

SSD-Resnet50: Research on Pedestrian Detection Technology

Baolong Xu*, Changyu Zhao, Junhua Zhao

School of Computer Science, Qufu Normal University, Rizhao, China, 276826

*Corresponding author: 18363000207@163.com

Abstract. In the era of artificial intelligence, with the rapid development of computer hardware, the defects of neural network algorithm have been further improved, the technology of artificial intelligence is progressing, the algorithm and technology of target detection have been further improved, and pedestrian detection technology is an important research direction of target detection technology. The main research content of this paper consists of three parts: the improvement of the original SSD algorithm, and the fundamental composition and characteristics of the neural network; Changing VGG16, the basis of the original SSD algorithm, to Resnet50 can improve the detection speed and accuracy. A feature fusion module is introduced to fuse the feature map with large receptive field and small receptive field, so as to improve the overall expression ability of the feature map to the image.

Keywords: Pedestrian Detection, SSD, Resnet50, Feature Fusion.

1. Introduction

Pedestrian detection has very common applications in many scenarios, such as intelligent monitoring, intelligent traffic, driverless cars, and human behavior analysis and prediction. In 2016, Alphago was born unexpectedly, making the development of artificial intelligence attract people's attention again, ushering in another boom. The application of deep learning to detect pedestrians is a very important direction of the development of artificial intelligence. Pedestrian detection and feature extraction is a very difficult technology in computer vision processing. In the earliest target detection methods, feature selection is mostly completed based on visual features. In addition, the algorithm training is limited by the lack of data set. Background interference is also a common phenomenon in detection, such as traffic lights, trees, etc., may be wrongly detected as pedestrians.

The task of pedestrian detection is to use the pedestrian detection algorithm to judge how many pedestrians exist in a picture, and return to the position space of the pedestrians if there are any. In 2012, Alex Net appeared in the application of image classification and won the first prize in ILSVRC visual recognition competition. Since then, deep convolutional neural network has been widely used in the research of object detection technology. Since then, the processing of computer vision problems has also made great progress, target detection technology has also been a breakthrough development, just as said before, the main application of target detection and target positioning is pedestrian detection. R-CNN algorithm also plays an indispensable role in pedestrian detection technology. Its algorithm process can be divided into four steps: the first step is regional nomination. The first step is to generate 1000~2000 candidate regions. The second step is the normalization of the region size: scale the 2000 candidate region frames to 227×227 , and then extract the features of the candidate frames obtained in the first step using the deep network, which is the picture classification network. The third step is to transfer the feature vector obtained in the second step to the SVM classifier. The function of this step is to eliminate the stack suggestion box for each category by using non-maximum suppression. The fourth step is to use the regressor to refine the candidate frame, that is, to adjust the position of the candidate frame, because the regional candidate frame obtained by ss algorithm is not so accurate [1].

SSD algorithm is a very effective detection algorithm proposed after the emergence of deep learning. SSD is only detected on its top layer and is a single-level target detection algorithm [2]. SSD algorithm refers to a variety of data sets, and at the same time carries out detection tests on a variety of images with different resolutions, effectively improving the accuracy of target detection.

Moreover, it is a very useful and simple single-stage target detector. Regarding the detection of some small targets, SSD has some highlights in the detection accuracy and detection speed. Most of the data sets used in the detection are PASCAL voc, MS-COCO. PASCAL voc not only contains data for object detection, but also image classification, semantic segmentation, and motion detection. In addition, MS-COCO also includes many small objects and more dense positioning objects.

To sum up, deep learning algorithm is used for pedestrian detection in this paper, and SSD pedestrian detection algorithm based on convolutional neural network is mainly used.

2. Research method

2.1. Overview of convolutional neural networks

After the back propagation algorithm was proposed, neural network has been widely used in the study of machine learning. Since neural network contains a large number of parameters, it often has defects of overfitting and too long training time. However, traditional methods such as Boosting, Logistic regression and SVM still have some great advantages. The three main operations of convolutional neural network are local sensitivity field, weight sharing and pooling layer. Through these operations, the number of some network parameters can be effectively reduced, and the network has a certain stability, and the influence caused by overfitting defects can be reduced. FIG. 1 shows a simple convolutional neural network structure [3].

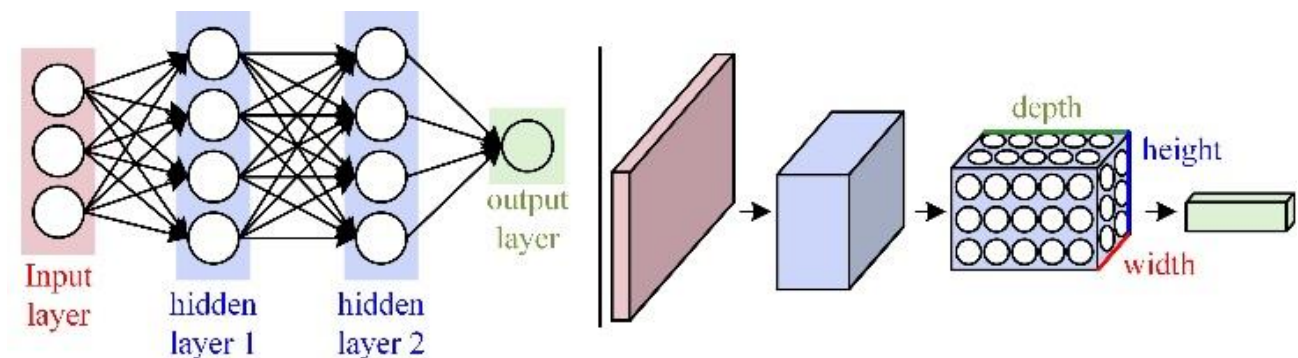


Figure 1. Simple convolutional neural network

Local receptive field: The connection between all things in each picture is not integral and inseparable, and the spatial connection between them is only part of each other. The overall information can be obtained at last without feeling processing all images. This method can effectively reduce the number of parameters required to be connected, greatly reduce the amount of computation and improve the speed of computation [4].

2.2. Setup of SSD network

SSD algorithm has relatively fast detection speed and accuracy, and its specific structure is shown in Figure 2. The SSD algorithm runs regression operations on multiple predictive feature maps, so much of the subtle information included in the underlying feature maps is also taken into account. The most prominent operation of the SSD is to predict the categories of many boundary boxes and calculate the category score through the IOU, as well as the offset that the boundary boxes may produce, and to apply a very small convolution filter to process the predicted feature map. The original SSD algorithm is based on VGG16 as the basic network architecture. Later, several convolutional layers are added on the basis of VGG16 neural network to obtain more predictive feature maps. In this network, the Conv5_3 of VGG16 was used, and then the full connection layer 6 and 7 were processed to convert them into convolutional layers. Resampled parameters from the full connection layer 6 and 7, and the pooled layer 5 was changed from the convolution kernel of 2×2 steps to the convolution kernel of 3×3 steps to 1, and all the discard layer and the full connection layer 8 were deleted.

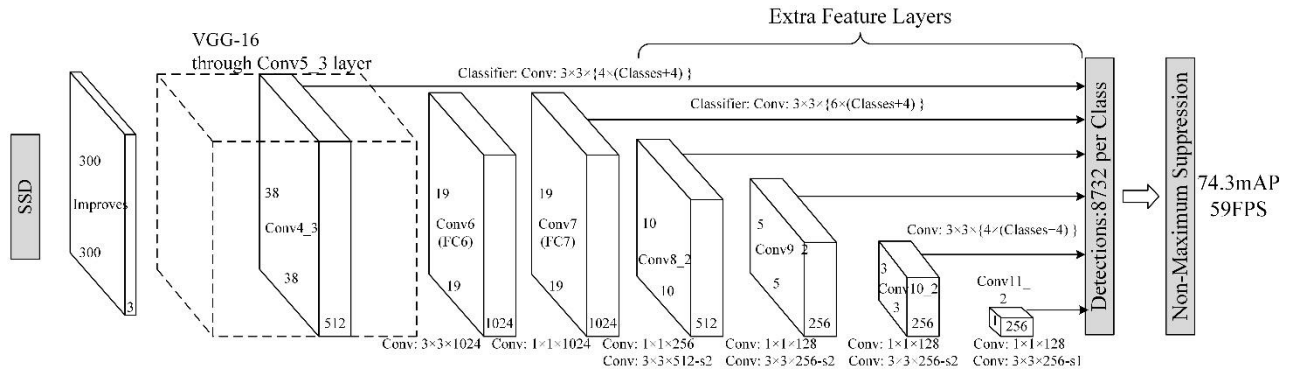


Figure 2. Original SSD network architecture

2.2.1 Generation of the default bounding box

At prediction time, the network generates a score for each category's default bounding box. Two eigenmatrices, one is an 8×8 matrix and the other is a 4×4 matrix. Compared with the 4×4 eigenmatrix, the 8×8 eigenmatrix has a lower degree of abstraction, so it retains more details. Then, the 8×8 feature layer is used to predict small targets on the feature matrix of the relative lower layer, and then the default boundary box can be better matched with the real boundary box. For 4×4 feature layer, it is suitable to detect relatively larger targets. Table 1 shows the dimensions and corresponding proportions of default boundary boxes at each level. There will be two values in the dimensions of each default boundary box, which is a case of our scale one. An extra default boundary box will be added on each feature layer. The reason why the network made them generate 4 default boundary boxes in conv4_3, conv10_2 and conv11_2 is that we removed the ratio of 1:3 and 3:1, and the remaining prediction feature layers all had 6 default boundary boxes, as shown in Table 1 and Figure 3.

Table 1. The default bounding box setting

Feature layer	Width and height of the feature layer	Default box size	Default number of boxes
Feature layer 1	38×38	$21\{1/2, 1, 2\}; \sqrt{21 \times 45}\{1\}$	$38 \times 38 \times 4$
Feature layer 2	19×19	$45\{1/3, 1/2, 1, 2, 3\}; \sqrt{45 \times 99}\{1\}$	$19 \times 19 \times 6$
Feature layer 3	10×10	$99\{1/3, 1/2, 1, 2, 3\}; \sqrt{99 \times 153}\{1\}$	$10 \times 10 \times 6$
Feature layer 4	5×5	$153\{1/3, 1/2, 1, 2, 3\}; \sqrt{153 \times 207}\{1\}$	$5 \times 5 \times 6$
Feature layer 5	3×3	$207\{1/2, 1, 2\}; \sqrt{207 \times 261}\{1\}$	$3 \times 3 \times 4$
Feature layer 6	1×1	$261\{1/2, 1, 2\}; \sqrt{261 \times 315}\{1\}$	$1 \times 1 \times 4$

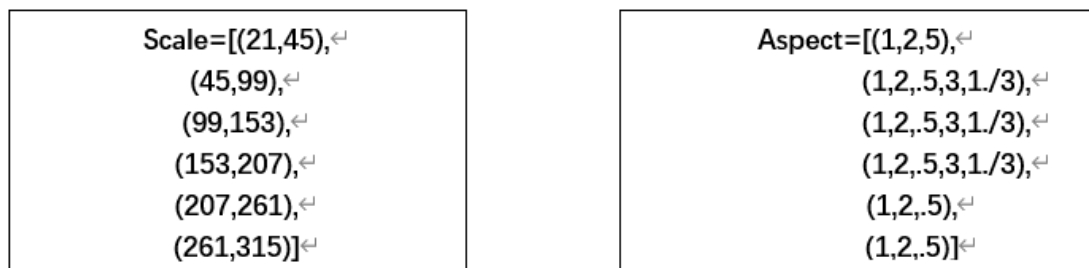


Figure 3. The size setting of the default bounding box

2.2.2 The implementation of a predictor

For the prediction feature layer with height and width $m \times n$ and depth p , the convolution kernel with depth p and convolution kernel with 3×3 can be directly used to realize the prediction of boundary box regression parameters. Then, the probability fraction and the coordinate offset relative to the default boundary box can be generated by 3×3 convolution. For each position in the feature map, k

default bounding boxes are generated to compute c class scores and 4 coordinate offsets. So you have to convolve the picture with c plus 4 times k convolution kernels. If $(c+4) \times k \times m \times n$ output values are generated relative to an m by n size prediction feature layer [5].

2.2.3 The selection of positive and negative samples

There are two matching criteria for the selection of positive samples. The first matches every actual boundary box and selects the default boundary box with the largest IOU value. This is the first matching criterion, and the second matching criterion is for any default boundary box. As long as the IOU value is greater than 0.5 with any actual boundary box, it is also considered as a positive sample [6]. In addition to positive samples, other samples can be classified as negative samples. However, in the training process, it will appear that the number of positive samples of the actual boundary box matched by the default boundary box is very small, which is basically several to a dozen. Therefore, if the rest of the default bounding box is used as a negative sample for training, it will lead to sample imbalance. And one of the choices with negative samples is to use such a strategy. First of all, for all the remaining negative samples, is its maximum confidence loss. If the confidence is lost, the greater the value, the greater the probability that the negative sample error will be detected as the target, which is unacceptable. Therefore, some negative samples in the front line are selected according to the calculated confidence loss [7].

2.2.4 Setup of SSD network

There are two losses, the first is category loss, the second is location loss. N is the number of positive samples that are matched, and α has a fixed value of 1.

1) Class loss

$$L(x, c, l, g) = \frac{1}{N} (L_{\text{conf}}(x, c) + L_{\text{loc}}(x, l, g)) \quad (1)$$

$$L_{\text{conf}}(x, c) = -\sum_{i \in \text{Pos}} x_{ij}^p \log(\hat{c}_i^p) - \sum_{i \in \text{Neg}} \log(\hat{c}_i^0) \text{ where } \hat{c}_i^p = \frac{\exp(c_i^p)}{\sum_p \exp(c_i^p)}; \quad (2)$$

$\hat{c}_i^p$ is the probability p of the predicted i th default boundary box corresponding to the actual boundary box;

$x_{ij}^p = \{0, 1\}$ is the JTH actual boundary box matched by the i th default boundary box (category is p);

2) Positioning loss

The positioning loss is only for the positive sample. Since the negative sample has no actual boundary frame, it is impossible to calculate its positioning loss, and it is meaningless. x_{ij}^k is meaningless, as 1 is OK.

$$L_{\text{loc}}(x, l, g) = \sum_{i \in \text{Pos}} \sum_{m \in \{cx, cy, w, h\}} x_{ij}^k \text{smooth}_{L1}(l_m^i - \hat{g}_j^m) \quad (3)$$

$$\hat{g}_j^{cx} = (g_j^{cx} - d_i^{cx})/d_i^w \quad \hat{g}_j^{cy} = (g_j^{cy} - d_i^{cy})/d_i^h \quad \hat{g}_j^w = \log\left(\frac{g_j^w}{d_i^w}\right) \quad \hat{g}_j^h = \log\left(\frac{g_j^h}{d_i^h}\right) \quad (4)$$

$$\text{smooth}_{L1}(x) = \begin{cases} 0.5x^2 & \text{if } |x| < 1, \\ |x| - 0.5 & \text{otherwise} \end{cases}; \quad (5)$$

l_m^i is the regression parameter corresponding to the i th positive sample regression parameter;

d_i^{cx} corresponds to the x -coordinate of the center point of the i -th default bounding box;

$\hat{g}_j^m$ is the regression parameter of the JTH actual boundary frame matched by positive sample i ;

2.3. IOU

The non-maximum suppression culls the overlapping suggestion box by means of IOU, which is an intersection ratio of the prediction probabilities of the two target boxes. The mathematical expression is $\frac{A \cap B}{A \cup B}$, as shown in Figure 4. For each category, it is necessary to first find a target with a

high score, and then calculate the IOU value of other target boxes and this target box after finding the target with a high score. Then a judgment is made on the IOU value of each boundary box and the boundary box with the highest score. If the value is greater than a given threshold, it will be deleted. Then the target with the highest score is saved, and then the target with the highest score is found in the remaining boundary boxes. Then the calculation is carried out in accordance with the process just described, and the deletion operation is performed. Finally, all boundary boxes are iterated [8].

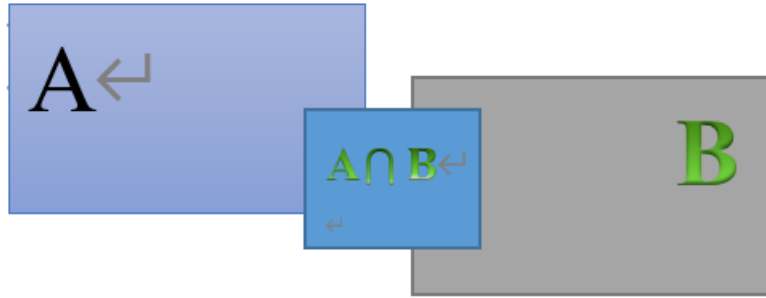


Figure 4. IOU calculation area diagram

3. SSD Pedestrian detection based on Resnet50

Resnet50 was used as the foundation to build the SSD network. The network structure was shown in Figure 5, which improved the accuracy of the detection of small pedestrians and only retained the layer structure before conv4. For the conv4 layer structure, the step spacing of the convolution is changed to a size of 1 by 1. The experimental results show that this design improves the speed and accuracy of pedestrian detection.

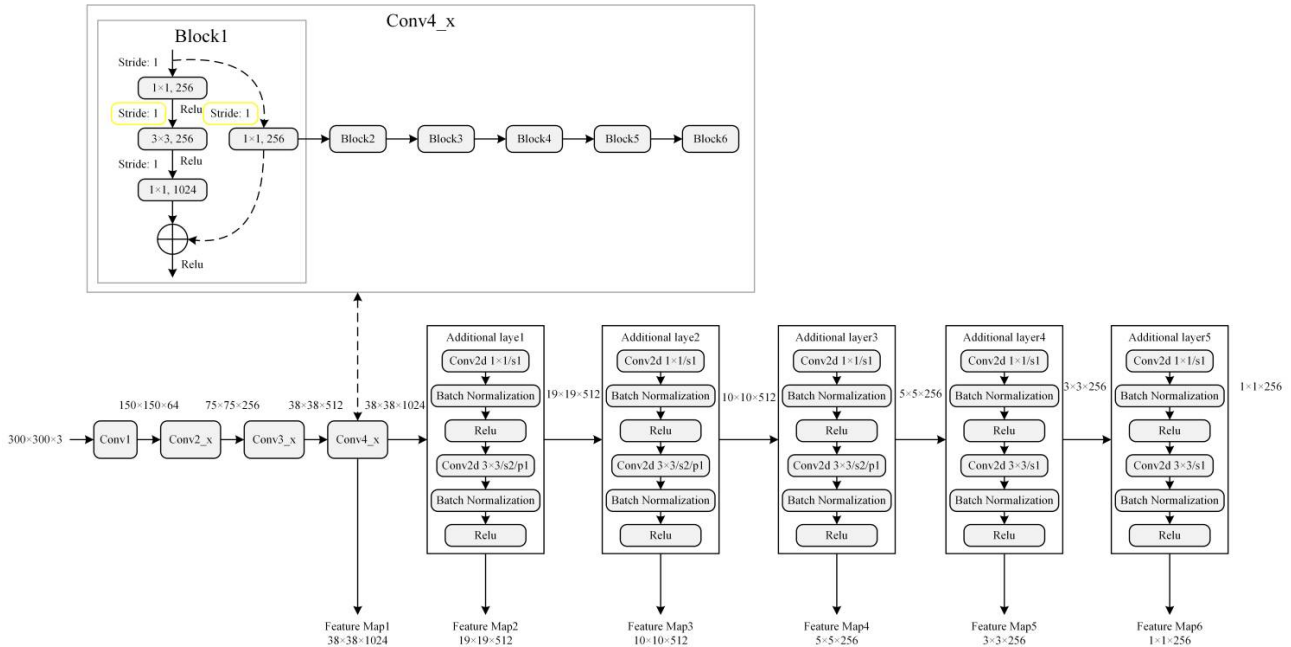


Figure 5. Improved SSD network architecture

3.1. Algorithm design background

Hierarchical feature extraction is the main idea of SSD network, and then pedestrian detection and border regression. The basic SSD network application VGG16 extracts pedestrian characteristic information. VGG16 uses a smaller convolution kernel, which is conducive to stacking more layers, establishing a deep network structure, and improving the visual performance. With the rapid development of deep learning, the construction of deep convolutional neural network has become the main direction of deep learning. Studies have proved that deepening the network structure is effective

in improving the detection of pedestrian features. However, vgg16 belongs to a relatively shallow network, and the feature extraction of pedestrians is not sufficient. In this paper, Resnet50 is used to replace vgg16 in order to improve the overall performance of the algorithm.

3.2. Feature extraction network

Resnet50 adopts residual structure to increase the number of layers in the network without overfitting. In the process of stacking of convolutional layers, 1×1 convolution check is used to effectively reduce the dimension of the original image, and then 3×3 convolution kernel is used for feature extraction. The purpose of this is to effectively reduce the number of network parameters.

3.3. Feature map fusion module

Traditional feature fusion: only the features of the last layer of the network are adopted;

Image pyramid: transform the original image into differentiated image, and then carry out feature extraction detection;

Multi-scale feature fusion: Feature fusion can be carried out during detection. Just like SSD used in this paper, it makes use of such multi-scale feature fusion mode, without using the step of up-sampling. Instead, it selects features with different scales from the prediction feature map generated in the network and then carries out fusion.

The SSD algorithm generates prediction feature maps of six scales at the sixth layer. The deep feature map extracts a lot of semantic information from the original image and has strong feature interpretation ability, but lacks target details. Due to the small sensitivity field of the shallow feature map to the original image, it can effectively detect some small targets in the original image, but the shallow convolutional layer is less, resulting in less semantic information. It is easy to cause underfitting, so multi-scale fusion is carried out between shallow feature map and deep feature map to improve the detection accuracy of small targets.

4. Analysis of experimental results

4.1. Experimental conditions and data set

The training of the model requires the use of a lot of sample data. When we began to conduct in-depth research on pedestrian detection, a series of public data sets of algorithm performance kept appearing, such as PASCAL VOC, COCO and Caltech. Caltech Pedestrian database: This database is a large scale pedestrian database at present, which effectively marks the relationship between the time between rectangular borders and their occlusion. coco has five types of annotations, each of which corresponds to a json file. This paper adopted some COCO and Caltech, PASCAL VOC data sets [9].

Windows10, Microsoft's operating system, was used in the experiment. The graphics card model was GTX1650 and the memory was 16g. Pre-training weight is used by Nvidia SSD FP32. Training method: Prepare the data set in advance, download the weight of the corresponding pre-training model, the total training times are 15000 times, divided into three stages, the 0~4999 times of training for the first stage, set the learning rate of 0.001, the 5000~9999 times for the second stage, also set the learning rate of 0.001, the 10000~14999 times for the third stage, Continue to set the learning rate at 0.001.

4.2. Model construction

Firstly, the CNN convolutional network and pedestrian detection before deep learning are introduced, and some networks constructed with deep learning methods are used for pedestrian detection. The SSD network architecture based on Resnet50 is proposed, and the feature fusion method is used to make some improvements to the original SSD algorithm. It can be seen from Table 2 and Figure 6 that these improvements are effective in the detection and marking of small pedestrians. Meanwhile, the detection rate does not decrease and the missed detection rate decreases, as shown in Table 2[10].

Table 2. Improved comparison of average missed rate of SSD algorithm on Caltech pedestrian data set

Pedestrian detection algorithm	Input size	Average missed rate of all (%)	Decrease (%)	Small average omission rate (%)	Decrease (%)
SSD	300×300	57.42	-	80.78	-
SSD+Resnet50	300×300	55.62	1.8	78.41	2.37
SSD+Resnet50+merge	300×300	54.92	2.5	76.56	4.22

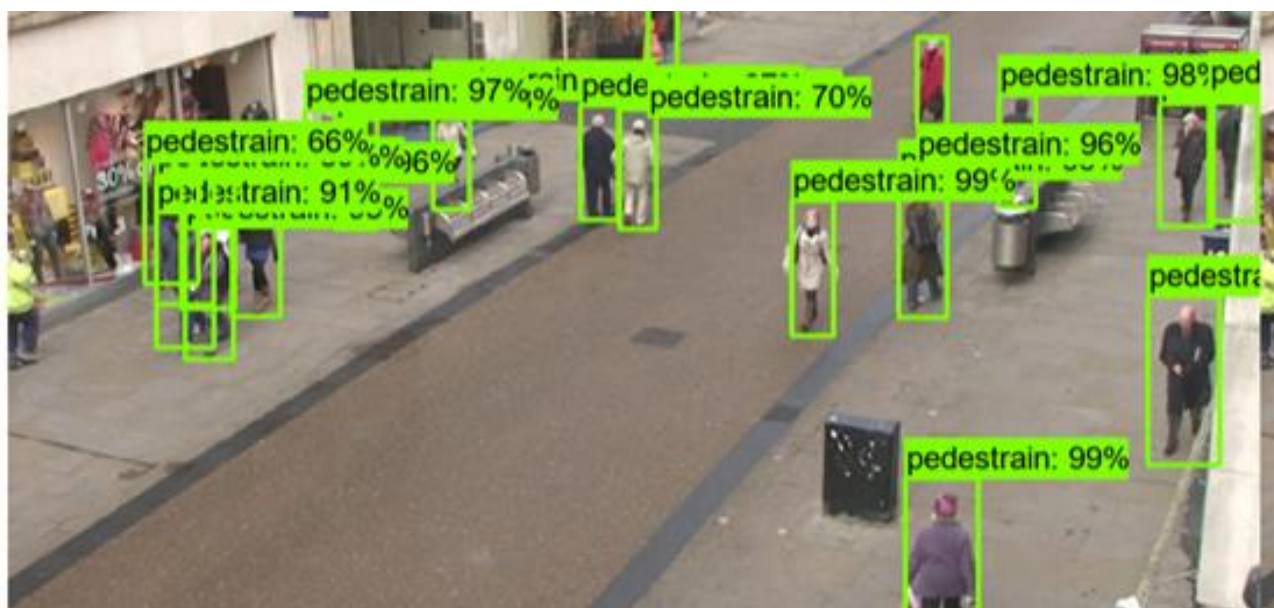


Figure 6. Experimental result diagram

5. Conclusion

This paper mainly introduces the problem of pedestrian detection algorithm in target detection in computer vision. RCNN series is a Two-stage target detection algorithm. Compared with the One-stage target detection algorithm like SSD, the detection speed is slower and the detection accuracy is relatively higher. The SSD algorithm selects the target box directly in the original image, and then extracts the feature map, which greatly improves the detection speed, but the detection effect for small pedestrians is not very satisfactory. Therefore, I made some improvements on the basis of the original SSD algorithm. Resnet50 is the SSD network architecture based on the network, and uses the feature fusion method. In this way, the detection speed and detection accuracy have been greatly improved.

The prospect of pedestrian detection algorithm is to start from some problems, make use of its advantages, solve the problems and optimize the algorithm. Firstly, pedestrian detection needs to process and compare the original image, and then extract the feature of the target. The running speed of SSD algorithm is higher than most algorithms, but the effect of SSD algorithm in small target detection is still not satisfactory. Compared with other algorithms, there is no obvious advantage and room for further optimization. Therefore, this paper explores and improves the SSD algorithm to optimize the problem of poor detection effect on small targets. Now a series of pedestrian detection algorithms, for the computer hardware and software requirements are relatively high, because pedestrian detection needs to use a large number of data sets, training data set required time is also relatively long, some data sets can reach the maximum tens of G, running speed is relatively slow, so the graphics card requirements of the equipment is relatively high. We need to optimize the algorithm better, so that the accurate detection of pedestrians can be achieved without using too much equipment and a lot of time.

References

- [1] Wang Lu, Wang Shuai, Zhang Guofeng, Xu Lisheng. Pedestrian detection method based on Semantic segmentation attention and visible area prediction [J]. Journal of Northeastern University (Natural Science),2021,42(09):1261-1267.
- [2] Xiong Shouyu, Tao Qingchuan, Dai Yafeng. A lightweight Convolutional Neural Network Method for Pedestrian Detection [J]. Computer Applications and Software,2021,38(09):220-225+231.
- [3] Kang Shuai, Zhang Jianwu, Zhu Zunjie, Tong Guofeng. Pedestrian detection method in complex visual scene based on improved YOLOv4 algorithm [J]. Telecommunication Science,2021,37(08):46-56.
- [4] Shan Zhiyong, Zhang Zhongyue. Pedestrian Detection Based on Improved Faster R-CNN Algorithm [J]. Modern Computer,2021,27(23):124-128.
- [5] Zhang Xiangqing. Pedestrian Small Target Detection from UAV Perspective Based on YOLO Algorithm [J]. Information and Computer (Theoretical Edition), 201,33(15):76-78. (in Chinese).
- [6] Chen Min, Wang Chishe, Hao Dahui. Pedestrian detection algorithm based on SSD multi-mode and multi-scale feature Fusion [J]. Journal of Jinling University of Science and Technology,2021,37(02):33-38.
- [7] Bao Wenbin, Zhang Dongquan. XSSD-P: Improved SSD Pedestrian Detection Algorithm [J/OL]. Computer Engineering and Applications :1-11[2021-10-04].
- [8] Yu Bo, Liu Chang. Pedestrian Detection Method Based on Improved SSD Algorithm [J]. Electronic Measurement Technique,2021,44(12):24-28.
- [9] Kuang Xianyan, Yang Jiangbo, Zhang Jianhua. Pedestrian Detection Based on Improved DSSD Algorithm [J]. Chinese Scientific Instrument,2021(05):21-27.
- [10] Qian Wenqian, Wang Jun. Pedestrian Target Detection Based on Lightweight SSD Algorithm [J]. Computer Simulation,202,39(09):487-491.