSD: A Deep Learning Based Framework for Robust Traffic Sign Detection under Challenging Weather Conditions. *IEEE Trans. Intell. Transp. Syst.* **2020**, *23*, 5150–5162.
3. Cao, J.; Zhang, J.; Jin, X. A Traffic-Sign Detection Algorithm Based on Improved Sparse R-cnn. *IEEE Access* **2021**, *9*, 122774–122788.
4. Shin, H.-C.; Roth, H.R.; Gao, M.; Lu, L.; Xu, Z.; Nogues, I.; Yao, J.; Mollura, D.J.; Summers, R.M. Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. *IEEE Trans. Med. Imaging* **2016**, *35*, 1285–1298.
5. Girshick, R.B.; 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.
6. Wei, L.; Dragomir, A.; Dumitru, E.; Christian, S.; Scott, R.; Cheng-Yang, F.; 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.
7. Redmon, J.; Divvala, S.K.; Girshick, R.B.; Farhadi, A. You only look once: unified, real-time object detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
8. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6517–6525.
9. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
10. Bochkovskiy, A.; Wang, C.-Y.; Liao, H.-Y.M. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004.10934.
11. Zhu, X.; Lyu, S.; Wang, X.; Zhao, Q. TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, Montreal, BC, Canada, 11–17 October 2021; pp. 2778–2788.
12. 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:abs/2209.02976.
13. Wang, C.-Y.; Bochkovskiy, A.; Liao, H.-Y.M.J.A. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv* **2022**, arXiv:abs/2207.02696.
14. Chen, C.; Liu, M.Y.; Tuzel, O.; Xiao, J. R-CNN for small object detection. In Proceedings of the Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, 20–24 November 2016; Springer: Cham, Switzerland, 2016.
15. Lim, J.-S.; Astrid, M.; Yoon, H.; Lee, S.-I. Small object detection using context and attention. In Proceedings of the 2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), Jeju, Republic of Korea, 13–16 April 2019; pp. 181–186.
16. Liu, H.; Duan, X.; Chen, H.; Lou, H.; Deng, L. DBF - YOLO: UAV Small Targets Detection Based on Shallow Feature Fusion. *IEEJ Trans. Electr. Electron. Eng.* **2023**, *18*, 605–612.
17. Yang, R.; Li, W.; Shang, X.; Zhu, D.; Man, X. KPE-YOLOv5: An Improved Small Target Detection Algorithm Based on YOLOv5. *Electronics* **2023**, *12*, 817.

18. Zhang, X.; Feng, Y.; Zhang, S.; Wang, N.; Mei, S. Finding Nonrigid Tiny Person With Densely Cropped and Local Attention Object Detector Networks in Low-Altitude Aerial Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* **2022**, *15*, 4371–4385.
19. Jin, R.; Lin, D. Adaptive Anchor for Fast Object Detection in Aerial Image. *IEEE Geosci. Remote. Sens. Lett.* **2020**, *17*, 839–843.
20. Liu, P.; Wang, Q.; Zhang, H.; Mi, J.; Liu, Y. A Lightweight Object Detection Algorithm for Remote Sensing Images Based on Attention Mechanism and YOLOv5s. *Remote. Sens.* **2023**, *15*, 2429.
21. Du, D.; Zhang, Y.; Bo, L.; Shi, H.; Wang, X. VisDrone-SOT2019: The vision meets drone single object tracking challenge results. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), Seoul, Republic of Korea, 27–28 October 2019.
22. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the 2018, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 8759–8768.
23. Zhang, Y.F.; Ren, W.Q.; Zhang, Z.; Jia, Z.; Wang, L.; Tan, T.N. Focal and efficient IOU loss for accurate bounding box regression. *Neurocomputing* **2022**, *506*, 146–157, doi:10.1016/j.neucom.2022.07.042.
24. Zheng, Z.; Wang, P.; Ren, D.; Liu, W.; Ye, R.; Hu, Q.; Zuo, W. Enhancing Geometric Factors in Model Learning and Inference for Object Detection and Instance Segmentation. *IEEE Trans. Cybern*. **2020**, *52*, 8574–8586.
25. Yang, L.; Zhang, R.Y.; Li, L.; Xie, X. SimAM: A simple, parameter-free attention module for convolutional neural networks. In Proceedings of the International Conference on Machine Learning, online, 18–24 July 2021.
26. Sunkara, R.; Luo, T. No more strided convolutions or pooling: A new cnn building block for low-resolution images and small objects. In Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, Grenoble, France, 19–23 September 2022.
27. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015.
28. Elfwing, S.; Uchibe, E.; Doya, K. Sigmoid-Weighted Linear Units for Neural Network Function Approximation in Reinforcement Learning. *Neural Netw.* **2017**, *107*, 3–11.
29. Lin, T.Y.; Dollar, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
30. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
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, Munich, Germany, 8–14 September 2018.
32. 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.
33. Tan, M.; Pang, R.; Le, Q.V. EfficientDet: Scalable and efficient object detection. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 10778–10787.