information

Article

MCF-YOLOv5: A Small Target Detection Algorithm Based on Multi-Scale Feature Fusion Improved YOLOv5

Song Gao, Mingwang Gao * and Zhihui Wei

The School of Mechanical Engineering, Shandong University of Technology, Zibo 255000, China; 13396445036@163.com (S.G.); sdutwzh@163.com (Z.W.)
* Correspondence: gaomingwang@sdut.edu.cn

Abstract: In recent years, many deep learning-based object detection methods have performed well in various applications, especially in large-scale object detection. However, when detecting small targets, previous object detection algorithms cannot achieve good results due to the characteristics of the small targets themselves. To address the aforementioned issues, we propose the small object algorithm model MCF-YOLOv5, which has undergone three improvements based on YOLOv5. Firstly, a data augmentation strategy combining Mixup and Mosaic is used to increase the number of small targets in the image and reduce the interference of noise and changes in detection. Secondly, in order to accurately locate the position of small targets and reduce the impact of unimportant information on small targets in the image, the attention mechanism coordinate attention is introduced in YOLOv5's neck network. Finally, we improve the Feature Pyramid Network (FPN) structure and add a small object detection layer to enhance the feature extraction ability of small objects and improve the detection accuracy of small objects. The experimental results show that, with a small increase in computational complexity, the proposed MCF-YOLOv5 achieves better performance than the baseline on both the VisDrone2021 dataset and the Tsinghua Tencent100K dataset. Compared with YOLOv5, MCF-YOLOv5 has improved detection AP_{small} by 3.3% and 3.6%, respectively.

Keywords: YOLOv5; Mixup; coordinate attention; small target detection layer

Citation: Gao, S.; Gao, M.; Wei, Z. MCF-YOLOv5: A Small Target Detection Algorithm Based on Multi-Scale Feature Fusion Improved YOLOv5. *Information* **2024**, *15*, 285. https://doi.org/10.3390/info15050285

Academic Editor: Gabriel Luque

Received: 21 March 2024
Revised: 7 May 2024
Accepted: 10 May 2024
Published: 17 May 2024

1. Introduction

As one of the important branches of computer vision tasks, in recent years, target detection algorithms based on deep learning have been rapidly developed and target detection has been more and more widely used in fields such as automatic driving, robot vision, military surveillance, and medical image analysis and so on [1]. At present, there are two mainstream deep learning-based target detection methods. One is a one-stage target detection method represented by the You Only Look Once (YOLO) series and the other is a two-stage target detection method represented by the region selection-based convolutional neural network (R-CNN) series [2]. Although the accuracy and efficiency of these target detection methods have been greatly improved, they still perform poorly in detecting small targets [3]. Unlike the rapid development of general target detection, small target detection has not been well addressed, and thus it has been a research hotspot in the field of target detection.

The characteristics of small targets themselves lead to the following problems when detecting small targets. (1) In complex background environments, the number of small target samples is small and small target detection is easily affected by noise and changes, which can lead to overfitting problems in the model. (2) Due to the small volume of small targets, it is difficult to locate and re-identify them during the target detection process, resulting in detection errors in the model. (3) The model extracts fewer discriminative features for small targets and even suffers from feature information loss after multiple downsampling.

In order to solve the problems and challenges in small target detection, we propose a new detection model (named MCF-YOLOv5) to improve the detection efficiency.

Our main contributions are as follows. (1) Using an augmentation strategy combining Mixup [4] and Mosaic to reduce the interference of noise and changes on small targets, increase the number of specific targets, and thus reduce the overfitting problem of the model [5]. (2) The use of coordinate attention modules to embed position information into feature layer channels solves the problem of small target localization. (3) By adding two cross-layer connections to improve the Feature Pyramid Network (FPN) structure and introducing a small object detection layer on this basis, the network can obtain richer feature information and improve the detection performance of small objects.

2. Related Work

Early researchers achieved target detection by designing artificial features and using various acceleration methods but traditional target detection has poor generalization ability and low robustness. Since the great success of the AlexNet model in the ImageNet Large Scale Visual Recognition Challenge in 2012 [6], research on target detection techniques has been devoted to deep learning-based approaches.

2.1. Target Detection

At this stage, there are two types of target detection: a two-stage target detection algorithm based on candidate regions and a one-stage target detection algorithm based on regression. The two-stage target detection algorithm originated from R-CNN [7] proposed by Girshick et al. in 2014 for image target detection and segmentation, which achieved optimal results on the VOC2007 and VOC2010 datasets. Since then, CNN-based target detection methods have become a hotspot for researchers [8,9]. In 2015, Girshick et al. improved Fast R-CNN on the basis of R-CNN. In 2017, Girshick et al. and others proposed Faster R-CNN [8], which uses a Region Proposal Network (RPN) region-generating network instead of the traditional sliding window and Selective Search (SS) methods to improve the detection speed. In 2017, He et al. proposed Mask R-CNN [9], which further improves the detection accuracy by embedding a Fully-Convolutional Network (FCN) semantic segmentation module and using a RoIAlign strategy. Although the two-stage algorithm is more accurate than the traditional algorithm, the high complexity of the model and the number of parameters make it difficult to deploy and detect in real time.

The one-stage target detection algorithm is the YOLO algorithm series proposed by Redmon [10–12]. The YOLOv1 algorithm proposed in 2015 merges the extraction and detection of candidate boxes into a single stage, which greatly improves the detection speed by obtaining the specific location information and category classification information of the target detection through direct regression. The YOLO9000 network was proposed in 2016, which introduces the Anchor Box, which utilizes the K-means clustering method to calculate better a priori box parameters and improve the detection performance of the network. In the same year, under the influence of the YOLO algorithm, Liu et al. proposed the SSD algorithm [13], which utilizes multi-scale feature maps for target detection and effectively improves the detection accuracy of targets of different sizes. In 2018, Redmon again proposed the YOLOv3 algorithm, which introduces the residual network module and replaces Darknet-19 in the backbone network with Darknet-53. Borrowing the idea of a feature pyramid (FPN) [14], the prediction is performed at three different sizes separately, which improves both detection accuracy and speed. After that, the YOLOv4 algorithm was proposed in 2020 [15], which optimizes the network to varying degrees in terms of data processing, backbone network, network training, activation function, and loss function and does not drastically alter the network. In the same year, Glenn-jocher et al. proposed the YOLOv5 algorithm, which uses the focus structure and Cross Stage Partial (CSP) structure in the backbone network and the FPN + PAN structure in the neck end; at the same time, some small strategies are used in the training process to improve the detection speed and accuracy. The YOLO series of models are continuously improved and optimized with the development of deep learning, such as YOLOX and YOLOv7 [5,16].

These models have improved the performance and speed of target detection and made important breakthroughs in the field of computer vision.

2.2. Small Target Detection

The poor performance of small target detection is mainly due to the limitations of the structure of the model itself and the characteristics of the small targets themselves. The low resolution and background information of small objects make it difficult for models to extract features and small object recognition is highly susceptible to background interference, which poses challenges to the localization and recognition of small objects [17]. In order to obtain richer feature information on small targets, Kisantal et al. use the copy-and-paste method to increase the number of training samples and expand the area covered by small targets in order to significantly increase the diversity of small target locations [18]. Chen et al. scaled and spliced target images of different scales in the dataset so as to make the small-sized targets contain richer information, thus obtaining a better enhancement effect. In order to overcome the problem of low resolution of small targets [19], Romano et al. proposed a super-resolution method, which enables the neural network to learn the mapping relationship between the low-resolution image and the equivalent high-resolution image [20]. In recent years, with the development of Generative Adversarial Networks, Bai et al. proposed a multitask GAN based on recovering clear super-resolution objects from blurred small objects to recover clear super-resolution objects [21]. Li et al. introduced a perceptual GAN method for recognizing small objects [22]. Pang et al. proposed JCS-Net for reducing the image difference between small-scale targets and large-scale targets. Multilayer channel features are constructed based on HOG+LUV and JCS-Net to train small-scale pedestrian detectors [23]. For the problem of unsatisfactory small target detection, although some techniques have been developed to improve the performance of small target detection, none of them are very effective. Therefore, this paper proposes the MCF-YOLOv5 algorithm. Experiments prove that the algorithm does improve the small target detection.

3. Methods

YOLOv5 is a widely used target detection model at this stage, which is applied in various fields of people's lives and it has higher real-time accuracy than the newer version of yolov7 in some scenarios [24]. It consists of five versions: YOLOv5n, YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x. The overall architecture of the network is the same for all five versions, only different depths and widths are used in each sub-module to meet the needs of different scenarios [25]. In order to strike a balance between speed and accuracy, the latest version 6.1 of YOLOv5s is used in this paper. Figure 1 shows the structural diagram of YOLOv5s. YOLOv5 consists of four parts: Input, Backbone, Neck, and Detection. In the input section, YOLOv5 mainly enhances the dataset through Mosaic data augmentation. The backbone consists of Focus, CBS (Conv BN Silu), C3 (CSPDarkNet53), and SPP (Spatial Pyramid Pool) modules for feature extraction. The Neck section adopts the structure of FPN and PANet to enhance the feature fusion ability of the network. The Detect section is used for object detection at different scales. YOLOv5-6.1 replaces the Focus module with a 6x6 convolutional layer, which is theoretically equivalent, but for some existing GPU devices (and corresponding optimization algorithms), using a 6x6 convolutional layer is more effective than the Focus module, and replaces the SPP (Spatial Pyramid Pool) module with an SPP (Spatial Pool Pool) module. The Pooling module has been replaced by the SPPF module, which has more than doubled the processing speed.

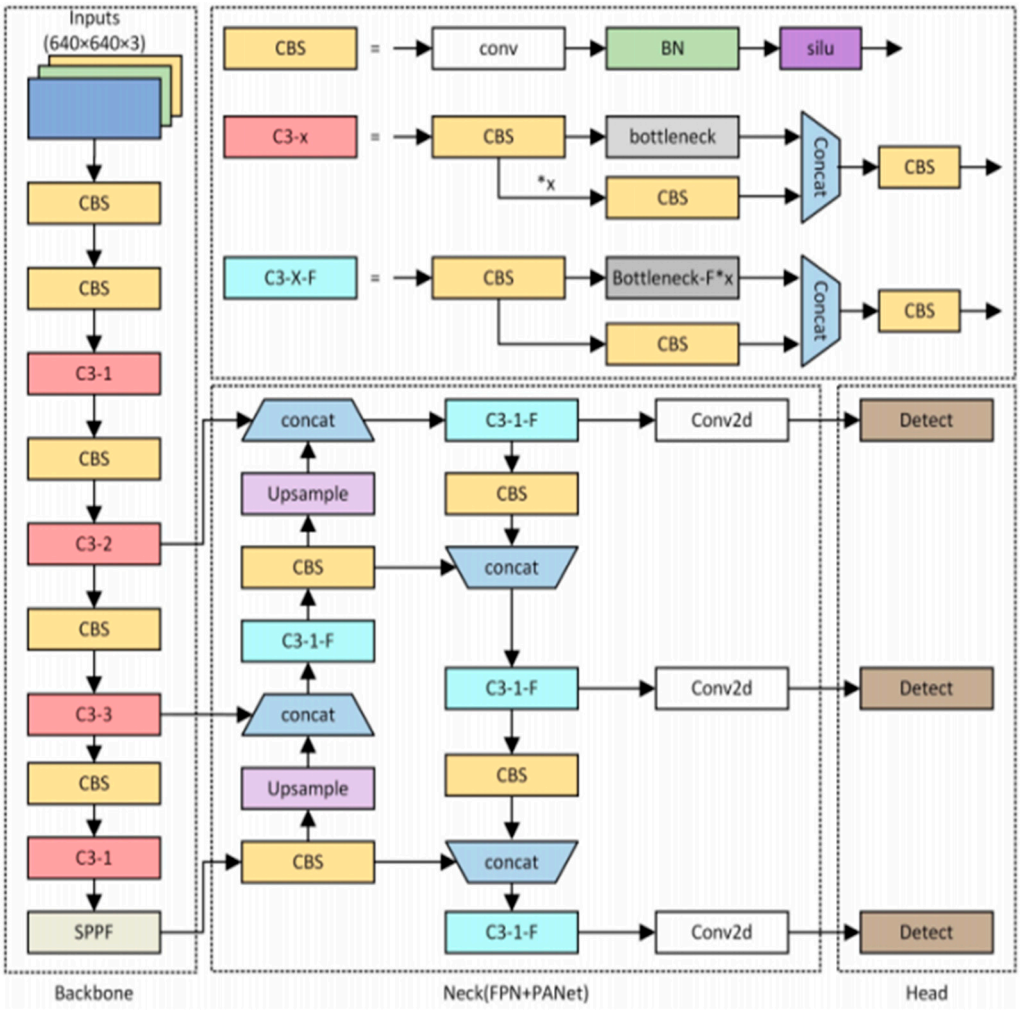

Figure 1. Original YOLOv5 architecture.

3.1. Data Augmentation

In the field of object detection, especially small object detection, data augmentation technology is particularly important. Effective data augmentation not only increases data diversity but also helps models better generalize to new and unprecedented data. This paper adopts a data augmentation method combining Mixup and Mosaic. On the one hand, using the Mixup enhancement strategy to linearly model samples and the parts between samples helps the model learn smoother decision boundaries. This method enables the model to better understand the training samples, thereby reducing the interference of noise and changes on small object detection and minimizing overfitting problems of the model. On the other hand, using Mosaic data augmentation to enhance the random cropping, rotation, and connection of any four images in the dataset increases scene complexity and the number of small targets in the sample, thereby improving the model's generalization ability and robustness in detecting small targets.

Mixup directly interpolates the two training samples linearly at the pixel level, while the labels corresponding to the synthesized images are likewise linear combinations of the original sample labels. Specifically, given two input samples and their corresponding labels, the new image is generated and labeled as in Equation (1), as follows:

$$\begin{cases} img\_c = \lambda img\_a + (1-\lambda) img\_b \\ label\_c = \lambda label\_a + (1-\lambda) label\_b \end{cases} \quad (1)$$

where λ is the interpolation coefficient satisfying the β distribution $\lambda \sim \beta(a, a)\alpha \in (0, \infty)$, derive $\lambda \in [0, 1]$.

Figure 2 shows the effect image of Mosaic and Mixup data augmentation. The specific steps are as follows. Firstly, randomly select one image as the background image in the dataset, then randomly select four images from the dataset for Mosaic data augmentation, and scale and concatenate the four images to form a composite image. Secondly, perform Mixup data augmentation on the background image and composite image in a certain proportion to generate a new image fused with five original data pixels. Finally, the new image and new labels are fed into the algorithm model for training. Since each image in the dataset contains small targets, $\lambda = 0.5$ is set to enable the network to detect each small target on the new image. After using composite data augmentation in the model, turn off the data augmentation strategy for the last 15 epochs [5].

Figure 2. Mixup and Mosaic data augmentation effect image.

3.2. Attention Module

The attention mechanism is a method that helps network models detect specific targets importantly; it has been widely studied and applied in many fields including target detection. For example, Squeeze-and-Excitation (SE) [26] and Convolutional Block Attention Module (CBAM) [27] both act as efficient attention mechanisms, which make them bring improvement to the performance of joined models. However, neither of them can bring good improvement for small target detection; the Squeeze-and-Excitation (SE) module only considers the encoding of inter-channel information and ignores the importance of location information, which is very important for small target detection. The Block Attention Module (CBAM) module, although it considers channel information and location information, uses large-scale pooling to utilize the location information to capture only local correlations, which is not able to solve the remote dependencies in vision tasks. Difficulty in localization has been one of the difficulties in small target detection and position information in the image is the key to detection. Therefore, we introduce the coordinate attention mechanism [28] into our model to improve the ability of capturing position information and improve the detection of small targets.

Coordinate attention (CA) is a lightweight, efficient, plug-and-play attention mechanism that embeds position information into the channel attention to enable the network

model to accurately localize to the information that is more critical to the current task, reduce the attention to other information, and improve the efficiency and accuracy of task processing.

Coordinate attention(CA) is divided into two steps: coordinate information embedding and coordinate attention generation, as shown in Figure 3.

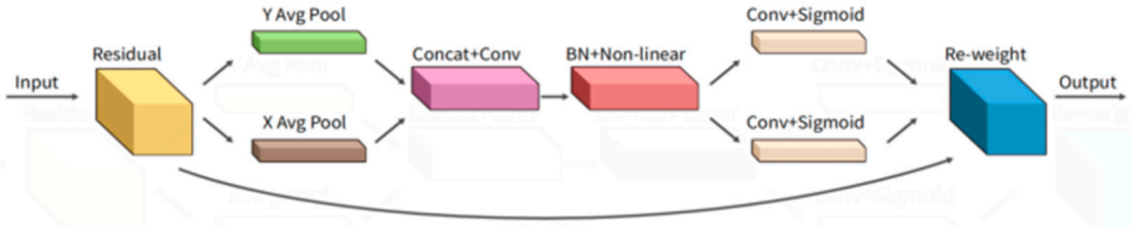

Figure 3. Coordinate attention.

Specifically, given an input X of size $C \times H \times W$, pooling kernels of size $(H, 1)$ and $(1, W)$ are used to encode information from different channels along the horizontal and vertical directions, respectively. For the feature of the c-th channel, the pooled output formula of the feature with height h is given as Equation (2), as follows:

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

Similarly, the output formula for feature pooling with width w is expressed as Equation (3), as follows:

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

These two transformations yield a pair of orientation-aware feature maps and implement coordinate information embedding. The next thing to do is to generate the attention weight matrix. The two coded features Z^h and Z^w are connected and transformed using the 1×1 convolutional transform function F_1 to perform the transform operation on them as in Equation (4):

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

where δ is a nonlinear activation function, $f \in R^{C/r \times (H+w)}$ is an intermediate feature map that encodes spatial information in the horizontal and vertical directions, and r is the downsampling rate.

f is decomposed into two independent tensors along the spatial dimension defined by $f^h \in R^{C/r \times H}$ and $f^w \in R^{C/r \times W}$, using two 1×1 convolution operations F_h and F_w to recover the number of channels of the tensor to C, and then processed using the Sigmoid activation function to obtain g^h and g^w. The matrix after multiplying and unfolding of the two is the weight matrix M and multiplying the input feature X by M yields the final output of the coordinate attention module Y.This is as shown in Equation (5), as follows:

$$\begin{gathered} g^h = \delta\left(F_h\left(f^h\right)\right), g^w = \delta(F_w(f^w)) \\ M_c(i, j) = g_c^h(i) \times g_c^w(j) \\ y_c(i, j) = x_c(i, j) \otimes M_c(i, j) \end{gathered} \tag{5}$$

In this article, coordinate attention (CA) is used to extract attention regions, which helps MCF-YOLOv5 resist confusing information and focus attention on small target objects, while also avoiding a lot of computational overhead to more accurately locate the exact position of the object of interest. The following experimental results also proved that the addition of this module indeed improved the accuracy of the model.

3.3. Construction of a Bidirectional Feature Fusion Network (FPN)

Feature fusion is a crucial step in small object detection as it combines information from different network layers to enhance the model's overall understanding of the image. In object detection tasks, deep features typically contain rich semantic information, which helps identify object categories in the image. Shallow features, on the other hand, are rich in high-resolution details and positional information, which helps to accurately locate the boundaries of objects. By fusing deep and shallow features, the model's understanding and abstraction ability of images can be improved, better expressing the complexity and diversity of images. However, many existing methods improve the performance of small object detection algorithms by transmitting deep features upwards to shallow layers. Although this one-way upstream feature fusion can enhance semantic information, it often overlooks the transmission of shallow detail features to deep layers, which limits further improvement in model performance. Therefore, this paper improves the original Feature Pyramid Network (FPN) structure and proposes a bidirectional feature fusion Feature Pyramid Network (FPN) structure, adding two cross-layer connections (B1, B2), as shown in Figure 4. Assuming the input image size is 640 × 640, the B1 layer fuses the 160 × 160 shallow feature map generated by the P1 layer with the 40 × 40 deep feature map generated by the P3 layer. This makes the B1 layer not only contain high-resolution details from shallow layers but also semantic information from deep layers; B2 integrates the 20 × 20 deep feature map generated by the P4 layer with the 80 × 80 feature map generated by P2. The B2 layer also ensures that the feature map retains details while also possessing rich semantic information. Through this bidirectional feature fusion, the newly generated feature map contains both detailed features and semantic information, greatly improving the network's ability to analyze complex scenes and effectively improving the model's performance in small object detection tasks.

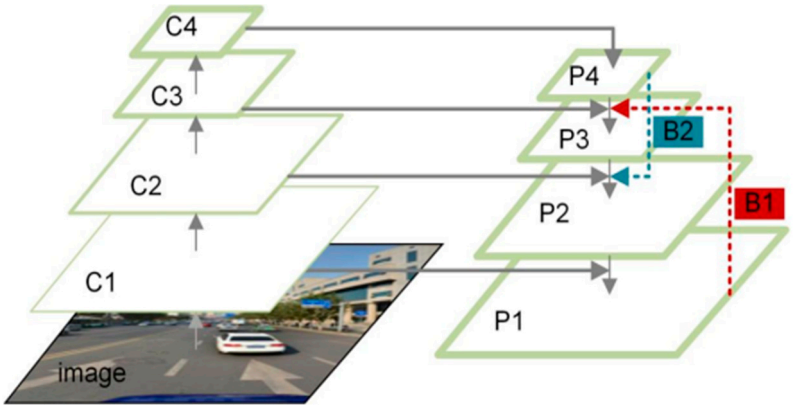

Figure 4. Improved neck structure.

Small targets will lose some feature information in the continuous downsampling process, resulting in unsatisfactory small target detection. The shallow layer contains more location and detail information and the shallow features have higher resolution, so whether the shallow features can be fully utilized is crucial for detecting small targets. After improving the Feature Pyramid Network (FPN) structure, we introduce the small target detection layer, as shown in Figure 5. This layer is more sensitive to small targets and after the deep layer and shallow layer features are fused and connected, the feature expression ability of small targets is enhanced, thus improving the detection accuracy of the model for small targets.

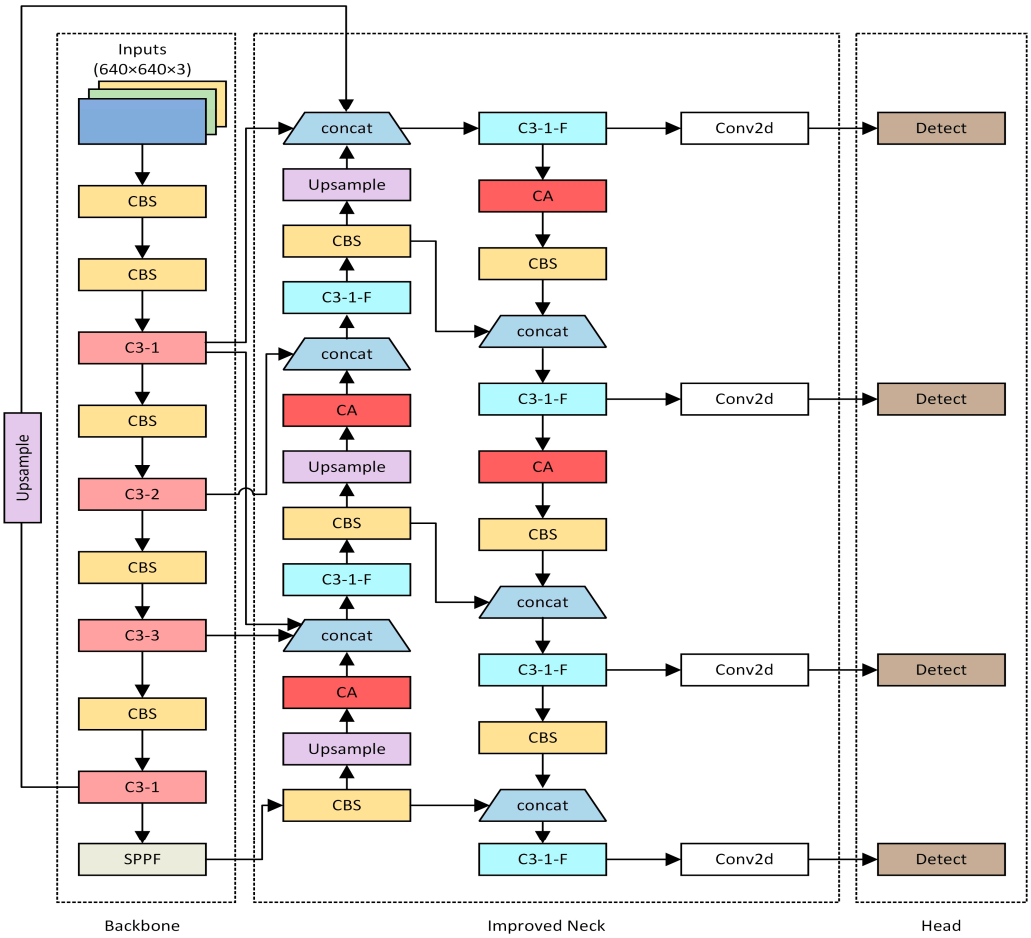

Figure 5. Improved YOLOv5 architecture.

4. Experiments

4.1. Datasets

In order to minimize the effect of sample imbalance, we choose to verify the performance of the proposed method on the VisDrone2021 dataset [29] and the Tsinghua-Tencent100K dataset [30], as shown in Figure 6. The figure depicts the distribution of large, medium, and small targets in the two datasets. Small targets have pixels smaller than 32×32, medium targets have pixels between 32×32 and 96×96, and large targets have pixels larger than 96×96.

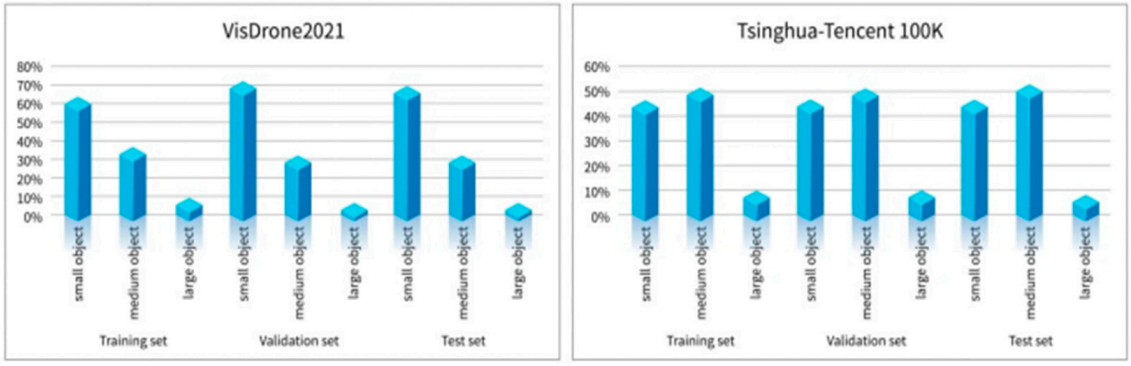

Figure 6. Visualization of target distribution for the VisDrone2021 and Tsinghua-Tencent 100k dataset.

The VisDrone2021 dataset is a dataset of unmanned aerial vehicle (UAV) aerial photography. This dataset defines 10 categories with rich scenarios, including 6471 training images,

548 validation images, and 3190 test images, in which the proportion of small targets is about 65%, which is a more suitable dataset for small target detection.VisDrone2021 is also a very challenging dataset, with uneven class distribution, many small targets, scarce large targets, small variance between some classes, serious category confusion, and some different degrees of occlusion and deformation in all types of targets.

The Tsinghua-Tencent100K dataset is produced by a joint lab between Tsinghua University and Tencent and this dataset contains a more comprehensive set of traffic sign categories, with 221 different categories appearing in the entire dataset. Most of the traffic signs in this dataset are very small and the lighting and weather conditions of the shooting locations are different, which provides stronger generalization ability in practical applications, so Tsinghua-Tencent100K is used as the dataset for training models. However, this dataset has serious category imbalance and some categories do not even appear in the training set. We select 45 categories from it that have more than 100 instances. There are 7260 sheets in the training set, 1908 sheets in the validation set, and 845 sheets in the test set, totaling 19,369 targets.

4.2. Training Setup

The hardware configuration of this experiment is NVIDIA RTX3080 GPU and Intel i7-12700 2.70 GHz CPU; meanwhile, the software environment is the Pytorch deep learning framework under the PyTorch v1.8.0 system on the Ubuntu 18.04.5 operating system. The input image size is 640 $\times$ 640, the weight descent coefficient is 0.0005, the initial learning rate is 0.01, and a total of 200 epochs are iterated with the StochasticGradientDescent (SGD) gradient descent optimizer. The one-cycle learning rate decay is used to ensure that the model can converge more stably in the later stages of training and other default settings are used.

4.3. Evaluation Indicators and Model Validity

Evaluation metrics are important measures of various aspects of a model's characteristics. Mean average precision (mAP), precision (P), recall (R), and average precision (AP) are four important metrics for assessing a model's performance with respect to the following questions: is the probability of an actual positive sample among all samples predicted to be positive, defined in Equation (6); is the probability of an actual positive sample among all samples predicted to be positive, defined in Equation (7); is the average of Precision Recall (PR) curves at different Recall values obtained for a given threshold, defined in Equation (8); and is the mean value of the of all categories in the whole dataset, defined in Equation (9).

$$\text{Precision(P)} = \frac{TP}{TP + FP} \tag{6}$$

$$\text{Recall}(R) = \frac{TP}{TP + FN} \tag{7}$$

where TP denotes the number of positive samples that the model is able to correctly classify as positive. TP denotes the number of negative samples that the knot model incorrectly identifies as positive. FN denotes the number of positive samples that the type incorrectly identifies as negative samples.

$$\text{Average Precision(AP)} = \int_0^1 P(R)dR \tag{8}$$

$$\text{mean Average Precision(mAP)} = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{9}$$

where N is the number of classes and AP_i is the AP of class i.

To evaluate the effectiveness of the proposed method, we conducted experiments on the VisDrone2021 and Tsinghua Tencent100K datasets. The experimental results of this model on the VisDrone2021 dataset are shown in Table 1. Compared with the model

performance in the baseline, our proposed model achieved a 5.2% improvement in mAP_{50} values and improved detection average precision for small, medium, and large targets by 3.3%, 5.6%, and 4%, respectively. Figure 7 shows the detection results of MCF-YOLOv5 on some samples on the VisDrone2021 dataset., Visualization refers to the output results of a model after processing a dataset. It can be seen that compared to the baseline, our model can detect more small targets with higher accuracy. Figure 8 shows the confusion matrix of the improved model and baseline model on the VisDrone2021 dataset. Compared with the baseline model, the confusion rate of the improved model decreased and the classification accuracy of each category significantly improved.

Table 1. Performance evaluation of MCF-YOLOv5 on the VisDrone2021 dataset.

Model	mAP	mAP_{50}	mAP_{75}	AP_{small}	AP_{medium}	AP_{large}
YOLOv5s	18.4	33.1	15.6	10.5	26.5	36.1
MCF-YOLOv5	22.1	38.3	21.5	13.8	32.1	40.1

YOLOv5s Ours

Figure 7. Visualization of the detection results for YOLOv5s and MCF-YOLOv5 on the VisDrone2021 dataset.

Table 2 presents the experimental results of the model on the Tsinghua Tencent100K dataset, where the improved model increased the mAP_{50} value by 5.7% and the average detection precision of small, medium, and large targets increased by 3.6%, 2.5%, and 3.3% compared to the baseline model, respectively. Figure 9 shows the mean average precision, curves of the improved model, and the baseline model. It can be seen that in the first 60 epochs, the mean average precision (mAP) values of the two curves increase at a similar rate but the improved model has higher detection accuracy. After the 60th epoch, the two models begin to converge and the curves gradually become smoother. Figure 10 shows some detection results on the dataset and it can be seen that MCF-YOLOv5 has higher detection accuracy and can also detect targets missed from the baseline.

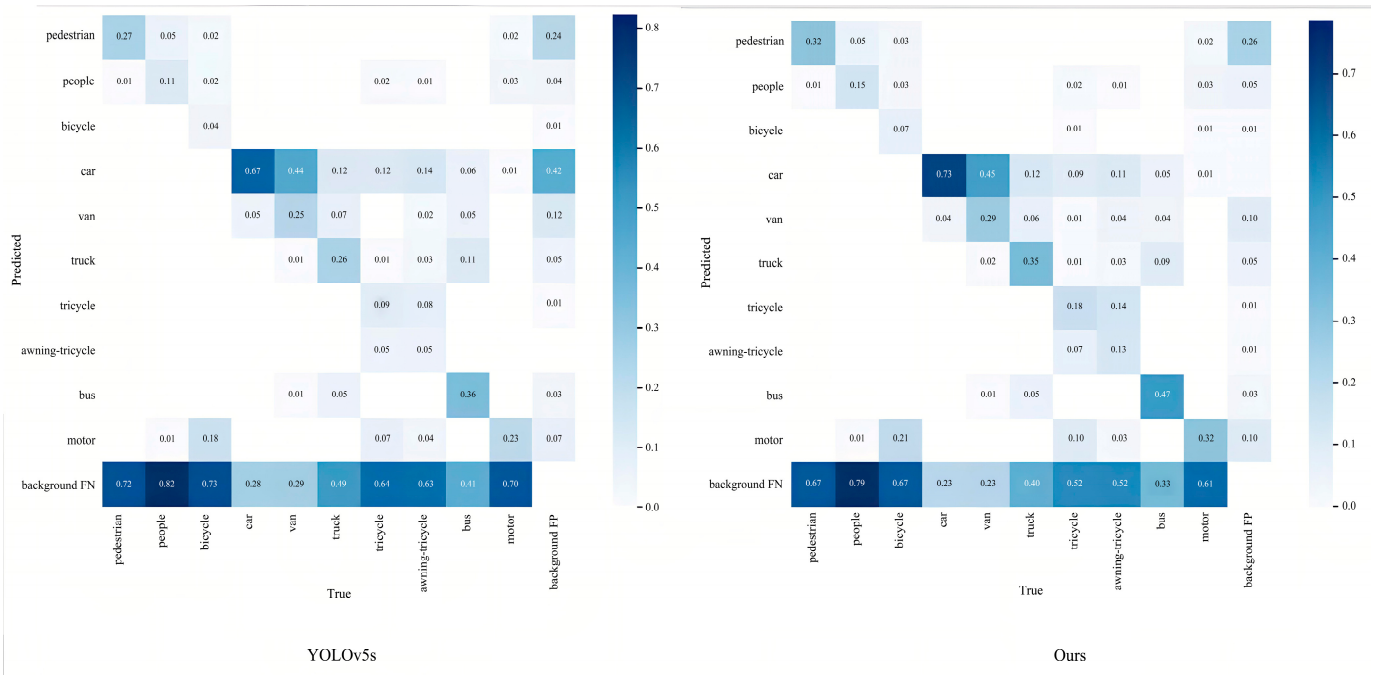

Figure 8. Confusion matrix of YOLOv5s and MCF-YOLOv5.

Table 2. Performance evaluation of MCF-YOLOv5 on the TT100K dataset.

Model	mAP	mAP_{50}	mAP_{75}	AP_{small}	AP_{medium}	AP_{large}
YOLOv5s	62.5	79.4	60.4	48.2	69.7	76.5
MCF-YOLOv5	63.8	85.1	62.5	51.8	72.2	79.8

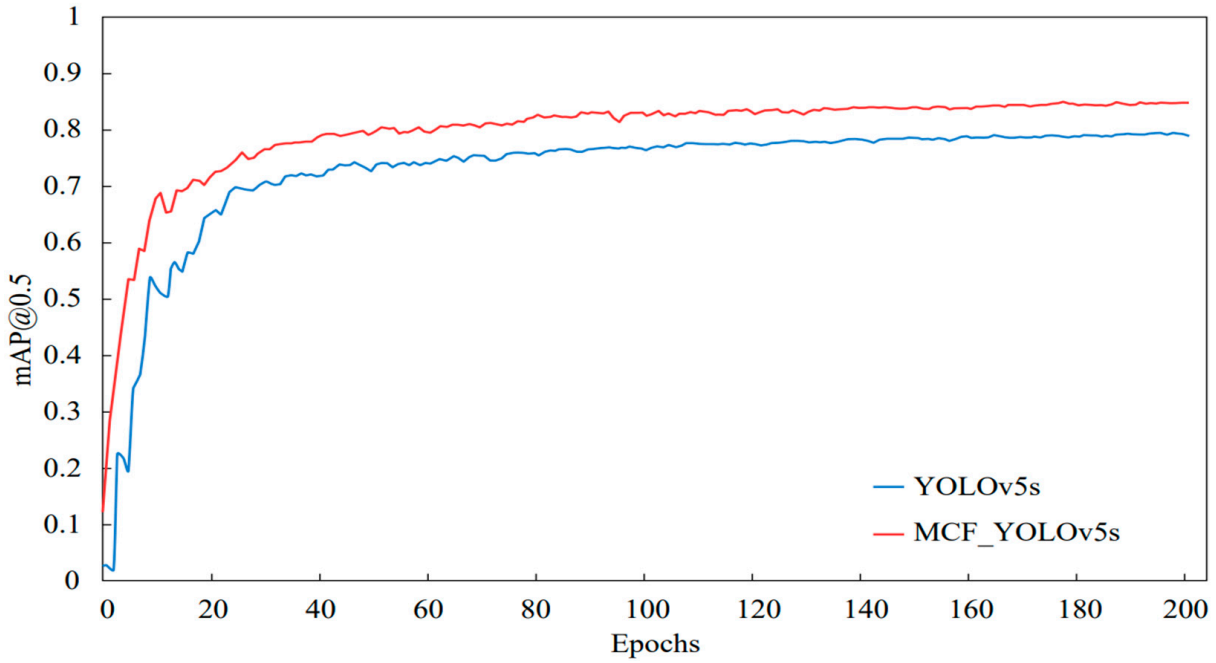

Figure 9. Mean average accuracy curve of YOLOv5s and MCF-YOLOv5.

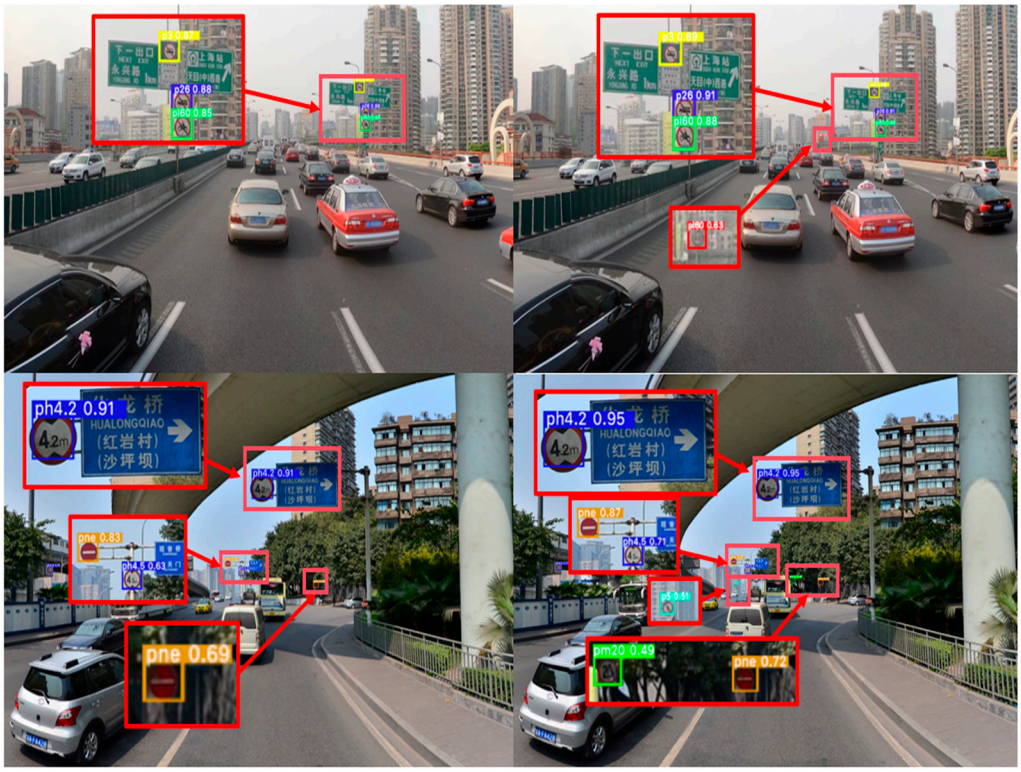

Figure 10. Visualization of YOLOv5s and MCF-YOLOv5 detection results on the Tsinghua Tencent100K dataset.

4.4. Ablation Study of the Proposed Model

We analyzed the contributions of different components in the MCF-YOLOv5 model and conducted ablation experiments on the VisDrone2021 dataset. The application of Mixup+Mosaic, coordinate attention (CA), and small object detection layer (SODL) improved the precision of small object detection by 0.4%, 1.3%, and 2.1%, respectively. Table 3 lists the effects of each component. The results of ablation research indicate that by adding a combination enhancement strategy to the basic network of YOLOv5, the mAP_{50} of the model can reach 33.4% and the AP_{small} can reach 10.9%. After adding the coordinate attention module, the value of mAP_{50} increased from 33.4% to 34.5% and AP_{small} can reach 11.7%. This improvement also reflects that the positioning ability of the coordinate attention module has improved the performance of small object detection. Finally, after adding the small object detection layer again, the value of mAP_{50} increased from 34.5% to 38.3% and the value of AP_{small} increased from 11.7% to 13.8%. This indicates that by improving the FPN structure and introducing a small object detection layer, the model can learn more small object features, greatly improving the detection accuracy of the model for small objects.

Table 3. Ablation study of different components in our MCF-YOLOv5 model.

YOLOv5	Mixup+Mosaic	CA	SODL	mAP	mAP_{50}	mAP_{75}	AP_{small}	AP_{medium}	AP_{large}
✓				18.4	33.1	15.6	10.5	26.5	36.1
✓	✓			18.9	33.4	15.7	10.9	26.5	36.4
✓		✓		19.5	33.8	16.5	11.7	27.2	37.1
✓			✓	20.3	34.7	18.1	12.6	28.6	37.8
✓	✓	✓		19.9	34.5	17.2	12.1	29.5	38.3

Table 3. *Cont.*

YOLOv5	Mixup+Mosaic	CA	SODL	mAP	mAP_{50}	mAP_{75}	AP_{small}	AP_{medium}	AP_{large}
✓	✓		✓	20.8	36.0	19.5	12.9	30.4	38.7
✓		✓	✓	21.7	37.9	20.8	13.3	31.2	39.3
✓	✓	✓	✓	22.1	38.3	21.5	13.8	32.1	40.1

4.5. Comparison with Other Detection Models

To verify the superiority of the proposed MCF-YOLOv5 model, we compared it with other advanced models on the VisDrone2021 dataset, such as YOLOX, YOLOv7s, and TPH-YOLOv5. Table 4 presents comparative data on the precision, recall, mean average precision, and detection precision of large and small targets for different models under the same settings. It is evident that our MCF-YOLOv5 model is significantly superior to the benchmark model in all aspects, as well as some advanced models. This indicates that our model has achieved satisfactory results. Table 5 presents the comparison data of the improved model, baseline model, YOLOv5m parameter count, floating point operations (FLOPs), and frames per second (FPS). It can be seen that with a small increase in computational cost, better detection performance was achieved than YOLOv5m.

Table 4. Performance comparison of advanced detection models.

Model	P	R	mAP	mAP_{50}	mAP_{75}	AP_{small}	AP_{large}
SSD	38.7	30.3	17.6	27.9	11.4	9.5	33.4
YOLOv5s	40.5	33.2	18.4	33.1	15.6	10.7	36.1
YOLOX [5]	41.9	34.5	19.0	34.8	16.5	10.9	37.8
YOLOv7s [16]	43.5	35.4	19.5	36.4	17.3	-	-
TPH-YOLOv5 [31]	44.7	36.6	21.4	37.6	19.7	12.9	-
YOLOv5m	44.5	36.3	20.9	36.9	19.3	12.3	41.5
MCF-YOLOv5(ours)	45.2	37.0	22.1	38.3	21.1	13.8	40.1

Table 5. Comparison of parameter quantity, FOLPs, and FPS for YOLOv5s, YOLOv5m, and MCF-YOLOv5 at the same input size.

Model	Input Size	Params (M)	FLOPs (G)	FPS
YOLOv5s	640 × 640	7.2	16.21	95.2
MCF-Yolov5(ours)	640 × 640	10.31	25.52	88.5
YOLOv5-M	640 × 640	21.2	48.7	67.1

5. Conclusions

In the field of object detection, how to quickly and accurately detect small targets in images has always been a major challenge. To address this issue, this paper proposes an improved MCF-YOLOv5 model based on the YOLOv5 algorithm. In terms of data augmentation, we use a combination data augmentation method to increase the specific sample size while reducing the risk of overfitting in the model, suppressing noise interference in detection, and improving the model's generalization ability. In terms of network structure, by improving the Feature Pyramid Network (FPN) part and adding a small object detection layer, the model obtains more feature information, solving the problem of insufficient small object features or loss of some feature information during continuous downsampling. Introducing a coordinate attention (CA) module into the neck network to focus more on areas of interest reduces the impact of irrelevant information on detection. The MCF-YOLOv5 model has shown excellent detection performance on the Tsinghua-Tenent100K dataset and the VisDrone2019 dataset, outperforming the baseline model with minimal computational cost. The MCF-YOLOv5 model is suitable for detecting long-distance small target scenarios such as traffic signs and drone aerial photography and it provides a new solution for small

target detection. In the future, our work will be able to make the model more lightweight while ensuring its accuracy.

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

Funding: This study was not funded.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Please contact author for data requests.

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

References

1. Tong, K.; Wu, Y.; Zhou, F. Recent advances in small object detection based on deep learning: A review. *Image Vis. Comput.* **2020**, *97*, 103910. [CrossRef]
2. Zhao, Z.Q.; Zheng, P.; Xu, S.T.; Wu, X. Object detection with deep learning: A review. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *30*, 3212–3232. [CrossRef] [PubMed]
3. Zaidi, S.S.A.; Ansari, M.S.; Aslam, A.; Kanwal, N.; Asghar, M.; Lee, B. A survey of modern deep learning based object detection models. *Digit. Signal Process.* **2022**, *126*, 103514. [CrossRef]
4. Zhang, H.; Cisse, M.; Dauphin, Y.N.; Lopez-Paz, D. mixup: Beyond empirical risk minimization. *arXiv* **2017**, arXiv:1710.09412. [CrossRef]
5. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. Yolox: Exceeding yolo series in 2021. *arXiv* **2021**, arXiv:2107.08430. [CrossRef]
6. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
7. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587. [CrossRef]
8. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 1137–1149. [CrossRef] [PubMed]
9. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2980–2988. [CrossRef]
10. Redmon, J.; Divvala, S.; Girshick, R.; 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. [CrossRef]
11. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6517–6525. [CrossRef]
12. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
13. 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; pp. 21–37. [CrossRef]
14. Lin, T.Y.; Dollár, 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; pp. 936–944. [CrossRef]
15. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934. [CrossRef]
16. 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 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 7464–7475. [CrossRef]
17. Zhang, J.; Meng, Y.; Chen, Z. A small target detection method based on deep learning with considerate feature and effectively expanded sample size. *IEEE Access* **2021**, *9*, 96559–96572. [CrossRef]
18. Kisantal, M.; Wojna, Z.; Murawski, J.; Naruniec, J.; Cho, K. Augmentation for small object detection. *arXiv* **2019**, arXiv:1902.07296. [CrossRef]
19. Chen, C.; Liu, M.Y.; Tuzel, O.; Xiao, J. 2017. 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; pp. 214–230. [CrossRef]

20. Romano, Y.; Isidoro, J.; Milanfar, P. RAISR: Rapid and accurate image super resolution. *IEEE Trans. Comput. Imaging* **2016**, *3*, 110–125. [CrossRef]
21. Zhang, Y.; Bai, Y.; Ding, M.; Ghanem, B. Multi-task generative adversarial network for detecting small objects in the wild. *Int. J. Comput. Vis.* **2020**, *128*, 1810–1828. [CrossRef]
22. Li, J.; Liang, X.; Wei, Y.; Xu, T.; Feng, J.; Yan, S. Perceptual Generative Adversarial Networks for Small Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1951–1959. [CrossRef]
23. Pang, Y.; Cao, J.; Wang, J.; Han, J. JCS-Net: Joint classification and super-resolution network for small-scale pedestrian detection in surveillance images. *IEEE Trans. Inf. Forensics Secur.* **2019**, *14*, 3322–3331. [CrossRef]
24. Olorunshola, O.E.; Irhebhude, M.E.; Evwiekpaefe, A.E. A comparative study of YOLOv5 and YOLOv7 object detection algorithms. *J. Comput. Soc. Inform.* **2023**, *2*, 1–12. [CrossRef]
25. Wang, M.; Yang, W.; Wang, L.; Chen, D.; Wei, F.; KeZiErBieKe, H.; Liao, Y. FE-YOLOv5: Feature enhancement network based on YOLOv5 for small object detection. *J. Vis. Commun. Image Represent.* **2023**, *90*, 103752. [CrossRef]
26. 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. [CrossRef]
27. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. 2018. CBAM: Convolutional Block Attention Module. In Proceedings of the Computer Vision—ECCV 2018: 15th European Conference, Munich, Germany, 8–14 September 2018; pp. 3–19. [CrossRef]
28. 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. [CrossRef]
29. 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 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Montreal, BC, Canada, 11–17 October 2021; pp. 2847–2854. [CrossRef]
30. Zhu, Z.; Liang, D.; Zhang, S.; Huang, X.; Li, B.; Hu, S. Traffic-Sign Detection and Classification in the Wild. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2110–2118. [CrossRef]
31. 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 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Montreal, BC, Canada, 11–17 October 2021; pp. 2778–2788. [CrossRef]