ISBN: 978-981-18-5852-9
WCSE 2022 Spring Event: 2022 9th International Conference on Industrial Engineering and Applications
doi: 10.18178/wcse.2022.04.065

Object Detection of Roadside Pedestrians and Vehicles Based on Improved YOLO v5

Qirui Li [1], Wanyu Deng [2], Huijiao Xu [3] and Xiaoting Feng [4]

1 College of Computer Science
Xi'an University of Posts and Telecommunictions

Abstract. Target detection algorithms based on deep learning are more and more widely used in the field of autonomous driving. At present, the target detection of roadside pedestrians and vehicles in the field of autonomous driving mainly faces the problems of too many types of vehicles, complex detection target backgrounds, overlapping of people and vehicles, and too many small targets. In view of the existing problems, this paper improves the YOLO v5 algorithm, optimizes the feature extraction network of YOLO v5, adds a small target detection head, improves the ability of the backbone network to extract target features, and strengthens the detection of small targets; adds the Selayer attention mechanism to improve the model sensitivity to channel features and enhances the network's ability to detect occluded targets. The experimental results show that the improved YOLO v5 model can achieve an average P of 91.3% for the detection of 10 kinds of roadside pedestrians and vehicles, and the mAP@0.5 of 80.1%. Compared with the original YOLO v5, the P is increased by 8.4%, and the mAP@0.5 is increased by 5.2%.

Keywords: YOLO v5, target detection, attention mechanism, small target detection.

1. Introduction

With the development of autonomous driving technology and the increase of car ownership in our country, the roadside perception system has also become a necessary technical means to assist various Internet of Vehicles application scenarios [1]. As the key link in the field of autonomous driving, roadside pedestrian and vehicle target detection is of great significance to accurately and quickly detect pedestrians and vehicles. In physic traffic, there are many types of vehicles occluding each other, people and vehicles overlap, pedestrian distribution is complex, weather conditions are changeable, and light changes are sharp and other problems will cause interference to the detection of vehicles and pedestrians. To solve these problems, it is necessary to design a target detection algorithm that is suitable for the detection. The target detection algorithm based on deep learning has strong learning ability, and the most widely used in target detection algorithms mainly include the one-stage and two-stage [2]. Two-stage means that the target detection algorithms need to be completed for two steps. The first step is to obtain the selected area, and the second step is to classify, including R-CNN algorithm, SPP-Net algorithm, Fast R- CNN algorithm , Faster R-CNN algorithm, et al. [3]-[6].

The one-stage target detection algorithm is based on the regression frame, and does not need to find candidate regions specially, and can detect the target object only once, including SSD algorithm, YOLO v1 algorithm, YOLO v2 algorithm, YOLO v3 algorithm, YOLO v4 algorithm, YOLO v5 algorithm, et al [7]-[11]. Different from the two-stage detection algorithm that represented by R-CNN, the YOLO network has a simple structure, and is about 10 times faster than Faster R-CNN, and has better real-time performance. These target detection algorithms have a high accuracy in vehicles and pedestrians detection, but when the traffic is congested, the vehicles and pedestrians occlusion, and the detection target size is too small, the detection accuracy will decrease. In order to solve this problem, the optimization of the YOLO v5 network structure mainly include two aspects. To solve the problems of mutual occlusion between vehicles and pedestrians with complex backgrounds, the attention mechanism is added into the network to improve the feature extraction and feature fusion capabilities, so that the network can select the information that is more critical to the current task goal from a large amount of information. To detect the problem that the size of the target is too small and the small targets are relatively dense, the main purpose is to add a small target detection head to the network to improve the network's detection ability for densely distributed small targets.

2. YOLO v5 Algorithm Model Framework

Compared with YOLO v3 and YOLO v4 algorithms, YOLO v5 has improved the backbone, neck part, activation function and loss function, et al. The network structure of YOLO v5 is shown in the Fig.1.

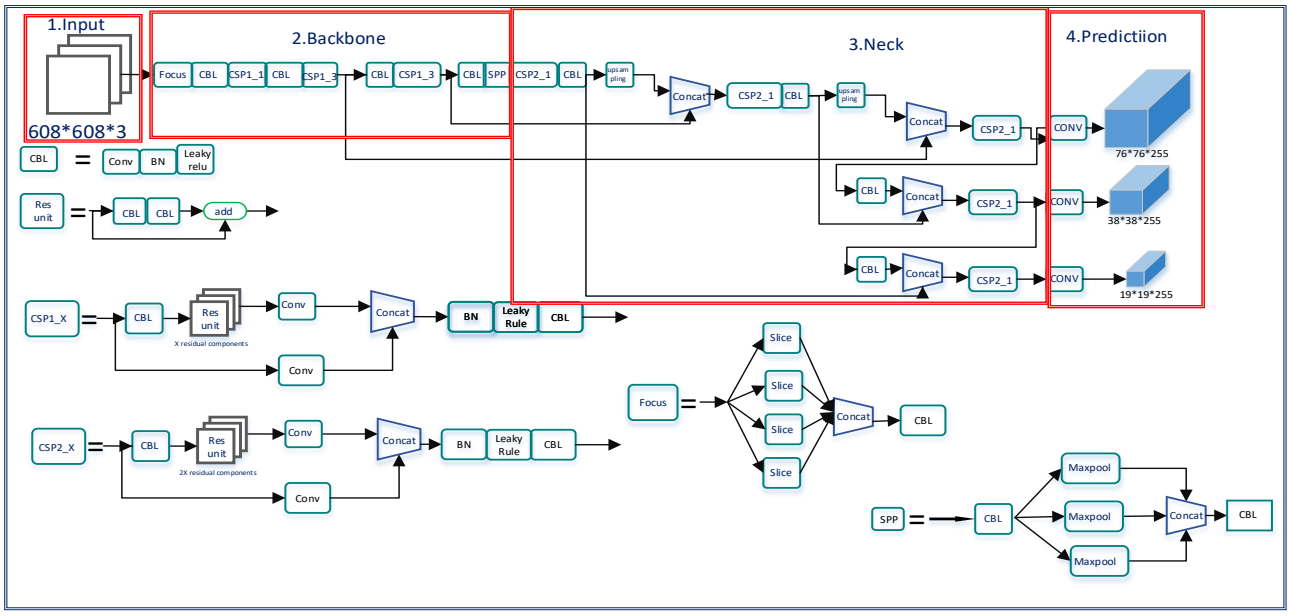

Fig. 1: The network structure of YOLO v5.

The input of the YOLO v5 network contains an image preprocessing stage, which uses adaptive image scaling to scale and normalize the input image,and uses Mosaic data augmentation to improve the accuracy of the network [13]. Backbone of YOLO v5 not only uses the CSPDarknet53 structure, but also take the Focus structure as the benchmark network. The Neck network is located between the benchmark network and the head network, and uses the SPP module and the FPN+PAN module. Using the Neck network can improve the robustness of the algorithm and the diversity of learning features. The head output contains a classification branch and a regression branch for the output of target detection results.

At the input end of YOLO v5, adaptive image scaling is performed in the model inference stage, and the original input image is scaled to a fixed size by black border filling, and then sent to the detection network. The default in YOLOv3 and YOLO v4 the padded value is (0,0,0), while the default padded value in YOLO v5 is (114,114,114). Not only the CSPDarknet53 structure, but also the Focus structure is used in the benchmark network of YOLOv5. Before the picture enters the Backbone, the input picture is sliced by the slice operation. Taking the YOLO v5s model as an example, the image with the original size of 640×640×3 is input into the Focus structure, and the slicing operation is used first to turn it into a feature map of 320×320×12, and then after a convolution operation, it finally becomes 320×320×32.The feature map of 32, the Focus structure of YOLO v5s finally uses 32 convolution kernels, and the slice operation of Focus is different from that of YOLO v4. Two CSP structures, CSPNet (Cross Stage Partial Network), are designed in YOLO v5. CSP1-X structure is used in Backbone network, and CSP2-X structure is used in Neck. CSPNet can ensure the accuracy and reduce the amount of computation by integrating gradient changes into the feature map from start to end.

The learning ability of CNN is enhanced, and the computational bottleneck and memory cost are reduced. The Neck part of YOLO v5 adopts the structure of FPN+PAN. FPN is top-down, and the entire feature pyramid is enhanced by passing down the high-level semantic information. PAN is bottom-up, and the strong localization features of the bottom layer are passed on. The structure of FPN+PAN is shown in Fig.2.

In the Neck structure of Yolo v5, the CSP2 structure is adopted to enhance the ability of network feature fusion. The output of Yolo v5 uses GIoU_Loss as the loss function of the Bounding box:

$$\mathrm{GIoU} = \mathrm{IoU} - \frac{|A_c - U|}{|A_c|} \tag{1}$$

The meaning of the above formula is that first calculate the minimum closure area of the two boxes, then calculate the IoU, then calculate the proportion of the area that does not belong to the two boxes in the closure area to the closure area, and finally subtract this proportion from the IoU is the GIoU.

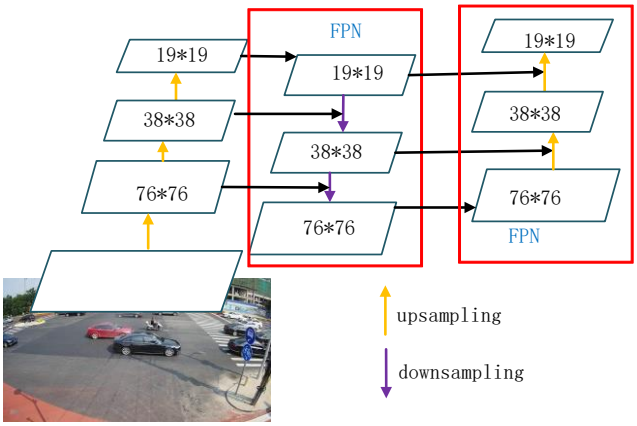

Fig. 2: The structure of FPN+PAN.

3. Improvements of YOLO v5 Algorithm

3.1. Add Small Target Detection Head

The data set used in this paper is a roadside data set produced by Wanji Technology and Beijing Artificial Intelligence Research Institute for the field of intelligent driving. It is collected by a roadside smart base station that independently developed by Wanji Technology. There are a large number of small target pedestrians and vehicles, so on the basis of the original YOLO v5 network, a small target detection head for small targets is added. Combined with the other three prediction heads, the detection accuracy of small targets in traffic scenarios is further improved. In the data set, because the distance between vehicles and pedestrians is different away from the acquisition base station, there are large-scale vehicles, and there are also densely distributed vehicles and pedestrians. Adding one more prediction head on the basis of the original three prediction heads can effectively alleviate the negative impact caused by drastic target scale changes, and is more sensitive to small objects. Although the computational cost and storage cost are increased, the detection performance of small objects has been greatly improved. The improved Neck and prediction part is shown in the Fig.3.

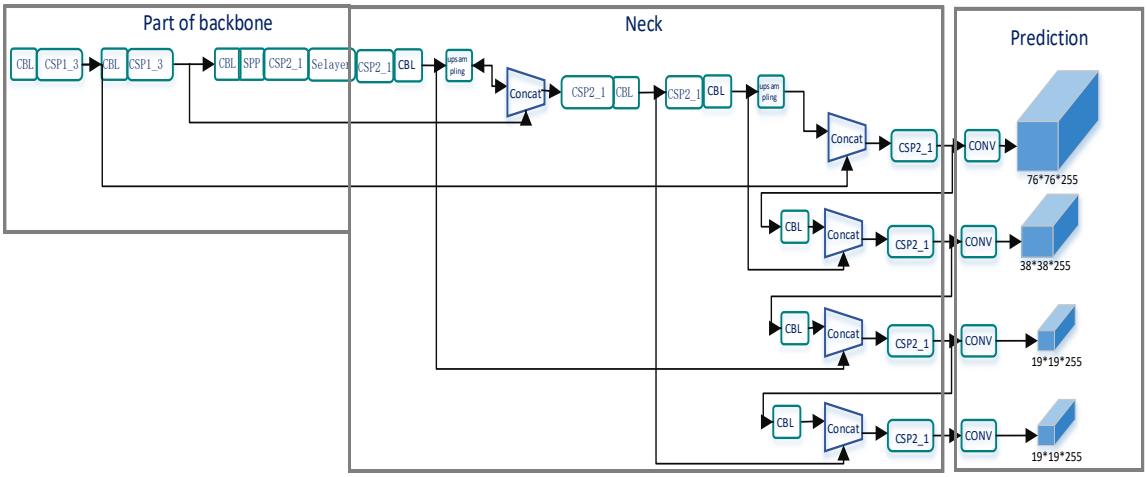

Fig. 3: The improved Neck and Prediction structure.

3.2. Add Attention Mechanism Selayer

In essence, the attention mechanism is similar to the human observation mechanism.The basic idea of the attention mechanism in target detection is to make the model more focused, ignore secondary information,

and focus on the most important information. The essence is to use the relevant feature map to learn the weight distribution, and then apply the learned weight into the original feature map and finally perform a weighted summation,assist the model to assign different weights to each part of the input and extract more important information, which is helpful for the model to make more accurate judgments, and the calculation and storage of the model will not cost more.

According to the different model structures of the attention mechanism, there are three kinds of attention domains of the attention mechanism in computer vision: the spatial domain, the channel domain and the mixed domain. The attention mechanism added into YOLO v5 in this paper is Selayer,an attention mechanism whose attention domain is channel domain [14]. The weights are the same in the plane dimension. In each channel dimension, different weights are learned. The larger the weight, the higher the correlation. Therefore, Selayer ignores the local information of each channel and directly performs global average pooling on the information in one channel. The main implementation step of Selayer is to perform global average pooling on the input feature layer, and then perform two full connections, first connect a smaller number of neurons, and then to fully connect the same number of neurons as the input layer, finally take Sigmoid and fix the value between 0-1 to get the weight of the input feature layer to buy a channel, this weight is between 0-1, and the weight after the final processing can be the original input feature layer.

In this paper, the attention mechanism Selayer is added into the last layer of the Backbone of YOLO v5 , as is shown in the figure, the main function is to classify the target in the process of vehicle pedestrian detection, using different labels, the vehicles and pedestrians in the dataset are divided into 12 types, the YOLO v5 with the attention mechanism has significantly improved the detection effect of vehicles and pedestrians. The improved Backbone structure is shown in the Fig.4.

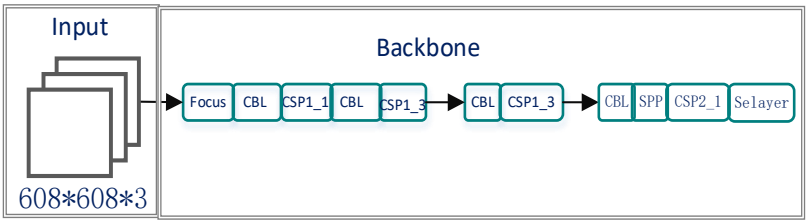

Fig. 4: The improved Backbone.

4. Experiment and Result Analysis

4.1. Data Set

The data set used in this paper is the world's first public large-scale roadside data set released by Wanji Technology and Beijing Artificial Intelligence Research Institute. The data set contains a total of 2,000 photos of real traffic in my country, covering 12 categories of targets [15]. The categories of labels are: Pedestrain, Bicycle, Motorcycle, Tricycle, Car, Van, Cargo, Truck, Bus, Semi-trailing Tractor, Special-vehicle, RoadblockVehicle. 1900 images were randomly selected as the training set and validation set, and the remaining 100 images were used as the test set.

4.2. The Metrics of Model Identification Accuracy

In this paper, P, R, mAP are selected as the measurement metrics of the model recognition accuracy, and the calculation of these metrics are introduced next.

With a confusion matrix, the predicted values were divided into TP (true positives), TP (true negatives), FP (false positive), and FN (false negatives). TP can be understood as the correct prediction for the positive class, TN is the correct prediction for the negative class, FP is the wrong prediction for the positive class, and FN is the wrong prediction for the negative class. According to these indicators, P, R, mAP@0.5 can be calculated.

The precision rate (Precision, P) is defined as the proportion of the amples whose true value is positive among all the targets whose predicted value is positive. It is calculated as shown below.

$$\mathrm{P} = \frac{TP}{TP + FP} \tag{2}$$

The recall rate (Recall, R) is defined as the proportion of the samples whose predicted value is also positive among all the targets whose true value is positive. The recall rate can intuitively show the comprehensive degree of the model.It is calculated as shown below.

$$\mathrm{R} = \frac{TP}{TP + FN} \tag{3}$$

The PR curve is to select different thresholds, and then obtain different combinations of P and R, and then draw the P in each pair of combinations as the ordinate and R as the abscissa. When the difference between the positive and negative samples is relatively large, the performance of the classifier can be reflected. AP@0.5 is the area enclosed by the PR curve and the coordinate axis when the threshold of the confusion matrix is 0.5. It is calculated as shown below.

$$\mathrm{AP@0.5} = \frac{1}{n}\sum_{n}^{1} P_i = \frac{1}{n}P_1 + \frac{1}{n}P_2 + \cdots + \frac{1}{n}P_n \tag{4}$$

The mAP@0.5 is defined as the mean value of AP@0.5 of all categories, which reflects the trend of the model's precision rate with the recall rate. Assuming that the number of categories is C. It is calculated as shown below.

$$\mathrm{mAP@0.5} = \frac{1}{nc}\sum_{n}^{1} P_i = \frac{1}{nc}P_1 + \frac{1}{nc}P_2 + \cdots + \frac{1}{nc}P_n \tag{5}$$

4.3. Experimental Results Analysis

The experimental environment of this paper is as follows: the experimental platform is 64-bit Win10 system, 32G memory, Cuda 10.1, GPU is NVIDIA 1080ti, PyTorch1.8.1.

The results analysis of small target detection. Aiming at the problem of too many small targets in roadside pedestrians and vehicles target detection, a small target detection head is added on the basis of the original 3 detection heads of YOLO v5. There are mainly three types of small targets in the data set, named Pedestrian, Car and Roadblock. For these three small targets, the comparison of metric P is shown in the Table 1. As can be seen from the Table 1,compared with the original YOLO v5, after adding the small target detection head and attention mechanism, the P of Car is increased by 6%, the P of Pedestrian is increased by 8.6%, and the P of Roadblock is increased by 6%.

Table 1: The Comparison of Metric P

	YOLO v5	YOLO v5 with detecte head	YOLO v5 with Selayer	YOLO v5 with Selayer and detect head
Car	85%	90.2%	88.3%	91%
Pedestrian	85.1%	86.4%	81.3%	93.7%
Roadblock	85.1%	86.4%	96.2%	98.4%

The results analysis of all targets with complex background. In view of the problem of inaccurate detection caused by the complex background of occluding targets in the detection of pedestrian and vehicle targets on the side of the road, this paper adds the attention mechanism Selayer to YOLO v5 to solve the problem. In addition to the experimental comparison between several improved YOLO v5, it was also compared with YOLO v3 and YOLO v4. After data cleaning, a total of 10 categories were included. The average value of detection indicators for these 10 categories by different algorithm models are shown in the Table 2.

Table 2: Results of Different Algorithms

Algorithms	Metrics		
	P	Rcall	mAP@0.5
YOLO v3	63.5%	52.4%	54.8%
YOLO v4	72.1%	61.3%	67.5%
YOLO v5	82.9%	70.5%	74.9%
YOLO v5 with detecte head	85.4%	73.9%	79.1%
YOLO v5 with Selayer	85.9%	71.6%	78.9%
YOLO v5 with Selayer and detect head	91.3%	69.6%	80.1%

According to the data in the TABLE II, the improved algorithm is greatly improved compared with YOLO v3 and YOLO v4. Compared with YOLO v5 before the improvement, P and mAP @0.5 are improved by 8.4% and 5.2% respectively.

For heavily occluded vehicles and pedestrians, the improved YOLO v5 model can also accurately detect them.The Fig.5. is the original image, and the target detection result is shown in the Fig.6.

Fig. 5: The image of pedestrians and vehicles blocking each other.

Fig. 6: The detection result of pedestrians and vehicles blocking each other.

5. Conclusion

Aiming at the problem of inaccurate detection caused by too many small targets and occlusion of detection targets in roadside pedestrian and vehicle target detection, this paper improved on the basis of YOLO v5 by adding a small target detection head to alleviate the influence caused by excessive change of target size, so as to improve the sensitivity to small targets. A channel-based attentional mechanism, Selayer, was added to YOLO v5 to make the model pay more attention to primary information and ignore secondary

information. And the pre-training model yolov5s.pt was used to train the network. The improved algorithm is very practical. In the follow-up research, the data set will be improved to be more complex with real traffic scenarios. The trained model will be deployed on the flip-flop driving vehicle to promote the development of intelligent transportation.

6. References

[1] Xuyan Bao, Bingyan Yu, and Wan Jing, “Research on the development status and testing methods of the roadside perception system of the Internet of Vehicles[J], “ Mobile Communication,2021.45(06):43-47.

[2] Yangwei, Xuefeng Du, Zhang Yong, “A Survey of Vehicle Object Detection Algorithms Based on Deep Learning [J], “ Automotive Practical Technology,2022.47(02):24-26.

[3] 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 recog- nition.2014:580-587.

[4] He K,Zhang X,Ren S,et al.Spatial pyramid pooling in deep convolutional networks for visual recognition[J].IEEE tran- sactions on pat- tern analysis and machine intelligence,2015, 37(9):1904-1916.

[5] GIRSHICK R.Fast R-CNN[C]/Proceedings of IEEE Interna-tional Conference on Computer Vision. Washington:IEEE Computer Soc- iety Press.2015:1440-1448.

[6] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J].IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(6) : 1137－1149.

[7] Liu W,Anguelov D,Erhan D,et al.SSD:single shot mul tibox detector[C]/European Conference on Computer Vision-ECCV. Springer:In- ternational Publishing,2016:21-37.

[8] Redmon J,Divvala S,Girshick R,et al.You only look once: unified,real-time object detection[C]/Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016:779-788.

[9] REDMON J, FARHADI A.YOLO9000: better, faster, stronger[C]/ / IEEE Conference on Computer Vision & Pattern Recognition. Las Vegas: CVPR, 2016: 6517－6525.

[10] Redmon J., Farhadi A. Yolov3: An incremental improvement[J]. arXiv Computer Vision and Pattern Recognition, 2018: 2121-2126.

[11] BOCHKOVSKIY A, WANG C Y, LIAO H . YOLOv4: optimal speed and accuracy of object detection [Z/OL].(2020－04－23) [2021－05－20]. https: / / arxiv.org / abs /2004. 10934.

[12] Qian Kun, Chenxun Li,Meishan Chen, Wang Yao, “Ship target and key position detection algorithm based on YOLO V5[J] “, Systems Engineering and Electronics, 2020.13(05):13-17.

[13] Xiaochun Zhu, Zitao Chen, “Helmet wearing detection based on YOLO V5[J], “ Journal of Nanjing Institute of Technology,2021.19(04):7-11.

[14] Jie H , Li S , Gang S , et al. Squeeze-and-Excitation Networks[J]. IEEE TrXansactions on Pattern Analysis and Machine Intelligence, 2017, PP(99).

[15] Yongqiang D , Dengjiang W , Gang C , et al. BAAI-VANJEE Roadside Dataset: Towards the Connected Automated Vehicle Highway technologies in Challenging Environments of China[J]. 2021.