A Small-Scale Object Detection Algorithm in Intelligent Transportation Scenarios

Junzi Song , Chunyan Han * , Chenni Wu

Posted Date: 3 September 2024

doi: 10.20944/preprints202409.0224.v1

Keywords: intelligent transportation; small object detection; YOLOv4 tiny; feature pyramid; information entropy

Article

A Small-Scale Object Detection Algorithm in Intelligent Transportation Scenarios

Junzi Song, Chunyan Han * and Chenni Wu

School of Software, Northeastern University (NEU), Shenyang 110169, China
* Correspondence: hancy@swc.neu.edu.cn

Abstract: In response to the problem of poor detection ability of object detection models for small-scale targets in intelligent transportation scenarios, a fusion method is proposed to enhance the features of small-scale targets, starting from feature utilization and fusion methods. The algorithm is based on the YOLOv4 tiny framework and enhances the utilization of shallow and mid level features on the basis of FPN, improving the detection accuracy of small and medium-sized targets; In view of the problem that the background of the intelligent traffic scene image is cluttered and there is more redundant information, the CBAM attention module is used to improve the attention of the model to the traffic target; To address the problem of data imbalance and prior bounding box adaptation in custom traffic datasets that expand traffic images in COCO and VOC, we propose a Copy Paste method with improved generation method and a K-means algorithm with improved distance measurement to enhance the model's detection ability for corresponding categories. Comparative experiments were conducted on a customized 260 thousand traffic dataset containing public traffic images, and the results showed that compared to YOLOv4 tiny, the proposed algorithm improved mAP by 4.9% while still ensuring the real-time performance of the model.

Keywords: intelligent transportation; small object detection; YOLOv4 tiny; feature pyramid; information entropy

1. Introduction

For object detection models applied to intelligent transportation, not only should accuracy be emphasized, but also the speed of model detection, requiring a balance between accuracy and speed. In the intelligent traffic scenario, vehicles, pedestrians and other targets tend to have smaller scales. Especially when the vehicle is traveling too fast, if these targets cannot be detected in time and accurately, it will have a serious impact on the accurate operation of the subsequent intelligent traffic system. In recent years, although the overall detection performance of object detection has been greatly improved, the research progress of small object detection is relatively slow, and models in intelligent transportation scenarios require real-time performance. Therefore, further exploration is still needed for small object detection methods in intelligent transportation.

The limitation of small object detection capability is partly due to the imbalance in target scale in training data, and partly due to the limitations of the detection network itself [1]. For most datasets, medium to large scale targets account for the majority, while small scale targets only account for a small proportion. For the model, good detection of medium and large scale targets will bring more gains, so the detection of small scale targets will be ignored. For the structural part of the model itself, in order to obtain more deep-seated semantic information, most detection networks use more convolutional and pooling layers to stack, and multi-layer stacking will cause the information of small targets to gradually disappear as the network layer propagates [2], resulting in the inability to detect small targets well. The FPN [3] proposed by Lin T Y et al. and the PAN [4] used in YOLOv4 alleviate the problem of information loss to some extent by fusing shallow and deep feature maps. However, their utilization and fusion of shallow and deep information, as well as their complexity, still need further improvement. On the basis of FPN and PAN, a group of feature utilization and

 doi:10.20944/preprints202409.0224.v1

fusion methods with more complex structures have emerged. The common problem is that improving accuracy increases model complexity, which can affect the running speed of the model.

Based on the above analysis, considering that the object detection model in intelligent transportation scenes needs to ensure real-time performance, the main methods to solve the problem of small object detection include data processing and multi-scale feature fusion [5]. This article mainly improves the data processing and detection model structure to improve the object detection effect in intelligent transportation scenes. In terms of model structure, for the detection of small-scale targets, the feature fusion method is enhanced on the basis of FPN to enhance the model's detection ability for small-scale targets. The attention mechanism CBAM module is also used to further enhance detection accuracy, while ensuring that the model still has real-time performance after the above improvements. In terms of data processing, to address the problem of imbalanced datasets with small samples and targets, an improved Copy Paste method is used for corresponding feature enhancement, effectively enhancing the model's detection ability for these targets; Subsequently, in response to the adaptation problem between the model's prior bounding boxes and the traffic dataset, an improved K-means algorithm was used for prior bounding box clustering to obtain prior bounding boxes that fit the custom traffic dataset and improve the model's detection accuracy for each category.

Finally, we designed a series of experiments to prove our conclusion using a customized 300000 traffic dataset as the training and testing set. The improved model based on PF Net feature fusion structure proposed in this article has increased mAP by 2.01%; After adding three CBAM modules, the mAP of the model increased by 4.03%. For small targets of concern, taking the reflector cone as an example, the final improved model PF-YOLOv4 tiny CBAM can increase by 1.69 percentage points. After using the improved Copy paste data augmentation method for small-scale targets, the detection accuracy has improved by at least 1%; On the basis of the above, K-means was used for prior bounding box clustering, which improved the detection accuracy of some categories by 3%.

In summary, our main contributions are:

- We propose an improved feature fusion structure PF Net based on FPN, which ensures real-time performance while further improving accuracy.
- An improved model PF-YOLOv4 tiny CBAM with added CBAM attention module was proposed, which makes the model pay more attention to the targets in the image, further improves the detection accuracy of the model, and ensures that the improved model can meet the real-time requirements of intelligent transportation scenarios.
- A data augmentation method based on Copy Paste improvement has been proposed to enhance the detection ability of small targets in custom traffic datasets.
- A K-means method for improving distance measurement was proposed, which was applied to custom traffic datasets to obtain more suitable prior bounding box and further improve the detection performance of targets.

2. Related Work

2.1. Object Detection Model Based on Deep Learning

With the development of deep learning, a large number of object detection models based on deep learning have been proposed. Object detection based on deep learning mainly obtains the object category and object position in the image through convolutional network. Generally, it can be divided into one-stage model and two-stage model. The first model that uses deep learning to solve the object detection problem is R-CNN (region-based Convolutional Neural Network)[6], which firstly obtains the candidate Region through the traditional extraction method, and then obtains the object category and location through the convolutional network. Compared with other traditional motion modeling object detection methods and machine learning object detection methods, it has a great breakthrough and great performance improvement. After that, SPP-Net[7] continued to use the convolutional structure of deep learning and innovatively proposed the space pyramid pool. Later, Ross Girshick proposed the Fast R-CNN network [8] and introduced ROI Pooling of areas of interest. The ROI

 doi:10.20944/preprints202409.0224.v1

Pooling optimized the problem of R-CNN repeatedly extracting multiple candidate areas in an image and improved the efficiency of image processing. On the basis of Fast R-CNN network, Faster R-CNN[9] is proposed, which uses RPN structure to generate candidate regions and eliminates the traditional selective search, thus improving the execution efficiency. At the same time, the concept of anchor box was proposed for the first time in Faster R-CNN to realize end-to-end object detection and further improve the precision of small object detection. At the same time, Faster R-CNN has a great improvement in speed compared with the previous models. All of the above models belong to the two-stage object detection model. The two-stage model first extracts the candidate region, and then uses the convolutional network to obtain the approximate position and final position of the target and other relevant information respectively.

In the field of intelligent transportation, although the two-stage model has high accuracy, its running speed is limited and it cannot meet the requirements of real-time performance. Therefore, the one-stage object detection model without additional extraction of candidate regions is proposed, which only requires a convolutional network to obtain the location and category of the object. The representative models of single - stage object detection model include YOLO series model and SSD, etc. Based on the YOLO model, Liu W et al. proposed the SSD[10] model. The SSD model follows the concept of anchor box. CNN structure is used for direct detection during detection, and multi-scale feature map is not only used for the feature map of the last convolutional layer. Multiscale detection of SSD improves the detection ability of the model for small targets. Subsequently, YOLOv3[11], a very representative model in the YOLO series, adopted the structure of feature pyramid and three detection heads, corresponding to the detection of large, medium and small targets respectively, significantly improving the detection ability of the model for small targets. In 2020, the paper of YOLOv4 was published. YOLOv4 combined a variety of new techniques in the field of deep learning, especially proposed Mosaic, self-antagonistic training data enhancement method and innovative feature fusion structure of PAN. These data enhancement methods amplified the target features, and PAN structure further improved the feature fusion method. The shallow and deep information fusion is strengthened to improve the detection effect of targets at various scales. YOLOv4 papers also published a variety of models of different complexity, such as YOLOv4 tiny. The structure of YOLOv4 tiny is more simplification, and the relatively simple FPN structure is used to facilitate the selection of appropriate models according to the required precision and real-time requirements. The strategies and techniques used in the later YOLOv5 and YOLOv4 are roughly the same, more engineering optimization logic is introduced, and a variety of lightweight models with different complexity are designed, such as YOLOv5s.The above model is anchor based object detection model . With its development, anchor free object detection models have been proposed continuously in recent years. For example, CornerNet[12] and FSAF[13], which are the first work of anchor free, provide new ideas for object detection while avoiding some disadvantages brought by anchor mechanism. It is another branch of object detection one-stage model.

The performance of the above representative one-stage and two-stage object detection models is comprehensively compared. Currently, the one-stage model mentioned above is still the main model used for object detection in the intelligent transportation scene. At the same time, although the object detection model has been constantly developed and replaced, and even new detection ideas have appeared, there has been no great breakthrough in small object detection. This paper takes YOLOv4 tiny, which is widely used in industry, as the benchmark model. YOLOv4 tiny is a lightweight model, which can fully guarantee real-time performance, so it is improved based on it.

2.2. Data-Based

Data-based methods solve problems from the Data set itself, and such methods tend to be effective only for specific data sets, such as COCO. There are two imbalance problems in COCO data sets: image-level and instance-level imbalance. image-level imbalance means that only 51.8% of pictures contain small objects in COCO. The representative method to solve this problem is Stitcher. Stitcher[14] takes the proportion of small object loss in the total loss as the feedback signal. When the proportion is less than a certain threshold, the four pictures will be combined into one picture as the

 doi:10.20944/preprints202409.0224.v1

input of the next iteration, which is equivalent to increasing the number of small objects. instance-level imbalance means that the pixel area of small and medium-sized objects occupies 1% in COCO, so data enhancement is usually adopted. The data enhancement method can specifically improve the number of features of the specified target, not only enhance the generalization of the model, but also balance the data set, so as to improve the detection ability of the corresponding target, and does not affect the real-time application of the model.

One of the reasons why small targets are difficult to detect is the unbalance of large and small target samples. Generally, medium and large scale targets occupy a large proportion in the public dataset of general scenes or the data sets of intelligent transportation scenes, which leads to more attention to the detection ability of medium and large scale targets in the process of model learning. At the same time, when the artificially set prior bounding box size is significantly different from the real bounding box, it will lead to fewer positive training samples for small targets and more unbalanced anchors matching medium and large targets, thus making the model ignore the detection of small targets. Therefore, data enhancement is widely used in object detection models to improve the detection ability of small targets. YOLOv4 proposed Mosaic data enhancement model, which used four images to splice a new image sample, enriched the background of the detected target, transformed the large target of the original image into small target, expanded the number of small targets in the data set, and improved the detection ability of the model on small targets. Kisantal M et al. [15] raised oversampling of small target training samples, copy and paste small targets in the COCO sample of the public data set, use this method to provide enough of the small target and anchor matching, in turn, promote small object detection ability, It is proved that the Copy-Paste method is effective in improving small object detection. Reference [16] has proved the effectiveness of Copy-Paste in instance segmentation and the universal validity of the data enhancement method. However, Copy-Paste in literature [15] copies and pastes back the original image to achieve small target data enhancement, which will cause imbalance of data set to a certain extent. Because the procedure only increases the number of small targets, it does not increase the number of images containing small targets.

Our method uses an improved Copy-Paste idea, which increases the number of small targets and the number of pictures containing small targets at the same time, thus improving the detection ability of the model on small targets. As described in the previous data enhancement section, copying a small target to multiple positions in the picture can increase the number of anchor matched by small targets, increase the training weight of small targets, and reduce the bias of the network to large targets. Similarly, in reverse thinking, if the data set has been determined, the setting strategy of anchor responsible for small targets can also be added to make the learning of small targets more adequate in the training process. For example, in FaceBoxes[17], one of the contributions is the anchor densification strategy, which enables different types of anchors to have the same density on the image, significantly improving the recall rate of small-scale faces. S3FD[18] reduces the threshold of IoU for positive samples of small objects, and relieves the threshold of IoU to 0.1 for a small number of positive samples. Dot Distance[19] designed a new index DotD to replace IoU for the allocation standard of positive and negative samples. Simply put, IoU was replaced by the distance function of two bounding box centers. Therefore, in the data set after data enhancement, consideration about anchor is further added in this paper. The K-means algorithm is used to conduct scale clustering on the marked targets in the data set to obtain the anchor proportion suitable for the data set, so as to improve the adaptation ability of the model's prior bounding box to small targets during model training and ensure more adequate model learning.

2.3. Multi-Scale Feature Fusion

Another effective way to improve model detection capabilities is to make full use of the multi-scale features generated during the model of model convolution. This method is not only based on the principle of intuitive image pyramids, but also the key concepts in information theory, such as information entropy and cross-entropy, and through reasonable characteristic coding and decoding processes to improve the accuracy and robustness of target detection.

 doi:10.20944/preprints202409.0224.v1

In information theory, information entropy is an important indicator of the amount of information, indicating the size of random variable uncertainty. In target detection, the amount of information of different scale features contains different amounts of information. The shallow layers of features usually include more details (such as edges and textures), while deep features pay more attention to semantic information (such as object categories). Through the calculation of information entropy, we can quantify the importance of the characteristics of each scale, thereby optimizing the feature selection strategy to improve the effectiveness and efficiency of the model[20].

Cross entropy is used as a common loss function in machine learning, and the difference between the real distribution and the prediction distribution is measured. In the target detection model, by designing a cross-entropy-based loss function, it can effectively train the model to make it output closer to the real target position and category. In addition, for the combination of multi-scale features, the cross-level entropy loss can be used to calculate the loss for the characteristic diagrams of each scale, and then weighted the summary to guide the model more accurately to learn and use multi-scale features[21].

The characteristic fusion process can be regarded as a process of information coding and decoding. The encoder (convolutional layer) is responsible for extracting multi-level feature information from the input image, and this information encodes in a specific way (such as the form of the feature diagram). The decoder (usually a sampling or feature fusion layer) is responsible for decoding these encoded feature information back to the space domain to form the final detection results. In the multi-scale feature fusion, the encoder is responsible for extracting the characteristics of multi-scale, while the decoder effectively integrates these features through specific fusion strategies (such as FPN, PAN[22]) to improve the model's detection capacity of the model.

SSD is the first object detection model that attempts to use multi-scale feature maps for prediction. It performs detection on 6 feature maps of different scale sizes, which improves the detection ability of the model for multi-scale targets including small targets. FPN proposed the concept of feature pyramid for the first time, and fused the feature information of different downsampling rates to improve the feature expression ability. On the basis of FPN, YOLOv4 improved the structure and designed PAN. PAN added a branch on FPN to obtain better detection effect. In addition, there are many other kinds of feature utilization and fusion methods in the object detection model. For example, the simple bidirectional fusion represents the BiFPN[23], and the BiFPN is an improvement on the PAN bidirectional foundation. Multiple BiFPN can be used in series, and the BiFPN will increase the amount of computation compared with PAN. In addition to simple bidirectional fusion, there are more complex bidirectional fusion, such as ASFF[24] and Recursive-FPN[25]. Based on FPN, ASFF studies the effects of each stage and further integrates the effects of the three stage features. The integration of different stage features uses the attention mechanism, so that the contribution of other stages to the stage features can be controlled. Recursive-FPN refers to the output of the fusion of traditional FPNS, which is then input to the Backbone for a secondary cycle. To sum up, the above methods are to carry out repeated flow and fusion of characteristic information in different directions to achieve the purpose of improving accuracy. However, such methods will greatly improve the complexity of the model. The above feature fusion methods with better accuracy are often difficult to ensure the real-time performance of the model, so they cannot be easily used in intelligent transportation scenes. In our method, instead of blindly improving the complexity of feature fusion structure to improve the model accuracy, we combined FPN and PAN structure to design a real-time PF-Net, which can improve the object detection accuracy, especially the detection accuracy of small targets.

3. Methods

In this paper, we focus on improving the detection capability of the object detection model for small-scale targets in intelligent transportation scenarios. Our model is improved based on the YOLOv4 tiny structure, which is mainly divided into structural optimization and data-based processing. In terms of data, data enhancement is adopted. The improved Copy-Paste data enhancement method is mainly used to increase the number of samples and features of small targets.

 doi:10.20944/preprints202409.0224.v1

At the same time, in order to make the prior bounding box better match with the small target in the custom data set, K-means clustering method with improved distance measure is used to get a more appropriate scale of the prior bounding box. In terms of structure, the original feature pyramid of YOLOv4 tiny was changed, and a new feature utilization and fusion mode between FPN and PAN was proposed, which was named PF-Net. While ensuring real-time performance, the model's detection ability for multi-scale targets, especially small targets, was improved. Finally, CBAM attention module is added into the network to improve the model's attention to the target in the image and further improve the detection ability of small target. In the next few sections, we'll cover the structure of the model (Section 3.1) and data-based processing (Section 3.2).

3.1. Model Structure

This section mainly introduces the improved multi-scale feature fusion of PF-Net based on FPN and the use of attention module CBAM.

3.1.1. Model Multi-Scale Feature Fusion Method

In order to control the complexity of the model to ensure real-time performance and improve the detection accuracy of the model, especially the detection accuracy of small targets, the PF-YOLOv4 tiny model is proposed in this paper. On the basis of the original two detection heads of YOLOv4 tiny, one detection head is added to the lower sampling layer. Three detection heads can enhance the detection ability of the model to multi-scale targets. Since shallow features are beneficial to small object detection, this operation can enhance the detection ability of the model to small targets. In addition, a new multi-scale feature fusion structure PF-Net is designed based on the feature utilization and fusion modes of FPN and PANet. On the basis of YOLOv4 tiny, the FPN structure is replaced and an improved multi-scale feature fusion structure PF-Net is used as the feature utilization and fusion mode of PF-YOLOv4 tiny. The overall structure and feature utilization and fusion of PF-YOLOv4 tiny are shown in Figure 1. PF-YOLOv4 tiny increases the application of network to shallow features, which can improve the accuracy of object detection by this model.

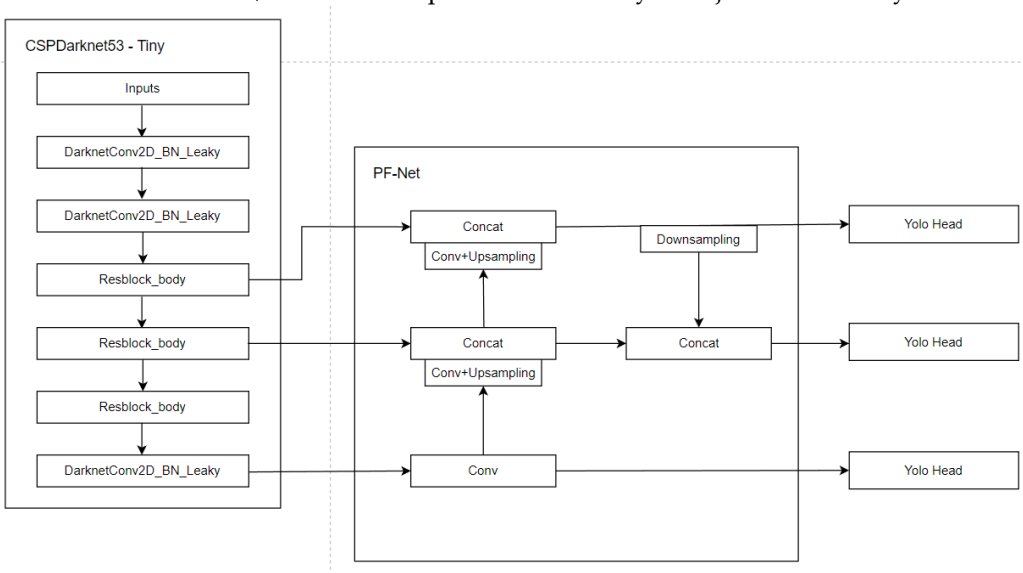


Figure 1. PF-YOLOv4 tiny Network structure.

3.1.2. Attention Mechanism

In the object detection task, each part in any picture has different importance. As shown in Figure 2, what we need is vehicle-related information to realize the vehicle detection task, so we want to pay more attention to the vehicle-related part. Each part of an image is assigned a weight equal to the amount of attention people pay to each part of the image. In this way, the weight can simulate the different focus of people's attention when they see the picture, that is, to achieve an attention

 doi:10.20944/preprints202409.0224.v1

mechanism. We can use the attention mechanism to improve the model's attention to traffic targets or small targets, so as to improve the detection ability of the object detection model.

Figure 2. Vehicle detection diagram in data set.

In our model, we use three CBAM structures to improve the model's focus on important targets, and obtain the improved PF-YOLOv4 tiny-CBAM based on YOLOv4 tiny. CBAM is a lightweight structure that can usually be added in any layer of convolution. Using CBAM attention module in object detection tasks can make the model suppress invalid information areas and pay more attention to areas containing key information in the image. PF-YOLOv4 tiny-CBAM adds a CBAM module after the last convolutional layer of the backbone network, between the first residual module and PF-Net connection, and between the second residual module and PF-Net connection. Since the last convolutional layer corresponds to the largest target scale and has a large number of channels, it is easy to mix in invalid information, so it is necessary to use the attention mechanism to make the model focus on the feature graph containing effective information. After the residual module, CBAM module is added because the feature maps corresponding to these two layers are of large scale, and attention mechanism is needed to make the model pay more attention to the features of the target region. At the same time, adding CBAM modules to these layers does not affect the backbone network, so the weight of the original YOLOv4 tiny can be used for initial training, which makes the model convergence easier and the model with better precision performance can be obtained. See Figure 3 for the network structure diagram of PF-YOLOv4 tiny-CBAM.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 September 2024 doi:10.20944/preprints202409.0224.v1

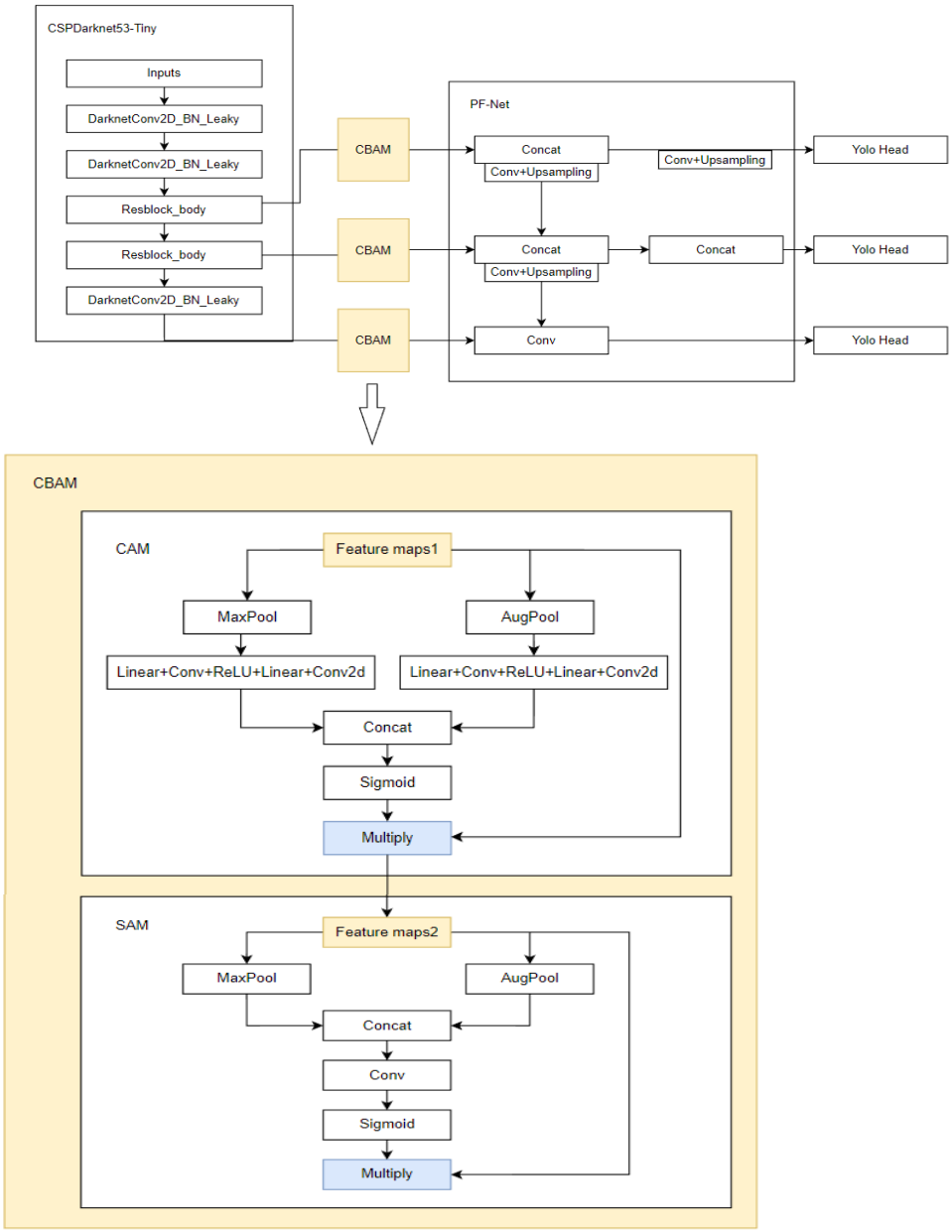


Figure 3. Structure diagram of PF-YOLOv4 tiny-CBAM.

For each CBAM module in Figure 3, the feature figure $F \in R^{C\times H\times W}$ output from different feature layers of the backbone network is taken as input. After CAM processing, the feature figure F_1 is obtained, and after SAM processing, the final feature figureis F_2 taken as output. $M_{CAM} \in R^{C\times 1\times 1}$is a channel attention diagram generated by the channel attention mechanism, and $M_{SAM} \in R^{1\times H\times W}$ is a spatial attention diagram generated by the spatial attention mechanism.

$$F_1 = M_{CAM}(F)\otimes F \quad (1)$$

$$F_2 = M_{SAM}(F_1)\otimes F_1 \quad (2)$$

In CAM module, the feature maps of backbone network are averaged and maximized respectively, then shared MLPS and a series of activation operations are used to get channel attention diagram M_{CAM}, in which MLP composed of Liner+Conv, etc. are shared parameters. The calculation process of CAM is shown in Formula 3.

$$M_{CAM}(F) = \text{sigmoid}(\text{MLP}(\text{AvgPool}(F)) + \text{MLP}(\text{MaxPool}(F))) \quad (3)$$

 doi:10.20944/preprints202409.0224.v1

In SAM module, the spatial attention diagram M_{SAM} is obtained after averaging and maximum pooling and splicing, and then after convolution and activation. The calculation process of SAM is shown in Formula 4.

$$M_{SAM}(F_1) = \text{sigmoid}(\text{Conv}([\text{AvgPool}(F_1); \text{MaxPool}(F_1)])) \quad (4)$$

3.2. Data Processing

3.2.1. Data Enhancement

We use the data enhancement method to balance the number of small and medium-sized targets in the data set from the perspective of amplifying small target features. Instead of just following the three methods of enhancement used in the paper that proposed the Copy-paste method, we make enhancement by selecting a small target in a single image and then copying and pasting several times at random locations in the image. Select multiple small targets in a single image and copy and paste them anywhere in the image. Select all the small targets in a single image and copy and paste them multiple times at any location in that image. We choose to select a number of small targets from the whole data set as the material library of small targets, and then select a number of pictures in the data set as the background library, and paste random positions on the pictures in the background library by using the targets in the randomly selected material library.

Taking the reflective cone and throwing objects of small target objects in the data set as an example, the data of the reflective cone and throwing objects are generally less, and for the camera perspective, the reflective cone belongs to the small-scale target, as shown in Figure 4.

However, in real life, it is not easy to collect the data of reflective cones and throwing objects on the pavement. One is because there are fewer reflective cones and throwing objects, only in the occurrence of accidents, maintenance and other situations to be collected. Second, due to road safety and other issues, it is impossible to carry out more artificial creation, such as the placement of reflective cones on the road surface and throwing objects. Generally speaking, the data of reflective cones and throwing objects are suitable for amplification through data enhancement. In this paper, the Copy-Paste method mentioned above is used to enhance the data of reflective cones and throwing objects. The enhanced effect diagram is shown in Figure 5. The enhanced data set contains more target numbers of reflective cones and throwing objects, which can effectively improve the detection ability of the model for such targets.

Figure 4. Example of a reflective cone from a camera perspective.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 September 2024 doi:10.20944/preprints202409.0224.v1

Figure 5. An example diagram of a reflective cone and sprinkles enhanced with Copy-Paste.

3.2.2. Prior Bounding Box Clustering

In our method, K-means algorithm is used to cluster prior bounding boxes, but the calculation method of the distance between the two targets is modified. Euclidean distance is no longer used, but IoU, which is more consistent with the target box, is used for definition. Using Euclidean distance to measure the distance between each target and clustering center, the measurement errors may be related to the size of bounding boxes, and large bounding boxes usually have more errors than small bounding boxes. Therefore, for the target frame, IOU is more appropriate for distance measurement, assuming $anchor = (w_a, h_a), box = (w_b, h_b)$, where w represents the width of anchor and h represents the height of anchor. See Formula (5) and Formula (6) for the specific calculation of the intersection ratio of two anchors. In calculation, it is assumed that the center points of all target frames coincide with each other, and only the width and height of the target frames are needed, which can further simplify the calculation. Our complete clustering procedure is shown in Algorithm 1.

$$\mathrm{d}(\mathrm{box}, \mathrm{anchor}) = 1 - \mathrm{IOU}(\mathrm{box}, \mathrm{anchor}) \tag{5}$$

$$\mathrm{d}(\mathrm{box}, \mathrm{anchor}) = 1 - \frac{\min(\mathrm{w_a}, \mathrm{w_b}) * \min(\mathrm{h_a}, \mathrm{h_b})}{\mathrm{w_a h_a} + \mathrm{w_b h_b} - \min(\mathrm{w_a}, \mathrm{w_b}) * \min(\mathrm{h_a}, \mathrm{h_b})} \tag{6}$$

Algorithm 1 K-means clustering process

Input: image1… imageN annotated data

Output: 9 anchors of different widths and heights

1: Initially, 9 anchors given in COCO dataset were selected as the clustering center, and the number of clustering centers was set as k=9.

2: Calculate the distance between each target a in the data set and each cluster center b:

$$d = 1 - \frac{\min(w_a, w_b) * \min(h_a, h_b)}{w_a h_a + w_b h_b - \min(w_a, w_b) * \min(h_a, h_b)}$$

3: The class is divided according to the value of d.

 doi:10.20944/preprints202409.0224.v1

4:**repeat**

5:**until** ($iters \geq 150$)

The clustering results are as follows: 10, 10, 24, 16, 21, 40, 52, 28, 48, 84, 114, 48, 120,117, 280,145, 519,278.

4. Experiment

4.1. Introduction to Data Sets

The object detection algorithm in this paper is applied to the traffic camera perspective to provide real-time target categories and positions required in camera images for the subsequent determination of traffic congestion, traffic violations and other tasks. The dataset is based on traffic images from the public dataset COCO and VOC datasets, and mainly includes road monitoring data from several cities in China, including scenes such as highways and intersections, which are amplified by camera images.

In order to make the model provide upstream detection results for more tasks in the future, the data set contains more categories, a total of 31 targets. Before data enhancement, it mainly contains about 200,000 data, and the resolution of most images is 1920×1080, about 32G, which meets the basic data requirements of object detection model training. During the experiment, we mainly focus on the categories of person, car, reflective cone and sprinkles, which contain many small scale objects. The model training size is selected to conform to 608×320 of 1920×1080.

4.2. Model Structure Comparison Experiment

In this paper, a new multi-scale feature fusion method, PF-Net, is proposed, and the attention module CBAM is added. The experimental results of these two structures will be compared in the following sections.

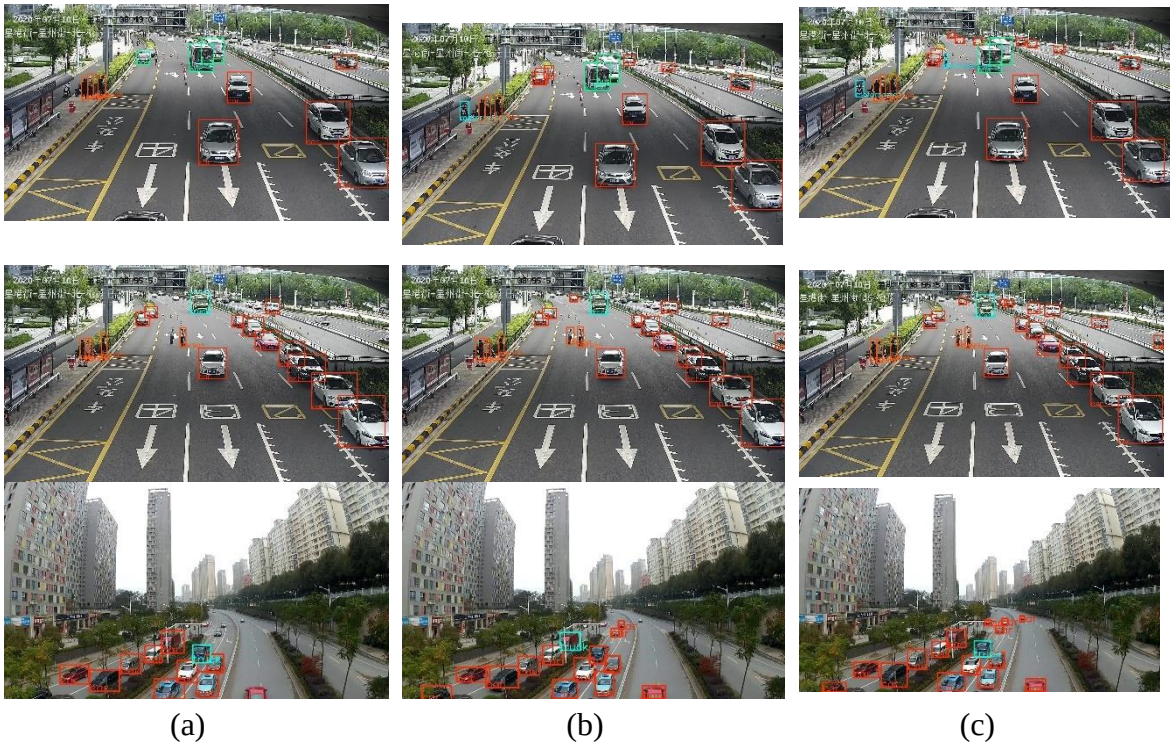
It can be seen from Table 1 that the modified PF-Net structure can effectively improve the detection accuracy of the model for categories of focus, such as person, car, etc. At the same time, the model with the addition of the attention-mechanism CBAM module, Compared with both YOLOv4 tiny and PF-YOLOv4 tiny, the detection effect of these categories is further improved, and can reach up to 3 to 4 percentage points of the improvement of individual categories.

Table 1. Evaluation results on test data of custom traffic data set (AP/%).

Class / Model	person	car	reflector cone	throwing objects
YOLOv4 tiny	69.40%	82.28%	66.92%	90.47%
PF-YOLOv4 tiny	70.59%	84.27%	67.61%	91.49%
PF-YOLOv4 tiny-CBAM	73.79%	86.07%	68.61%	91.09%

0.2 was selected as the confidence threshold for display, and images that did not contain scenes in the training set were selected for the test of YOLOv4 tiny, PF-YOLOv4 tiny and PF-YOLOv4 tiny-CBAM models, and representative test diagrams were selected for analysis, as shown in Figure 6.

 doi:10.20944/preprints202409.0224.v1



(a) (b) (c)

Figure 6. Comparison of detection results of various algorithms. (a) YOLOv4 tiny test results; (b) PF-YOLOv4 tiny test results; (c) PF-YOLOv4 tiny-CBAM test results.

It can be seen from the detection diagram that the detection effect of the improved PF-YOLOv4 tiny and PF-YOLOv4 tiny-CBAM model is better than that of the original YOLOv4 tiny model, and the detection effect of PF-YOLOv4 tiny-CBAM is better. The improved model can detect some small targets better than the original model, such as people, distant cars and reflective cones, which can prove the effectiveness of the improved model. Part of the false detection and missing detection still need to be further optimized from the aspect of data, which will be elaborated and tested in Section 4.3.

The experiments in Table 2 prove that adding an additional anchors head, increasing 6 anchors to 9 anchors allows the model to better adapt to multiple anchors at multiple scales, and the model's ability to examine multiple categories in the dataset will be enhanced to varying degrees. Using the improved PFNet structure instead of the original FPN in the model allows the model to expand the size span of detection targets, while using the repeated feature fusion structure for small and medium targets to ensure that the accuracy of these relatively small targets can be improved. And improved mAP by 2.01%.The added channel attention module increases mAP by 1.35 percentage points over the improved PF-YOLOv4 tiny, compared to YOLOv4 tiny, it has increased by 4.03%. Taking the reflector cone as an example, the final improved model PF-YOLOv4 tiny CBAM can increase by 1.69 percentage points,and also ensures that the accuracy of most categories is improved relative to the PF-YOLOv4 tiny. PF-YOLOv4 tiny-CBAM is equivalent to further improve the detection performance on the basis of guaranteeing the detection capability of PF-YOLOv4 tiny.

Table 2. Comparison of mAP and real-time results of the model.

Item / Model	mAP	FPS
YOLOv4 tiny	60.68%	93
PF-YOLOv4 tiny	62.69%	87
PF-YOLOv4 tiny-CBAM	64.04%	81

 doi:10.20944/preprints202409.0224.v1

According to the real-time experiment Table 2, although PF-YOLOv4 tiny and PF-YOLOv4 tiny-CBAM with modified structure detect fewer images per second than YOLOv4 tiny, they still have the characteristics of real-time and can be applied to intelligent transportation system. The improvement is a kind of precision improvement at the expense of a small amount of real-time performance.

4.3. Comparative Experiment Based on Data

The effectiveness verification experiments of Copy-paste mainly show the effects of reflecting cones and throwing objects.

As can be seen from Table 3, the model after data enhancement has stronger detection ability for several categories of enhancement, and can achieve a maximum accuracy improvement of 3 to 4 percentage points. Compared with PF-YOLOv4 tiny, the improved model with CBAM module added at the same time has higher detection accuracy. The attention mechanism of PF-YOLOv4 tiny-CBAM plays a role, making the model pay more attention to the target in the image, so as to obtain better detection effect.

Table 3. Detection effect of the corresponding category after data set amplification (AP/%).

Model \ Class	reflector cone	throwing objects
YOLOv4 tiny	66.92%	90.47%
PF-YOLOv4 tiny	67.61%	91.49%
PF-YOLOv4 tiny-CBAM	68.61%	91.09%
PF-YOLOv4 tiny+Copy paste	68.41%	92.54%
PF-YOLOv4 tiny-CBAM+Copy paste	69.32%	91.98%

Take reflection cone detection as an example, select scene pictures not included in the training set for testing, and the representative test results are shown in Figure 7. Figure (a) of Figure 7 uses PF-YOLOv4-tiny-CBAM model without data enhancement. It can be seen that although the structure has been modified and the overall detection accuracy has been improved, the amount of data for reflection cone is small. Without sufficient training data, it still cannot be detected. For this category, the generalization of the model cannot achieve good results. Figure 7b,c show the detection effect of PF-YOLOv4-tiny and PF-YOLOv4-tiny-CBAM, which are trained with data enhanced using Copy-Paste. In Figure 7d–f, the same is true. It can be seen that the model with enhanced data has better detection ability for reflective cones, and the PF-YOLOv4-tiny-CBAM with enhanced data has better generalization.

In the unamplified data set, there was a single scene of reflective cones and sprinkles, a small number of targets for this category, and poor generalization. As shown in Figure 7 and Table 3 above, the generalization ability of the model trained with the data set enhanced by data has been enhanced for the reflective cone. Through data enhancement, the model's learning of the features of this category has been improved, and thus its detection and generalization ability for this category has been improved.

After K-means was used for clustering, the improved version of YOLOv4 tiny model was used, and the detection effects of various categories were shown in Table 4.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 September 2024 doi:10.20944/preprints202409.0224.v1

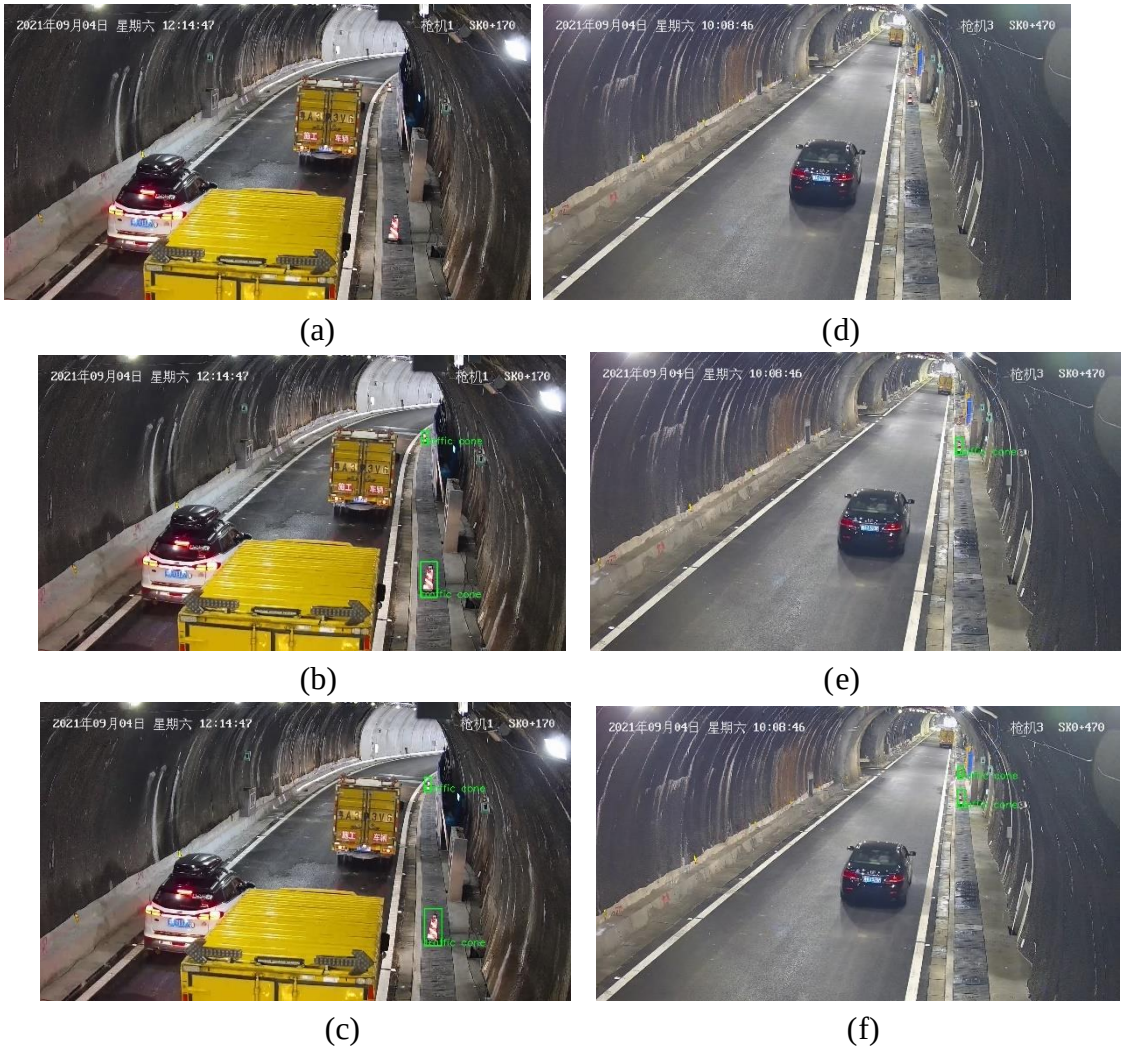


Figure 7. Reflection cone detection effect. (a) Example1: Using PF-YOLOv4-tiny-CBAM model without data enhancement; (b) Example1: Using PF-YOLOv4-tiny model with data enhancement; (c) Example1: Using PF-YOLOv4-tiny-CBAM model with data enhancement; (d) Example2: Using PF-YOLOv4-tiny-CBAM model without data enhancement; (e) Example2: Using PF-YOLOv4-tiny model with data enhancement; (f) Example2: Using PF-YOLOv4-tiny-CBAM model with data enhancement.

Table 4. Detection effect of corresponding categories after anchor clustering by K-means (AP/%).

Class / Model	person	car	reflector cone	throwing objects
YOLOv4 tiny	69.40%	82.28%	66.92%	90.47%
PF-YOLOv4 tiny	70.59%	84.27%	67.61%	91.49%
PF-YOLOv4 tiny-CBAM	73.79%	86.07%	68.61%	91.09%
PF-YOLOv4 tiny +Copy paste+ K-means	73.79%	86.73%	69.41%	92.14%
PF-YOLOv4 tiny-CBAM +Copy paste+ K-means	76.99%	89.05%	70.32%	91.58%

It can be found through the experiment that both the improved PF-YOLOv4 tiny and PF-YOLOv4 tiny-CBAM can improve the AP value of most categories after K-means clustering for anchor. It is worth noting that the improvement span of detection accuracy of categories such as car and person is higher than that of some other categories, which increases by about 3%. This may be

 doi:10.20944/preprints202409.0224.v1

because these categories such as car have a large number of targets. K-means clustering can have a greater influence on the clustering center, and anchors more suitable for these categories can be obtained. It is further explained that selecting the prior bounding box ratio of the appropriate data set is helpful to improve the detection ability of the model. And the final improved PF-YOLOv4 tiny CBAM+Copy paste+K-means model increased mAP by 4.9% compared to the original YOLOv4 tiny.

5. Conclusions

In this paper, an improved model based on YOLOv4 tiny is proposed to address the issue of small pedestrian targets in some vehicles in intelligent transportation scenarios. Based on the FPN structure, the number of detection heads has been increased, and a top-down feature fusion path has been added for small and medium-sized targets. At the same time, a CBAM module has been added to assist in enhancing the model's ability to detect small targets and ensuring its real-time performance. Tested on a 260000 custom traffic dataset containing some public traffic images, this improvement improved the model's mAP by 4.03%, and the detection accuracy for small targets was also correspondingly improved. To address the issue of imbalanced data and features in custom traffic datasets containing public VOC and COCO partial traffic images, an improved Copy Paste is used to enhance the features of some categories, ensuring that the AP values of the corresponding categories have at least one point of improvement. Using K-means with improved distance measurement to solve the mismatch problem between the dataset and prior bounding boxes, some categories can achieve a 3 percentage point improvement.

Author Contributions: Analysis, C.H. and J.S.; methodology, J.S.; software, J.S.; supervision, C.H.; writing—original draft preparation, J.S.; writing—review and editing, C.H. and C.W. ; data processing, C.W.. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Not applicable.

Acknowledgments: The authors thank all the anonymous reviewers for their constructive comments and also thank all the editors for their careful proofreading..

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

References

1. Kisantal M, Wojna Z, Murawski J, et al. Augmentation for small object detection[J]. arXiv preprint arXiv:1902.07296, 2019.
2. Liu Y, Sun P, Wergeles N, et al. A survey and performance evaluation of deep learning methods for small object detection[J]. Expert Systems with Applications, 2021, 172: 114602.
3. Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2117-2125.
4. Bochkovskiy A, Wang C Y, Liao H Y M. Yolov4: Optimal speed and accuracy of object detection[J]. arXiv preprint arXiv:2004.10934, 2020.
5. Cheng G, Yuan X, Yao X, et al. Towards large-scale small object detection: Survey and benchmarks[J]. arXiv preprint arXiv:2207.14096, 2022.
6. Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 580-587.
7. He K, Zhang X, Ren S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2015, 37(9): 1904-1916.
8. Girshick R. Fast r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.
9. Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[J]. Advances in neural information processing systems, 2015, 28.
10. Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]//Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. Springer International Publishing, 2016: 21-37.
11. Redmon J, Farhadi A. Yolov3: An incremental improvement[J]. arXiv preprint arXiv:1804.02767, 2018.

 doi:10.20944/preprints202409.0224.v1

12. Duan K, Bai S, Xie L, et al. Centernet: Keypoint triplets for object detection[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 6569-6578.
13. Zhu C, He Y, Savvides M. Feature selective anchor-free module for single-shot object detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 840-849.
14. Chen Y, Zhang P, Li Z, et al. Feedback-driven data provider for object detection. arXiv 2020[J]. arXiv preprint arXiv:2004.12432.
15. Kisantal M, Wojna Z, Murawski J, et al. Augmentation for small object detection[J]. arXiv preprint arXiv:1902.07296, 2019.
16. Ghiasi G, Cui Y, Srinivas A, et al. Simple copy-paste is a strong data augmentation method for instance segmentation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 2918-2928.
17. Zhang S, Zhu X, Lei Z, et al. Faceboxes: A CPU real-time face detector with high accuracy[C]//2017 IEEE International Joint Conference on Biometrics (IJCB). IEEE, 2017: 1-9.
18. Zhang S, Zhu X, Lei Z, et al. S3fd: Single shot scale-invariant face detector[C]//Proceedings of the IEEE international conference on computer vision. 2017: 192-201.
19. Xu C, Wang J, Yang W, et al. Dot distance for tiny object detection in aerial images[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 1192-1201.
20. Shannon C E. A mathematical theory of communication[J]. The Bell system technical journal, 1948, 27(3): 379-423.
21. Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2980-2988.
22. Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2117-2125.
23. Tan M, Pang R, Le Q V. Efficientdet: Scalable and efficient object detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 10781-10790.
24. Liu S, Huang D, Wang Y. Learning spatial fusion for single-shot object detection[J]. arXiv preprint arXiv:1911.09516, 2019.
25. Qiao S, Chen L C, Yuille A. Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 10213-10224.