A novel small-scale pedestrian detection method base on residual block group of CenterNet

Shuangcai Liu, Hui Ma*, Yan Lv, Zedong Yang, Chunxin Fang

College of Electronic Engineering, Heilongjiang University, Harbin 150080, China

Abstract: Small-scale pedestrian detection is beneficial to improve the accuracy of operation in practical scenarios. However, the sample characteristics of small-scale pedestrians are too few to be accurately captured. Therefore, this paper designs a Three ResNet Blocks based on CenterNet (TRBBC-Net) detection model. Aiming at the limited ability of a single feature extraction block to extract semantic information at different levels in the network, this paper proposes Three ResNet Blocks (TRB), which is a simple and effective multi-block group. This block group integrates three different basic blocks, each of which extracts pedestrian information separately to enhance information flow in the network structure and make detection results more accurate. In addition, combined with the advantages of activation function in the model expression, a maximum suppression activation function is introduced to improve the performance of the detector by preventing numerical explosion being sensitive to decimal. Comprehensive experiments on pedestrian detection datasets (Caltech and ETH) show that proposed method TRBBC-Net exhibits excellent performance, compared with the most advanced AdaptFaster RCNN method, the average miss rate of pedestrian pixel height greater than 20 is higher than 10.87% in Caltech dataset.

Keywords: Small-scale pedestrian, TRB, pedestrian detection.

1. Introduction

Pedestrian detection is a technology used to detect and locate pedestrians, especially in the field of automatic driving [1], video surveillance [2] and robot vision. There are different scales for pedestrian detection. Among them, small-scale pedestrian detection is widely applied in practical scenarios. The successful prior detection of small-scale pedestrians can leave a lot of operation time for automatic driving. However, there are too few sample features of small-scale pedestrians, which will always lead to misclassification and loss, especially in the case of large environmental impact, such as occlusion [3] and light impact. Therefore, a method for detecting small-scale objects under complex circumstances is urgently needed.

Small-scale pedestrian detection is now a difficulty in the studies of pedestrian detection. As shown in Fig.1, there are only 13 pedestrians with a pixel height between 0-10 in the Caltech training set, there are 183 pedestrians in the test set. The imbalance between test set and training set is an important reason for poor detection results. Among pedestrians whose height is less than 50 pixels, as the pixel height of pedestrians decreases, the proportion of pedestrians in the test set to those in the training set gradually increases. Therefore, a detection method is needed to conduct in-depth data mining in case of insufficient training set data. Most of the papers analyzed the results of pedestrian with a pixel height of greater than 50, which accounted for only 20.49% of the pedestrians, so it is necessary to improve the detection results of small-scale pedestrians.

The traditional manual feature extraction method can be used to more accurately capture local information such as the appearance and form of pedestrians. As shown in Fig.2,

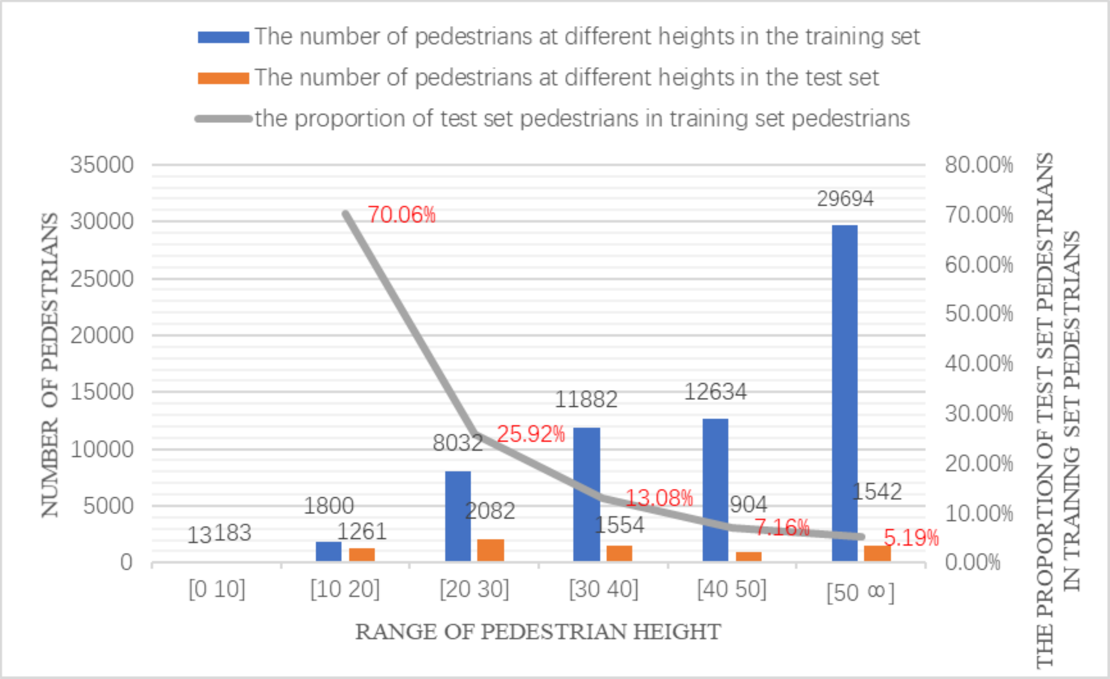


Fig.1. Proportion of small-scale pedestrians' test set

small-scale pedestrians have few features, so the detection results of small-scale samples under complicated circumstances are not sufficiently robust, and they are easy to overfit with datasets. With the rapid development of deep learning, many deep learning methods have been widely used for pedestrian detection. RCNN series algorithm is the first convolution algorithm applied to pedestrian detection. The Faster RCNN [4] algorithm has achieved high accuracy in the directional pedestrian detection by proposing the RPN, but considering its slow speed, this paper also proposes the Cascade RCNN [5] by optimizing its neural network, which applies multi-step ROI consolidation and prediction together with RPN to further increase the recognition rate of pedestrian detection. Recently, many high-precision pedestrian detection methods based on deep learning have been proposed, such as OR-CNN [6], AdaptFaster RCNN [7] and other models, and these models have made outstanding contributions to improve the recognition rate of pedestrian detection. These two-stage algorithms obtain high accuracy at the expense of speed prediction.

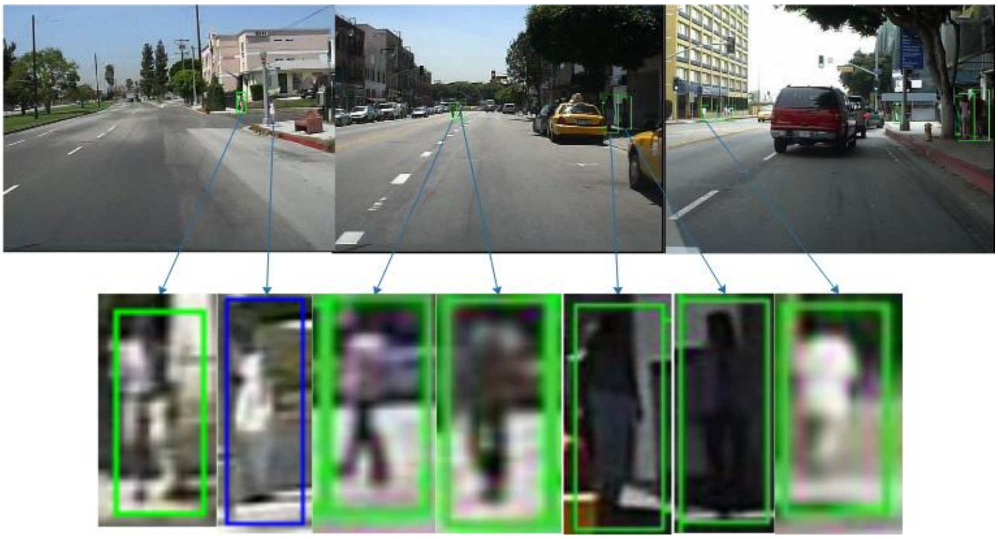


Fig.2. Small-scale pedestrian in different scenes

This paper proposed a pedestrian detection scheme based on small-scale sample mining to better detect pedestrians at the expense of computational speed while reducing impact caused by environmental factors. The method proposed in this paper was a detection network based on the central point that adopts different methods from most successful detection methods. Most detectors firstly list the potential locations of the target, and then classify these locations. When the network model was built in this paper, the target center was taken as a point, the problem of target detection was transformed into that of critical point estimation, and then other target attributes were returned. The network is an end-to-end network, simpler in model and greater in speed. Firstly, in the context of network, activation functions can bring nonlinear capabilities into neural networks. This paper proposed two kinds of activation functions used for the network. Relu6 can restrict the maximum value of the activation function to prevent numerical explosion and decimal sensitization, and h-switch restricts the activation function again on relu6 to reduce the operation cost of the activation function. Secondly, it was observed that pedestrian characteristics in the neural network were affected by many activation functions at each stage of the network. Consequently, a lot of pedestrian information could not be accurately extracted. Therefore, this paper proposed the Three ResNet Blocks architecture, which used different TRB at different locations in each stage of the network, especially at the initial stage of network sampling where the information should be retained to the maximum extent to reduce the information loss when the activation function was set to negative signal 0. Finally, by adjusting the proportions of scale loss, center offset loss and center loss of the overall network loss function was adjusted so that a suitable model for testing could be more easily obtained.

The main contributions of this study are as follows:

(1) Aiming at the insufficient feature extraction for small-scale pedestrian detection, this paper proposed a TRBBC-Net detection structure based on CenterNet, which has achieved better effects for small-scale pedestrian detection in Caltech and ETH datasets.

(2) The traditional single feature extraction block has limited ability to extract semantic information at different levels. This paper proposed TRB to reduce the use of activation functions in the early stage of down sampling phase and enhance the information flow in the network structure, thus improving the detection results of the network.

(3) Considering that the activation function of the network cannot avoid numerical loss, this paper introduced an activation function to suppress the maximum value so as to prevent the numerical explosion and decimal sensitization and to improve the capability of the detector.

2. Related Work

In the early stage of pedestrian detection, most classifiers are used to extract manual features [8] such as HAAR [9] and SIFT [10] features through sliding windows for the classification and regression of pedestrian targets. Integrated channel feature [11] (ICF) is a relatively effective method for extracting manual features. It mainly uses directional gradient feature (HOG [12]), color feature (LUV) and overall channel feature to achieve good results in pedestrian detection. And a lot of modified methods have been proposed such as ACF [13] and LDCF [14]. But in recent years, the pedestrian detector based on CNN has demonstrated great potential in pedestrian detection. RPN+BF [15] method uses an enhanced random forest for learning on the basis of the generation region of Faster-RCNN and improves the performance of small-scale pedestrian detection by means of difficult sample mining strategy. SA-Fast RCNN [16], an expansion based on Fast RCNN, uses multiple sub-networks to detect pedestrians of different scales, and has achieved good results in small-scale pedestrian detection. Similarly, MS-CNN [17] also conducts multi-scale detection based on Faster-RCNN, aiming to improve the results of small-scale pedestrian detection. Compared with the effect of a single classifier on small-scale pedestrian detection, multiple classifiers are more likely to mine the potential information about pedestrians. On the basis of using Faster-RCNN, F-DNN+SS [18] uses multiple parallel classifiers, including GoogleNet [19] and ResNet [20], to greatly improves the detection of small-scale pedestrians compared with the previous methods. In addition, it suppresses background through semantic segmentation to enhance the ability of pedestrian recognition. In SDS-RCNN [21], semantic segmentation is used to strongly improve the result of pedestrian detection, and the segmentation mask is placed at the top of the feature map as an enhancement of pedestrian detector. TLL [22] adopts a new marking method for datasets in the network. Compared with frame marking method, central axis marking method is more sensitive to small-scale pedestrians while causing a small amount of deviation to the training.

The study of pedestrian detection usually focuses on the performance of small-scale pedestrian detection, but ignores the advantage in speed. In the Cascade RCNN, the anchor conclusion that can train SSD [23] can be obtained by applying multi-step ROI pooling and forecast in the RPN [24], so ALFNet [25] adopts multi-step prediction and progressive positioning, which breaks through the limitations of single-stage network in pedestrian detection and achieves high accuracy while significantly improving the speed. In the CSP [26], in order to further improve the speed of the network, the method proposed by the CornerNet is selected and the use of anchor is discarded. By predicting the size and center of the image, the superiority of the Anchor-Free [27] method in pedestrian detection has been verified. In this paper, the information of small-scale samples was further mined in the basic CenterNet to further observe whether speed and accuracy could be improved for detecting small-scale samples by making hierarchical improvement on the basic block and inhibiting the activation function.

3. Detection method for small-scale pedestrians

In this paper, a new model was created on the basis of CenterNet by modifying the basic block and suppressing maximum activation function. The modified basic block did not discard the feature information of pedestrians and the numerical accuracy could be maintained by suppressing the maximum value of inactivation function. The model proposed in this paper was more inclined to extract the features of small-scale pedestrians.

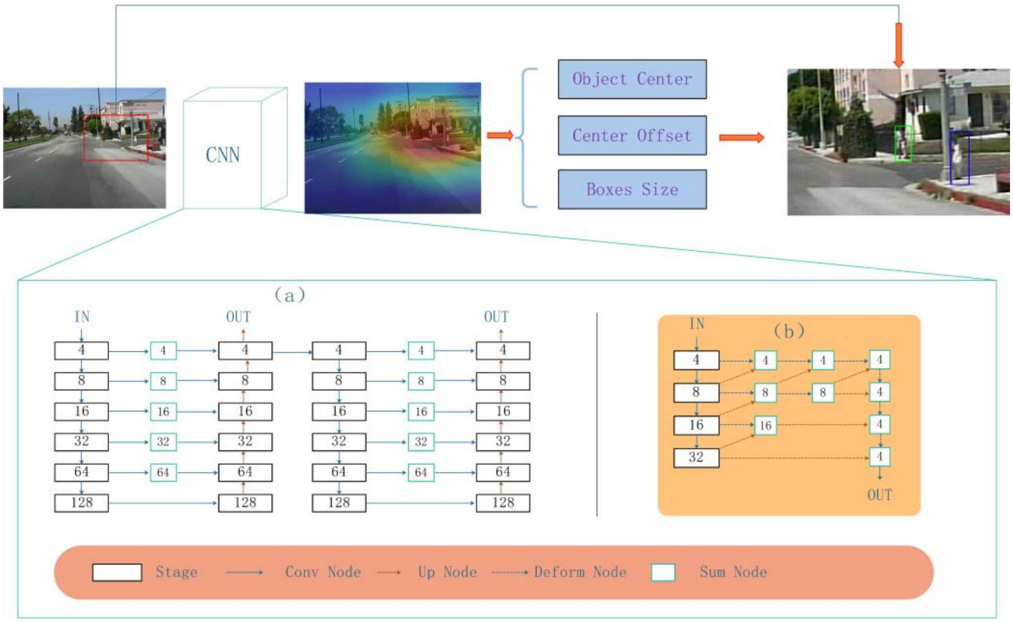


Fig.3. CenterNet

3.1 Network framework

CenterNet is a member of the Anchor-Free family network and is a simple and fast end-to-end center-based detection algorithm. The CenterNet framework is mainly composed of feature extraction network, detection head and loss function. The framework of CenterNet is shown in Fig.3.

Extracting feature network is an important part of semantic analysis. This paper mainly selected Hourglass-104 network and DLA-34 network. The Hourglass-104 network was originally used for posture recognition and was derived from multi-scale information mining of network structure. As shown in Fig.3 (a), when images entered the network, multiple down-sampling was conducted to recognize the information of pedestrians of different scales and sizes. In Fig.3 (b), DLA-34 was adopted. This network was mainly selected in this paper because of its great running speed and deformation convolution, which replaced the convolution of fixed box thinking to simulate how human brains judged things and enhanced the semantic analysis of small-scale pedestrians.

The detection head was mainly composed of three parts, namely pedestrian center point, the offset value of pedestrian center point and the size of pedestrian target. A 128×128×5 heatmap was obtained by extracting features, in which 128×128×1 was used to detect the pedestrian center point of Gaussian distribution, and the regression on the height and width of pedestrians was carried out for each Gaussian point, and thus the concept of anchor was discarded. The loss function of the detection head is as follows:

$$L = L_K + \lambda_{size} L_{size} + \lambda_{off} L_{off} \tag{1}$$

where $\lambda_{size} = 0.1$, $\lambda_{off} = 1$, L_k is aimed at calculating the loss function of the center point of the heatmap. L_{size} is the loss value of the height and width of pedestrians, and L_{off} is the partial offset prediction loss of each center point due to the deviation of data dispersion caused by down-sampling under the network. The specific steps for training this algorithm are demonstrated in Algorithm 1.

Algorithm 1. The algorithm flow of CenterNet

```
Input：
    images and ground truth of marked pedestrian detection frame; Parameters that need to be configured in the network.
Output:
    The model predicts the target detection box of the test set.
1. Network reasoning is conducted to extract the maximum 100 peak values of pedestrians on the heatmap;
2. By determining the peak point of the pedestrian, the corresponding pedestrian size value and the offset are obtained;
3. The obtained value is placed in the loss function to generate a more reasonable model through continuous iterative reasoning;
4. The data of correction data is put on the model for testing. In case of underfitting, the model needs to continue to increase the number of iterations; in case of overfitting, the training number should be appropriately reduced to obtain a reasonable network model;
5. The data of test set is put on the model to accurately locate and identify pedestrians.
```

3.2 TRB

When relu sets negative signals to 0, it affects network propagation, which is especially serious at the beginning of training. Therefore, this paper proposed TRB composed of three different basic blocks at different stages of network (as shown in Fig.3, a network stage is a down-sampling or up-sampling stage). In the early stage, compared with the original basic network structure block, the number of activation functions was reduced and the information flow in the network was promoted. As shown in Fig.4, TRB in the network structures were placed in different positions, respectively named TRB-I, TRB-II, and TRB-III.

In this paper, three kinds of basic building blocks were used in each phase. At the beginning of each stage, TRB-I was used as the basic block in this paper to reduce the use of activation function and greatly promoted information flow, enabling originally lost information to promote better fitting learning of the network. In the end of each stage, TRB -III block was used as the basic block to pass the fitted information to the next stage by activation function. And TRB-II block was used as the basic block in other stages. The activation function was directly removed from the end of the TRB-II block to reduce the information flow of the network in this stage.

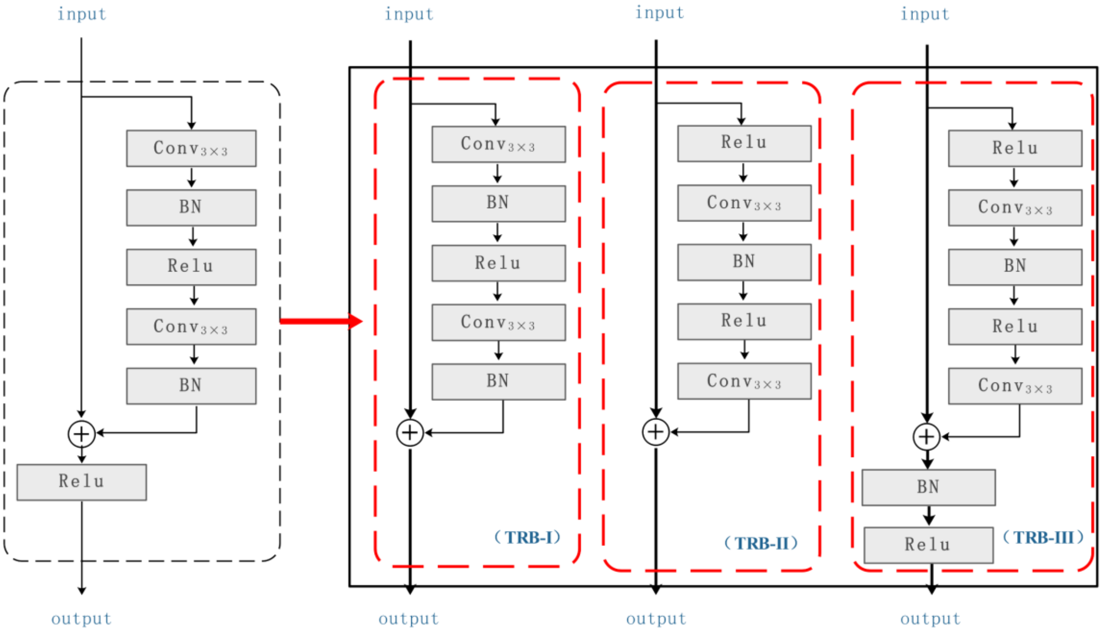


Fig. 4. Three ResNet Blocks (TRB)

3.3 Lead up the activation function to suppress the maximum value

Relu was used for the activation functions in both Hourglass-104 network and DLA-34 network in the backbone network in this paper. With the continuous improvement of the activation function in recent years, relu has solved the problems of gradient disappearance and gradient explosion, but it cannot significantly improve the accuracy of the network. Therefore, the maximum value of the activation function relu used in the network structure was limited in this study, was experimentally evaluated by comparing with other methods.

Relu6 is modified on relu, which is the original activation function relu, and its maximum output value is limited to 6, The specific formula is as follows:

$$relu6(x)=\begin{cases} relu(x) & if\ \ x\le 6 \\ 6 & if\ \ x>6 \end{cases} \tag{2}$$

If the activation range of relu is not limited, numerical explosion will be generated and distributed in a large range, which will have a negative impact on precision. At the same time, the activation function relu6 enables the network to learn the sparse feature as early as possible. Moreover, since the maximum value is limited, the decimal expression ability will be enhanced, which is of great benefit to the datasets with occlusion or small samples.

H-switch makes some improvements on relu6 to reach the limit, increase the number of network layers and reduce the cost of nonlinear activation function, the specific formula is as follows:

$$h-switch(x)=x\frac{relu6(x+3)}{6} \tag{3}$$

3.4 TRBBC-Net detection model

In order to better detect small-scale pedestrians, the TRB and improved activation function were combined with CenterNet framework in this paper, and our end-to-end detection model (TRBBC-Net) was proposed. For better experimental comparison, we marked the part that only changed the basic block as CenterNet + TRB, and the activation function that only suppressed the maximum value as CenterNet + relu6 and CenterNet + h-switch.

As shown in Fig.5, we changed the size of input image to 512×512×3 and entered the network for feature extraction. We changed the original feature extraction basic block to a multi-block group with small-scale features and suppressed the maximum value of the activation function. Aiming at the network loss of small-scale pedestrian features, we suppressed the maximum value of the activation function and maintained the accuracy of feature extraction.

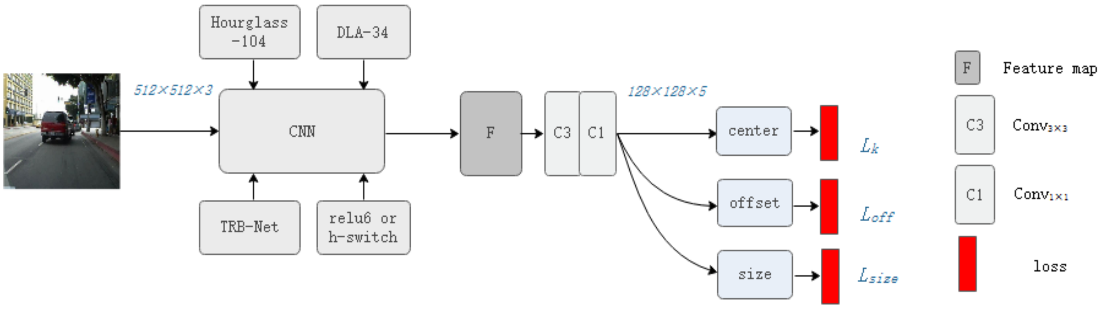


Fig. 5. Detection model proposed in this paper

4. Experiments

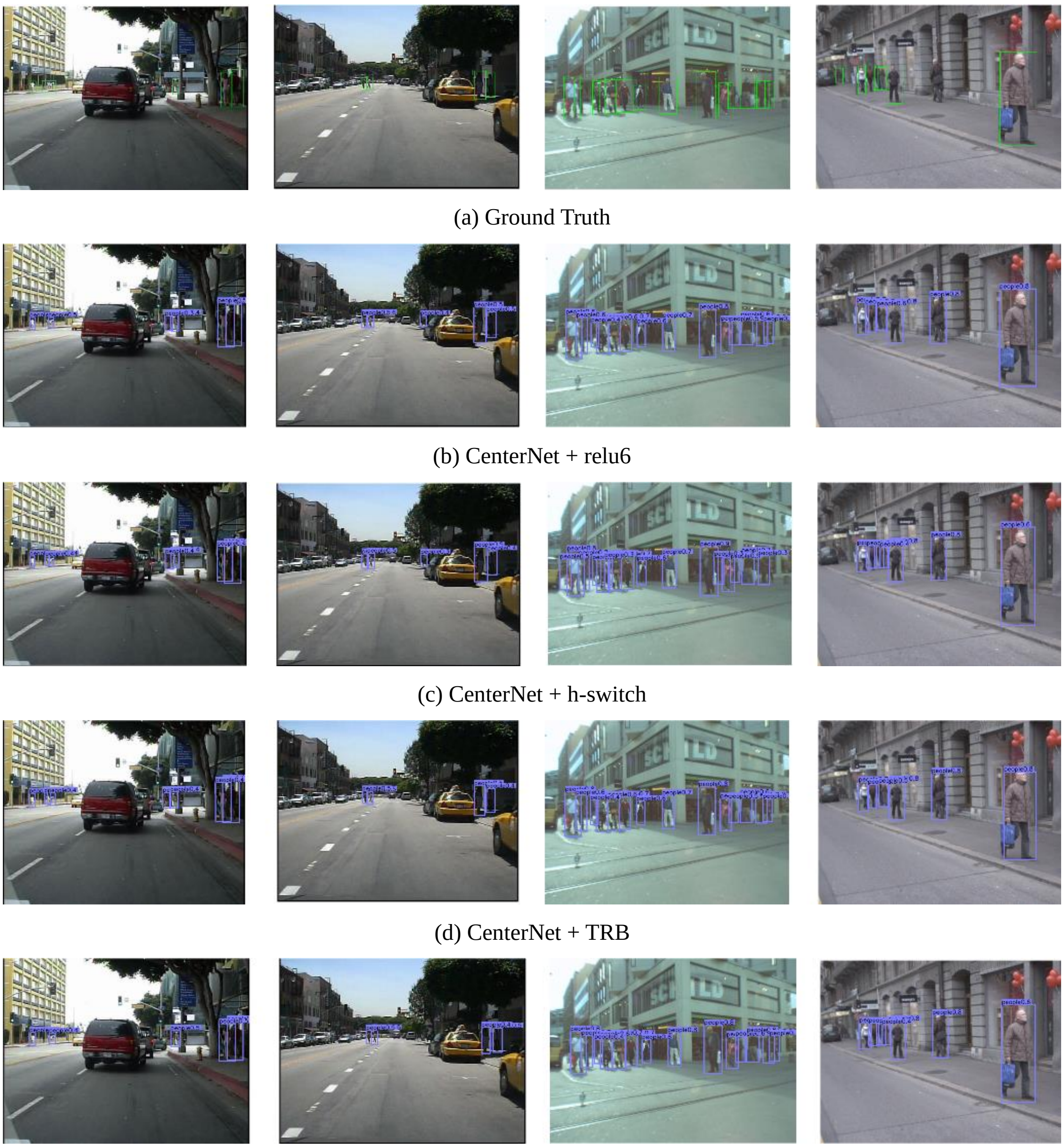
4.1 Experimental Settings

Two datasets were used for experiment. One was the Caltech pedestrian database [28], which was currently the largest vehicle-mounted pedestrian database. The duration of the video was about 10 hours, and its resolution was 640×480, 30 frames per second. The model in this paper was trained through new annotations provided by [29]. There were 42,788 images in the training set and 4,024 images in the standard test set. The Caltech evaluation standard was adopted. Between [10^{-2},10^{0}], the logarithmic average false negative rate MR^{-2} value of each image was taken, that is, each image was false positive (FPPI) as the evaluation index. The other dataset was the ETH pedestrian database, which used a pair of vehicle-mounted AVT Marlins F03 3C cameras for shooting, with a resolution of 640×480 and a frame rate of 13~14 frames per second. There were 1,450 images for training in Group 3, 4, 570 images for correction in Group 5 and 7, and 800 images in Group 8 for testing. In addition to FPPI, average accuracy (AP) was used as a supplement. The higher the AP value, the more accurately pedestrians were detected.

4.2 Result Display

As shown in Fig.6, four samples were randomly selected from the test set of two sample sets. Green box in Fig.6 (a) represents the ground truth for pedestrian detection, from which it can be seen that the features of pedestrians and surrounding environment was too complicated to distinguish. The first and second pedestrians on the left of the first sample were too small and had low pixels, and the first and second pedestrians on the right of the second sample were too dim, making it more difficult to test. The pedestrians in the third sample were seriously blocked, which could be still detected by the algorithm proposed in this paper.

In Fig.6, missing and false detection occurred to samples 1 and 2. After the network was modified, missing detection no longer occurred to sample 2 though false detection still occurred. Therefore, limiting the range of the original network activation function would improve the accuracy of the network to a certain extent. The TRB proposed in this paper was to allow the sample information to circulate as widely as possible in the network. As can be seen from Fig.6, the false detection of the sample could be removed by adding detection frame, indicating that it is necessary to increase network circulation and reuse information.



(a) Ground Truth

(b) CenterNet + relu6

(c) CenterNet + h-switch

(d) CenterNet + TRB

(f) TRBBC-Net

Fig.6. Pedestrian detection results: Rows 1 and 2 are Caltech data sets, and Rows 3 and 4 are ETH data sets. (a) Ground truth of data set; (b) (c) (d) Results of pedestrian detection after network modification; (f) Results of dataset detection visualized by the method proposed in this paper.

4.3 Test results at Caltech

4.3.1. Comparison with the latest methods

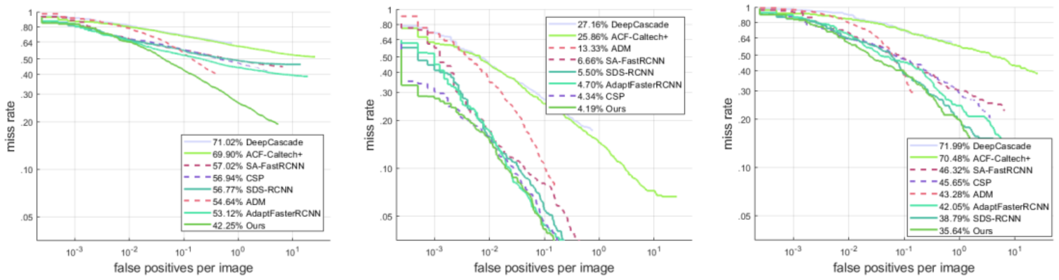
This paper analyzed the Caltech dataset and replaced the activation function relu of the basic block in the network with h-switch. The h-switch activation function was firstly proposed in MobileNetv3, which is also an improved version on relu6, but h-switch has a better effect on the deep network as it can reduce cost of fractional linear activation function. In this paper, the result of the model calculated by the activation function h-switch was 1.09% worse than that calculated by relu6. Relu made the network more sensitive to decimals by restricting its maximum value. Therefore, small targets had a higher recall rate, as shown in Table 1.

TABLE 1

Network improvement results in pedestrian detection

Backbone	Change	Reasonable (MR^{-2}(%))	AP	AR	ARsmall
Hourglass-104	-	9.27%	33.4	-	-
DLA-34	-	7.26%	38.8	24.6	19.1
	CenterNet + relu6	5.30%	39.5	25.3	21.0
	CenterNet + h-switch	6.39%	39.0	25.9	23.0
	CenterNet + TRB	6.14%	41.2	26.0	21.2
	TRBBC-Net	4.19%	42.4	27.5	23.6

As shown in Table 1, the effect of the basic block in the TRB was significantly reduced by 1.12%. The DLA-34 network became more precise at the premise of a small amount of speed loss achieved by modifying its blocks. The MR^{-2} (%) of the proposed network was 3.07, lower than that of the DLA-34 network. Meanwhile, FPPI results in (a) under all conditions are shown in Fig.7. Compared with other methods, the method proposed in this paper could improve small-scale pedestrian detection results. The pedestrians in Fig.7 (a) were all the pedestrians with a pixel greater than 20. In addition to 50-∞ pedestrians, the detection results of small-scale pedestrians in 20-50 was the key of the FPPI value. The detection result calculated by the method proposed in this paper was 10.87% lower than that calculated by AdaptFaster RCNN, indicating that the proposed method is true and effective.



(a) All (b) Reasonable (c) Heavy Occlusion

Fig.7. Comparisons with others on Caltech using new annotations.

TABLE 2

Comparison of detection results for pedestrians at different heights

Method	Occ∈[0.2 0.65]	Test pedestrian height range MR^{-2}(%)				IoU
		[10 20]	[20 30]	[30 40]	[40 50]	

CSP	-	-	-	-	22.02	0.5
SDS-RCNN	-	-	-	-	20.02	0.5
AdaptFaster RCNN	-	-	-	-	15.91	0.5
CenterNet+ relu6	-	82.60	55.89	39.41	16.70	0.5
CenterNet+ h-switch	-	91.76	57.64	47.13	18.49	0.5
CenterNet+ TRB	-	91.20	56.33	39.42	18.02	0.5
TRBBC-Net	-	**80.98**	**54.47**	**38.71**	**15.71**	0.5

4.3.2. Contrast experiment

Height comparison: The pixel results of the pedestrian in the height of 40-50 calculated by classic algorithms such as CSP, SDS-RCNN and AdaptFaster RCNN are shown in Table 2. Their MR^{-2} values were 6.31%, 4.31% and 0.2% higher than that calculated by the method proposed in this paper. But in terms of the detection results for the pixel of 40, MR^{-2} value calculated by the method proposed in this paper was much lower than those calculated by the classic algorithms mentioned earlier.

Occ contrast: This paper compared the results of small-scale pedestrians blocked by Occ∈[0.2 0.65], as shown in Table 3. The method proposed in this paper could capture 43.41% of pedestrians within the pedestrian height range of 40-50, 10% lower than that captured by AdaptFaster RCNN. But classic algorithms could not well capture the blocked pedestrians with a pixel of 30-40, while MR^{-2} calculated by the algorithm proposed in this paper could be up to 67.44%.

IoU comparison: In Table 4, this paper changed the IOU value of pedestrian detection from 0.5 to 0.75, which required the higher accuracy for pedestrians. In this case, MR^{-2} calculated by the algorithm proposed in this paper could be up to 48.79%, indicating that it is aimed at small-scale pedestrians. This paper has achieved a win-win situation in terms of both accuracy and recall rate.

TABLE 3

Comparison of detection results for occlusion pedestrians under different heights

Method	Occ∈[0.2 0.65]	Test pedestrian height range MR-2(%)				IoU
		[10 20]	[20 30]	[30 40]	[40 50]	
ADM[30]	√	-	-	-	56.08	0.5
SDS-RCNN	√	-	-	-	57.02	0.5
AdaptFaster RCNN	√	-	-	-	53.90	0.5
CenterNet+ relu6	√	-	-	69.05	54.91	0.5
CenterNet+ h-switch	√	-	-	73.35	58.64	0.5
CenterNet+ TRB	√	-	-	70.01	54.48	0.5
TRBBC-Net	√	-	-	**67.44**	**43.41**	0.5

TABLE 4

Comparison of detection results of pedestrians at different heights and IoU

Method	Occ∈[0.2 0.65]	Test pedestrian height range MR-2(%)				IoU
		[10 20]	[20 30]	[30 40]	[40 50]	
CenterNet+ relu6	-	-	-	-	51.96	0.75
CenterNet+ h-switch	-	-	-	-	52.04	0.75
CenterNet+ TRB	-	-	-	-	51.47	0.75
TRBBC-Net	-	-	-	-	**48.79**	0.75

4.4 Detection results on the ETH dataset

In the ETH dataset, the pedestrian data was mostly relatively balanced pedestrians, and the proportion of small-scale pedestrians was relatively small. As shown in Table 5, the detected AP value was 79.9 in an unmodified network, but after it increased to some extent in the network modified by the method proposed in this paper. In the proposed TRB, it increased by 0.6%. However, combined with the activation function relu6, the AP value of the method proposed in this paper was up to 81. The algorithm proposed in this paper could still be used for mining the information of 1,450 images in the ETH training set, so this algorithm can be extensively applied.

TABLE 5

Network improvement results in ETH dataset

Backbone	Change	AP
Hourglass-104	-	79.3
DLA-34	-	79.9
	CenterNet + relu6	80.9
	CenterNet + h-switch	79.8
	CenterNet + TRB	80.5
	TRBBC-Net	**81.0**

4.5 Importance of Speed

In the network of this paper, by comparing the DLA-34 network in the backbone with the Hourglass-104 network, it was concluded that the AP values of DLA-34 network accuracy on Caltech and ETH datasets increased by 5.4 and 0.6, respectively, as shown in Table 6. In addition, the DLA-34 network model was faster in terms of the predicted speed of the model, which met the requirements of the speed and accuracy of pedestrian detection. Under the condition of multiple scales, this paper selected 1, 1.25, 1.5, 0.5 and 0.75. As can be clearly seen in Table 6, the accuracy of DLA-34 network was improved by 0.7 and 0.2, but at the expense of about five times the time. Therefore, all the prediction experiments in this paper were discussed on a single scale.

TABLE 6

Results of two different datasets (AP-1: Caltech, AP-2: ETH) corresponding to two backbones in single-scale and multi-scale

	AP-1	AP-2	Time(ms/img)
Hourglass-104 (Single Scale)	33.4	79.3	77
DLA-34(Single Scale)	38.8	79.9	28
Hourglass-104(Multi Scale)	35.1	80.1	365
DLA-34(Multi Scale)	39.5	80.2	127

5. Conclusion

This paper proposed a TRBBC-Net detection model for small-scale pedestrians. The multi-block group TRB was used for feature extraction, which integrated three different basic blocks to enhance the information flow in the network structure. In addition, an activation function that restrained the maximum value was introduced to retain the information of small-scale features. A large number of experiments on two pedestrian datasets demonstrate the effectiveness of the model. In the proposed model, small-scale feature information is largely dependent on the mining of feature samples by network, and there is a lack of specific means, so we should improve the design of the network structure.

References:

[1] A. Geiger, P. Lenz, R. Urtasun, Are we ready for autonomous driving? the kitti vision benchmark suite, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 3354-3361.

[2] X. Wang, M. Wang, W. Li, Scene-specific pedestrian detection for static video surveillance, IEEE Trans. Pattern Anal. Mach. Intell. 36 (2) (2014) 361-374.

[3] W. Y. Hsu, Automatic pedestrian detection in partially occluded single image, Integr. Comput.-Aided Eng. 25 (4) (2018) 369-379.

[4] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, IEEE Trans. Pattern Anal. Mach. Intell. 39 (6) (2017) 1137-1149.

[5] Z. Cai, N. Vasconcelos, Cascade R-CNN: Delving into high quality object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6154-6162.

[6] S. Zhang, L. Wen, X. Bian, S. Z. Lei, Occlusion-aware R-CNN: Detecting pedestrians in a crowd, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 637-653.

[7] T. Wang, X. Zhang, L. Yuan, J. Feng, Few-shot adaptive faster R-CNN, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 7166-7175.

[8] P. Viola, M. Jones, Robust real-time face detection, Int. J. Comput. Vision 57 (2) (2004) 137-154.

[9] R. Lienhart, J. Maydt, An extended set of Haar-like features for rapid object detection, in: Proceedings of the International Conference on Image Processing, 2002, pp. I-I.

[10] K. Liao, G. Liu, Y. Hui, An improvement to the sift descriptor for image representation and matching, Pattern Recognit. Lett. 34 (11) (2013) 1211-1220.

[11] P. Dollár, Z. Tu, P. Perona, S. Belongie, Integral channel features, Proc. British Machine Vision Conf. 2009.

[12] X. Wang, T.X. Han, S. Yan, An HOG-LBP human detector with partial occlusion handling, in: Proceedings of the IEEE 12th International Conference on Computer Vision, 2009, pp. 32-39.

[13] P. Dollár, R. Appel, S. Belongie, P. Perona, Fast feature pyramids for object detection, IEEE Trans. Pattern Anal. Mach. Intell. 36 (8) (2014) 1532-1545.

[14] W. Nam, P. Dollár, J.H. Han, Local decorrelation for improved pedestrian detection, in: Proceedings of the 27th International Conference on Neural Information Processing Systems, 2014, pp. 424–432.

[15] L. Zhang, L. Lin, X. Liang, K. He, Is faster R-CNN doing well for pedestrian detection?, in: Proceedings of the European Conference on Computer Vision, 2016, pp. 443-457.

[16] J. Li, X. Liang, S. Shen, T. Xu, J. Feng, S. Yan, Scale-aware fast R-CNN for pedestrian detection, IEEE Trans. Multimedia 20 (4) (2018) 985-996.

[17] R. Girshick, Fast R-CNN, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1440-1448.

[18] X. Du, M. El-Khamy, J. Lee, L. Davis, Fused Dnn: A deep neural network fusion approach to fast and robust pedestrian detection, in: Proceedings of the IEEE Winter Conference on Applications of Computer Vision, 2017, pp. 953-961.

[19] S. I. Jung, K. S. Hong, Deep network aided by guiding network for pedestrian detection, Pattern Recognit. Lett. 90 (2017) 43-49.

[20] Z. Wu, C. Shen, A. van den Hengel, Wider or deeper: Revisiting the ResNet model for visual recognition, Pattern Recognit. 90 (2019) 119-133.

[21] G. Brazil, X. Yin, X. Liu, Illuminating pedestrians via simultaneous detection and segmentation, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 4960-4969.

[22] T. Song, L. Sun, D. Xie, H. Sun, S. Pu, Small-scale pedestrian detection based on somatic topology localization and temporal feature aggregation, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 536-551.

[23] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, SSD: Single shot multibox detector, in: Proceedings of the European Conference on Computer Vision, 2015, pp. 21-37.

[24] P. Ballester, R.M. Araujo, On the performance of GoogLeNet and AlexNet applied to sketches, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2016, pp. 1124–1128.

[25] W. Liu, S. Liao, W. Hu, X. Liang, X. Chen, Learning efficient single-stage pedestrian detectors by asymptotic localization fitting, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 643-659.

[26] W. Liu, I. Hasan, S. Liao, Center and scale prediction: A box-free approach for pedestrian and face detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019.

[27] C. Zhu, Y. He, M. Savvides, Feature selective anchor-free module for single-shot object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 840-849.

[28] P. Dollar, C. Wojek, B. Schiele, P. Perona, Pedestrian detection: An evaluation of the state of the art, IEEE Trans. Pattern Anal. Mach. Intell. 34 (4) (2012) 743-761.

[29] S. Zhang, R. Benenson, M. Omran, J. Hosang, B. Schiele, How far are we from solving pedestrian detection?, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1259-1267.

[30] X. Zhang, L. Cheng, B. Li, H. Hu, Too far to see? not really!—pedestrian detection with scale-aware localization policy, IEEE Trans. Image Process. 27 (8) (2018) 3703-3715.