Research on Improved YOLOv4-based Pedestrian Detection Model

Ruiqi Hu*, Yujun Li

School of Information Engineering, Nan Chang University, NanChang, China.

* Corresponding Author Email: 735048720@qq.com

Abstract. Pedestrian safety affects many fields such as driverless vehicles. In order to improve the precision of the pedestrian detection method and distinguish people-like objects from pedestrians, an improved YOLOv4 pedestrian detection method is proposed. Firstly, RepVGG Block is introduced into the feature extraction layer and feature fusion layer of YOLOv4 to improve the feature extraction ability of the network and reduce the loss of feature information. Then the SENet attention mechanism is introduced to make the algorithm focus more on the useful information. Finally, the SIoU loss function is introduced to the regression of the pedestrian target frame, which improves the convergence speed and reduces the blindness of the target frame. Experimental results show that, compared with the original YOLOv4 algorithm, this improved algorithm has higher detection precision, and it can also distinguish people from people-like objects on the published Pedestrian Detection Data set, with a detection precision of 83.7%.

Keywords: Pedestrian detection; YOLOv4; RepVGG Block; SENet attention; SIoU loss function.

1. Introduction

In recent years, with the development of the social economy, pedestrian detection plays an important role in various monitoring fields, such as the field of driverless vehicles. The safety of passengers and pedestrians is an important factor that restricts the development of this field. Through cameras that obtain various images and detection of pedestrians in the images, the safety of unmanned driving can be improved. [1]. For example, in places with large traffic such as shopping malls and squares, pedestrian detection plays an important role in traffic detection and safety monitoring.

Target detection can be divided into traditional detection methods and the current mainstream methods based on deep learning. Traditional detection methods need artificial feature extraction and machine learning to realize target detection. Due to the influence of prior knowledge by artificial feature extraction, traditional target detection methods have poor robustness and detection speed [2]. Compared with the traditional methods, the target detection method based on deep learning has better detection performance. At present, the two main target detection algorithms are the Two-stage algorithm and the One-stage algorithm, both of which show various advantages, but also face various difficulties such as occlusion of people and confusion of people and portraits. The Two-stage algorithm is a method based on candidate areas. First, it is necessary to select the candidate areas that may contain the target to be measured, then identify and further correct the candidate areas, and acquire the final result. The main representative algorithm is R-CNN [3], Fast R-CNN [4], Faster R-CNN [5], and so on. The One-stage algorithm combines candidate frame and classification into one stage, and the whole stage is transformed into a regression problem, which improves the speed of target detection. The main representative algorithm is YOLO [6], SSD [7], YOLOv2 [8], YOLOv3 [9], YOLOv4 [10], Retina Net [11], and so on.

In order to realize pedestrian detection in a better way, many people have improved and optimized YOLO series algorithms. Zhang Mengge et al. [12] proposed the MobileNet-Yolov3 model and adopted the video adaptive inference algorithm based on the three-frame difference method and particle filter, which greatly improved the video detection rate. Hao Xuzheng et al. [13] proposed a pedestrian detection method based on a depth residual network, which made the network better adapt to pedestrian characteristics. Shi Ruijiao et al. [14] designed a quadruple downsampling branch to solve the problem of pedestrian feature loss after multiple downsampling. Channel-space attention

mechanism was introduced in the feature fusion stage to reduce the background noise interference. Meanwhile, CIoU loss function was introduced to solve the problem of inconsistent optimization of the mean square error loss function. Li Xiaoyan et al. [15] put forward the G-MBNet detection algorithm, which integrated CSPNet and ResNet18 to construct a lightweight feature extraction network. ECA channel attention module was introduced to improve the attention of the network to important features. Finally, the parameters and volume of the MBNet feature extraction network are reduced by combining the Ghost convolution module. To solve the problem of low precision of pedestrian detection in complex visual scenes, Kang Shuai et al. [16] introduced the hybrid hole convolution into the YOLOv4 backbone network, and the spatial sawtooth hole convolution structure was proposed to replace the spatial pyramid pool structure, which had a better effect than the original YOLOv4 network. This paper is based on the YOLOv4 model, and makes a series of improvements, with the help of the RepVGG network [20] to replace the original conv3×3 convolution layer with the RepVGG Block module, which broadens the width of the network. In the feature fusion stage, SENet [21] is embedded in the Neck network, which makes the network pay more attention to key information and improves the learning ability of the network. In the original YOLOv4, CIoU is used as the prediction frame loss. In the present paper, SIoU loss function is introduced [22] ss the loss function of the preselected frame, which redefines the distance loss, considers the vector angle between frames, and reduces the degree of freedom of regression. SIoU also accelerates the convergence of the network and improves the precision of training.

2. YOLOv4 Target Detection Algorithm

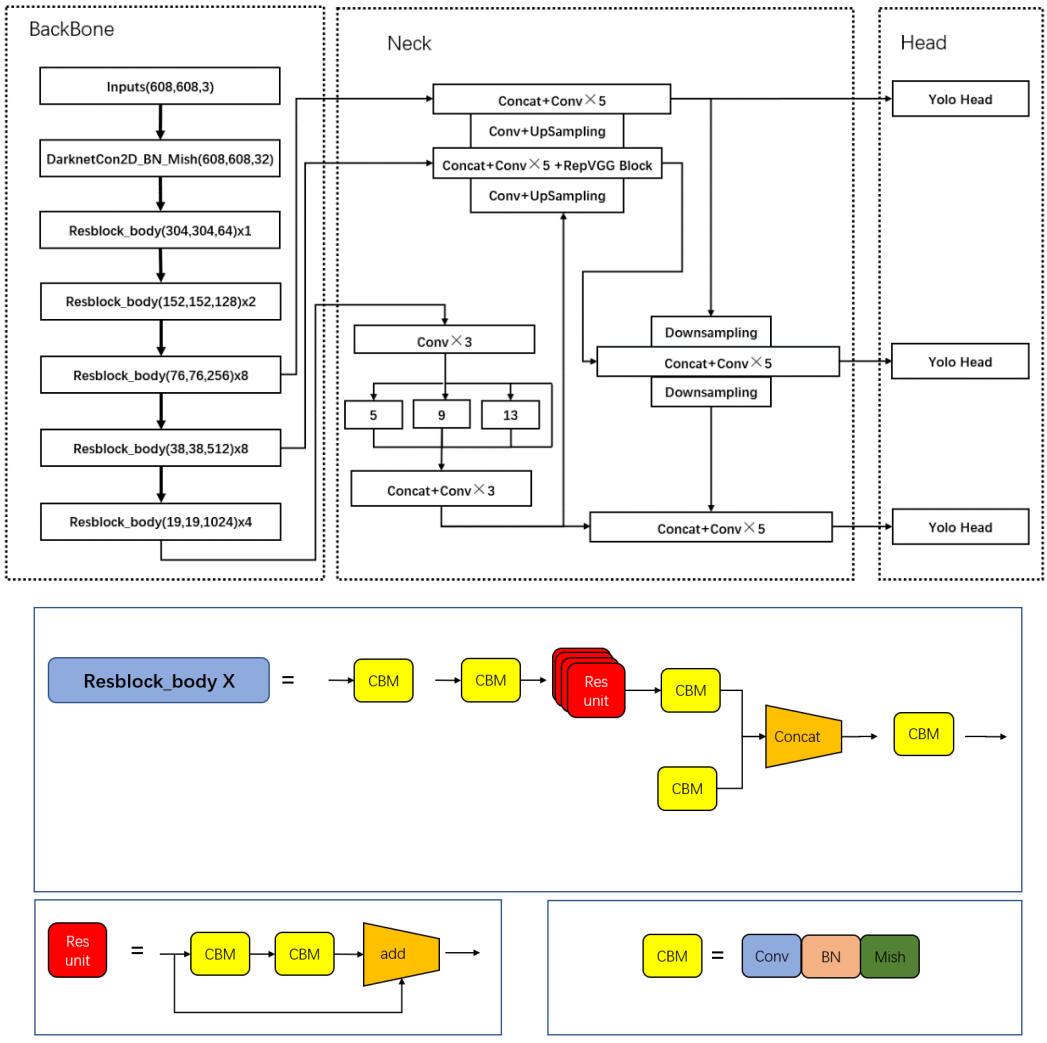

Figure 1. YOLOv4 Network Structure Diagram

YOLOv4 algorithm is an all-around improvement based on YOLOv3, and its performance has improved a lot. The network structure is shown in Figure 1. The input is enhanced by Mosaic data, and four different images are spliced to form a new image. On the Backbone of the feature extraction network, Darknet53 of YOLOv3 is combined with CSPNet [17] to form CSPDarknet53, which solves the problem of gradient disappearance and improves the learning ability of the network at the same time. In the feature fusion stage, SPPNet is added to the YOLOv3 feature pyramid FPN.[18], which constitutes PANet [19]. It can carry out a multi-scale fusion of features and improve the ability of feature extraction. YOLOv4 has three outputs, that is, 76x76, 38x38, and 19x19, which correspond to small, medium, and large targets. Meanwhile, many techniques such as CIoU loss function, Mish activation function, and Label Smoothing are introduced.

3. The Improvement of YOLOv4

3.1 Introducing the RepVGG Block

RepVGG [20] is an efficient VGG convolution neural network, and its structural re-parameterized is used to improve its performance and its precision can be compared with that of SOTA networks such as EfficientNet and RegNet. Affected by the residual structure of Resnet, the identity residual difference branch and 1x1 volume integral branch are added to each 3x3 convolution structure, which constitutes the basic structure of RepVGG Block. The branches are not connected across layers like Resnet but in a direct way. The addition of the two branches greatly improves the performance of the model. The structure of the block is shown in Figure 2. When the input channel of RepVGG Block is not equal to the output channel, set stride=2, and the dimension of feature space decreases, corresponding to RepVGG Block1 on the left of Figure 2. When the input channel of RepVGG Block is equal to the output channel, set stride=1, and then the identity residual differential branch can be added, as shown in RepVGG Block 2 on the right of Figure 2.

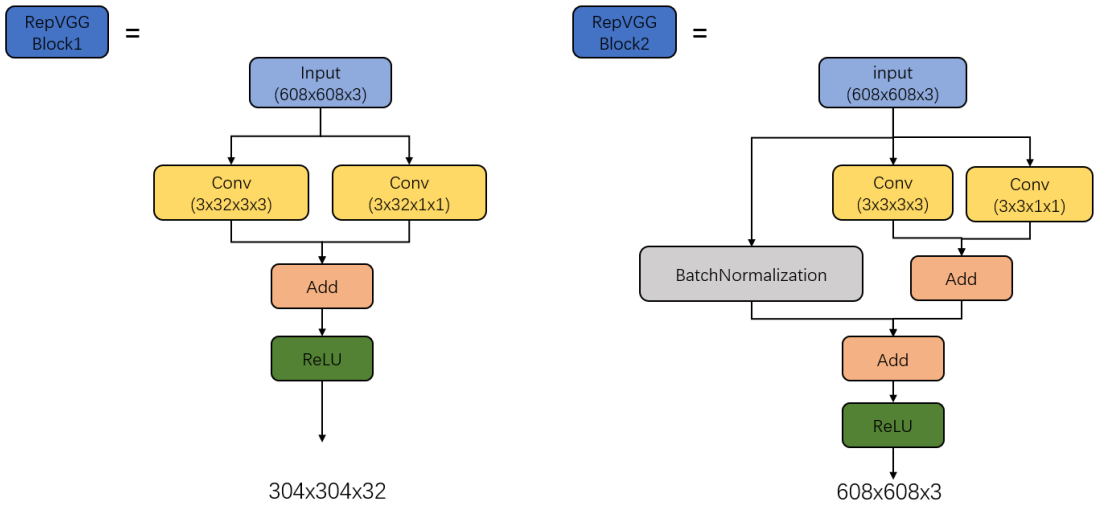

Figure 2. Two Kinds of Structure Diagrams of Repvgg Block

Based on the excellent performance of the Rep VGG network, this paper replaces part of the 3×3 convolution in the feature extraction layer and feature fusion layer of the YOLOv4 network with RepVGG Block, which broadens the network width and has a better performance in pedestrian detection.

3.2 Adding the SENet Module

SENet [21] is a CNN structure composed of Squeeze, Excitation, and Reweight, which can adjust features by using global information and make the network pay attention to important feature information while suppressing unimportant feature information. The structure diagram of the SENet

module is shown in Figure 3. Firstly, the input image is pooled globally to obtain a 1x1xC characteristic image, and then the dimension is reduced by the full connection for the first time. Then, the original dimension is raised by the full connection for the second time, and the weight of 1x1xC is obtained by the Sigmoid activation function. Finally, the result is obtained by multiplying the weight by the corresponding channel of the input image.

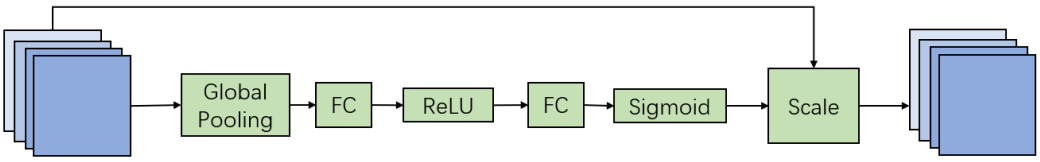

Figure 3. SENet Network Structure Diagram

At present, a large number of scholars have demonstrated the effectiveness of SENet in the field of target recognition. However, faced with different scenarios and different network structures, they do not know which part of the network structure should be added to SENet. In order to play a better role, in the later experimental stage, we add SENet to three different positions in the YOLOv4 feature fusion layer, and the specific positions are shown in Figure 4:

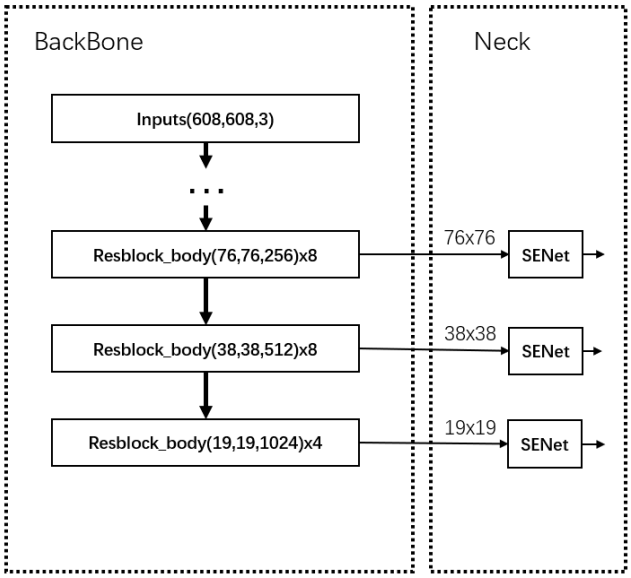

Figure 4. Three Different Positions to Add Attention Mechanism

After the output of the three characteristics of BackBone, that is 76×76，38×38，19×19, SENet is connected, resulting in three different network structures, including, YOLOv4-se1, YOLOv4-se2, YOLOv4-se3.

3.3 SIoU Loss

Traditional target detection loss function depends on the aggregation of boundary regression indicators, such as GIoU, DIoU, CIoU, etc. However, the loss functions used so far have not considered the mismatch between GT and the direction of the prediction frame, which leads to slow convergence speed and low efficiency, because the prediction frame does not know which direction to converge. Therefore, this paper chooses a new loss function SIoU Loss [22], which takes into account the vector angle between the required regressions and redefines the penalty index.

SioLoss consists of four cost functions: angle cost, distance cost, shape cost, and IoU cost.

3.3.1 Angle Cost

The model first tries to make predictions on the nearest X-axis or Y-axis and then approaches along the nearest axis. As shown in Figure 5, if $\alpha \leq \frac{\pi}{4}$, the convergence process is minimized as α, otherwise, it is minimized as $\beta = \frac{\pi}{2} - \alpha$.

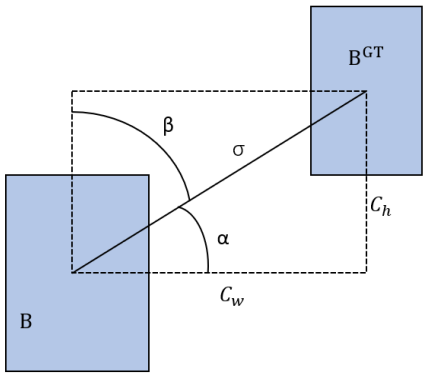

Figure 5. Angle Cost Calculation Process

In order to realize the above process, the following definitions are introduced:

$$\Lambda = 1 - 2 * \sin^2\left(\arcsin(x) - \frac{\pi}{4}\right)$$

Specifically:

$$x = \frac{c_h}{\sigma} = \sin(\alpha)$$

$$\sigma = \sqrt{\left(b_{c_x}^{gt} - b_{c_x}\right)^2 + \left(b_{c_y}^{gt} - b_{c_y}\right)^2}$$

$$c_h = max\left(b_{c_y}^{gt}, b_{c_y}\right) - \min\left(b_{c_x}^{gt}, b_{c_x}\right)$$

3.3.2 Distance Cost

The definition of distance cost takes into account the above angle cost.

$$\Delta = \sum_{t=x,y} (1 - e^{-\gamma \rho_t})$$

Specifically:

$$\rho_x = \left(\frac{b_{c_x}^{gt} - b_{c_x}}{c_w}\right)^2, \rho_y = \left(\frac{b_{c_y}^{gt} - b_{c_y}}{c_h}\right)^2, \gamma = 2 - \Lambda$$

The distance calculation is shown in Figure 6.

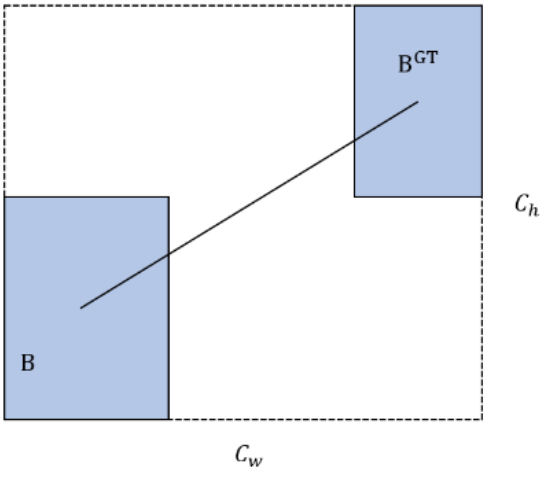

Figure 6. Distance Calculation between Prediction Box and GT

3.3.3 Shape cost

The shape cost is defined as follows:

$$\Omega = \sum_{t=w,h} (1 - e^{-\omega_t})^{\theta}$$

Specifically:

$$\omega_w = \frac{|w - w^{gt}|}{\max(w, w^{gt})}, \omega_h = \frac{|h - h^{gt}|}{\max(h, h^{gt})}$$

The value of θ controls the size of the shape cost, generally between 2 and 6.

3.3.4 IoU Cost

IoU cost is defined as follows:

$$IoU = \frac{|B \cap B^{GT}|}{|B \cup B^{GT}|}$$

The final loss function is defined as follows:

$$L_{box} = 1 - IoU + \frac{\Delta + \Omega}{2}$$

SIoU adds the matching direction penalty on the basis of considering the traditional penalties (distance, shape, and IoU) of GT and prediction box, which greatly helps the prediction box to move to the nearest axis quickly, reduces the degree of freedom of the prediction box and improves the precision of training.

3.4 The Improved Network Model

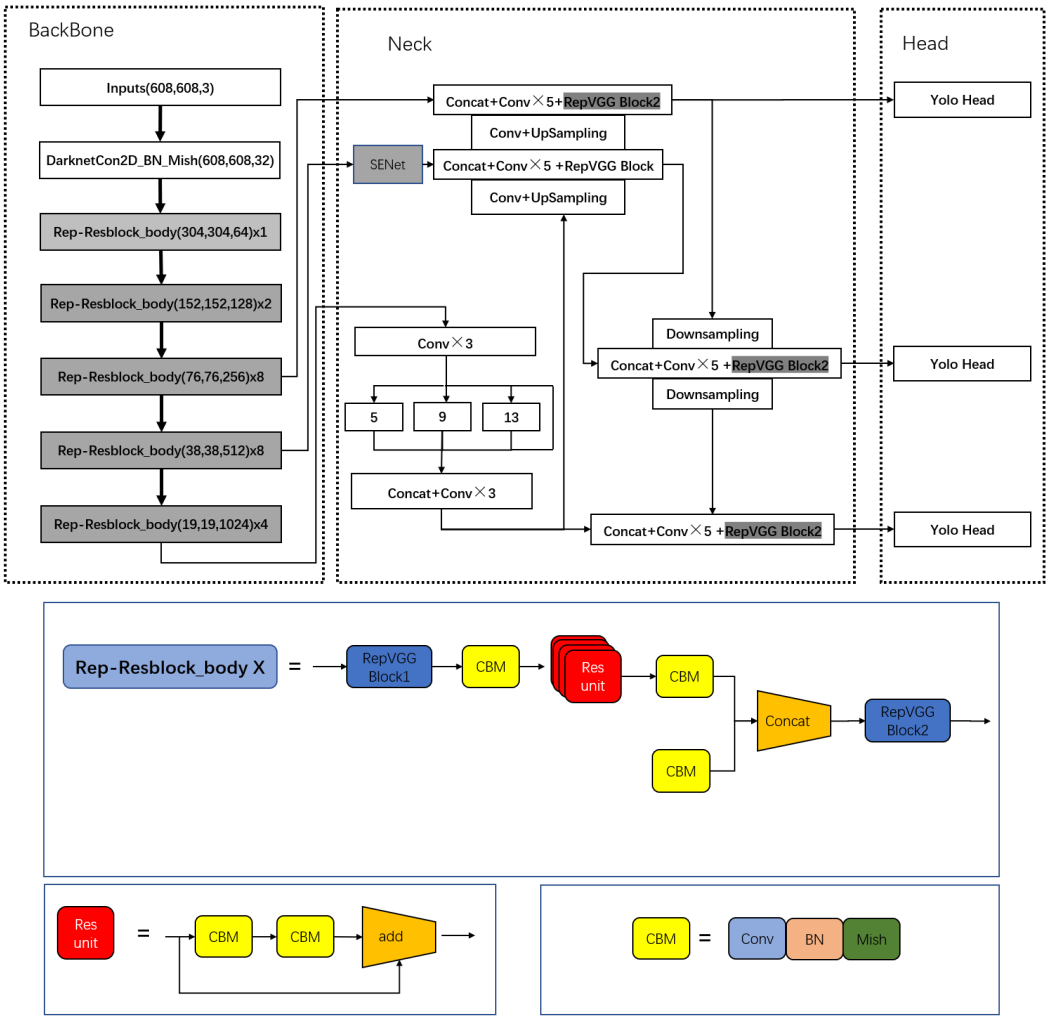

Figure 7. Improved YOLOv4 Network Structure Diagram

The following experiments demonstrate that adding SENet module to the output of the BackBone 38×38 is the most effective way to improve the network, namely the YOLOv4-se2 network structure. Adding RepVGG Block to BackBone and Neck, then changing the original CIoU loss function of YOLOv4 into SIoU loss function, and finally acquiring the network structure improved by three methods, as shown in Figure 7.

4. Data and Analysis of Experimental Results

4.1 Experimental Data Set

Almost all pedestrian detection data sets only contain such images as pedestrians, but when the images contain many objects close to human characteristics, training with these data sets will lead to some false alarms. The data set selected in this experiment is from Pedestrian Detection Data set published by Kaggle [23] (PDDS, henceforth), including person and person-like objects (PnPLO, henceforth). In the data set, person-like objects include statues, mannequins, scarecrows, and robots. The number of person and person-like objects is shown in Table 1.

Table 1. Number of Various Targets in PDDS

Category name	Person	Person-like
Quantity/piece	1623	1368

PDDS contains a total of 1339 annotated images. The division of training set, verification set, and test set in this paper is shown in Table 2.

Table 2. Division of PDDS

data set	Training set	Verification set	Test set	total
Quantity/sheet	964	241	134	1339

The display of PDDS is shown in Figure 8. It can be seen that there are many confusing person-like objects among people.

Figure 8. The Display of PDDS

4.2 Experimental Environment and Settings

The experimental environment: ubuntu18.04 operating system is adopted, the deep learning framework is Pytorch1.6, cuda10.2, cudnn7, and the GPU are NVIDIA Tesla V100 with 32G video memory.

4.3 The Network Parameters

In this paper, precision (P), recall (R), and mean Average Precision (mAP) are used to evaluate the performance of the algorithm. The three indicators are as follows:

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

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

$$\text{mAP} = \frac{\sum_{i=1}^{n} AP_i}{n}$$

TP (True Positives) in the above formula indicates the number of correctly detected targets, FP (False Positives) indicates the number of incorrectly detected targets, FN (False Negatives) indicates the number of undetected targets, n represents the total number of categories, and AP (Average Precision) represents the average precision of a certain category.

The experiment basically adopts YOLOv4 official parameter settings, and mosaic data enhancement. The input image size is640×640, the batchsize of training and testing is 8, the epoch is 100, the initial learning rate is 0.01, and the learning rate momentum is 0.937.

4.4 Experiment and Result Analysis

4.4.1 Adding RepVGG Block to Experimental Analysis

In order to verify whether replacing the partial convolution of YOLOv4 with RepVGG Block can improve the performance of the network, this section compares YOLOv4 network with the improved YOLOv4-Rep network, and the rest are consistent. The experimental results are shown in Table 3.

Table 3. Verification Experiment of REPVGG Block Module

Network structure	P/%	R/%	mAP@.5/%
YOLOv4	41.7	82.1	75.2
YOLOv4-Rep	**46.0**	81.9	**75.7**

From the experimental results, it can be seen that the precision (P) and mAP@.5 of the improved network are improved in different degrees, but only the recall ratio (R) is reduced by 0.2%, which is enough to verify the effectiveness of the improvement.

4.4.2 Experimental Analysis by Adding SENet Module

In order to verify the effectiveness of the attention mechanism in different positions of three different characteristic scales of BackBone, three different network structures in Section 2.2 are compared experimentally, and the experimental results are shown in Table 4.

Table 4. SENet Module Verification Experiment

Network structure	P/%	R/%	mAP@.5/%
YOLOv4	41.7	82.1	75.2
YOLOv-se1	44.6	84.2	77.3
YOLOv-se2	**45.4**	83.7	77.0
YOLOv-se3	45.1	84.6	74.7

It can be seen from the results that the precision (P) of the network structure YOLOv4-se2 is the highest. Although R and mAP@.5 are not the highest, they are improved and have good performance compared with YOLOv4.

4.4.3 Experimental Analysis of SIoU Loss Function

In order to verify the effectiveness of SIoU loss function, this section compares SIoU with four common loss functions of DIoU, CIoU, EIoU, and GIoU (CIoU was used in the YOLOv4 network), and only changes the loss function of YOLOv4 to the above ones, leaving the rest unchanged. The experimental results are shown in Table 5.

Table 5. The Verification Experiment of the SIoU Loss Function

Loss function	P/%	R/%	mAP@.5/%
CIoU	41.7	82.1	75.2
GIoU	44.9	82.3	75.5
DIoU	40.9	82.0	73.9
EIoU	38.4	79.1	66.8
SIoU	**45.4**	82.0	75.7

The experimental results in the table show that SIoU's precision (P) and mAP@.5 are the highest. Except for EIoU's low recall ratio (R), other recall ratios (R) are relatively similar. Therefore, the improvement of SIoU's loss function is really effective.

4.5 Analysis of the Improved YOLOv4 Experimental Results

As the initial prior box of the original YOLOv4 is obtained by clustering on the VOC data set, it is quite different from the pedestrian detection data set in this paper. In order to reduce the missed detection rate of the prior frame, this paper adopts the K-means clustering algorithm to re-cluster the labeled size of the training set and obtains the optimal prior frame size. Adding all the above three improved methods, and when the epoch is 300, the improved YOLOv4 is obtained. The experimental results are shown in Table 6.

Table 6. The Experimental Results of the Improved YOLOv4

Network structure	P/%	R/%	mAP@.5/%
Improved YOLOv4	0.713	83.1	83.7

From the experimental results, it can be seen that the improved YOLOv4 network has a good performance, the percentage of mAP reaches 83.7%, and it can also distinguish people and people-like objects in the data set. Compared with the original YOLOv4 network, it has better feature extraction ability and generalization ability.

5. Conclusion

In this paper, based on the network framework of YOLOv4, firstly, the network is improved by using RepVGG Block multi-branch module, and then the expressive force of network features is enhanced by adding SENet. Finally, the original CIoU loss function of YOLOv4 is replaced by the SIoU loss function. Through the three methods, YOLOv4, a classic target detection algorithm, is optimized.

The experimental results show that the precision of the three improved methods is increased by 4.3%, 3.7%, and 3.7%, respectively compared with YOLOv4. Although this paper uses PDDS which is easy to mix pedestrians and humanoid objects, these methods can distinguish easily mixed objects on the YOLOv4 network at the same time, with the highest precision of 71.3% and the highest mAP of 83.5%. This paper only discusses the preliminary exploration of the pedestrian detection algorithm, and how to ensure real-time performance and further improve the generalization ability and precision of the algorithm are the next research directions.

References

[1] Xin Yu., Gao Hongbo., Zhao Jianhui., & Zhou Mo. (2018). Overview of deep learning intelligent driving methods. Journal of Tsinghua University (Natural Science Edition) 58(04), 438-444.

[2] Geng Yining., Liu Shuaishi., Liu Taiting., Yan Wenyang., & Lian Yufeng. (2021). Survey of pedestrian detection technology based on computer vision. Journal of Computer Applications 41(S1), 43-50.

[3] Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 580-587).

[4] Girshick, R. (2015). Fast r-cnn. In Proceedings of the IEEE international conference on computer vision (pp. 1440-1448).

[5] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems, 28.

[6] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 779-788).

[7] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016, October). Ssd: Single shot multibox detector. In European conference on computer vision (pp. 21-37). Springer, Cham.

[8] Redmon, J., & Farhadi, A. (2017). YOLO9000: better, faster, stronger. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7263-7271).

[9] Redmon, J., & Farhadi, A. (2018). Yolov3: An incremental improvement. arXiv preprint arXiv: 1804. 02767.

[10] Bochkovskiy, A., Wang, C. Y., & Liao, H. Y. M. (2020). Yolov4: Optimal speed and precision of object detection. arXiv preprint arXiv: 2004.10934.

[11] Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In Proceedings of the IEEE international conference on computer vision (pp. 2980-2988).

[12] Zhang Mengge., Li Jian., & Chen Huiwen. (2022). Real-time video pedestrian detection algorithm based on YoloV3. Microcontrollers & Embedded Systems, 22(06), 29-32.

[13] Hao Xuzheng., & Chai Zhengyi. (2019). Improved pedestrian detection method based on depth residual network. Application Research of Computers, 36(05), 1569-1572+1584.

[14] Shi Ruijiao., Chen Houjin., Li Jupeng., Li Yanfeng., Li Feng., & Wan Chengkai. (2022). Small-scale pedestrian detection algorithm in railway scene based on attention and multi-level feature fusion. Journal of the China Railway Society, 44(05),76-83.

[15] Li Xiaoyan, Fu Huitong, Niu Wentao, Wang Peng, Lv Zhigang & Wang Weiming. (2022). Multimodal pedestrian detection algorithm based on deep learning. Journal of Xi'an Jiaotong University, (10), 76-83.

[16] Kang Shuai, Zhang Jianwu, Zhu Zunjie & Tong Guofeng. (2021). Improved YOLOv4 algorithm for pedestrian detection in complex visual scene. Telecommunications Science, 37(08), 46-56.

[17] Wang, C. Y., Liao, H. Y. M., Wu, Y. H., Chen, P. Y., Hsieh, J. W., & Yeh, I. H. (2020). CSPNet: A new backbone that can enhance learning capability of CNN. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops (pp. 390-391).

[18] He, K., Zhang, X., Ren, S., & Sun, J. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE transactions on pattern analysis and machine intelligence, 37(9), 1904-1916.

[19] Liu, S., Qi, L., Qin, H., Shi, J., & Jia, J. (2018). Path aggregation network for instance segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 8759-8768).

[20] Ding, X., Zhang, X., Ma, N., Han, J., Ding, G., & Sun, J. (2021). Repvgg: Making vgg-style convnets great again. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 13733-13742).

[21] Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7132-7141).

[22] Gevorgyan, Z. (2022). SIoU Loss: More Powerful Learning for Bounding Box Regression. arXiv preprint arXiv:2205.12740.

[23] Karthika, N. J., & Chandran, S. (2020). Addressing the False Positives in Pedestrian Detection. In Electronic Systems and Intelligent Computing (pp. 1083-1092). Springer, Singapore.