Dense Crowd Detection based on Deep Learning

Bin Li[a], Yuhui Yang

Logistics Engineering College, Shanghai Maritime University, Shanghai 201306, China.

[a]137395011@qq.com

Abstract

In response to the recurrence of new coronavirus pneumonia, governments of various countries have proposed a variety of intervention measures, including strict restrictions on crowd gathering activities. Based on this, this article proposes a flow calculation system based on deep learning, which can calculate the flow of people in large areas such as scenic spots and shopping malls. In the existing solutions, when the crowd density reaches a certain value, people will block each other, which reduces the detection accuracy, and the data has no reference value. For this reason, this paper designs an interleaved feature extraction network, which uses the attention mechanism connection scheme to process the feature information. In addition, this paper chooses a faster-RCNN detector, whose RPN network structure helps eliminate false detections and improve detection accuracy. The experimental results show that the feature extraction network based on the faster-RCNN detector proposed in this paper has a recognition rate of 89.8% under low crowd density and 83.7% under high crowd density and severe occlusion, which has certain practical value.

Keywords

Interspersed Network; Attention Mechanism; Faster RCNN; People Flow Detection.

1. Introduction

At the beginning of this year, the COVID-19 epidemic swept the world. With strong government intervention, the epidemic was effectively controlled. In the daily basic epidemic prevention work, controlling the population density of the area is an important task, and obtaining the number of pedestrians in a certain area is a key reference index for control. Although there are many existing crowd counting schemes, such as channel-type induction counting, video tracking counting, etc., when the flow of people is large, people block each other, and the difficulty of detection increases and the counting accuracy decreases. In recent years, with the development of computer hardware technology and in-depth research on deep learning application technology, the application pain points of counting the appellants have been effectively solved.

In the field of computer vision, for pedestrian detection schemes, the early scheme was to extract pedestrian features from the whole and part of the image by designing a feature device, and then train the classifier based on the extracted information to achieve the purpose of detection, such as a typical Hog feature device and SVM classifier. However, when the crowd density is high, the occlusion phenomenon is serious, the feature device cannot accurately extract the features, and the classification effect of the classifier is not good. At present, the detection scheme for pedestrians is more carried out through deep convolutional neural networks. Deep convolutional neural networks also need to extract the features of pedestrians. However, unlike the Hog operator, feature operators need to be artificially designed. The product neural network automatically optimizes and optimizes the parameters of each layer of convolutional network by converting the input picture into a matrix through forward and backward iterative calculations, that is, the network can automatically generate a dedicated feature extraction network based on the training samples. The application range of convolutional neural network is wider, and the robustness and feature fit are stronger. In the field of computational vision, more and more scene solutions rely on convolutional neural networks to land. For example, in pedestrian detection, Yun I, Jung C, etc. designed alignment networks based on saliency and bounding box information to cooperate with convolution Neural network solves the problem of pedestrian detection when the head or other limbs of the pedestrian are occluded [1].Fei C, Liu B and others have designed an instance-level context prediction module in the convolutional neural network to solve the problem of occlusion in pedestrian detection. The features of different levels of the upper and lower networks are merged and then transferred to the deep network, which improves the diversity of feature information [2].Liu C and Lu J et al. use multi-scale feature extraction and simultaneously introduce amplification- Reduce the module to enhance the accuracy of pedestrian detection by merging the local details of the special detection image and the context information of the convolutional neural network [3]. Li G, Yang Y and others are based on the YOLO network, using a deep separable convolution to reduce network parameters to reduce computational costs, and at the same time adding the Squeeze-and-Excitation module to weight the network convolutional layer to improve the network's ability to perform in hazy weather Pedestrian detection [4]. Based on SSD detection network, Yang J, He WY and others have integrated the local feature extraction method to obtain features of different positions, different aspect ratios and sizes in the feature map, which improves the performance in complex environments. Pedestrian detection capability and applied it to pedestrian detection in subway stations [5].Yu X, Si Y et al. based on Faster RCNN algorithm, optimized the network structure by combining feature cascade and hard negative mining strategy, and applied it to complex Pedestrian detection task in natural environment [6].

In this paper, a human flow calculation system is designed based on convolutional neural networks. Its general workflow is to monitor the timing of taking pictures and send them to the convolutional neural network to count the number of people, and then store the results in the database. The accumulated data can be obtained[7]. Real-time information on the number of people in the area and the flow of people in the interval. Aiming at the situation that when there is a large flow of people, mutual occlusion between people leads to false detections, this paper designs an interspersed feature extraction network, based on the ResNet feature extraction network added an interspersed module, the function of this module is to collect a wide area With local feature information, the collected feature information is de-redundant through the spatial attention mechanism, and then passed to the rest of the network to broaden the feature width of the network, and then provide the detector with richer feature information to improve detection accuracy .

2. Related research

The advantage of the target detection algorithm based on deep learning is that the convolutional neural network can automatically complete the feature extraction of the detected target. Convolutional neural networks are mainly composed of convolutional layers, and their combination has a greater impact on the extraction performance of the network. The early convolutional neural network structure is dominated by the string type[8]. The convolutional layers are combined into a network by end-to-end connections. For example, VGG-Net has only 19 layers. Due to the shallower network, the extracted feature information is limited, so the early volume Product neural networks are more commonly used for classification tasks. In order to improve the feature extraction ability of convolutional neural networks, subsequent network structure types have a variety of variants, such as GoogleNet, which uses the Inception module, which widens the width of each layer of the network (that is, in the same layer of the network) Including multiple convolution kernels of different scales) to improve feature extraction capabilities[9].At present, the mainstream feature extraction network is the residual network, and its overall type is still dominated by string, but the residual layer structure design allows the network to contain more convolutional layers, compared with VGG-Net such networks [10]. In other words, when the number of network layers is larger, the depth is also deeper. The extracted feature information contains more semantic information, and the residual structure avoids the problem of gradient disappearance during training due to too many network layers[11].The residual unit in the network can perform identity mapping between input and output, avoiding the problem of feature information disappearing during transmission[12]. The structure of the residual unit and its working principle are as follows:

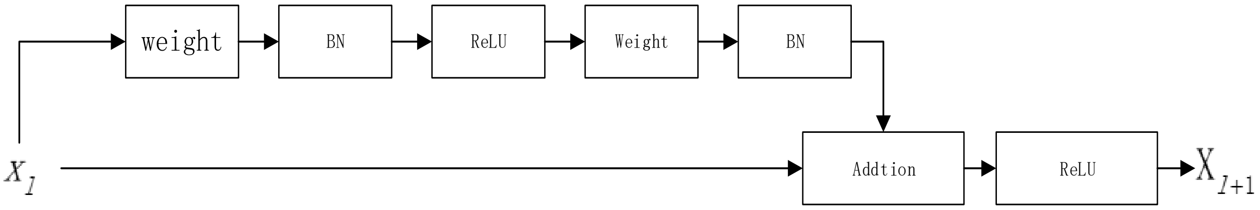


Fig. 1 Residual unit structure

As shown in Figure 1, the residual unit can be divided into two ways in structure, the mapping link (left side of Figure 1) and the residual link (right side of Figure 1). When the input x_l of this unit passes through the mapping link, it can be expressed as $h(x_l)$. The residual link usually contains 2 to 3 convolutional layers. The link can be expressed by the formula $F(x_l, W_l)$, and the output of the residual unit can be expressed as:

$$y_l = h(x_l) + F(x_l, W_l) \tag{1}$$

The output of this unit after the activation function is used as the input of the next residual unit can be expressed as:

$$x_{l+1} = f(y_l) \tag{2}$$

The relationship between x_l and x_{l+1} can be simply understood as $x_{l+1} = x_l + F(x_l, W_l)$, when the number of network layers is deeper, the relationship between x_l and x_{l+1} can be expressed as:

$$x_L = x_l + \sum_{i=1}^{L-1} F(x_i, W_i) \tag{3}$$

In the process of network training, the value of $\frac{\partial}{\partial x_l}\sum_{i=1}^{L-1} F(x_i, W_i)$ does not equal -1 continuously, which means that the problem of gradient disappearance will not occur during training. In summary, this paper selects the residual network with better performance as the feature extraction network, and optimizes the network based on ResNet52 to further improve the network performance[13].

In addition, this paper chose the Faster-RCNN network as the detector module of this design. The reason for the choice is that the RPN (RegionProposal Network) network contained in the Faster RCNN network can more accurately fit the target location information. The RPN network is provided with a set of anchors with different sizes and proportions, and each pixel in the feature map is represented as an anchor point as the center of the anchor. First traverse all anchor points. When the intersection ratio of the group of anchors and Ground Truth is greater than the threshold, the area framed by the anchor point and the anchor is a candidate area. After that, the extracted candidate area is sent to the classifier and regressor of the detection network module[14].The classifier determines the category of the target in the candidate area, and the regressor is responsible for adjusting the position coordinates of the candidate area to make it more suitable for the actual target. location information.

3. Feature extraction operation based on interspersed network

Through comparative experiments, this paper uses VGG19 and ResNet50 as the feature extraction network of Faster-RCNN respectively. The experimental comparison results show that the feature extraction operation based on ResNet50 is significantly better than VGG19 in detection results. However, there are still obvious flaws in the detection results. When two people block each other in the image, the detector will mistake the two for one person, or miss one of them. Through network analysis, it is found that the problems mainly exist in the following points: First, when the crowd is occluded, the target pixel area of the occluded pedestrian is small. As the network depth increases, the downsampling operation makes the occluded target pixel effective The information is assimilated by the surrounding information, and the final effective information is less; the second is that the occlusion is diverse, the ResNet network structure is relatively simple, and the generalization ability is poor[15].For some special occlusion situations, the target cannot be effectively detected. In order to improve the robustness of the algorithm, this paper optimizes the structure on the basis of ResNet, and enhances the use of shallow features to improve the feature extraction ability of the network for small targets. At the same time, it enriches the shape of the network to make the network fit different occlusion states[16].

3.1 Feature aggregation delivery module

First of all, this article adds a feature aggregation transfer module to each residual unit of the ResNet network. The internal results of the module are shown in the following figure:

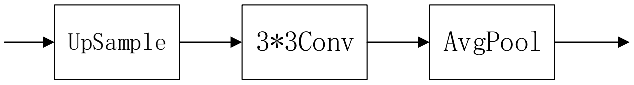


Fig. 2 Feature aggregation delivery module structure

The feature map is passed to the module first through the maximum up-sampling operation. The purpose of up-sampling is to assimilate the pixel area occupied by the small target to the surrounding pixels to enlarge the proportion of the pixel area occupied; then the 3x3 convolutional layer is used to compare the enlarged The feature map is used for feature extraction; after the extraction operation is completed, the size of the feature map needs to be reduced so that it can be passed to the next residual unit of the ResNet network. This article uses mean down-sampling to save the feature information of small targets to a greater extent. And the down-sampling operation can filter some redundant feature information. The combination of the feature aggregation transfer module and the ResNet network is that the first input image in the initial stage of the network will be passed to the first feature combination transfer module, and then the module will simultaneously transfer the extracted feature information to the next feature aggregation transfer The residual unit of the module and the following ResNet network. The subsequent modules repeat the operation and transfer the extracted feature information to the next module and the next residual unit. In terms of the overall network type, the feature aggregation transfer module is interspersed and embedded in the ResNet network, so the new network is named interspersed network. In the network, the connection of the various feature aggregation and transfer modules can also be understood as a feature extraction network, but compared with the ResNet network, it contains fewer convolutional layers, only 15, and it is judged as a shallow network from the depth of the network. The extracted feature information is more morphological features, that is, it contains the boundary information and texture information of the target. The feature aggregation transfer module transfers the extracted shallow features to the ResNet network to supplement the feature information richness. The feature information of small targets can be effectively extracted through the feature aggregation transfer module. The ResNet network combined with shallow features can provide more information about the shape and location of the target for the subsequent RPN network. The RPN network depends on the boundary information and texture information of the target[17]. Accurately locate the position of each target, and improve the ability to judge the foreground and background.

3.2 Attention mechanism channel based on spatial domain

In addition, in this article, the ResNet network as a whole contains four different scale residual unit groups of 200x200, 100x100, 50x50 and 25x25. The network is divided into two parts (200x200, 100x100 scale is the upper part corresponding to the shallow network, 50x50, 25x25 scale

Corresponding to the deep network for the next part). [18]In order to enhance the utilization of the target's shallow morphological feature information, this paper adds a spatial domain-based attention mechanism channel to the first feature aggregation transfer module in the two scales of 200x200 and 100x100, and transmits the shallow data to the deep network after processing. The first feature aggregation delivery module in two scales of 50x50 and 25x25. Through the spatial domain, the spatial information of the target in the feature map can be transferred to another space, and the key position information can be extracted. The module structure of the spatial domain attention mechanism channel is as follows:

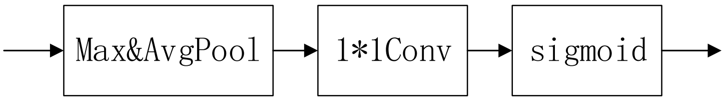


Fig. 3 Spatial domain attention mechanism channel

The feature map is first transferred to the spatial module consisting of the maximum pooling layer and the average pooling layer, and then input to the 1x1 convolutional layer. The 1x1 convolutional layer is selected to avoid the large change in the feature during the large convolution calculation[19]. Map information, and 1x1 convolutional layer can play the role of smooth transition of features between layers. In addition, the special setting for the 1x1 convolutional layer is to set the number of channels to only 1, and then use the sigmoid function to activate the feature map. Since the feature map has only one channel, the sigmoid function is equivalent to assigning a value to each pixel block on the feature map. The pixel area containing the target has a larger weight, while other areas have a smaller weight. Finally, the weighted feature map is passed to the deep network. In summary, the interleaved network structure based on ResNet optimization in this paper is shown in the following figure:

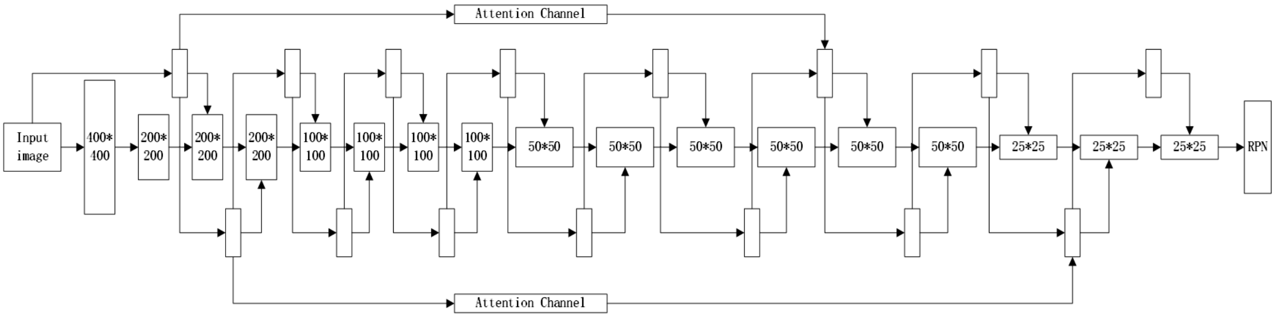


Fig. 4 Interspersed network structure diagram

3.3 Design and test of human flow calculation system

This paper is based on the Faster-RCNN model to design a pedestrian flow calculation system for the designated area. At present, there are two ways to calculate, one is the tracking count based on the video stream, and the other is the detection count based on the picture. Based on the counting of video streams, due to the need to re-identify pedestrians, when the degree of crowd occlusion is high, although Faster-RCNN can detect the target, the re-identification network is more difficult to extract the features of the occluded target[20]. The same target is misidentified as a new target, resulting in double counting. Therefore, the scheme adopted in this paper is the detection and counting based on pictures. Set the frame rate of the surveillance video stream, send the extracted frame images to the network for detection, and store the number of detected targets in the database, which can calculate and analyze the flow of people in each period, or Calculate the current number of pedestrians in real time[21].

4. Experiment and data analysis

4.1 Data set and evaluation standard setting

The data set used in the training model in this article is mainly composed of the cityPerson outdoor pedestrian data set and the self-collected and labeled indoor pedestrian data set. It contains 25,000 pictures, 15,000 randomly selected as the training set, 3000 as the validation set, and 7000 As a test set. In addition, the problem to be solved in this paper is to detect pedestrians under occlusion. In order to measure the performance of the interspersed network designed in this paper under occlusion, this paper creates an occlusion test set according to different degrees of occlusion. The degree of occlusion is defined by the occlusion rate Occ (Occlusion), and the calculation formula of the occlusion rate is as follows:

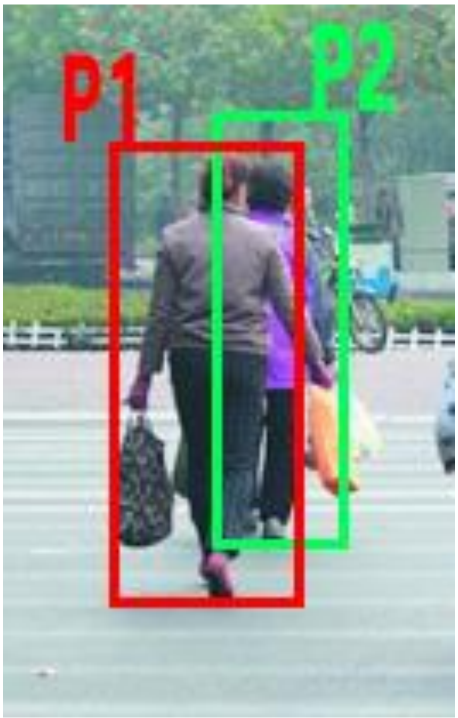


Fig. 5 Occlusion

In the above figure, the two pedestrians are represented by $P_1(x_{min}^1, y_{min}^1, x_{max}^1, y_{max}^1)$ and $P_2(x_{min}^2, y_{min}^2, x_{max}^2, y_{max}^2)$ respectively, where P_2 is the occluded target, and the area of the occluded area $area_o$ is:

$$w = min(x_{max}^1, x_{max}^2) - max(x_{min}^1, x_{min}^2) \tag{4}$$

$$h = min(y_{max}^1, y_{max}^2) - max(y_{min}^1, y_{min}^2) \tag{5}$$

$$area_o = w \times h \tag{6}$$

The occlusion rate is

$$Occ = \frac{area_o}{area_{P_2}} \tag{7}$$

According to the occlusion rate, this article divides the occlusion degree into three categories,

Slight occlusion: $0\% < Occ \leq 25\%$;

Ordinary occlusion: $25\% < Occ \leq 50\%$;

Severe occlusion: $50\% < Occ$

This article uses the following indicators to evaluate the excellent detection performance of each model: Accuracy rate, used to evaluate the accuracy of the model for the target category classification:

Recall rate, used to evaluate the model's ability to recall the target quantity: Cross-combination ratio is used to evaluate the accuracy of the model predicting the target position: The average accuracy is used to evaluate the comprehensive detection performance of the model:

4.2 Comparative test results analysis

In this paper, ResNet-Faster RCNN and CrossNet-Faster RCNN are set as the control group, and ResNet-SSD, CrossNet-SSD, and YOLO v4 are added for auxiliary verification. First, compare the learning effects before and after the feature extraction network optimization according to the loss function graph of the model training. The comparison group uses the Faster RCNN series algorithm and the SSD series algorithm. The loss function graphs of the two are as follows:

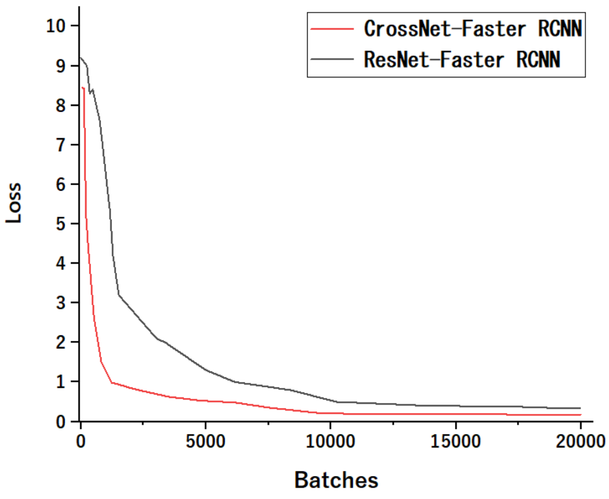


Fig. 6 faster-Rcnn

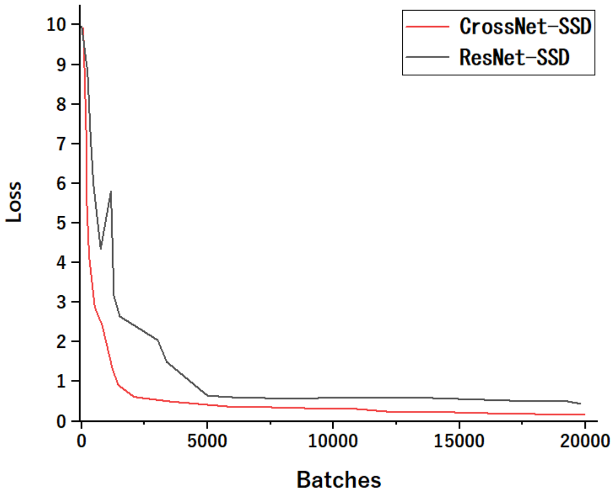


Fig. 7 SSD

In the two figures a and b, the Faster RCNN and SSD detection algorithm based on the CrossNet network has a smoother overall curve, a faster decline, and a lower loss function value than the ResNet network. Based on the model trained by the ResNet network, the average SSD model has a large amplitude fluctuation in the early stage of network training, and the Faster RCNN and SSD model have small fluctuations in the later stage of the network training, and the network is relatively unstable. The comparison of loss function curves shows that the CrossNet network has a stronger learning ability than the ResNet network. It also proves that its ability to extract target features is excellent and improves the efficiency of model learning.

The trained model adopts the evaluation index for comprehensive test comparison, and the results are shown in the following table:

Table 1. Comparison of detection performance test results

	Precision	Recall	IoU	AP
ResNet-Faster RCNN	85.4%	87.3%	75.8%	82.7%
CrossNet-Faster RCNN	90.7%	92.6%	77.5%	86.3%
ResNet-SSD	80.3%	84.1%	70.1%	77.6%
CrossNet-SSD	84.4%	88.9%	74.6%	83.2%
YOLO v4	86.6%	86.8%	73.9%	84.4%

It can be seen from the test data in the above table that the data obtained by using CrossNet as the feature extraction network test has improved compared with ResNet in all indicators. And after the SSD network is equipped with CrossNet, its detection performance is similar to the current single-stage detection algorithm YOLO v4 with the best detection performance. Among all models, the CrossNet-Faster RCNN model has the highest recall rate and cross-over ratio. Its high recall index indicates that the model has a low probability of false detection, and its overall robustness in actual use is better; its cross-over ratio is better than the test. The value proves that the CrossNet network strengthens the extraction and utilization of shallow features, so that the network has a stronger ability to perceive targets and more accurate positioning.

In addition, this paper tests the detection ability of each model under different degrees of occlusion. The test results are shown in the following table:

Table 2. The average detection accuracy of each model under different occlusion conditions

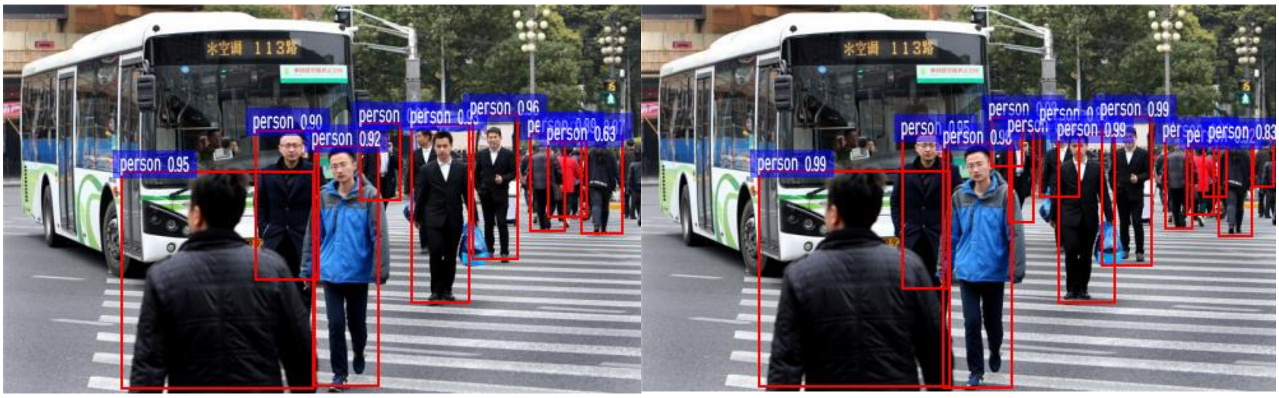
	Light occlusion	Normal occlusion	Severe occlusion
ResNet-Faster RCNN	85.6%	82.7%	77.3%
CrossNet-Faster RCNN	89.8%	86.3%	83.7%
ResNet-SSD	80.3%	77.6%	71.8%
CrossNet-SSD	85.1%	83.2%	75.0%
YOLO v4	87.7%	84.4%	80.5%

Table 3. The recall rate of each model under different occlusion conditions

	Light occlusion	Normal occlusion	Severe occlusion
ResNet-Faster RCNN	89.6%	86.8%	83.8%
CrossNet-Faster RCNN	95.2%	93.2%	90.3%
ResNet-SSD	85.9%	84.8%	79.0%
CrossNet-SSD	90.6%	88.4%	85.9%
YOLO v4	89.7%	83.5%	80.8%

From the data in Table 2 and Table 3, it can be seen that the CrossNet-Faster RCNN model has the best test data in various degrees of occlusion.

A set of actual detection results of ResNet-Faster RCNN and CrossNet-Faster RCNN are used as a visual comparison. The test comparison results are shown in the following figure:



a. ResNet-Faster RCNN b. CrossNet-Faster RCNN

Fig. 8 Test effect diagram

From the above figure, it can be observed that CrossNet-Faster RCNN is better than ResNet-Faster RCNN in the detection results of the following two points: First, the confidence score of the detection results CrossNet-Faster RCNN is higher than ResNet-Faster RCNN; In terms of the number of detection targets, CrossNet-Faster RCNN detects more comprehensively. Take the crowd in the upper right corner of the figure as an example, which has a higher degree of occlusion. ResNet-Faster RCNN only detects three people and misses three people, while CrossNet-Faster RCNN detects Out of 5 people, 1 missed inspection. The visual comparison results directly indicate that CrossNet-Faster RCNN has better detection effects in complex environments.

5. Conclusion

In this paper, aiming at the problem that the mutual occlusion cannot be accurately detected in the high-density situation, the CrossNet-Faster RCNN network is optimized and designed based on the ResNet-Faster RCNN network. The optimized scheme is aimed at the ResNet50 network, involving the following two points in total. One is to design a feature aggregation transfer module and cross-embed it with the ResNet network. The feature aggregation module enhances the extraction ability and reuse efficiency of shallow features , And pass it to the ResNet network to enrich the feature richness of the backbone network. In addition, the attention mechanism channel based on the spatial domain is added to strengthen the feature information interaction between the shallow network and the deep network. Based on the above optimization scheme, the feature information extracted in the backbone network contains more shallow morphological features than before. The Faster RCNN network can better locate and capture the target based on the texture features and boundary features of the extracted target. Improve detection accuracy. After experimental testing, the recall rate of CrossNet-Faster RCNN is 92.6%, which is an increase of 5.3% compared to before optimization, and the average detection accuracy is increased by 3.6% to 86.3% compared to before optimization. Under the occlusion test set made in this article, the test results of CrossNet-Faster RCNN are also optimal.

References

[1] Yun I, Jung C, Wang X, et al. Part-Level Convolutional Neural Networks for Pedestrian Detection Using Saliency and Boundary Box Alignment[J]. IEEE Access, 2019:1-1.

[2] Fei C, Liu B, Chen Z, et al. Learning Pixel-Level and Instance-Level Context-Aware Features for Pedestrian Detection in Crowds[J]. IEEE Access, 2019, PP (99):1-1.

[3] Lin C, Lu J, Wang G, et al. Graininess-Aware Deep Feature Learning for Pedestrian Detection[J]. IEEE Transactions on Image Processing, 2020, 29:3820-3834.

[4] Li G, Yang Y, Qu X. Deep Learning Approaches on Pedestrian Detection in Hazy Weather[J]. IEEE Transactions on Industrial Electronics, 2019, PP (99):1-1.

[5] Yang J, He W Y, Zhang T, et al. Research on Subway Pedestrian Detection Algorithms Based on SSD Model[J]. IET Intelligent Transport Systems, 2020(7553).

[6] Yu X, Si Y, Li L. Pedestrian detection based on improved Faster RCNN algorithm [C]// 2019 IEEE/CIC International Conference on Communications in China (ICCC). IEEE, 2019.

[7] Hosang J H,Omran M,Benenson R, et al. Taking a deeperlook at pedestrians [C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition, Boston,2015:4073-4082.

[8] Lin T Y, Goyal P, Girshick R, et al.Focal loss for dense object detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020,42(2):318-327.

[9] Hinton G E, Osindero S, Teh Y. A fast learning algorithm for deep belief nets[J].Neural Computation,2006, 18:1527-1554.

[10] Ouyang W, Wang X. A discriminative deep model for pedestrian detection with occlusion handling [C]// 2012 IEEE International Conference on Computer Vision and Pattern Recognition,2012:3258-3265.

[11] Ouyang W, Wang X. Joint deep learning for pedestrian detection[C]// 2013 IEEE International Conference on Computer Vision, 2013:2056-2063.

[12] Liu W, Anguelov D, Erhan D, et al. SSD:single shot multibox detector[C]// 14th European Conference on Computer Vision.Cham:Springer,2016:563-590.

[13] Fu C Y, Liu W, Ranga A, et al.DSSD:deconvolutional single shot detector[C]// 2017 IEEE International Conference on Computer Vision and Pattern Recognition, 2017:910-925.

[14] Redmon J, Divvala S, Girshick R, et al.You only look once:unified, real-time object detection[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition,2016:1103-1120.

[15] Wu B, Nevatia R. Detection of multiple, partially occluded humans in a single image by Bayesian combination of edgelet part detectors [C]// 10th IEEE International Conference on Computer Vision, 2005: 740-757.

[16] Wang X, Han T X, Yan S. An HOG-LBP human detector with partial occlusion handling [C]// 12th IEEE International Conference on Computer Vision,2009:1016-1033.

[17] Enzweiler M,Eigenstetter A,Schiele B,et al. Multi-cue pedestrian classification with partial occlusion handling[C]// 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2010:1366-1387.

[18] Wojek C, Walk S, Roth S, et al. Monocular 3D scene understanding with explicit occlusion reasoning[C]// 2011 IEEE Conference on Computer Vision and Pattern Recognition,2011:863-882.

[19] Mathias M, Benenson R, Timofte R, et al.Handling occlusions with Franken-classifiers[C]//2013 IEEE International Conference on Computer Vision,2013:1003-1015.

[20] Bell S,Zitnick C L,Bala K,et al. Inside-outside net:detecting objects in context with skip pooling and recurrent neural networks[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition, 2016:2874-2883.

[21] Zhou Chunluan,Yuan Junsong.Multi-label learning of part detectors for occluded pedestrian detection[J]. Pattern Recognition,2018,23(6):56-75.