American Journal of Electrical and Computer Engineering
2022; 6(1): 47-53
http://www.sciencepublishinggroup.com/j/ajece
doi: 10.11648/j.ajece.20220601.16
ISSN: 2640-0480 (Print); ISSN: 2640-0502 (Online)

Research on YOLOv5s Highway Pedestrian Detection Algorithm Integrating Attention Mechanism

Zhang Xue, Chang Li*

School of Information Science and Engineering, Shenyang University of Technology, Shenyang, China

Email address:
Zhang_xue1267@163.com (Zhang Xue), changlianli@163.com (Chang Li)
*Corresponding author

To cite this article:
Zhang Xue, Chang Li. Research on YOLOv5s Highway Pedestrian Detection Algorithm Integrating Attention Mechanism. *American Journal of Electrical and Computer Engineering*. Vol. 6, No. 1, 2022, pp. 47-53. doi: 10.11648/j.ajece.20220601.16

Received: May 8, 2022; **Accepted**: June 13, 2022; **Published**: June 16, 2022

Abstract: Pedestrian detection is widely used in daily life, but it is difficult to study in highway environment, such as occlusion overlap. In order to reduce error detection rate of highway pedestrian detection, an algorithm based on YOLOv5s was proposed. Since vehicle occlusion leads to the reduction of effective features of targets, CBAM attention and SE-NET mechanism module is introduced in the network of YOLOv5s to maximize the extraction of effective features. In order to prevent the spatial characteristic information in the trunk network from being damaged, CBAM module is added at the beginning and end of the structure, and SE-Net attention module is added in the neck network, that is, after the detection layer C3 module, the weight information obtained is connected with the subsequent Conv module, so that the model pays more attention to the pedestrian area. Due to low detection accuracy caused by pedestrian overlap. YOLOv5s was designed by combining DIOU_NMS candidate box screening mechanism. The results show that the mean average precision of YOLOv5s (IOU=0.5) increases by 0.48, and the value of Recall of the improved algorithm increases by 0.51 respectively. The improved pedestrian detection algorithm improves the accuracy of target box regression. Thus, the confidence of pedestrian detection is improved. Based on the improvement strategies mentioned above, the detection speed is 32fps, which meets the requirements of real-time detection.

Keywords: Pedestrian Detection, Overlapping Target Detection, YOLOv5s, Attention Mechanism, Non-maximum Suppression

1. Introduction

The most representative pedestrian detection algorithms are statistical learning and background modeling. Pedestrian detection based on statistical learning requires artificial design feature extraction process, and the calculation process is complicated with low accuracy, which does not meet the real-time requirements of pedestrian detection.

Background modeling is often used to extract moving foreground, but it is not applicable when pedestrian targets are dense and overlapping. For example, HOG was proposed by Dalal in 2005, generated HOG feature set and performed well in MIT pedestrian detection dataset combined with SVM classifier [1]. Nam realized pedestrian detection by calculating the channel feature map and eliminating the local correlation of each feature map [2]. Gavrila proposed a pedestrian template matching algorithm with a large amount of calculation and low detection accuracy [3]. Zhou focused on the problem of pedestrian re-recognition by strengthening the features of visible parts and introducing the Transform module to increase the gap between positive and negative samples, and adopted the two-stage algorithm with high detection accuracy but slow speed [4]. The single-stage detection algorithm directly regresses the position and confidence of the target at the prediction layer, such as YOLO series designed by Redmon [5] and SSD designed by Liu in 2016 [6]. An increasing number of researchers such as Deng Jie, Zhuang use single-stage algorithms to detect pedestrian [7, 8].

However, the one-stage algorithm is still weak in generalization and only applicable to general scenarios [9]. In the case of dim light, long distance, pedestrians gathering and occlusion, the missed detection rate is still very high, and the robustness of the model needs to be further improved to deal with complex real scenes.

2. Materials and Methods

2.1. Pedestrian Detection Model Based on Improved YOLOv5s

Convolutional Block Attention Module [10] is designed for Convolutional neural network, which is simple and efficient. CBAM computes adaptive learning features from both channel and spatial dimensions, multiplies attention graphs with feature graphs and reassigns weights. High-weight features are the focus of attention.

SE-Net was proposed by Hu et al. The SE-Net contains squeeze, excitation, and scale. SE-Net proves that feature channels are interdependent, and such cross-correlation information is applied to the distribution of feature weight to reduce the influence of unimportant features on network learning [11].

Figure 1 shows, CBAM attention mechanism module is introduced in the backbone network of YOLOv5s.

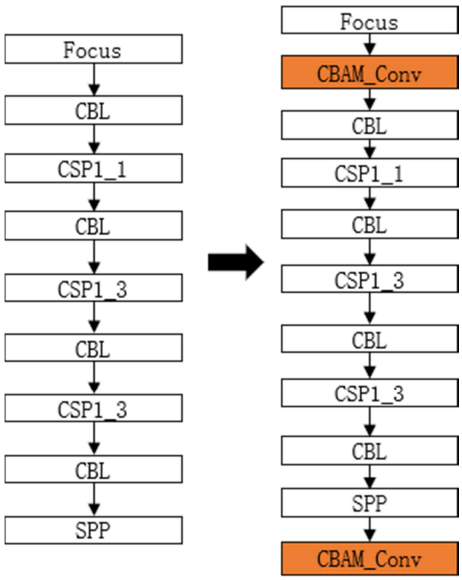

Figure 1. *Introduces CBAM into the backbone network.*

It is necessary to increase the weight of pedestrian features in the model for maximizing extraction from limited features. Therefore, deep learning and adding attention mechanism are used to make the model pay more attention to pedestrian areas and extract deeper pedestrian features. The specific method is to add CBAM convolution block attention module in the first and last layer of YOLOv5s backbone network, and introduce SE-Net separately after the C3 detection layer of YOLOv5s neck network.

Figure 2 shows that follow steps squeeze, excitation and scale, the SE-NET module was introduced into the neck network.

Firstly, feature vectors (256, 512, 1024) of the three channels output by the original YOLOv5s network detection layer are compressed by global pooling method, that is, feature graph U ($H \times W \times C$) is transformed into $1 \times 1 \times C$, and the global information between the features of each channel is obtained. Then the correlation model between channels is composed of two fully connected layers to establish the nonlinear transformation characteristics between and channels and then output weight information [12, 13].

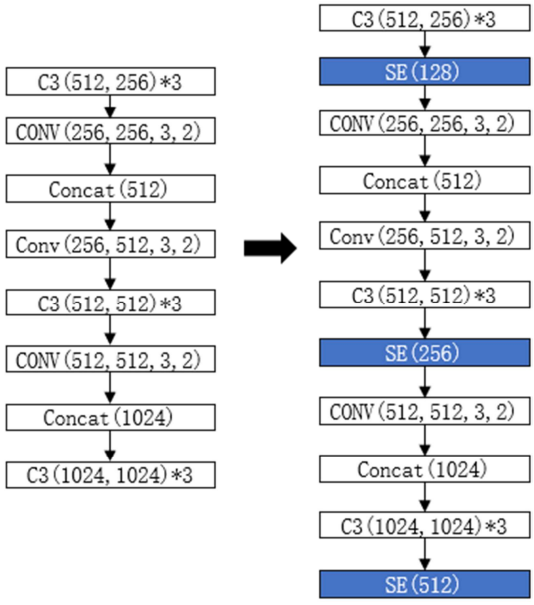

Figure 2. *Introduces SE-Net into the neck network.*

Figure 3 shows, in this paper, sample images of dense crowds under a complex background are selected, and Grad-CAM [14] is used to visualize the features of attention mechanism figure.

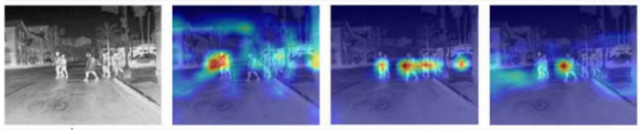

Figure 3. *Grad-CAM network visualization results.*

The high places represent the places where the network pays more attention, and vice versa. In Figure 3(b) is the original without the attentional mechanism, it can be found that its concerns are scattered and it contains more irrelevant and complex background information. Add attention in Figure 3(c) and 3(d), the network can effectively distinguish the surrounding complex background interference information and focus attention on the target. Visualization results show that, the attention mechanism is added to improve the attention of pedestrian target effectively.

2.2. Non-maximum Suppression

Weighted Non-Maximum Suppression is taken as the core part of post-processing in YOLOv5s, whose function is to screen the huge number of candidate boxes generated by the target detection model in the previous several stages. As shown in Figure 4 by sorting the confidence score of candidate frames, the frame with the highest confidence is selected and the rest of the frames are iterated. If the IOU

with the candidate frame with the highest confidence score is greater than a certain threshold, as shown in Figure 5, it will be deleted. Continue to select the one with the highest score among the unprocessed boxes and repeat the process until each target is selected by a candidate box [15].

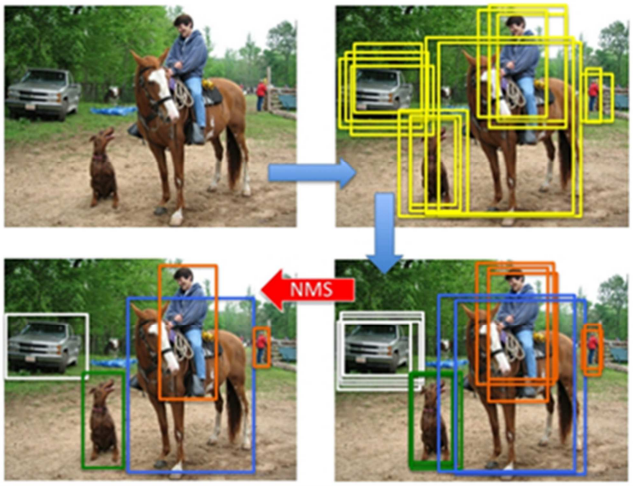

Figure 4. *Candidate box filtering process.*

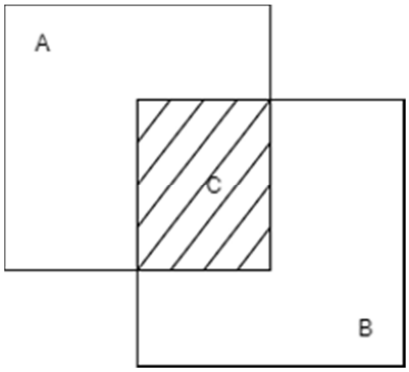

Figure 5. *IOU schematic.*

Where, A represents the rectangle at the upper left corner, B represents the rectangle at the lower right corner, C represents $A \cap B$, and D represents $A \cup B$, then the intersection ratio of A and B can be expressed as formula (1), namely.

$$IOU = \frac{C}{D} \tag{1}$$

When the value of intersection ratio is larger, which indicates that the area of intersection of candidate boxes is larger, the prediction box in the actual detection result is closer to the real mark box of the object. Generally speaking, with the continuous iteration of the algorithm, the intersection ratio of the algorithm gradually approaches 1, and even reaches 1.

However, for some pedestrian images with overlapping occlusion, when the IOU threshold is applied to traverse, due to the scale invariance of IOU, the maximum score box overlaps too much with another target candidate box and is deleted, which results in missing detection and affects the accuracy of the statistics of the number of people in the tunnel. Figure 6 is a schematic diagram of DIOU.

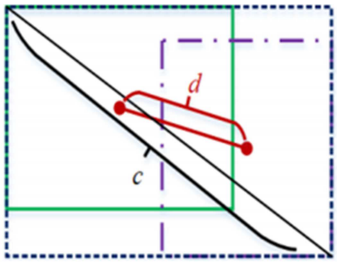

Figure 6. *Schematic diagram of DIOU.*

Where ρ represents the Euclidean distance between the central point coordinates of frame A and frame B, and c is the diagonal distance covering the minimum square.

$$DIOU = \frac{\rho^2(A,B)}{c^2} \tag{2}$$

DIOU_NMS takes DIOU as the criterion of NMS. For the prediction box with the highest confidence score Y.

$$s_i = \begin{cases} s_i, & IOU - R_{DIOU}(Y, B_i) < \varepsilon \\ 0, & IOU - R_{DIOU}(Y, B_i) \geq \varepsilon \end{cases} \tag{3}$$

The Euclidean distance between the IOU and two enclosures must be considered in the filtering process B_i. s_i represents the classification score and ε is the NMS threshold.

3. Details

3.1. Experimental Environment

All simulation experiments are based on Intel (R) Core (TM) i7-10750h CPU@2.60GHZ processor, 16GB memory, operating system WIN10, programming language Python3.6, using Pytorch1.9.0 framework. The maximum iteration is 200. Specific training parameters are shown in table 1.

Table 1. *Training parameters.*

Network parameters	Parameter value
momentum	0.949
Initial learning rate	0.0013
Maximum iteration	200
Attenuation coefficient	0.0005
Learning rate	240,270

3.2. Data Set Establishment

USC pedestrian detection test set is adopted in this paper. This data set consists of three parts, USC-A, USC-B and USC-C, with a total of 359 images and 816 pedestrian targets with different angles and degrees of occlusion. According to the ratio of 8:1:1, this data set is divided into training set, verification set and test set. Part of the USC data set is shown in Figure 7.

Figure 7. *Pedestrian detection data set.*

The specific distribution is shown in table 2.

Table 2. *USC data set overview.*

data type	USC-A	USC-B	USC-C
picture	205	54	100
Pedestrian	313	271	232
Angle	Front or back	Various angles	Multiple angles
Degree of occlusion	No occlusion	Mutual occlusion	No occlusion

VOC format is adopted for model training in this paper, and images are marked and xml files are generated, as shown in Figure 8.

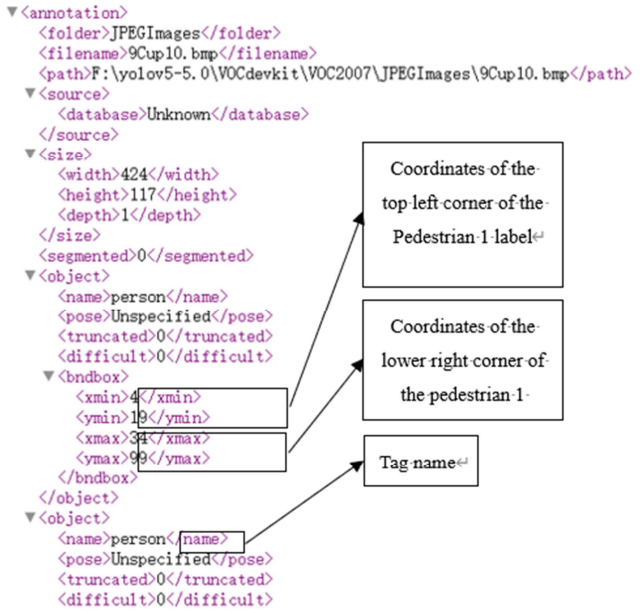

Figure 8. *Data annotation file.*

3.3. Pedestrian Detection Results

Figure 9 shows, YOLOv5s uses the result diagram of IOU_NMS and DIOU_NMS for pedestrian detection. In the position indicated by the yellow arrow in the figure, two pedestrians are highly overlapped, and DIOU_NMS can select both targets.

Figure 9. *Pedestrian detection results of DIOU_nms.*

Figure 10 shows some test results. The left figure is the original figure, the right figure is the result diagram of detecting the traveler's location and confidence, and the test result diagram other than USC data set.

Figure 10. *Pedestrian detection results.*

As shown in the table 3, the detection results of 7 images are compared. The number of original images represents the number of people obtained by observing images before detection, and the number of detected images represents the number of people detected by the algorithm. Among them, images 4, 5, 6 and 7 contain different degrees of occlusion and overlap.

Table 3. *Comparison of the number of test results.*

Image sequence number	original images	detected images
1	3	3
2	5	5
3	6	6
4	11	11
5	8	8
6	9	9
7	10	10

3.4. Discussions

Comparison was made between YOLOv5s before improvement and various improvement schemes. Recall, Precision and mAP were used as evaluation indexes of model detection in this research stage [16].

Figure 11 represents the Precision curve, and the horizontal coordinate represents the confidence of pedestrian detection. The left figure represents the Precision of YOLOv5s model, and right figure represents the Precision of the algorithm proposed in this paper. The figure shows that the Precision rate increases by 0.23 before and after improvement.

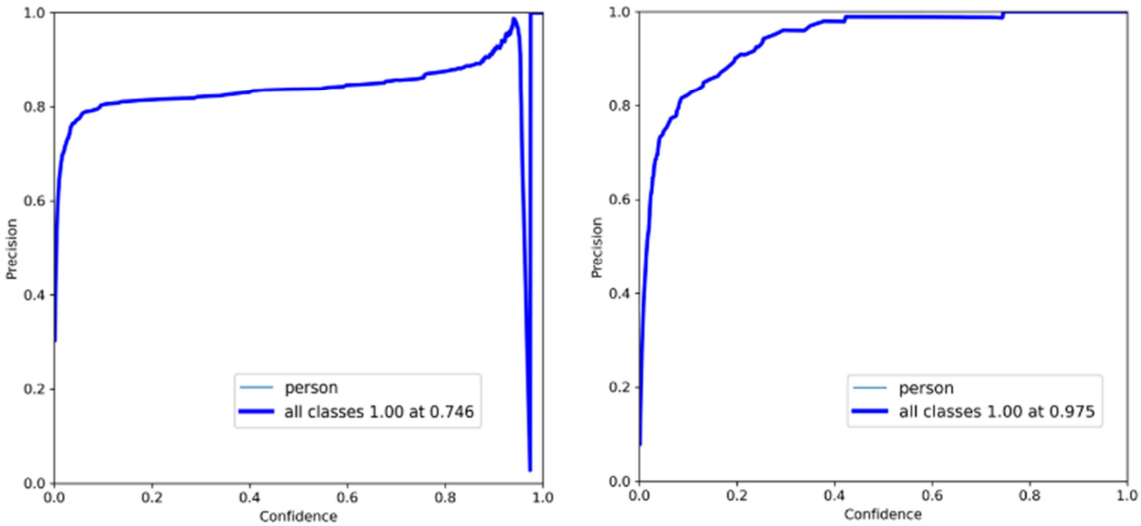

Figure 11. *Comparison of curves of precision.*

Figure 12 shows the comparison of Recall of the two algorithms. The figure shows that the Recall increases by 0.51 before and after improvement.

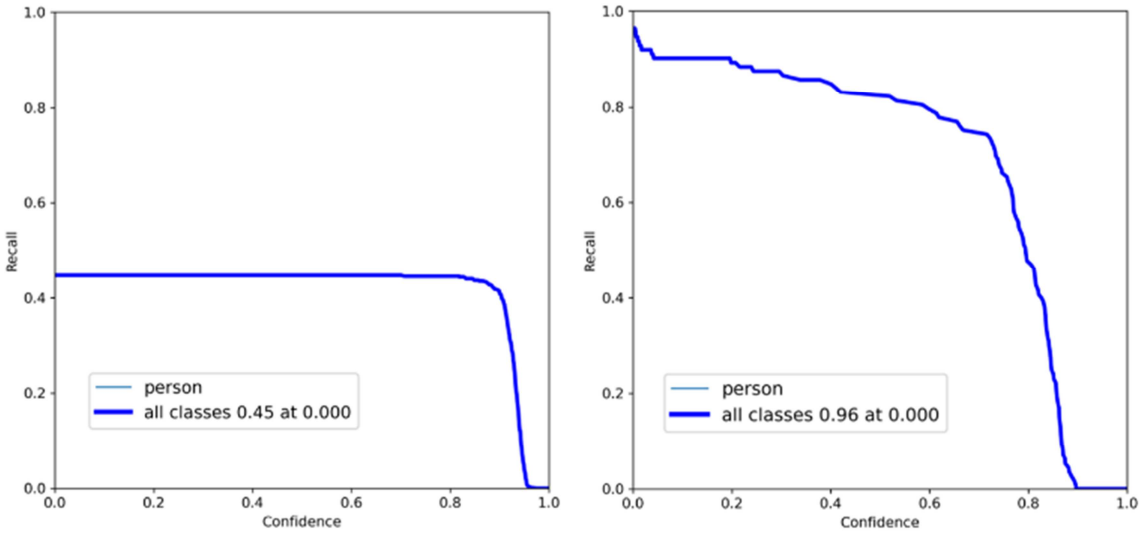

Figure 12. *Comparison of curves of Recall.*

P-R curve can be fitted according to Figure 11 and Figure 12, as shown in Figure 13.

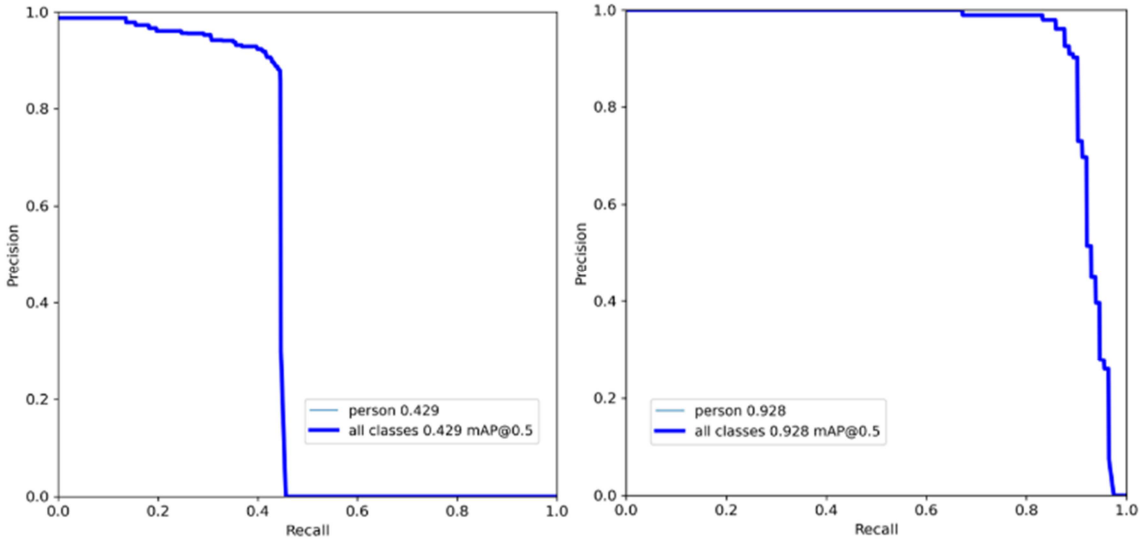

Figure 13. *Comparison of curves of P-R.*

It is clearly shown in the Figure 14 that mAP increases by 0.48 when the threshold value is 0.5 before and after improvement. In the threshold range of 0.5-0.9, the value increases by 0.12.

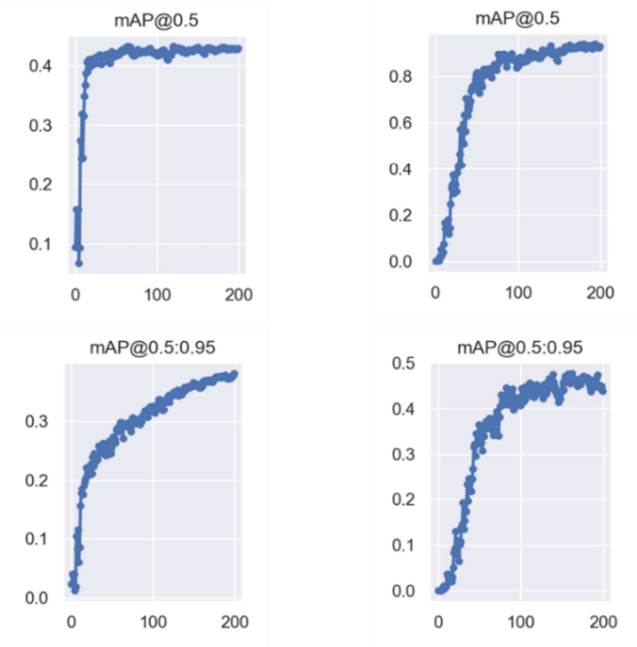

Figure 14. *Comparison of curves of mAP.*

Table 4 shows, in addition to the comparison with various performance indicators of the original YOLOv5s, the improved algorithm in this paper is also compared with common flame detection algorithms such as Faster R-CNN and YOLOv3 in other literatures. In a unified data set and the same operating environment, the detection accuracy of the proposed algorithm reaches 0.93, which is better than other algorithms. In terms of detection speed, it reaches 32fps.

Table 4. *Comparisons of results of five algorithms.*

Algorithm	mAP@0.5	Recall	FPS
Faster R-CNN	0.58	0.47	5
GMM	0.71	0.85	1
YOLOv3	0.38	0.46	29
YOlOv5s	0.43	0.45	30
This paper	0.93	0.96	32

4. Conclusion

This paper studies the detection of pedestrian targets by the target detection algorithm YOLOv5s in the highway environment. By introducing the attention mechanism CBAM and SE-Net, and combining DIOU to improve the non-maximum suppression candidate box screening mechanism, the experimental results show that the introduction of the attention mechanism makes the model pay more attention to the pedestrian area, reduces the interference of the background, and then reduces the error detection rate of the algorithm, The introduction of DIOU improves the robustness and accuracy of the algorithm to overlapping pedestrian targets. Because the algorithm is improved only for the slight occlusion and overlap of highway pedestrian detection, but in the face of the serious occlusion of pedestrians caused by complex environment, it involves small target multi-target detection. The algorithm studied in this paper still has some deficiencies, so the small target detection is the focus of the next research.

References

[1] Dalal N, Triggs B. Histograms of oriented gradients for human detection [C]//2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). IEEE, 2005, 1: 886-893.

[2] Nam W, P Dollár, Han J H. Local decorrelation for improved detection [J]. Advances in Neural Information Processing Systems, 2014, 1: 424-432.

[3] Gavrila D M, Munder S. Multi-cue pedestrian detection and tracking from a moving vehicle [J]. International Journal of Computer Vision, 2007, 73 (1): 41-59.

[4] Zhou C, Yang M, Yuan J. Discriminative feature transformation for occluded pedestrian detection [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 9557-9566.

[5] 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.

[6] Liu W, Anguelov D, Erhan D, et al. SSD: Single shot multibox detector [C]//European Conference on Computer Vision. Springer, Cham, 2016: 21-37.

[7] Deng J, Wan W G. Dense pedestrian detection based on improved YOLOv3 [J]. Electronic Measurement Technology, 2021, 44 (11): 90-95.

[8] Ahmed Z, Iniyavan R. Enhanced vulnerable pedestrian detection using deep learning [C]//2019 International Conference on Communication and Signal Processing (ICCSP). IEEE, 2019: 0971-0974.

[9] Zhuang C, Li Z, Zhu X, et al. SAD et: learning an efficient and accurate pedestrian detector [C]//2021 IEEE International Joint Conference on Biometrics (IJCB). IEEE, 2021: 1-8.

[10] 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.

[11] Hu J, Shen L, Sun G. Squeeze-and-excitation networks [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 7132-7141.

[12] Lin X Z, LIU J, Tian S, et al. Image Description Generation Method based on multi-space mixed attention [J]. Computer Applications, 2020, 40 (4): 985-989.

[13] Cheng M Y, Gai S Y, DA F P. Research on stereo matching network based on attention mechanism [J]. Acta Optica Sinica, 2020, 40 (14): 138-146.

[14] Selvaraju R R, Cogswell M, Das A, et al. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization [J]. International Journal of Computer Vision, 2020, 128 (2): 336-359.

[15] Felzenszwalb P F, Girshick R B, McAllester D, et al. Object detection with discriminatively trained part-based models [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 32 (9): 1627-1645.

[16] Yang H, Quan J C, LIANG X Y, et al. Progress in Object Detection Based on Weakly supervised Learning. Computer Engineering and Applications, 2021, 57 (16): 40-49.