Transactions on Computational and Scientific Methods | Vo. 4, No. 10, 2024
ISSN: 2998-8780
https://pspress.org/index.php/tcsm
Pinnacle Science Press

An Improved YOLO V5-s Algorithm for Real-Time Pedestrian Detection in Crowded Public Scenes

Ariadne McNulty[1], Anouk Elise[2]
Department of Electrical and Computer Engineering, University of Arizona, USA[1], Department of Electrical and Computer Engineering, University of Arizona, USA[2]
ariadne_mcnulty87@arizona.edu[1], Anouk.E98ol@gmail.com[2]

Abstract: Pedestrian detection is critical in fields such as autonomous driving, surveillance, and public safety, especially in crowded environments where occlusion and overlap of individuals pose significant challenges. While deep learning-based approaches have achieved notable progress, real-time detection on mobile devices remains difficult due to computational constraints. This paper proposes an improved YOLO V5-s algorithm designed for efficient pedestrian detection in crowded scenes on mobile platforms. To enhance performance, ShuffleNetV2 is introduced as a lightweight backbone, reducing model complexity while maintaining detection accuracy. The Convolutional Block Attention Module (CBAM) is integrated, employing channel and spatial attention mechanisms to optimize feature extraction. Additionally, data augmentation techniques and label smoothing are used to expand the training dataset, and the Complete Intersection over Union (CIoU) loss function is incorporated to improve detection precision and reduce missed detections. Experimental results on the MOT20det dataset show that the enhanced algorithm achieves a mean Average Precision (mAP) of 0.983 and a frame rate of 144 FPS, outperforming the original YOLO V5-s in both speed and accuracy. The lightweight model is suitable for real-time pedestrian detection on mobile devices in densely populated areas.

Keywords: ShuffleNetV2; Crowded Scene; YOLO V5-s; Mechanism of Attention.

1. Research Background and Significance

Pedestrian detection plays an important role in several fields, such as assisted driving system [1], vehicle monitoring system and early warning and protection system. As an important basic research topic in the field of target detection, pedestrian detection can provide effective information support for public places with high pedestrian density such as shopping malls and scenic spots and intelligent security fields [2].

With the rapid development of artificial intelligence industry and the improvement of computer hardware computing ability, scholars at home and abroad have carried out research on pedestrian detection schemes based on deep learning, and achieved some results[3]. However, in the application process of the current pedestrian detection algorithm in the actual large-scale crowded scene, there always exists the problem of high missed detection rate due to the overlap and occlusion of pedestrians, which still puzzles many researchers and is also a great challenge for pedestrian detection at present.

Many scholars have proposed different measures to improve the performance of algorithms based on deep learning theory. In 2019, Wojke [4] et al. proposed DeepSort algorithm, which used a residual network structure to extract the appearance information of the target, and associated the cosine distance of the appearance feature vector with the motion information using the Hungarian algorithm. However, its tracking effect depended on the accuracy and feature differentiation degree of the target detector. The tracking speed is closely related to the target detection speed. In 2019, Xu Chengji[5] et al. improved YOLO V3 by using the Attention mechanism and proposed the Attention-YOLO algorithm, which effectively improved the detection accuracy. However, its weakness lies in its inaccurate performance on a small range of discontinuous information.In 2021, taking the basic framework of RetinaNet, Zhou Dake[6] et al added spatial attention and channel attention subnetwork to the regression branch and classification branch respectively, and proposed a occlusion perception pedestrian detection algorithm combined with the dual attention mechanism, which effectively improved the performance of pedestrian detection algorithm in the case of severe occlusion. The impact of occlusion on detection is reduced, but the detection frame rate is only 11.8 FPS due to the additional computational amount brought by the dual attention mechanism subnetwork. Shen Junyu[8] et al. conducted end-to-end training based on the YOLO algorithm, quickly detected the number of targets in real-time video, and triggered the screenshot and video saving functions according to the preset threshold, realizing efficient detection and tracking of fish swarm. The system has strong robustness and high data processing and storage efficiency. However, due to the large number of fish swarm in video, Considering the special case of dense fish, the detection and counting of dense fish will have a high rate of missed detection.

In view of the problems and deficiencies in the above scholars' research, the author proposes an improved YOLO V5-s dense pedestrian detection algorithm based on literature research. In order to solve the problem of slow detection caused by insufficient computing power of mobile devices, shuffle lenetv2 is used as the lightweight backbone network to replace the original YOLOv5-s backbone network, because the lightweight network will reduce the network accuracy. Therefore, the channel attention module (CAM)[9] and spatial attention module (SAM) of CBAM are used to update the weights of different feature channels and spatial dimensions respectively. Improve the ability of network feature extraction and feature fusion, ensure the detection accuracy and improve the detection speed of mobile terminal at the same time, improve the generalization ability of the model through data enhancement and label smoothing, enrich the features of pedestrian samples, and use CIoU[10] to improve the original loss function of YOLO V5-s, improve the detection accuracy of the algorithm and reduce the rate of missed detection.

2. YOLO V5-s

YOLO V5 was proposed by Ultralytics LLC in May 2020 and is divided into YOLOV5-s, YOLO V5-m, YOLO V5-l and YOLO V5-x according to network depth and feature map width. In this paper, YOLO V5-s is adopted as the usage model, and its network model structure is shown in Figure 1. It can be seen from the network structure diagram that YOLO V5-s model is mainly divided into four parts, namely Input, Backbone, Neck and Prediction[11].

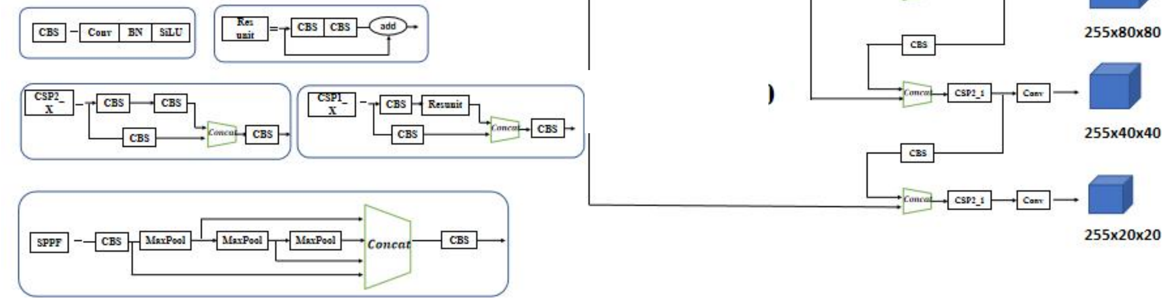


Figure 1. YOLO v5-s

3. CBAM Attention Mechanism

CBAM, a paper published in ECCV in 2018, not only considers the different importance of different feature channels, but also considers the importance degree of different locations of the same feature channel. In other words, it includes both the attention mechanism of the channel and the attention mechanism of the space.

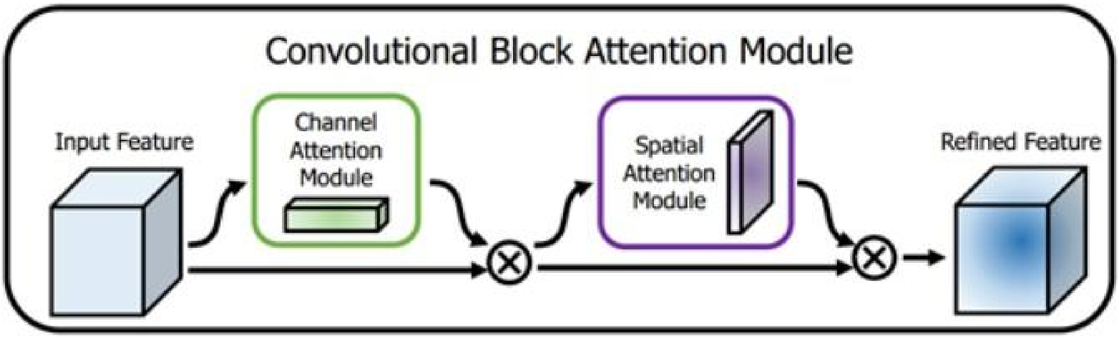


Figure 2. CBAM

The figure above is the schematic diagram of the whole CBAM, first through the attention mechanism module and then the spatial attention module. As for the impact of the two modules on the performance of the model in the sequence, the author of this paper also gives the experimental data comparison, which shows that the channel before space is better than the channel before space and the parallel mode of channel and spatial attention module.

So how is the channel attention module and the spatial attention module implemented?

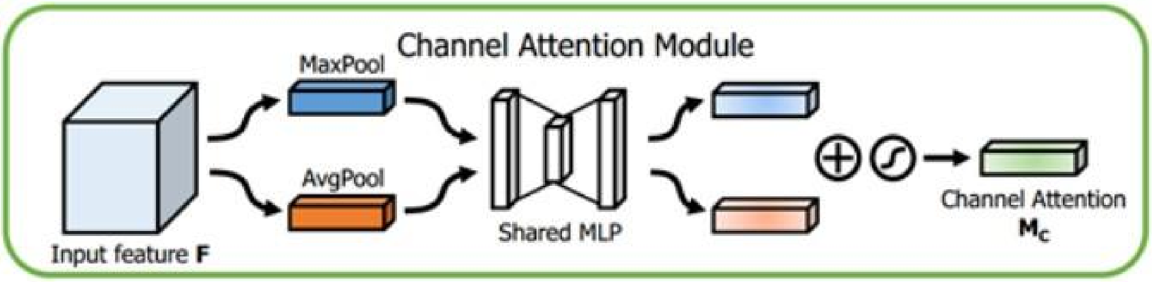


Figure 3. Channel Attention Module

CBAM adopts global average pooling and global maximum pooling, two different pooling means that the extracted high-level features are more abundant. Then, by modeling the correlation between the two fully connected layers and the corresponding activation function, the two outputs are combined to obtain the weights of each characteristic channel. Finally, after the weight of the feature channel is obtained, it is weighted to the original feature by channel multiplication to complete the re-calibration of the original feature in the channel dimension.

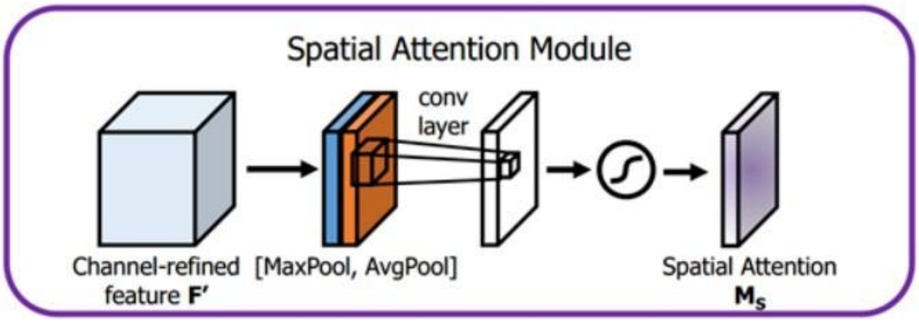


Figure 4. Spatial Attention Module

The first input is the feature that passes through the channel's attention module, and the global average pooling and global maximum pooling are also used. The difference is that the operation is carried out in the dimension of channels here, that is to say, all the input channels are pooled into two real numbers, and two (h*w*1) feature graphs are obtained from the input shape of (h*w*c). Then a 7*7 convolution kernel is used to form a new (h*w*1) feature graph after convolution. Finally, the same Scale operation, the feature of attention module is multiplied by the new feature map obtained to get the feature map adjusted by double attention.

4. Improved YOLO V5-s Model

4.1 YOLO V5-s Network Improvements

In order to further improve the pedestrian detection effect in dense scenes and improve the operation efficiency of mobile devices, an improved YOLO V5-s algorithm is proposed in this paper. Channel attention mechanism CBAM+ShuffleNetV2[12] is introduced to improve the backbone network of YOLO V5-s and improve the correlation expression of target information between different channels in the feature map. The network structure of YOLO V5-s after CBAM-Bloack[13] is added is shown in the figure below:

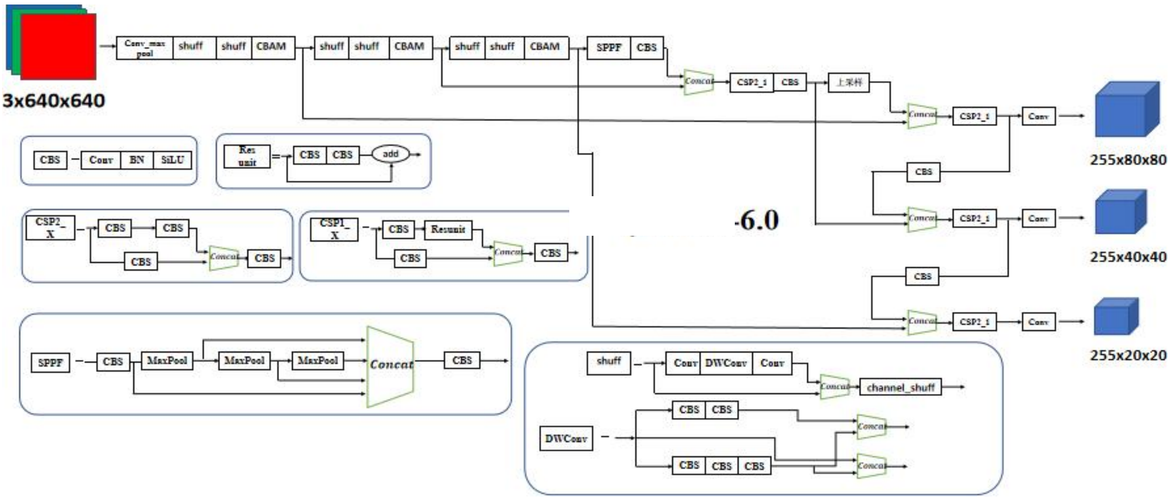
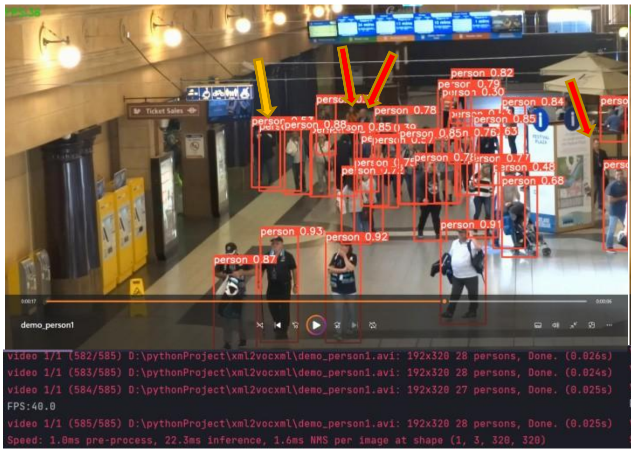
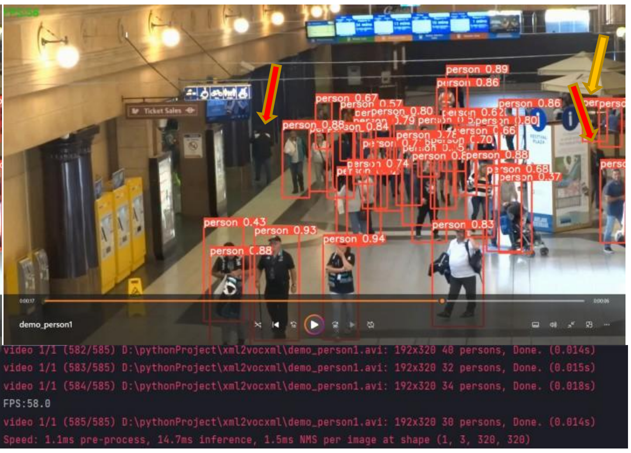


Figure 5. Improved YOLO V5-s network structure



```
video 1/1 (582/585) D:\pythonProject\xml2vocxml\demo_person1.avi: 192x320 28 persons, Done. (0.026s)
video 1/1 (583/585) D:\pythonProject\xml2vocxml\demo_person1.avi: 192x320 28 persons, Done. (0.024s)
video 1/1 (584/585) D:\pythonProject\xml2vocxml\demo_person1.avi: 192x320 27 persons, Done. (0.025s)
FPS:40.0
video 1/1 (585/585) D:\pythonProject\xml2vocxml\demo_person1.avi: 192x320 28 persons, Done. (0.025s)
Speed: 1.0ms pre-process, 22.3ms inference, 1.6ms NMS per image at shape (1, 3, 320, 320)
```



```
video 1/1 (582/585) D:\pythonProject\xml2vocxml\demo_person1.avi: 192x320 40 persons, Done. (0.014s)
video 1/1 (583/585) D:\pythonProject\xml2vocxml\demo_person1.avi: 192x320 32 persons, Done. (0.015s)
video 1/1 (584/585) D:\pythonProject\xml2vocxml\demo_person1.avi: 192x320 34 persons, Done. (0.018s)
FPS:58.0
video 1/1 (585/585) D:\pythonProject\xml2vocxml\demo_person1.avi: 192x320 30 persons, Done. (0.014s)
Speed: 1.1ms pre-process, 14.7ms inference, 1.5ms NMS per image at shape (1, 3, 320, 320)
```

Figure 6. YOLO-V5-s **Figure 7.** YOLO-V5-s(improve)

(The red arrow is the target of missed detection, and the yellow arrow is the target of miscalculation)

The figure above shows different reasoning results at the same moment when reasoning with the original YOLOV5-s and the improved YOLOV5-s using the same video on the GPU of the RTX2080TI. As shown in the reasoning results of the video in Figure 5, the reasoning speed reached 40FPS, and the detection results showed that 3 people were missed. As can be seen from the result of reasoning with the improved YOLOV5-s in Figure 6, the speed of reasoning reached 58FPS and only two people were missed at the same time.

It can be seen from the experimental results of the improved yolov5 that the introduction of CBAM module combined with loss function optimization can effectively filter out background interference in dense pedestrian detection, reduce the rate of missed detection and improve the detection accuracy. The introduction of shufflenetv2 as the backbone network for feature extraction can improve the detection speed of the model on the premise of ensuring the accuracy as much as possible.

4.2 Loss Function Improvement

4.2.1 Using CIoU

When detecting the target in the screen, the algorithm will generate more than one prediction box because there is more than one target in the field of view. Therefore, it is necessary to use the non-maximum suppression method to delete the redundant prediction boxes and select the prediction box closest to the real box. GIoU_Loss is used as the loss function in YOLO V5-s, and its principle is shown in Formula (1). GIoU added the measurement method of intersection scale, which effectively solved the problem when the boundary frame does not coincide. However, when the prediction box and the target box are mutually included, or the width and height are aligned, GIoU[14] will gradually degenerate into IoU in the process of regression, so that the relative position cannot be evaluated, and it is easy to increase the number of iterations and slow down the detection speed, and there is the risk of divergence.

$$\mathbf{RGIoU=RioU-|C-(A\cup B)|\ /\ |C|} \tag{1}$$

To solve the above problems, Zheng et al. took the central Distance between the center points of different target frames into consideration and proposed the distance-IOU (DIoU)[15] Loss with more stable regression, faster convergence and harder divergence. However, the consistency of the aspect ratio of the frame should be considered in the actual target detection. Therefore, based on the literature, the consistency of the bezel's aspect ratio is taken into consideration, and CIoU_Loss is introduced as a loss function to improve the YOLO V5-s algorithm. Compared with DIoU, the convergence speed of CIoU_Loss is faster and the regression effect is better[16].

The penalty item of CIoU_Loss is defined as follows:

$$\mathbf{RGIoU=[\rho^2(b,b^{gt})/c2]+\alpha v} \tag{2}$$

$$\mathbf{v=(4/\pi^2)(arctanw^{gt}/h^{gt}-arctanw/h)^2} \tag{3}$$

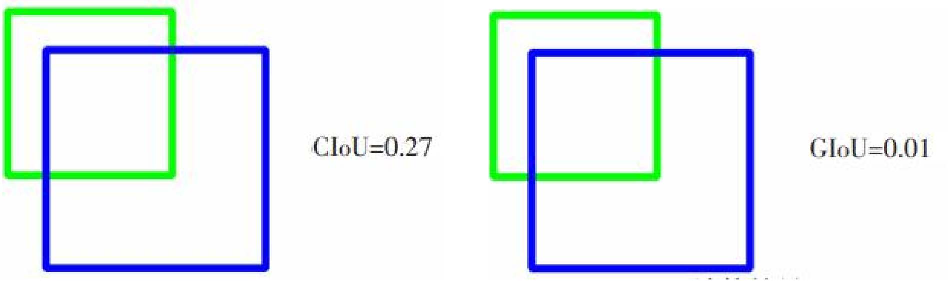
$$\mathbf{\alpha=v/[(1-IoU)+v]} \tag{4}$$

The final definition of CIoU_Loss is as follows:

$$\mathbf{L_{CioU}=1-IoU+[\rho^2(b,b^{gt})/c^2]+\ \alpha v} \tag{5}$$

In the above formula, **α** is a positive trade-off parameter and **v** is the consistency of the aspect ratio. In the above loss function, the center point of the detection frame and the target frame is represented by $\mathbf{b}$, $\mathbf{b}^{\mathbf{gt}}$, and the Euclidean distance is $\boldsymbol{\rho}$. $\mathbf{c}$ is the slant distance of the minimum rectangle covering the distance between the detection box and the target box.

As shown in FIG. 7, Opencv+numpy was used to draw the intersection between the predicted box and the actual box of the simulation algorithm of two rectangular boxes with different sizes and aspect ratios. GIoU was obtained by formula (1), and CIoU was calculated by formula (2)-(5). As can be seen from the calculation results of CIoU and GIoU in Figure 7, due to the degradation of GIoU loss at this time, it is very difficult to optimize when the prediction box bbox and ground_truth_bbox are included, especially in the horizontal and vertical direction convergence is difficult, but CIoU can still make the regression faster.



(a) CIoU Calculation result (b) GIoU Calculation result

Figure 8. Comparison of CIoU and GIoU when prediction box and target box overlap.

Therefore, CIoU was used in this paper to replace GIoU in the original YOLO V5 algorithm for anchor regression, so as to realize the gradient return when the real frame and the predicted frame did not coincide, and improve the convergence ability of the model. When the bounding box is adjacent to the real box, CIoU can optimize the disjoint bounding box, retain the bounding box with more accurate position, improve the accuracy of the model to predict the target position, and make the results obtained by non-maximum suppression screening more reasonable. At the same time, CIoU can reduce the training difficulty of the model and improve the accuracy of detection [17].

4.3 Training Data Processing Improvement

In most application scenarios, the original data set used by the training model cannot meet the ideal training needs, and obtaining more data sets will also increase the cost of training and bring more workload, so the better processing method is to conduct appropriate data preprocessing, including data enhancement and label smoothing processing.

The main purpose of applying data enhancement to data preprocessing is to expand the training set picture by means of data enhancement, which can make the data set samples used for training more diverse and reduce the influence of various additional factors on recognition. Adding random noise to the image can also effectively improve the generalization ability and robustness of the model. The single sample data enhancement methods that are used more frequently in the practical application process include scaling the image and twisting the length and width of the image, geometric transformation data enhancement to flip the image, and color transformation data enhancement to add noise to the image and modify the contrast, brightness, etc. The data enhancement method adopted in the experiment is based on the original traditional enhancement method, which adds the noise image random clipping and splicing method. That is, after adding noise to multiple images to be detected, a part of each image is cut out and a picture is synthesized for overall detection. This method can effectively improve the detection accuracy of the model under both small and large disturbance conditions. The essence of label smooth is a regularization process, which can reduce the possibility of overfitting training and make the probability distribution predicted by the model on the test set closer to the real distribution, thus improving the performance of the classifier. The label smoothing method adopted in the experiment in this paper is to randomly add false marks in the training set, and make it have a negative learning rate in the training process, so that the classification results of the model are closer to the correct classification results more quickly.

5. Research Scheme and Result Analysis

5.1 Experimental Platform and Data Set

In this paper, the hardware platform for model training and verification test is Core (TM) i7-9700k CPU@3.60GHz x8, with 16 GB memory and RTX 2080TI 12GB GPU, which is running on Linux operating system. According to the format requirements of the training set of YOLO series algorithms, the author selected 6126 images from the MOT20det data set of intensive pedestrian detection, converted all the annotation format of the data set into .txt format, and expanded the data set by using the data enhancement method mentioned above. Finally, a total of 9000 data sets were obtained. The training data set and the test data set were distinguished according to the ratio of 8:2.

5.2 Network Training

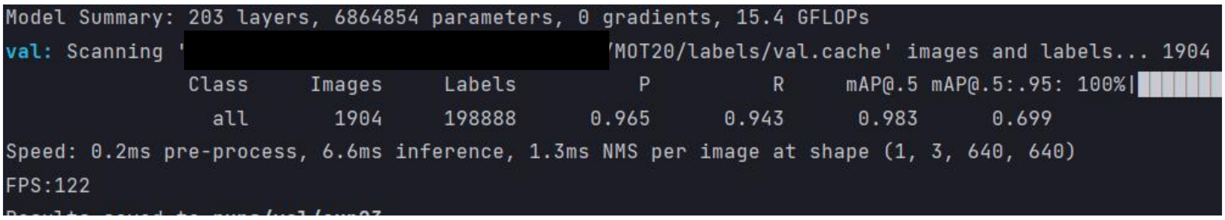
In this paper, Python language is used to establish the structure of the YOLO V5-s network model using the Pytorch deep learning framework. stochastic gradient descent (SGD) is used as the optimal algorithm during training. Optimize the parameters in the training process. In the training process, the momentum is set as 0.7, the weight attenuation is 0.0002, the initial learning rate is set as 0.01, the learning rate attenuation is 0.01 after every 10 training times, and the total training times is 100.

5.3 Model Evaluation and Comparison

In this paper, accuracy rate, recall rate, mean precision and harmonic mean are used as evaluation indexes during model training. Accuracy rate and recall rate are used as the criteria to distinguish the pedestrian detection and recognition effect, but they are negatively correlated. The average precision mean and harmonic mean are quantitative indexes considering both accuracy and recall rate. The larger their values are, the better the recognition effect will be.

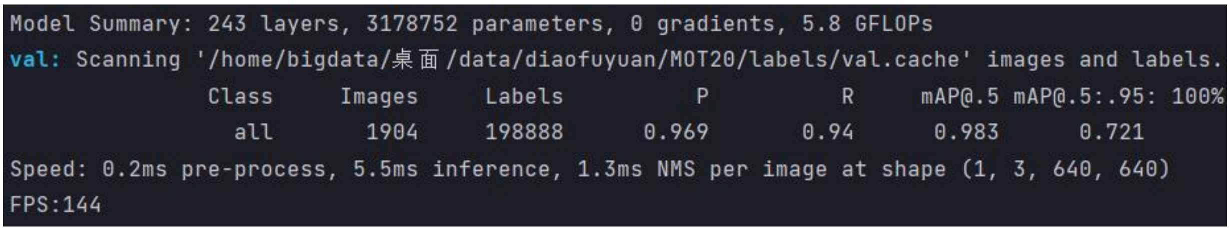
In order to verify the effectiveness of the modified network, the training time and actual detection effect of the original YOLO V5-s network and the improved YOLO V5-s network were compared in the same data set. After 100 training sessions in the same data set, the loss value in the training was compared with the convergence curve of mAP.

FIG. 8 shows the experimental results. FIG. a shows the results obtained by testing the original yolov5-s in the test set. The results showed that the average accuracy (mAP@0.5) was 0.983, recall ratio was 0.943, reference number was 6864854 and GFLOPS was 15.4. Figure b is the result of using the modified yolov5-s in the test set. The results showed that the average accuracy of the method (mAP.0.5) was 0.983, recall ratio was 0.94, reference number was 3178752 and GFLOPS was 5.8. It can be seen that the improved model based on yolov5-s in this experiment greatly reduces the number of parameters and computational complexity of the model and improves the detection efficiency on the premise of ensuring the accuracy as much as possible. Moreover, according to the reasoning results in the second part above, under the condition of the same misjudgment rate, the integrated attention mechanism can reduce the missed rate of blocked objects, and the reasoning speed of the improved yolov5-s is faster than that of the original. The comparison between Figure c and Figure d shows that CIOU loss function converges more rapidly than GIoU loss function for Bounding_Box regression.



```
Model Summary: 203 layers, 6864854 parameters, 0 gradients, 15.4 GFLOPs
val: Scanning '[redacted]/MOT20/labels/val.cache' images and labels... 1904
               Class     Images     Labels          P          R     mAP@.5 mAP@.5:.95: 100%|
                 all       1904     198888      0.965      0.943      0.983      0.699
Speed: 0.2ms pre-process, 6.6ms inference, 1.3ms NMS per image at shape (1, 3, 640, 640)
FPS:122
```

Fig (a). YOLOV5- s



```
Model Summary: 243 layers, 3178752 parameters, 0 gradients, 5.8 GFLOPs
val: Scanning '/home/bigdata/桌面/data/diaofuyuan/MOT20/labels/val.cache' images and labels.
               Class      Images      Labels           P           R      mAP@.5 mAP@.5:.95: 100%
                 all        1904      198888       0.969        0.94       0.983       0.721
Speed: 0.2ms pre-process, 5.5ms inference, 1.3ms NMS per image at shape (1, 3, 640, 640)
FPS:144
```

Fig (b). Improved YOLOV5-S

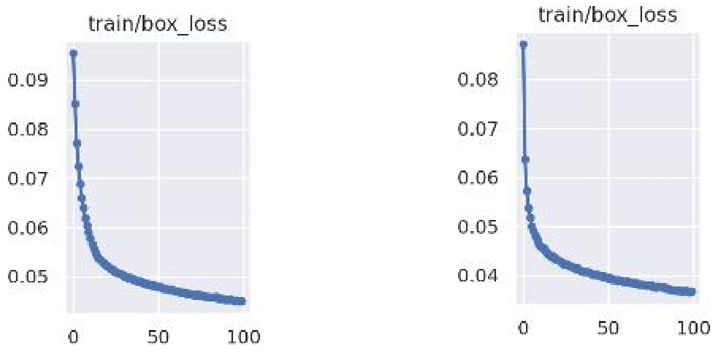


Fig (c). Test results for GIoU   Fig (d). Test results for CIoU

Figure 9. Test results of the YOLOV5-s and the improved yolov5-s on the test set

Table 1. Performance comparison of each algorithm

Network model	Precision	mAP	FPS
Faster_RCNN	0.989	0.992	32
YOLO v5-s	0.965	0.983	122
Improved YOLO V5-s(ours)	0.969	0.983	144

The table 1 shows the performance comparison of various algorithm models. The improved YOLOV5-S algorithm in this paper is faster than the original YOLOV5-s in speed, and the accuracy is basically unchanged. Compared with Faster_RCNN, although the average accuracy is not as high as Faster_RCNN, the detection speed is much higher than it. Therefore, the improved model based on YOLOv5-s in this paper is suitable for deployment on mobile devices.

6. Conclusion

Aiming at pedestrian detection in public scenes, the author studied the current mainstream YOLO V5-s algorithm and made the following improvements on the basis of the original YOLO V5-s algorithm:

(1) CBAM attention mechanism was introduced to improve the original YOLOV5-s network, and adaptive adjustment was made to the fused feature map;(2) Introducing shufflenetV2 module to build lightweight network and reduce model parameters and complexity;(3) By introducing data enhancement and label smoothing to expand the original data set, a large number of new training data was obtained, effectively improving the size of model training set and rapidly improving the effect of target detection; (4) CIoU parameters are introduced to improve the ability of network feature

extraction and feature fusion, while improving the detection accuracy and detection speed of the algorithm.

Compared with the original algorithm, on the MOT20det data set, the author proposed that the improved YOLO V5-based algorithm had better detection speed and missing rate than the original YOLOV5-s algorithm, while maintaining the accuracy of the original algorithm, and the mAP reached 0.983 (almost the same as the original YOLOV5-s). The frame rate reached 144 FPS, and the number of parameters and GFLOPs of the model were greatly reduced. It can realize the real-time and accurate detection task in the crowded scene on the mobile device.

References

[1] Qi Pengyu, Wang Hongyuan, Zhang Ji, et al. Crowded Pedestrian Detection Algorithm Based on Improved FCOS [J]. Journal of Intelligent Systems,2021,16(4):811-818.

[2] Liu Li, Zheng Yang, Fu Dongmei. Occlusion Pedestrian Detection Algorithm Based on Improved YOLO V3 Network Structure [J]. Pattern Recognition and Artificial Intelligence,2020,33(6):568-574.

[3] CAO J, SONG C, PENG S, et al. Pedestrian detection algorithm for intelligent vehicles in complex scenarios[J]. Sensors (Basel),2020,20 (13):3646(1-19).

[4] WOJKE N, BEWLEY A, PAULUS D. Simple online and real time tracking with a deep association metric[C]//IEEE International Conference on Image Processing. Beijing,2017:3645-3649.

[5] Xu Chengji, Wang Xiaofeng, Yang Yadong. Attention-yolo: A YOLO Detection Algorithm with Attention Mechanism [J]. Computer Engineering and Applications,2019,55(6):13-23,125.

[6] Zhou Da-ke, Song Rong, Yang Xin. Occlusion Perception Pedestrian Detection with Dual Attention Mechanism [J]. Journal of Harbin Institute of Technology,2021,53(9):156-163

[7] REZATOFIGHI H, TSOI N, GWAK J Y, et al. Generalized intersection over union: A metric and a loss for bounding box regression[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE,2019:658-666

[8] Shen Junyu, Li Linyan, Dai Yongliang, et al. Fish swarm Detection and Monitoring System Based on YOLO Algorithm [J]. Journal of Suzhou University of Science and Technology (Natural Science Edition), 2020, 37(3):68-73.

[9] Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 3-19.

[10] Cao Xuan, HAO Wanjun. Research on Dense Pedestrian Detection Algorithm with Improved YOLO V5 [J]. Journal of Suzhou University of Science and Technology (Natural Science Edition) ,2022,39(04):64-72.

[11] Zhu X, Lyu S, Wang X, et al. TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 2778-2788.

[12] Ma N, Zhang X, Zheng H T, et al. Shufflenet v2: Practical guidelines for efficient cnn architecture design[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 116-131.

[13] Fu H, Song G, Wang Y. Improved YOLOv4 marine target detection combined with CBAM[J]. Symmetry, 2021, 13(4): 623.

[14] Rezatofighi H, Tsoi N, Gwak J Y, et al. Generalized intersection over union: A metric and a loss for bounding box regression[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 658-666.

[15] ZHENG Z, WANG P, LIU W, et al. Distance-IoU loss: Faster and better learning for bounding box regression [C]//Proceedings of the AAAI Con-ference on Artificial Intelligence. 2020,34(7):12993-13000

[16] Janocha K, Czarnecki W M. On loss functions for deep neural networks in classification[J]. arXiv preprint arXiv: 1702.05659, 2017.

[17] Hou Z, Liu X, Chen L. Object detection algorithm for improving non-maximum suppression using GIoU [C]// IOP Conference Series: Materials Science and Engineering. IOP Publishing, 2020, 790(1): 012062.