2022 IEEE 5th Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC) | 978-1-6654-7968-4/22/$31.00 ©2022 IEEE | DOI: 10.1109/IMCEC55388.2022.10019992

Pedestrian detection based on multi-scale feature fusion

Lincai Huang [1] , Zhiwen Wang [2,*] , Xiaobiao Fu [1]
1. School of Automation, Guangxi University of Science and Technology, Liuzhou, China
2. School of Computer Science and Technology, Guangxi University of Science and Technology, Liuzhou, China
2267104777@qq.com, wzw69@126.com, f126youxing@126.com

Abstract—**To solve the problem of a large difference in target size in pedestrian detection, which leads to high pedestrian false detection rate and high miss detection rate of small-scale pedestrians, a multi-scale feature fusion method based on RetinaNet is proposed. After feature enhancement by extracting features from the backbone network, three branches of different scales are formed, which are effectively fused with the corresponding feature layer to further enrich the target information, and then detect pedestrians of different scales. Test on the open dataset shows that compared with the original RetinaNet algorithm, it can detect more pedestrians, especially small-scale pedestrians, and the model detection performance is better.**

Keywords—Pedestrian detection; RetinaNet; Multi-scale; Feature fusion

I. Introduction

With the development of the times, artificial intelligence is widely used in all aspects of life. As one of the important components of artificial intelligence, pedestrian detection technology plays an irreplaceable role in automatic driving, pedestrian motion analysis, etc. [1]. The traditional pedestrian detection method extracts pedestrian features manually and then detects them by feature classifier. This traditional detection method is not applicable in the case of variable scenes, which can not achieve rapid and accurate identification, and can not meet the detection requirements.

The deep learning algorithm can better solve this problem. According to the detection steps, the deep learning algorithm can be divided into two categories:

(1) Target detection algorithm based on two-phase: first generate candidate regions, and then classify and regression the candidate regions. For example, R-CNN and Faster R-CNN [2,3] are characterized by their high detection accuracy through candidate regions, resulting in slow detection speed and a large number of calculations.

(2) Target detection algorithm based on single stage: the image is directly sent to the neural network for feature extraction. In judging the category and position regression of objects, such algorithms include SSD, RetinaNet and YOLO series algorithms [4,5,6]. Without candidate boxes, the detection speed is rapidly improved. Many scholars have made a series of improvements to improve detection accuracy. The E-SSD algorithm proposed by Sun et al. [7] uses a deconvolution feature fusion module to enhance the detection of small-scale pedestrians and adds an attention mechanism to each feature layer to obtain more valuable target information. Cao et al. [8] improved the K-means clustering algorithm based on YOLO v3 to design anchor frame size and use the Soft NMS loss function. The model has strong anti-interference ability and generalization ability in complex scenes.

Although pedestrian detection technology has been widely improved, there are still some research difficulties unresolved [9]. In the face of complex and changeable environments, such as rain, fog, sand and dust, the picture quality is low and the pedestrians are ambiguous. Due to shooting, it is easy to have different sizes of pedestrian targets in a picture, which results in that pedestrians close to the shooting point, usually called large-scale pedestrians, being easy to be detected, while small-scale pedestrians far away, with low-resolution and few extracted features, lead to missed detection and false detection.

This paper studies the multi-scale problem of pedestrians and proposes a pedestrian detection method based on multi-scale feature fusion. By designing a multi-scale feature fusion module, the information of multiple feature layers is fused according to the corresponding scale, and then pedestrians of different sizes are detected.

II. Research on RetinaNet Pedestrian Detection Based on Multi-scale Feature Fusion

A. RetinaNet introduction

RetinaNet is a multi-scale prediction model. Its structure is shown in Fig.1. It consists of three main parts, namely feature extraction network, feature pyramid and boundary box prediction. ResNet50 is used as the feature extraction network. The residual network is a classic and effective feature extraction network. The gradient disappearance and gradient explosion caused by network deepening are effectively solved through residual connection. Feature Pyramid Networks (FPN) obtains the feature layers C3, C4 and C5 through the backbone network, after 1 × 1 convolutional horizontal connection converts the channel number to 256, fuses the shallow location information with the deep semantic information, and generates multi-scale feature maps of P3, P4, and P5. The P5 feature layer has a large receptive field, which is suitable for detecting large-scale targets, and P3 has a small receptive field, which is suitable for detecting small targets. The prediction part is connected after the FPN, including the target classification branch and the location branch, it's connected by four convolution numbers of 3 × 3, and finally connect the classified output and border output respectively.

Focal Loss function is used for classification loss function and the Smooth L1 loss function is used for regression loss function during training of RetinaNet model. Focal Loss function is calculated by cross-entropy loss function through formulas (1), (2) and (3). The original loss is multiplied by the adjustment factor, as shown in formula (4), α is the adjustment factor and the γ adjustment parameter.

$$\mathrm{CE}(\mathrm{p},\mathrm{y}) = \begin{cases} -\log(\mathrm{p}) & \text{if } \mathrm{y} = 1 \\ -\log(1-\mathrm{p}) & \text{otherwise} \end{cases} \qquad (1)$$

$$\mathrm{p_t} = \begin{cases} \mathrm{p} & \text{if } \mathrm{y} = 1 \\ 1-\mathrm{p} & \text{otherwise} \end{cases} \qquad (2)$$

$$\mathrm{CE}(\mathrm{p_t}) =- \alpha_\mathrm{t}\log(\mathrm{p_t}) \qquad (3)$$

$$\mathrm{FL}(\mathrm{p_t}) =- \alpha_\mathrm{t}(1-\mathrm{p_t})^{\gamma}\log(\mathrm{p_t}) \qquad (4)$$

y represents the value of the real tag, p is probability, p_t is the estimated probability, CE is the cross-entropy loss, FL is Focal Loss.

B. Multi-scale feature fusion module

Large-scale objects have clear outlines and obvious features, while small-scale objects have fuzzy outlines and unclear details in the image, which give people different visual features. How to consider the detection of both is challenging research. Usually, it is difficult to detect multi-scale targets using a single feature layer. The feature pyramid combines the high-level features with the low-level features through bottom-up, horizontal connection and top-down operations to obtain the feature layer information of different levels, which can significantly improve the multi-scale object detection of pedestrians and is widely used in multi-scale target detection.

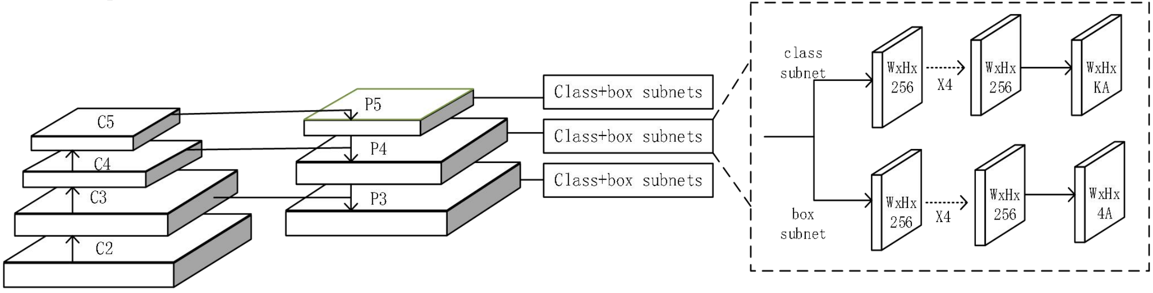


Fig. 1. RetinaNet network structure.

In order to further deal with the multi-scale problem of pedestrians, this paper designs a Multi-Scale Feature Fusion module (MSFF), whose structure is shown in Fig.2. Based on RetinaNet, after feature extraction of the backbone network, the multi-scale feature fusion module is introduced to enhance the extracted features, generate three branches of different sizes, and connect them to the corresponding feature pyramid layer for additive fusion.

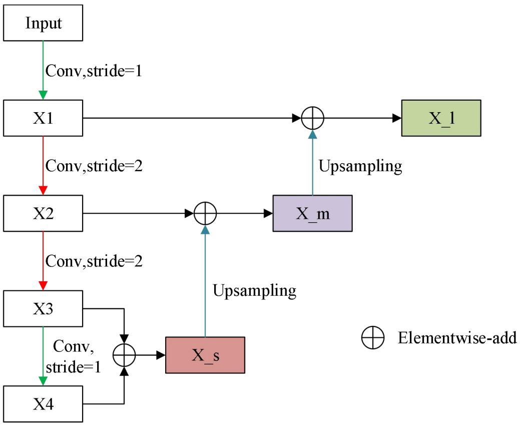


Fig. 2. Multi-Scale Feature Fusion module (MSFF)

Carry out multi-scale processing on the features extracted from the backbone network. First, in order to get more semantic information, the input feature layer is first subjected to a 3×3 convolution with a step size of 1 and Relu activation function, maintaining the resolution to get X1, and then convolution with two steps of size 2 and activation function, X2 and X3 are obtained. The resolution gradually decreases, and then goes through a convolution and activation function with a step size of 1 to get X4. X4 and X3 are added and fused to get X_s with smaller resolutions. X_s is added and fused with X2 through an up-sampling to get X_m with a medium resolution, and X_m is fused with X1 through an upsampling, to get the feature map with a higher resolution of X_l. In this way, three outputs with different sizes are obtained through a series of convolutions and fusion, and then the output is fused with the corresponding features with the same resolution.

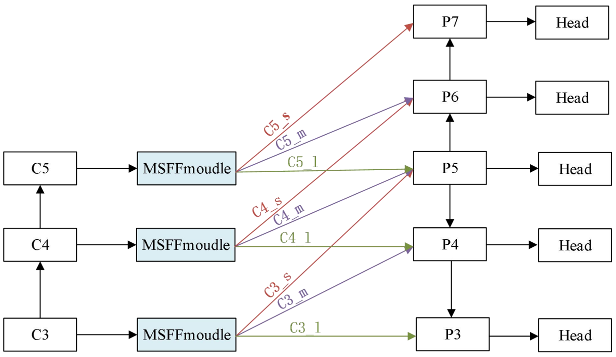


Fig. 3. Improved Retinanet

The overall structure is shown in Fig.3. Taking C5 as an example, after the multi-scale feature fusion module, three branches, C5_s, C5_m and C5_l, are generated with successively larger feature sizes. Among them, C5_l and P5 have the same feature resolution and can be added and fused. C5_m has the same resolution as P6, C5_s has the same resolution as p7, and performs fusion in turn. Therefore, each layer of the feature pyramid integrates spatial information and semantic information of different levels and scales. For example, the P5 layer integrates information from C5_l, C5_m and C3_s, compared with the original network, the feature layer has stronger feature extraction ability.

III. Experimental results and analysis

A. Datasets

This section conducts experiments on the public dataset to test the method proposed in this paper.

(1) PASCAL VOC dataset and COCO dataset are common datasets for target detection, which contain various common targets in life. In this paper, images containing the 'person' tag are extracted as a mixed data set. A total of 3288 photos contain various scenes in life, which is conducive to enriching the dataset and making the detection more authentic. 2630 training sets and 658 test sets are divided according to 8:2. The evaluation index adopts mAP to represent the average accuracy, F1 Score to represent the harmonic average of accuracy and recall, as well as Recall and Precision.

(2) Caltech pedestrian dataset is a video shot by the California Institute of Technology through the car camera, with a resolution of 640 × 480, containing about 250000 pictures and 350000 pedestrian frames, divided into Set00-Set10 sub-datasets, Set00-Set05 as the training set and Set06-Set10 as the test set. 4310 training sets and 4250 test sets were selected. The evaluation index adopts the missed detection rate index proposed by Dollar et al. [10], which refers to the false positive (FPPI) of each image on average between $[10^{-2},10^{0}]$. The lower the value, the better, expressed in MR^{-2}. The test subset is divided according to the height of pedestrians, as shown in Table I.

Table I. Subset division of multi-scale pedestrian test

Test Subset	Pedestrian height (pixels)
Reasonable	More than 50, visibility more than 65%
Large	more than 100
Near	[80，100]
Medium	[30，80]
Far	[20，30]

B. Experimental setup

The software environment used in this experiment is Windows, Pytorch framework, CUDA10.2, the hardware environment CPU is Intel i7-8700 processor, and the GPU is GTX1080. The input image size during experimental training is 512 × 512, the batch is 8, the optimizer is Adam, the initial learning rate is 1e-4, the learning rate is reduced by cosine annealing, and the momentum is 0.9.

C. Experimental analysis

With the original RetinaNet model as the test benchmark, the algorithm after adding the feature fusion module is tested on two datasets.

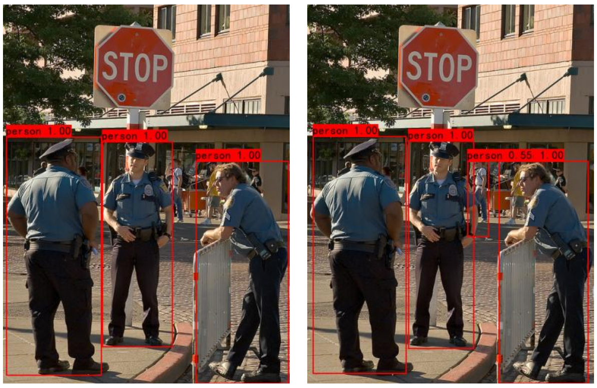
(1) In the mixed dataset, the experimental results are shown in Table II.

Table II. Test Results of Mixed Datasets

	mAP	F1	Recall	Precision
baseline	78.68%	0.77	71.61%	83.68%
ours	79.78%	0.78	72.29%	85.67%

It can be clearly seen from Table II that RetinaNet with feature fusion is better than the benchmark indicators in all indicators, in which mAP is 1.1% higher, F1 increased by 0.01, recall is 0.62% higher, and Precision is 1.99% higher. Therefore, from the perspective of evaluation indicators, the algorithm proposed in this paper shows certain advantages.

The mixed dataset detection effect is shown in Fig. 4.



(a) baseline (b)ours

Fig. 4. Test Results of Mixed Datasets

In Fig.4, in the original Retinanet, large-scale pedestrians are easy to detect, but small-scale pedestrians have missed detection. The algorithm proposed in this paper can accurately detect large-scale and small-scale pedestrians.

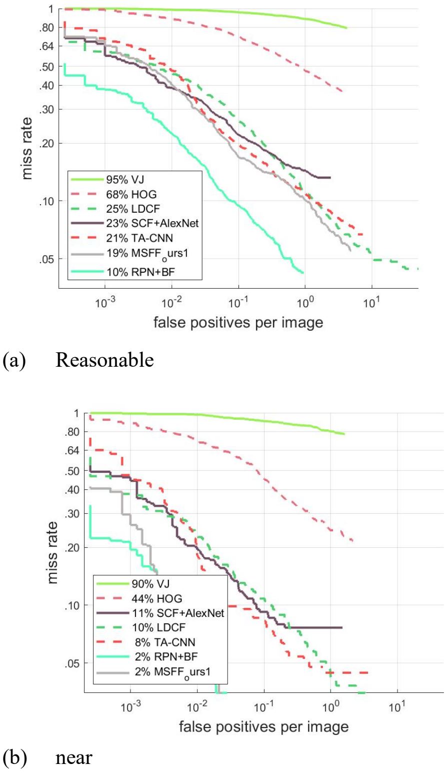
(2)The MR^{-2} detected on the Caltech dataset is shown in Table III.

TABLE III. TEST RESULTS ON CALTECH DATASET

	Reasonable	large	near	medium	far
baseline	19.98%	1.47%	2.77%	52.14%	96.23%
ours	19.14%	1.22%	1.52%	51.53%	93.49%

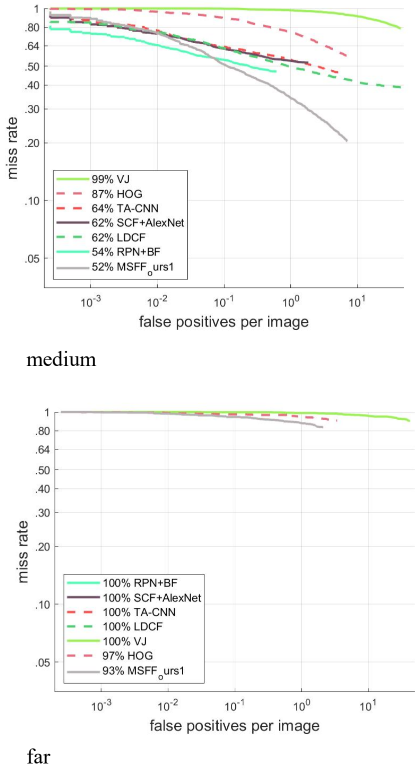
From the above results, it can be clearly seen that in the Reasonable subset, compared with the benchmark algorithm, the missing detection rate of the proposed algorithm is reduced by 0.84%. On the large-scale, it decreased by 0.25%, the near scale decreased by 1.25%, the medium scale decreased by 0.61%, the far scale decreased by 2.74%, and the missed detection rate of small scale pedestrians decreases significantly.

In this paper, LDCF[11], SCF+AlexNet[12], TA-CNN[13], RPN+BF[14] and other algorithms were selected to compare with MSFF-ours1 algorithm proposed by us. MR-FPPI curves of each subset are shown in Fig.5, and the algorithm in this paper has significant effects on subsets of all scales. Although there is a certain gap between the Reasonable subset and RPN+BF, the effect is better on subsets of other scales, especially on small-scale pedestrians of the far subset.



(a) Reasonable

(b) near



(c) medium

(d) far

Fig. 5. MR-FPPI curves for each subset

The detection effect of Caltech dataset is shown in Fig.6.

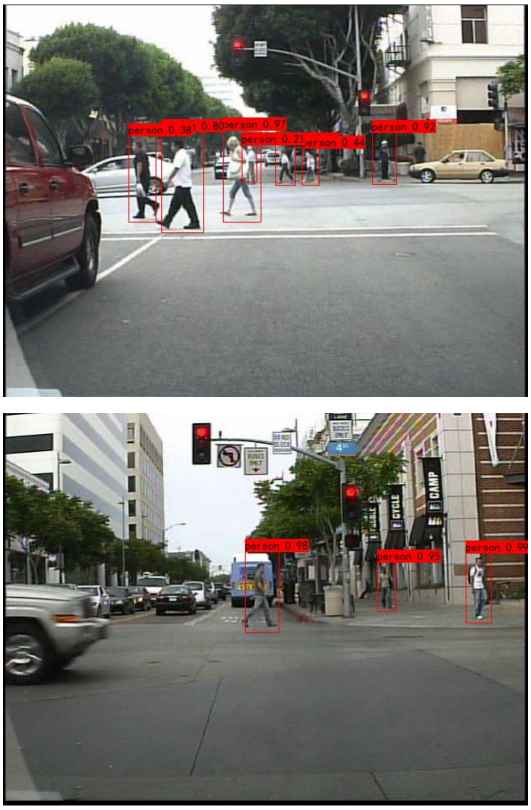


Fig.6. Caltech Dataset Test Results

In Fig.6, there are many small-scale pedestrians in Caltech dataset, and the algorithm proposed in this paper can accurately detect travelers.

The above experimental results show that the algorithm proposed in this paper can perform better than the baseline algorithm on different datasets, indicating the effectiveness of this algorithm and the strong robustness of this model on different datasets. After analysis, the proposed multi-scale feature fusion module can enhance the expression of features, and through multi-scale fusion, relatively complete feature information can be obtained, which is conducive to the detection of pedestrians at various scales.

IV. Conclusions

Aiming at the multi-scale problem of pedestrians, in this paper proposes a multi-scale feature fusion module based on the single-stage RetinaNet model. The fusion of feature layers of each scale can further enhance the feature expression ability, better identify pedestrians of different scales and improve the detection accuracy. Through the use of different datasets, the model is superior to the baseline model, which verifies that the model has good generalization.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (61962007, 61462008, 61751213, 61866004); Guangxi Natural Science Foundation Key Project (2018GXNSFDA294001, 2018GXNSFDA281009); Guangxi Natural Science Foundation (2018GXNSFAA294050); Guangxi Multi-source Information Mining and Security Key Laboratory Open Fund Project (MIMS19-04); Guangxi University of Science and Technology Innovation Team Project (gxkjdxcxtd201504); Guangxi Graduate Education Innovation Project (GKYC202106, GKYC202104, YCSW2021320); Guangxi University Student Innovation and Entrepreneurship Project (202110594133, 202110594134)

References

[1] Galvao L G, Abbod M, Kalganova T, et al. Pedestrian and Vehicle Detection in Autonomous Vehicle Perception Systems—A Review[J]. Sensors, 2021, 21(21): 7267.

[2] Wang M, Chen H, Li Y, et al. Multi-scale pedestrian detection based on self-attention and adaptively spatial feature fusion[J]. IET Intelligent Transport Systems, 2021, 15(6): 837-849.

[3] Ren J, Han J. A new multi-scale pedestrian detection algorithm in traffic environment[J]. Journal of Electrical Engineering & Technology, 2021, 16(2): 1151-1161.

[4] Yang S, Chen Z, Ma X, et al. Real-time high-precision pedestrian tracking: a detection–tracking–correction strategy based on improved SSD and Cascade R-CNN[J]. Journal of Real-Time Image Processing, 2022, 19(2): 287-302.

[5] Pei D, Jing M, Liu H, et al. A fast RetinaNet fusion framework for multi-spectral pedestrian detection[J]. Infrared Physics & Technology, 2020, 105: 103178.

[6] Hsu W Y, Lin W Y. Adaptive fusion of multi-scale YOLO for pedestrian detection[J]. IEEE Access, 2021, 9: 110063-110073.

[7] Cao J, Song C, Peng S, et al. Pedestrian detection algorithm for intelligent vehicles in complex scenarios[J]. Sensors, 2020, 20(13): 3646.

[8] Cao J, Song C, Peng S, et al. Pedestrian detection algorithm for intelligent vehicles in complex scenarios[J]. Sensors, 2020, 20(13): 3646.

[9] Han B, Wang Y, Yang Z, et al. Small-scale pedestrian detection based on deep neural network[J]. IEEE transactions on intelligent transportation systems, 2019, 21(7): 3046-3055.

[10] Dollar P, Wojek C, Schiele B, et al. Pedestrian detection: An evaluation of the state of the art[J]. IEEE transactions on pattern analysis and machine intelligence, 2011, 34(4): 743-761.

[11] Nam W, Dollár P, Han J H. Local decorrelation for improved pedestrian detection[J]. Advances in neural information processing systems, 2014, 27.

[12] Hosang J, Omran M, Benenson R, et al. Taking a deeper look at pedestrians[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 4073-4082.

[13] Tian Y, Luo P, Wang X, et al. Pedestrian detection aided by deep learning semantic tasks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 5079-5087.

[14] Zhang L, Lin L, Liang X, et al. Is faster R-CNN doing well for pedestrian detection?[C]//European conference on computer vision. Springer, Cham, 2016: 443-457.