



Pedestrian detection using RetinaNet with multi-branch structure and double pooling attention mechanism

Lincai Huang¹ · Zhiwen Wang² · Xiaobiao Fu¹

Received: 2 August 2022 / Revised: 13 April 2023 / Accepted: 15 May 2023 /
Published online: 1 June 2023

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

Abstract

Pedestrian detection technology, combined with techniques such as pedestrian tracking and behavior analysis, can be widely applied in fields closely related to people's lives such as traffic, security, and machine interaction. However, the multi-scale changes of pedestrians have always been a challenge for pedestrian detection. Aiming at the shortcomings of the traditional RetinaNet algorithm in multi-scale pedestrian detection, such as false detection, missed detection, and low detection accuracy, an improved RetinaNet algorithm is proposed to enhance the detection ability of the network model. This paper mainly makes innovations in the following two aspects. Firstly, in order to obtain more semantic information, we use a multi-branch structure to expand the network and extract the characteristics of different receptive fields at different depths. Secondly, in order to make the model pay more attention to the important information of pedestrian features, double pooling attention mechanism module is embedded in the prediction head of the model to enhance the correlation of feature information between channels, suppress unimportant information, and improve the detection accuracy of the model. Experiments were conducted on different datasets such as the COCO dataset, and the results showed that compared with the traditional RetinaNet model, the model proposed in this paper has improved in various evaluation indicators and has good performance, which can meet the needs of pedestrian detection.

Keywords pedestrian detection · RetinaNet · multi-branch construction · double pooling attention mechanism

1 Introduction

Pedestrian detection mainly studies the accurate detection of pedestrians and their positions from images or videos. Pedestrian detection technology can be applied to various fields of life and plays an important role in automatic driving and intelligent security [33]. With

✉ Zhiwen Wang
wzw69@126.com

¹ School of Automation, Guangxi University of Science and Technology, Liuzhou 545000, Guangxi, China

² School of Computer Science and Technology, Guangxi University of Science and Technology, Liu Zhou 545000, Guangxi, China

the development of artificial intelligence, deep learning technology is becoming more and more mature, and the research on pedestrian detection based on deep learning technology has also become a hot research topic, which has important application value and academic significance for today's social development [14, 28]. Through continuous innovation and development, pedestrian detection technology has achieved good results so far. According to research methods, it can be divided into two categories: pedestrian detection based on traditional feature extraction and pedestrian detection based on deep learning.

Traditional feature extraction methods mainly detect pedestrians through low-level feature extraction and feature classifiers. The underlying features, such as grayscale features, color features, and shape features, are typically HOG [51], LBP [17], etc. This method of using static features to describe specific targets often relies on people's experience to manually design features, and then use feature classifiers for information classification, such as SVM [3] and Ada-boost [20]. Mihçioğlu et al. [26] applied the detection based on the gradient histogram feature to pedestrian detection, which made significant progress in pedestrian detection technology. Kumar et al. [16] improved the accuracy of pedestrian detection by describing the shape and texture features of pedestrians and using a linear SVM classifier. Due to the complexity and change of the real scene, this manual feature method makes the generalization ability of the model weak and the robustness is poor and cannot meet realistic requirements [42].

With the rapid development and wide application of neural networks, the performance of object detection has been greatly improved. Target detection algorithms based on deep learning can be divided into one-stage and two-stage target detection algorithms, among which one-stage target detection algorithms such as SSD [35], YOLO series [4, 13, 44], RetinaNet [10, 50]. The detection algorithm based on logical regression is fast and the accuracy is reduced; while the two-stage algorithm R-CNN, Faster R-CNN [9, 43] and other algorithms based on candidate region selection are slow in detection, but have high accuracy. The two types of algorithms have their advantages and disadvantages. The continuous improvement of technology has led researchers to focus on the difficulties of pedestrian detection. For example, the application of pedestrian detection is easily affected by its own and external environment, and the occlusion problem of pedestrians in complex environments is not conducive to detection. In the application of autonomous driving, the detection speed should meet the requirements of real-time performance. In harsh weather environments, it is also important to ensure the accuracy of the detection. Therefore, in order to be more suitable for pedestrian detection, pedestrian detection technology still needs further improvement and optimization, and many scholars have proposed improved methods [19, 29]. Jiang et al. [15] proposed multi-spectral pedestrian detection, which complements the extracted feature information from two modes, effectively integrates deep features and thermal image features, and achieves multidimensional data mining, achieving a balance between speed and accuracy on the KAIST dataset; To accurately locate pedestrians at night, Li et al. [18] proposed using the YOLOv3 algorithm to detect pedestrians in infrared images at night. Yi et al. [46] used K-means clustering to find the most suitable anchor box on the pedestrian dataset, but there were still cases of small object missed detection. Li et al. [22] proposed a lightweight pedestrian detection network based on YOLO v5, using the Ghost module to reduce model parameters and computational complexity. Although there have been significant improvements in pedestrian detection technology, there are still cases of multi-scale pedestrian missed detections [30, 41, 49].

The multi-scale problem of pedestrians has always been a difficulty in detection, as in actual scenarios, the size of pedestrians often changes, accompanied by occlusion and image blurring, which is not conducive to machine detection. In response to the above issues, this article applies the RetinaNet model to pedestrian detection and proposes a

RetinaNet pedestrian detection based on a multi-branch structure and double pooling attention mechanism. The main contributions of this article are summarized as follows:

- (1) We propose adding a multi-branch structure to enhance the connection between the backbone network and the feature pyramid, expand the network width, extract multi-scale pedestrian feature information at different depths in multiple directions, and improve the expression ability of features.
- (2) We have improved the attention mechanism and proposed using double pooling attention mechanism module to increase the weight of important feature information, making the model more focused on key feature information.
- (3) Experiments and comparisons on multiple datasets have achieved excellent results, verifying the effectiveness of the model proposed in this paper.

The organizational arrangement of this article is as follows. The second part introduces the relevant work of multi-scale pedestrian detection, the third part elaborates in detail on the improved RetinaNet model, the fourth part conducts comparative experiments on different datasets such as the COCO dataset, and analyzes the experimental results, and the fifth part summarizes and prospects the future research of pedestrian detection technology.

2 Related work

Although current algorithms have achieved good results in general object detection, their application in pedestrian detection requires further improvement based on the characteristics and actual situation of pedestrians to achieve better results.

2.1 Multi-scale pedestrian detection

The multi-scale variation of pedestrians refers to the presentation of pedestrians of different scale sizes in the image, as shown in Fig. 1. This is due to the shooting distance. When the distance between pedestrians and the camera is different, the scale of the pedestrians

Fig. 1 Multi-scale pedestrian



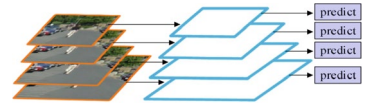
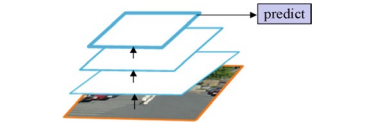
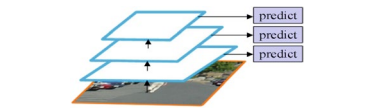
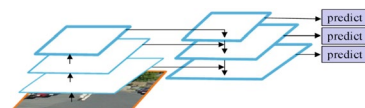
captured will vary. Large scale pedestrians are easy to detect, while small scale pedestrians have small pixels and blurred images, which can easily lead to missed detections. The multi-scale changes of pedestrians often result in unsatisfactory detection of small-scale pedestrians in images, which is a common and challenging problem in detection.

With the advancement of computer equipment and detection algorithms, there are also certain solutions to the multi-scale problem of pedestrians. He et al. [11] improved the anchor box and proposed an adaptive method for generating anchor boxes based on the multi-scale changes of pedestrians. To enhance the expressiveness of features, Wang et al. [38] embedded the non-local modules of the asymmetric pyramid into the backbone network and solved the problem of scale inconsistency with the method of adaptive spatial feature fusion (ASFF) [31]. Li et al. [21] believed that the design of anchor frames would introduce too many parameters, so they adopted an anchor-free frame detector and constructed a feature pyramid to dynamically transmit feature information at various scales, which has advantages in multi-scale pedestrian detection. Shao et al. [34] proposed that the double feature pyramid can effectively fuse multi-scale information in occluded pedestrians, making the extracted features clearer. Szegedy et al. [36] proposed the Inception network structure, which is characterized by the expansion between network layers, the introduction of multi-branch structure, and the use of convolution cores of different sizes to achieve diversity of network width, obtaining a receptive field of different sizes, and achieve multi-scale detection of images. In the traditional deep learning network model, it is usually used to increase the depth and width of the model to improve its performance of the model. With the deepening of the layers of the neural network, the model will appear overfitting. The Inception structure expands the width of the network by setting multiple convolutions of different sizes and depths so that the model can achieve better performance [32].

For multi-scale pedestrians, single scale feature layer detection cannot meet the requirements. Zhang et al. [47] believed that network detection of pedestrians at different scales has different understandings of different scale features and requires different expressions. Therefore, a prediction network about scale dependence was constructed. To achieve feature detection layer matching with multi-scale pedestrian size, the idea of multi-scale prediction is usually adopted to solve the problem. To achieve feature detection layer matching with multi-scale pedestrian size, the idea of multi-scale prediction is usually adopted to solve the problem. Multi-scale prediction represents the fusion of shallow position features and deep semantic features extracted, generating prediction layers of different resolutions to predict targets at corresponding scales [23], usually represented in the form of a pyramid. Table 1 shows four common types of pyramid based multi-scale prediction models.

The SSD [39] algorithm uses multi-scale feature representation methods to extract features from various convolutional layers and then detects targets to maximize the utilization of features at all levels. However, there is still limited utilization of deep features in the network. Each layer of the neural network has different information. In target location, the target is located according to the location information of the target, that is the shallow features. The type of target recognition needs to be based on the semantic information of the target, that is, the deep features. The feature pyramid combines high-level and low-level information through feature fusion and summarizes the local information, further improving the detection capability of the model. The RetinaNet algorithm adopts a feature pyramid and multi-scale prediction structure, and the Focal Loss function solves the problem of imbalanced positive and negative samples. These designs ensure the excellent detection performance of the RetinaNet model.

Table 1 Multiscale prediction structure

Network name	Network structure	Characteristic
Featurized image pyramid		<p>Realize multi-scale prediction of images, but there are drawbacks such as high computational complexity and memory occupation for each layer of images.</p>
Single feature map		<p>The structure is simple, but it is easy to lose the detailed information of the image, which is conducive to detecting large target objects, such as the Faster R-CNN algorithm.</p>
Pyramidal feature hierarchy		<p>Realize prediction for multiple feature layers, but fail to fully extract low-level features, such as SSD algorithm.</p>
Feature Pyramid Network		<p>Realize multi-scale prediction, integrate high-level and low-level information, fully acquire and utilize feature information, and be able to adapt to scale changes of targets, such as FPN and RetinaNet algorithms.</p>

2.2 Attention mechanism

In recent years, the attention mechanism has been widely used in object detection, image classification, and speech recognition in the field of deep learning. Attention mechanism refers to redistributing feature information resources according to the importance of attention objects, focusing on important information, and in this case suppressing other

unimportant information. In the application of pedestrian detection, to adapt the network receptive field to scale changes, Xue et al. [45] combined the global attention mechanism to effectively increase the receptive field on YOLO v5 so that the network has higher detection performance. Ma et al. [25] designed a multi-scale convolutional model to extract features at different scales and then used the attention mechanism module to obtain feature correlation information, achieving the goal of enhancing features. The experiment showed that the detection effect was significantly improved. Lv et al. [24] added a context attention module to feature extraction to better detect pedestrians through pedestrian context semantic information in order to obtain correlation information between various scales.

Common attention mechanisms include SE module [8], CBAM module [40], etc. The spatial attention module applies attention weights in space so that the model pays attention to the spatial information of the target, and the channel attention makes the model pay more attention to the information on the channel. The CBAM module allows the model to combine the spatial information and channel information of the image to conduct a more comprehensive analysis of the features so that the network model can pay more attention to the target features of the image. The attention operation makes the target get more attention by constructing the degree of association or spatial dependency between the channels. Generally, the method of reducing the dimension first and then increasing the dimension is adopted to ensure that it matches the original feature, in which the two branches are independent of each other. After the feature weights are learned, the connection operation will be performed.

3 RetinaNet algorithm with multi-branch structure and double pooling attention mechanism

3.1 Overall structure

The RetinaNet model is shown in Fig. 2, which is mainly divided into three parts: backbone network, feature pyramid, and classification prediction [7].

The backbone network feature extraction uses ResNet50. In the process of enhancing the network feature extraction ability, it is generally believed that the deeper the network, the more features it can learn. However, this is not absolute because deepening the network increases the number of model parameters, makes the model large, and slows down the training speed. Sometimes, problems such as vanishing gradients and exploding gradients can also occur, which can degrade the network and affect the training effect. The residual network can effectively solve the problem of network degradation with the deepening of network layers. ResNet network obtains residual neural networks of different depths by stacking a different number of basic residual blocks, plus pooling and activation functions. The backbone network ResNet50 selected in this article is composed of four stacked modules, as shown in Fig. 3. Each module is composed of a different number of residual blocks, with 3, 4, 6, and 3 connected in sequence.

After the main network extraction features, the characteristic fusion is performed through the horizontal connection of 1×1 convolution to build a characteristic pyramid so that the features of a single layer are fused into the information of different layers. Each layer has semantic information from the high level, which improves the feature learning ability of the network. The aspect ratio of the anchor box on each layer of the pyramid is $\{1:2, 1:1, 2:1\}$, based on the original three anchor boxes, adding anchor boxes of size $\{2^0,$

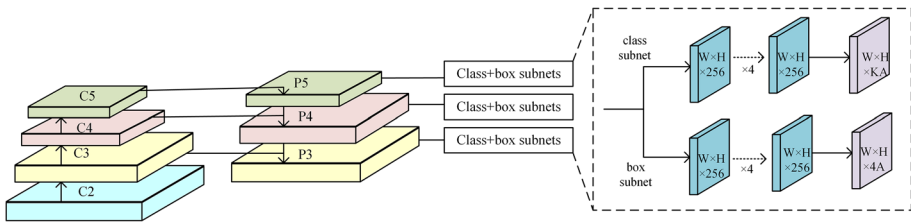


Fig. 2 RetinaNet model structure

$2^{1/3}, 2^{2/3}$ can generate 9 anchor boxes of different sizes, which can be adapted to multi-scale target detection. After the feature pyramid, classification prediction is a parallel classification and regression sub-network used for target classification and localization.

The multi-scale prediction method for multi-scale pedestrians can solve some of the difficulties. Large scale pedestrian features are obvious, and semantic information can still be extracted after multiple convolutions. Small scale pedestrians have low resolution and contain less feature information. There is a problem of insufficient feature extraction or semantic information loss after multiple convolutions, leading to missing detection. Therefore, it is necessary to use appropriate methods to fully extract multi-scale pedestrian information to ensure that semantic information is not lost. Therefore, this paper designs a network structure as shown in Fig. 4.

Convolution layer	Output layer	ResNet50
Conv1	112×112	7×7, 64, s=2 3×3 maxpool, s=2
Conv2_x	56×56	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} * 3$
Conv3_x	28×28	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix} * 4$
Conv4_x	14×14	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} * 6$
Conv5_x	7×7	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} * 3$
1×1 Average_pool,1000-dfc,Soft_max		

Fig. 3 ResNet50 network structure

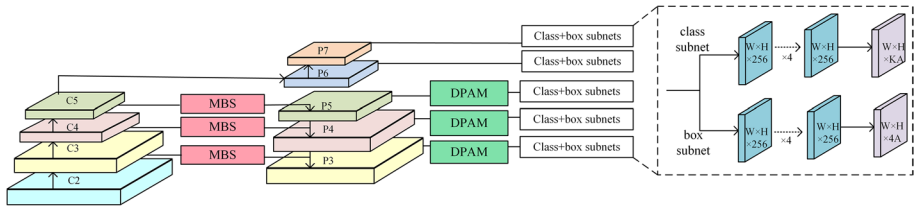


Fig. 4 RetinaNet structure with Multi-Branch Structure (MBS) and Double Pooling Attention Mechanism (DPAM)

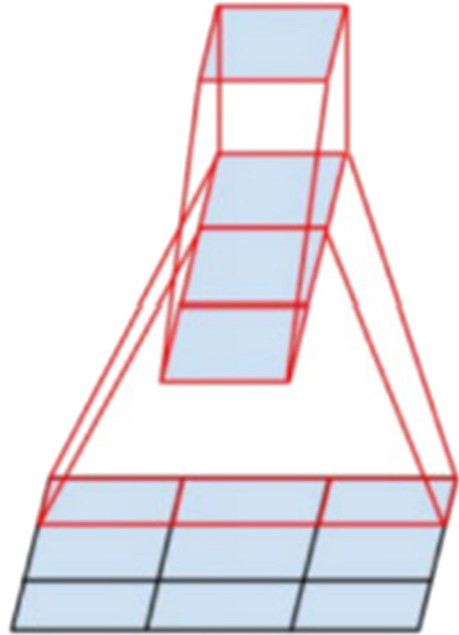
The multi-branch structure is applied between the backbone network and the feature pyramid in the RetinaNet model, that is, adding a multi-branch structure at the horizontal connection of C3, C4, and C5 layers, fully leveraging the extraction and utilization efficiency of feature information, and improving the diversity of features. Capture detailed texture information in shallow networks; In the middle layer network, multiple receptive fields and pooling operations can capture diverse feature information; In deep networks, extracting abstract features not only enhances the network's ability to analyze targets at different scales but also deepens and widens the network, effectively extracting multi-level information. Add double pooling attention mechanism module between the feature pyramid and the prediction branch to filter information for each fused branch, making the model focus on more useful pedestrian feature information, enhancing the detection performance of the network, and making predictions more accurate.

3.2 Multi-branch construction

In neural network models, generally speaking, the representation ability of the model is improved by expanding the depth and width of the model, but there are also side effects. The deeper the level of the neural network, the more parameters it generates, leading to overfitting. This not only results in high training costs but also low efficiency, and the Inception network structure can effectively solve this problem. Unlike most previous networks that directly stack convolutional layers to obtain deep networks, the Inception model proposed in reference [36] uses sparse connections, sets multiple convolutional kernels of different scales in parallel structure, and concatenates features to expand the width of the network and deepen the network, enabling the model to achieve better performance. Usually, using large convolutional kernels can extract information that is far from pixels, while using small convolutional kernels can extract information that is close to the pixel domain. The traditional large convolution has a better receptive field, but it is easy to lose some important information in the operation process. The Inception network structure uses decomposition convolution kernels to decompose a single large convolution into symmetric small convolutions or asymmetric convolutions, reducing the number of parameters as the network width and depth are increased. Decomposing convolution is adopted to reduce the parameters without changing the receptive field and improve the nonlinear demonstration ability of the model. As shown in Fig. 5, using 3×1 and 1×3 the receptive field after convolution sliding is equivalent to 3×3 Convolutions.

The multi-branch structure adopted in this article is shown in Fig. 6, which is divided into 4 parallel branches from the input, and each branch is applied with 1×1 .

Fig. 5 Sliding receptive field of asymmetric convolution



Convolution obtains the associated information of the image, reduces the number of feature channels, reduces dimensions, and aggregates information by reducing the number of channels so that features are superimposed at depth.

The small convolution size in the network can better capture the details of adjacent areas of the image, and the information is highly relevant. Then, the multi-scale convolution kernel is used to conduct convolution in the network with different depths to obtain a multi-scale receptive field, convert the details into advanced semantic features, and then splice the results from different branches according to the channels to aggregate the feature information of all branches to obtain a multi-channel feature map. Finally, output the results. The feature fusion method used in this article is concatenation, as more convolutions are used to extract image information. The addition operation on dimensions can effectively and completely fuse the information.

$$\begin{cases} X_1 = F_{(3,1)}(F_{(1,3)}(F_{(3,1)}(F_{(1,3)}(F_{(1,1)}(X)))) \\ X_2 = F_{(3,1)}(F_{(1,3)}(F_{(1,1)}(X))) \\ X_3 = F_{(1,1)}(F_{pool}(X)) \\ X_4 = F_{(1,1)}(X) \\ X_{out} = Concat(X_1, X_2, X_3, X_4) \end{cases} \tag{1}$$

The multi-branch structure can be represented by formula (1), where F represents convolution, X represents the input image, $X_i(i=1,2,3,4)$ represents the results of four branches, and X_{out} represents the output result. The design of a multi-branch structure is not blindly increasing depth and width, but rather maintaining information invariance through pooling operations to prevent information loss. By decomposing features,

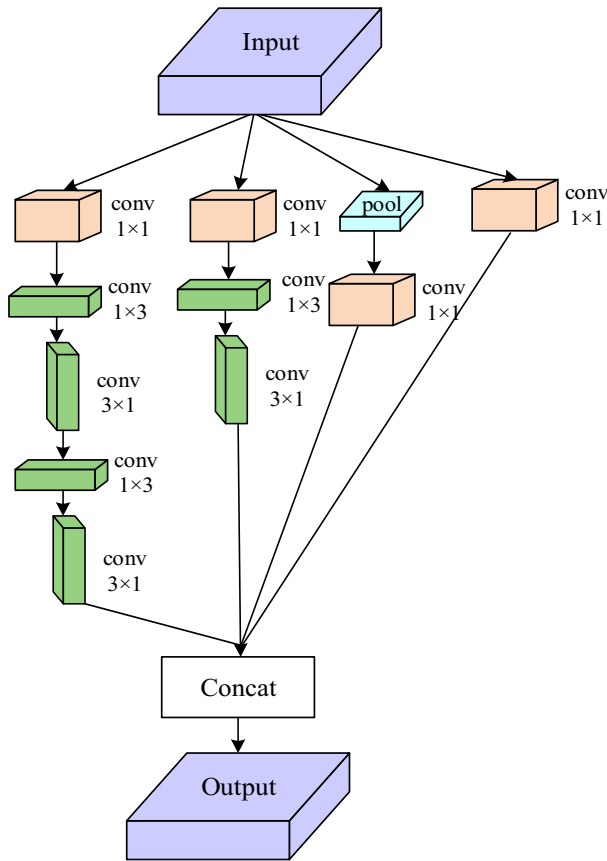


Fig. 6 Multi-branch construction

information is fully decomposed and utilized to improve the internal correlation of features. Utilizing smaller convolutional kernels to reduce dimensionality and achieve a balance between the depth and width of the network.

3.3 Double pooling attention mechanism

The information carried by the feature layer in different channels is different, and the channel correlation is also different. According to the structure of the channel attention SE module, the global average pooling is able to aggregate the information of each feature in the image. Inspired by this, combined with the role of global average pooling, double average pooling attention mechanism is selected on the channel to enhance feature information. First, a global average pooling is used to extract features, in order to prevent losing some important information, a global average pooling is used again to strengthen the extracted important features and named the double pooling attention module. The double pooling attention module is shown in Fig. 7.

The double pooling attention module consists of three parallel parts. The first branch completes feature mapping, and the second branch performs global average pooling.

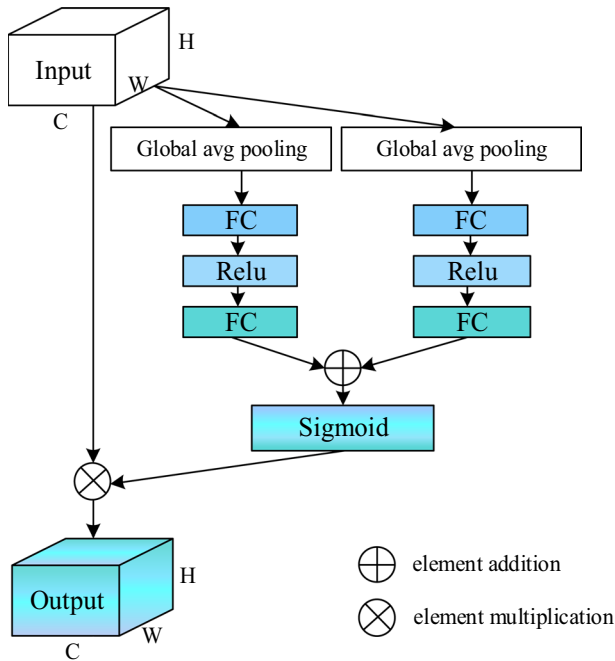


Fig. 7 Double pooling attention module

Global average pooling makes the convolution structure simpler, compresses the input features, reduces the number of parameters, and achieves the purpose of optimizing the network structure and preventing overfitting. Then, it passes through the full connection layer, activation function, and full connection layer in turn, which is expressed in formula (2).

$$X_2 = FC(\alpha(FC(P_{gav}(X)))) \tag{2}$$

Where X represents the input image and α represents the Relu activation function. The third branch is the same as the second branch, using global average pooling to extract feature information again and strengthen weight learning, represented by formula (3).

$$X_3 = FC(\alpha(FC(P_{gav}(X)))) \tag{3}$$

Finally, add and fuse the features extracted from the second branch and the third branch element by element, pass through the sigmoid activation function, and then combine the output with the original features of the first branch to multiply the elements to obtain the weighted features and output the results, which are expressed in formula (4), where β represents the sigmoid activation function and X_{out} represents the output results.

$$X_{out} = X_1 \otimes \beta(X_2 \oplus X_3) \tag{4}$$

The parameters of the double pooling attention module included in the model are shown in Table 2.

Table 2 Parameters of double pooling attention module

Layer	Size
Input	$H \times W \times C$
Global avg pooling	$1 \times 1 \times C$
FC	$1 \times 1 \times C/R$
Relu	$1 \times 1 \times C/R$
FC	$1 \times 1 \times C$
Sigmoid	$1 \times 1 \times C$
Output	$H \times W \times C$

Among them, H and W represent the height and width of the image, C is the channel, and R is the channel factor.

3.4 Loss function

The training process of the model is the process of continuously reducing the error. The loss function used for the training of the model is the positioning loss and the score loss, which are calculated by formula (5).

$$L = L_{reg} + L_{cls} \quad (5)$$

The smooth L1 loss function is used as the positioning loss function, which is calculated by formula (6).

$$smooth_{L1} = \begin{cases} 0.5x^2, & |x| < 1 \\ |x| - 0.5, & |x| \geq 1 \end{cases} \quad (6)$$

Among them, x is the difference between the predicted box and the actual box. The classification loss is calculated using the Focal Loss function using formula (7) and (8).

$$FL(p_t) = -\alpha_t(1 - p_t)^\gamma \log(p_t) \quad (7)$$

$$p_t = \begin{cases} p & \text{if } y = 1 \\ 1 - p & \text{otherwise} \end{cases} \quad (8)$$

In formula (7) and (8), p_t represents the probability of a positive sample, and y represents the value of the true label. α_t is a moderating factor that can be used to improve the distribution weight of positive and negative samples when there are many negative samples and only a few positive samples. γ is a focusing parameter that can effectively reduce the weight of easily detected samples and focus the training process on difficult to detect samples. As a result, Focal Loss solves the imbalance problem between easily detected and difficult to detect samples, thereby improving the accuracy of the model. After experiments, when $\alpha_t=0.25$ and $\gamma=2$ in this article, it can increase the weight of positive samples and reduce the weight of negative samples, achieving good detection results.

Table 3 Experimental environment

Lab environment	Version model
Operating system	Windows
CPU	Intel i7-8700 processor
GPU	GTX1080
Memory	32GB
Deep learning framework	Pytorch
CUDA version	10.2
Graphics memory	8G

4 Experimental results and analysis

4.1 Experimental environment

The experimental environment is shown in Table 3.

4.2 Experimental data

To enrich the pedestrian dataset and make the detection more realistic, this experiment uses different datasets for detection. Dataset 1 is a mixed dataset that selects images with "person" labels from the COCO dataset and the PASCAL VOC dataset. A total of 3288 images contain target images of various scenes and scales in daily life, and 2630 training sets and 658 test sets are divided according to 8:2.

Dataset 2 is the Caltech pedestrian dataset, which is a video captured by a car mounted camera at California Institute of Technology in the United States. It contains approximately 250000 images, with Set00-Set05 as the training set and Set06-Set10 as the test set. The Caltech pedestrian dataset has a large amount of data, including pedestrians of different scales. Therefore, this article selects images labeled "person" as the experimental data for this article, with 4310 training sets and 4225 test sets selected. According to reference [5], the test subsets are divided into multiple different scale levels based on the height of pedestrians in the image, and the division criteria are shown in Table 4.

4.3 Experimental parameters

The model used in this experiment is the improved RetinaNet pedestrian detection algorithm. In order to make the neural network achieve a better prediction effect, the input image size during training is 512×512, the batch is 8, the optimizer is adam, and the initial

Table 4 Subset division of multi-scale pedestrian testing

Test subset	Pedestrian height (pixels)
Large	Greater than 100
Near	80, 100
Medium	30, 80
Far	20, 30

learning rate is $1e^{-4}$. Cosine annealing is used to decrease the learning rate, the momentum is 0.9. Fig. 8 shows the changes in loss of the algorithm in this article during 150 epochs of training. By observing the changes in loss, the training effect can be judged. It can be seen that as the training epochs increase, the loss curve shows a downward trend and becomes smoother, stabilizing at around 0.05 and reaching a convergence state, indicating that the effect is optimal during the training process.

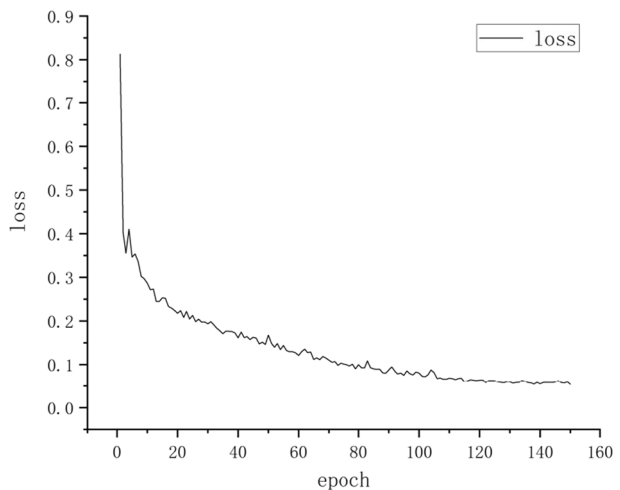
4.4 Evaluation indicators

To better evaluate the detection performance of network models, different algorithms and evaluation indicators were used to validate the model for different datasets. In a mixed dataset, the following indicators are selected as the evaluation indicators for this model: Precision, as shown in formula (9), represents the proportion of predicted pedestrians to the original sample of pedestrians. Recall, as shown in formula (10), represents the proportion of pedestrians in the pedestrian dataset who are correctly predicted; the F1 score is a weighted average of the precision rate and recall rate, which is the harmonic mean of the precision rate and recall rate, and is calculated by Formula (11); AP is the average accuracy under the IOU threshold, as shown in equation (12), representing the area included by the Precision and Recall curves; mAP is the average precision of all target categories. The higher the value, the higher the recognition accuracy of the model. The mAP is calculated by equation (13), indicating that AP is calculated first, and then the obtained value is divided by all categories N to obtain mAP. The detection target studied in this article is only people, so mAP is equal to AP. In formula (13), it was found that mAP combines precision and recall, taking into account false positives and true positives. Therefore, most detection models use this value as a reasonable evaluation indicator.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (9)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (10)$$

Fig. 8 Loss variation curve



$$F1 = 2 \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (11)$$

$$AP = \int_0^1 \text{Precision}(\text{Recall}) d\text{Recall} \quad (12)$$

$$mAP = \frac{\sum AP}{N} \quad (13)$$

TP means that the pedestrian is correctly detected, TN means that the negative sample of the pedestrian in the picture is correctly detected as a negative sample, FN means that the pedestrian sample is predicted as a negative sample, FP means that the negative sample in the data is predicted to be a positive sample.

In the Caltech dataset, the evaluation indicator uses the average logarithmic miss detection rate indicator, which represents the proportion of undetected pedestrians to the actual number of no pedestrians. MR^{-2} is used to represent the average of 9 values of the miss detection rate for each image uniformly sampled between logarithms $[10^{-2}, 10^0]$, and the miss detection rate is calculated using equation (14).

$$MR = \frac{FN}{TP + FN} = 1 - \text{Recall} \quad (14)$$

In N images, if the number of false positives is FP , then $FPPI$ represents the number of false positives per image, calculated using equation (15).

$$FPPI = \frac{FP}{N} \quad (15)$$

4.5 Experimental results on mixed datasets

4.5.1 Ablation experiment

To verify the effect of the multi-branch structure and double pooling attention module added in this article on the model, the following ablation experiments will be conducted, and the algorithm detection indicators are shown in Table 5.

When adding a multi-branch structure, mAP achieved 78.87% and improved Precision by 1.27%, indicating enhanced feature expression and a slight improvement in detection performance. When adding attention, the mAP of the algorithm reached 79.42%, an increase of 0.74% compared to the original, and Recall and Precision improved by 0.91% and 0.72%, indicating that the attention mechanism further strengthens the

Table 5 Ablation experiment detection data table

RetinaNet	Multi-branch construction	Double pooling attention	mAP	F1	Recall	Precision
√			78.68%	0.77	71.61%	83.68%
√	√		78.87%	0.78	71.49%	84.95%
√		√	79.42%	0.78	72.52%	84.40%
√	√	√	80.17%	0.78	72.29%	85.78%

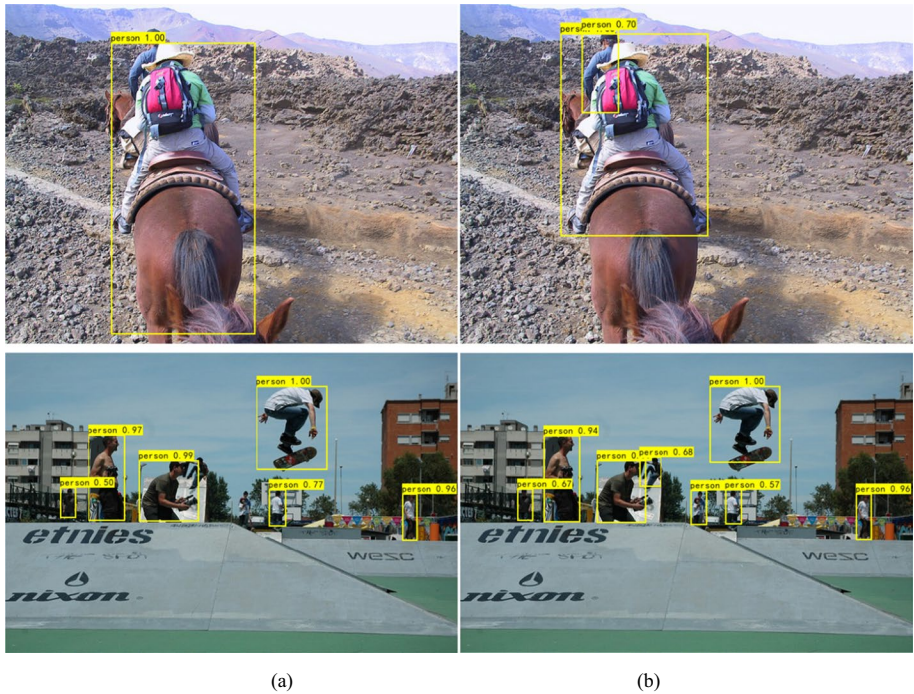


Fig. 9 Comparison of traditional RetinaNet and the algorithm in this paper. Among them (a) is the RetinaNet detection image; (b) is the detection image of the algorithm in this paper

important feature weights. When adding two modules simultaneously, the overall mAP index reached 80.17%, an increase of 1.49%, the Recall increase of 0.68%, and Precision increase of 2.1%. The above data proves that each module has improved the model to varying degrees, which is of great help in improving network performance.

The visual detection diagram of the traditional RetinaNet and the algorithm in this paper, as shown in Fig. 9, more intuitively illustrates the detection effect of this paper. In the first image, the algorithm proposed in this paper can effectively detect occluded individuals. In the second image, the algorithm can detect people of different scales, indicating the effectiveness of the algorithm proposed in this paper.

4.5.2 Comparative experiments

To demonstrate the effectiveness of the double pooling attention mechanism module proposed in this paper, SE and CBAM modules were selected for comparative experiments, and the experimental detection results are shown in Table 6.

Table 6 Selection of attention mechanism modules

Attention mechanism module	mAP	Recall	Precision
SE	79.18%	71.78%	83.99%
CBAM	79.02%	72.06%	84.77%
Double pooling attention	79.42%	72.52%	84.40%

Table 7 Comparison of detection data with other algorithms

Model	Backbone network	mAP	F1	Recall	Precision
SSD	VGG	76.36%	0.74	67.06%	82.99%
Faster R-CNN	Resnet50	75.28%	0.58	81.43%	44.70%
Improved RetinaNet	Resnet50	80.17%	0.78	72.29%	85.78%

From Table 6, it can be seen that the three attention mechanisms listed have improved network detection performance for basic RetinaNet. However, the mAP of the double pooling attention mechanism proposed in this article performs better than the SE and CBAM modules and has a better focus on pedestrian feature information.

To verify the superiority and reliability of the algorithm in this paper, the single-stage SSD algorithm and the two-stage Faster R-CNN algorithm are selected for comparison. The comparison results are shown in Table 7.

According to Table 7, the above results show that the improved RetinaNet is better than the other two algorithms in mAP, F1 value, and Precision. Among them, mAP is 3.81% higher than the SSD model, 4.89% higher than Faster R-CNN, and the Recall of Faster R-CNN reaches 81.43%, because Faster R-CNN is a two-stage algorithm, and the correct pedestrian detected number accounts for a large proportion of the number of pedestrians in the dataset. The F1 value is an evaluation indicator that combines Recall and Precision and is an indicator that can reflect the overall performance. The F1 value in the algorithm proposed in this paper has better results than SSD and Faster R-CNN.

The comparison of model parameters is shown in Table 8. After adding the multi-branch structure and double pooling attention mechanism, the model parameters only increased by 0.273M. Although compared with the SSD model, the parameters were higher, but the Precision of the model was improved.

The visual detection effect with Faster R-CNN is shown in Fig. 10.

It can be seen from the detection comparison between this algorithm and Faster R-CNN in Fig. 10, Faster R-CNN benefits from the two-stage detection performance and can detect most of the targets, but the detected targets also have redundant detection frames and a large number of false detection cases, and the detection algorithm in this paper can accurately detect the target. Pedestrian detection technology is widely used in various scenes in life, and there are certain requirements for real-time detection. FPS represents the number of frames processed per second. As can be seen from Fig. 11, the FPS of this paper reaches 30.01, about 3 times faster than Faster R-CNN, the improved RetinaNet can meet the requirements of real-time detection, comprehensively considered, the improved algorithm ensures accuracy and speed, and achieves an ideal and relatively balanced effect.

The visual detection effect of the SSD algorithm is shown in Fig. 12.

Table 8 Comparison of model parameters

Model	Params
RetinaNet	36.330M
Improved RetinaNet	36.603M
SSD	23.612M
Faster R-CNN	136.689M



Fig. 10 Comparison of Faster R-CNN and the algorithm in this paper. Among them (a) is the Faster R-CNN detection image; (b) is the detection image of the algorithm in this paper

From the comparison between SSD algorithm detection in Fig. 12 and the algorithm detection in this article, it can be seen that the SSD algorithm has poor performance in detecting small and medium size pedestrians, and the detection of small pedestrians is severe. However, the algorithm in this paper can accurately detect small targets. In the single-stage SSD algorithm, the algorithm proposed in this paper also has good results.

Fig. 11 Comparison of detection speed

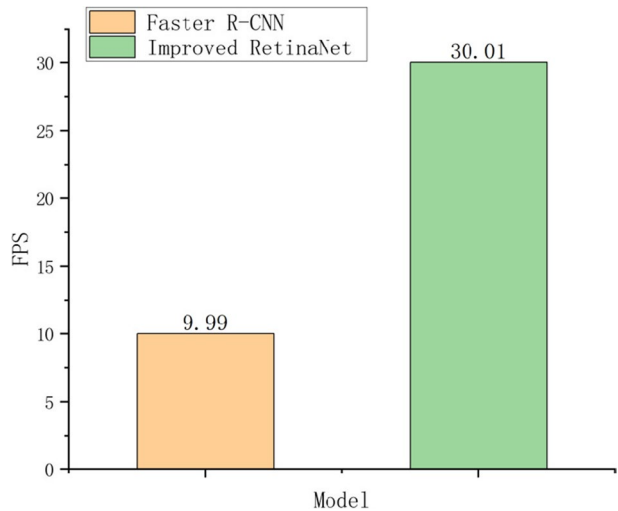




Fig. 12 Comparison of detection between SSD and the algorithm in this paper. Among them (a) is the SSD detection image; (b) is the detection image of the algorithm in this paper

4.6 Experimental results on Caltech dataset

To verify the generalization of the model, this paper uses the same method to conduct experiments on the Caltech dataset, and the detection results are shown in Table 9.

The Caltech dataset clearly distinguishes the scale range of pedestrians, allowing for a better understanding of pedestrian detection at various scales. The effect is good on the large scale, with a decrease in miss detection rate to 0, the MR^{-2} on the medium scale decreased by 0.72%, 1.44% on the far scale, and 1.18% on the all subset. The improved RetinaNet model has a varying degree of decrease in miss detection rates at various scales. The combination of multi-branch structure and attention mechanism enhances the feature extraction ability of the RetinaNet model, accurately capturing pedestrian feature information that is of concern and interest.

This article uses different algorithms for experimental comparison on the Caltech dataset, such as ACF [6], LDCF [27], which use traditional manual feature methods. Deep learning methods include RPN+BF [48], MS-CNN [2], etc. The comparison

Table 9 Caltech dataset detection results

Model	Large	Near	Medium	Far	All
RetinaNet	1.47%	2.38%	52.14%	96.23%	63.45%
Improved RetinaNet	0	2.05%	51.42%	94.79%	62.27%

results are shown in Table 10 the multi-branch structure and attention mechanism algorithm proposed in this article is represented by MBSAE, and the lower the MR^{-2} value, the better the effect.

From the comparison results, it can be seen that the detection performance of traditional manual feature extraction methods is significantly lower than that of deep learning methods. Compared with the current advanced algorithm RPN+BF, the algorithm proposed in this paper shows good detection performance at various scales. Compared with AR-Ped [1], the detection performance at the large and far scales has certain advantages, with differences of 1.28% and 5.21%, respectively. On other scales, although the algorithm presented in this paper performs poorly, the gap is relatively small, such as a difference of only 0.68% in the near subset. This is because the AR-Ped algorithm adopts a cascading method, which can extract more pedestrian context information. In general, the algorithm proposed in this article has made significant progress in multi-scale pedestrian detection.

In order to better evaluate the effectiveness of the model at different scales, based on the algorithm comparison results in Table 10, and taking the near and far scales as examples, we plotted MR-FPPI curves to more intuitively see the comparison results of each algorithm on pedestrians at different scales, as shown in Fig. 13. The curves use FPPI as the horizontal axis and MR as the vertical axis, showing a decreasing trend overall. The lower the curve, the better the effect, our proposed MBSAE algorithm has achieved good results.

The visualization of image detection results in the Caltech dataset is shown in Fig. 14.

Through the comparison in Fig. 14, it is found that when multiple small-scale pedestrians appear side by side in the distance, the RetinaNet algorithm suffers from serious missed and false detection of pedestrians. However, the improved algorithm in this paper can better detect pedestrians in the distance and near, indicating that the improved method proposed in this paper has stronger detection ability.

From the above detection results in different datasets, we can clearly see the detection effectiveness of this model. In the case of more targets, the algorithm in this paper can detect more targets, and the detection is correct; in the detection of small targets, this paper also has a good effect on small targets; in occluded pictures, the target can also be better detected. Therefore, through the above experiments, the advantages of this model can be fully demonstrated, it can be adapted to multi-scale pedestrian detection and has obvious effects on detection accuracy. To a certain extent, it can explain the applicability of pedestrian detection and the robustness of this method in pedestrian detection.

Table 10 Comparison with other algorithms on the Caltech dataset

Model	Large	Near	Medium	Far	All
VJ	86.21%	89.85%	99.38%	99.67%	99.53%
HOG	37.93%	44.04%	87.39%	97.15%	90.36%
ACF	23.48%	28.71%	76.44%	96.81%	81.83%
LDCF	8.21%	10.46%	61.82%	100%	71.25%
SCF+AlexNet [12]	7.01%	10.61%	62.34%	100%	70.33%
TA-CNN [37]	7.00%	7.96%	63.62%	100%	71.22%
MS-CNN	1.99%	2.60%	49.13%	97.23%	60.95%
RPN+BF	1.18%	2.26%	53.93%	100%	64.66%
AR-Ped	1.28%	1.37%	49.31%	100%	58.83%
MBSAE	0.00%	2.05%	51.42%	94.79%	62.27%

The indicators in the table represent the missed detection rate, and the lower the value, the better, so the better values are highlighted in bold

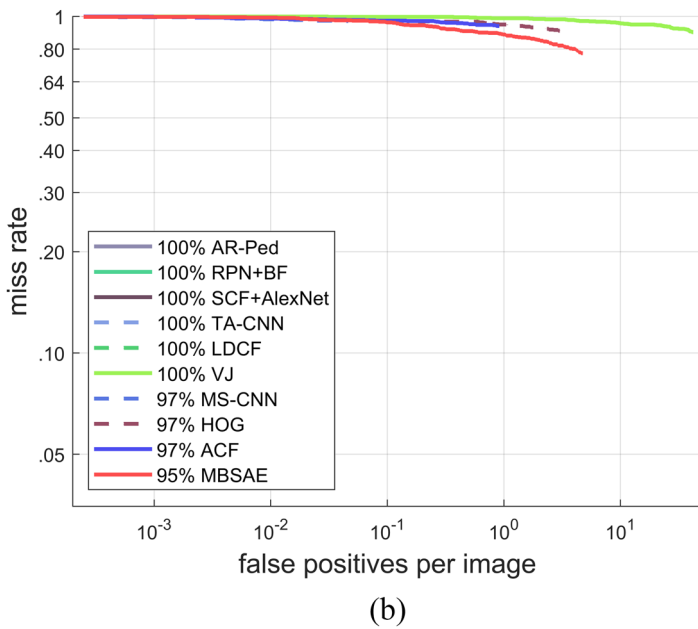
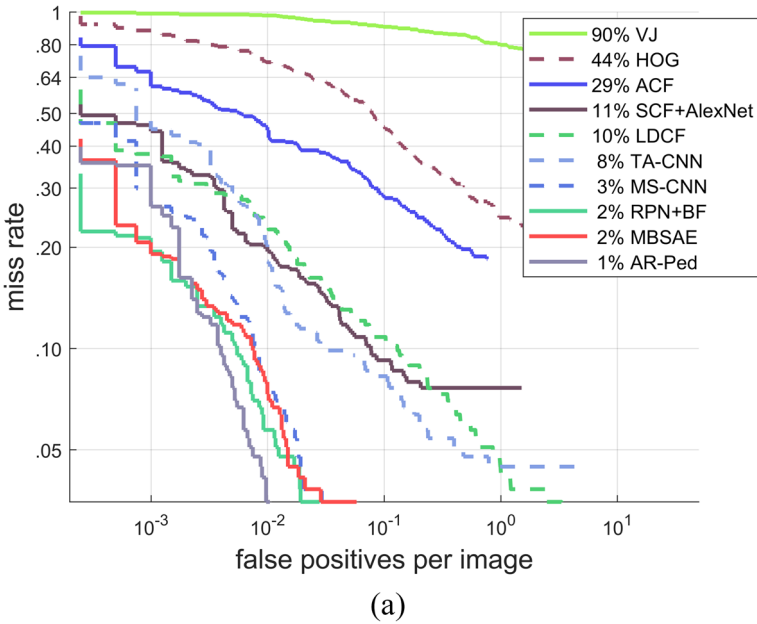


Fig. 13 MR-FPPI curves at different scales. Among them (a) is near scale; (b) is far scale

5 Conclusion

In this paper, the traditional RetinaNet model is improved and optimized for its shortcomings in multi-scale pedestrian detection. By adding a multi-branch structure as a feature enhancement module, the width of the network is increased, the nonlinear expression



Fig. 14 Comparison of detection images on the Caltech dataset. Among them (a) is the RetinaNet detection image; (b) is the detection image of the algorithm in this paper

ability of the model is improved, and rich multi-scale features are extracted; In order to make the model more focused on detecting important information, double pooling attention module was added to suppress unimportant information and further improve the model's detection accuracy. Verified in public datasets and compared with other algorithms, the model in this paper achieved a detection accuracy of 80.17%, fully demonstrating its detection ability and good generalization ability. In further research in the future, the model will continue to be optimized, and transfer learning and confrontation network methods can be used to enrich the dataset, and continue to solve the detection difficulties of model

application in harsh environments such as rain, fog, snow, etc., improve the detection accuracy and speed of the model, so that it can better complete the task of pedestrian detection.

Acknowledgments This work was funded by the National Natural Science Foundation of China, grant number 6192007, 62266009, 61462008, 61751213, 61866004; the Key projects of Guangxi Natural Science Foundation, grant number 2018GXNSFDA294001, 2018GXNSFDA281009; Guangxi Key Laboratory of Big Data in Finance and Economics, Grant No. FEDOP2022A06; 2020 Guangxi Education Department Degree and Postgraduate Education Reform Project, grant number 20210121210028; Research Fund of Guangxi Key Lab of Multi-source Information Mining & Security, grant number MIMS19-04; 2015 Innovation Team Project of Guangxi University of Science and Technology, grant number gxkjdx201504; College Students' Innovative Entrepreneurial Training, grant number 202210594036, 202210594037, 202210594038, 202210594041, 202110594133, 202110594134.

Author Contributions Conceptualization, L.c.H.; methodology, L.c.H.; software, X.b.F.; validation, X.b.F, Z.w.W.; formal analysis, L.c.H.; investigation, X.b.F; data curation, L.c.H.; writing— original draft preparation, L.c.H.; writing— review and editing, Z.w.W.; visualization, X.b.F.; supervision, X.b.F.; project administration, Z.w.W.; funding acquisition, Z.w.W. All authors have read and agreed to the published version of the manuscript.

Data availability The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Conflicts of Interest The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Brazil G, Liu X (2019) Pedestrian detection with autoregressive network phases. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 7231–7240
2. Cai Z, Fan Q, Feris R S, et al (2016) A unified multi-scale deep convolutional neural network for fast object detection. In Computer Vision—ECCV, pp 354–370
3. Calero MJF, Aldás M, Lázaro J et al (2019) Pedestrian detection under partial occlusion by using logic inference, HOG and SVM. *IEEE Latin Ame Transac* 17(09):1552–1559
4. Cao J, Song C, Peng S et al (2020) Pedestrian detection algorithm for intelligent vehicles in complex scenarios. *Sensors* 20(13):3646
5. Dollar P, Wojek C, Schiele B et al (2011) Pedestrian detection: An evaluation of the state of the art. *IEEE Trans Patt Analy Machine Int* 34(4):743–761
6. Dollár P, Appel R, Belongie S et al (2014) Fast feature pyramids for object detection. *IEEE Transac Patt Analy Machine Int* 36(8):1532–1545
7. Du K, Che X, Wang Y et al (2022) Comparison of RetinaNet-Based Single-Target Cascading and Multi-Target Detection Models for Administrative Regions in Network Map Pictures. *Sensors* 22(19):7594
8. Fukui H, Hirakawa T, Yamashita T, et al (2019) Attention branch network: Learning of attention mechanism for visual explanation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10705–10714.
9. Gawande U, Hajari K, Golhar Y (2022) SIRA: Scale illumination rotation affine invariant mask R-CNN for pedestrian detection. *Appl Int* 52(9):10398–10416
10. Ge Z, Wang J, Huang X et al (2021) Lla: Loss-aware label assignment for dense pedestrian detection. *Neurocomputing* 462:272–281
11. He Y, He N, Yu H et al (2023) From macro to micro: rethinking multi-scale pedestrian detection. *Multimed System*:1–13
12. Hosang J, Omran M, Benenson R, et al (2015) Taking a deeper look at pedestrians. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4073–4082
13. Hsu WY, Lin WY (2021) Adaptive fusion of multi-scale YOLO for pedestrian detection. *IEEE Access* 9:110063–110073

14. Ji QG, Chi R, Lu ZM (2018) Anomaly detection and localisation in the crowd scenes using a block-based social force model. *IET Image Proc* 12(1):133–137
15. Jiang Q, Dai J, Rui T et al (2022) Attention-Based Cross-Modality Feature Complementation for Multispectral Pedestrian Detection. *IEEE Access* 10:53797–53809
16. Kumar K, Mishra RK (2020) A heuristic SVM based pedestrian detection approach employing shape and texture descriptors. *Multimed Tools Appl* 79:21389–21408
17. Lei H, Yixiao W, Guoying C (2021) The Hierarchical Local Binary Patterns for Pedestrian Detection. In: 2021 5th CAA International Conference on Vehicular Control and Intelligence (CVCI), pp 1–8
18. Li W (2021) Infrared image pedestrian detection via YOLO-V3. In: 2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), pp 1052–1055
19. Li G, Yang Y, Qu X (2019) Deep learning approaches on pedestrian detection in hazy weather. *IEEE Transac Indust Electron* 67(10):8889–8899
20. Li G, Zong C, Liu G et al (2020) Application of Convolutional Neural Network (CNN)–AdaBoost Algorithm in Pedestrian Detection. *Sens. Mater* 32:1997–2006
21. Li Q, Qiang H, Li J (2021) Conditional random fields as message passing mechanism in anchor-free network for multi-scale pedestrian detection. *Inform Sci* 550:1–12
22. Li ML, Sun GB, Yu JX (2023) A pedestrian detection network model based on improved YOLOv5. *Entropy* 25(2):381
23. Lin TY, Dollár P, Girshick R, et al (2017) Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2117–2125
24. Lv H, Yan H, Liu K et al (2022) Yolov5-ac: Attention mechanism-based lightweight yolov5 for track pedestrian detection. *Sensors* 22(15):5903
25. Ma J, Wan H, Wang J et al (2021) An improved one-stage pedestrian detection method based on multi-scale attention feature extraction. *J Real-Time Image Proc*:1–14
26. Mihçioğlu ME, Alkar AZ (2019) Improving pedestrian safety using combined HOG and Haar partial detection in mobile systems. *Traffic Injury Prevent* 20(6):619–623
27. Nam W, Dollár P, Han J H (2014) Local decorrelation for improved pedestrian detection. *Advances in neural information processing systems* 27.
28. Nataprawira J, Gu Y (2021) Pedestrian detection using multispectral images and a deep neural network. *Sensors* 21(7):2536
29. Pei D, Jing M, Liu H et al (2020) A fast RetinaNet fusion framework for multi-spectral pedestrian detection. *Infrared Phys Technol* 105:103178
30. Qiu J, Wang L, Hu Y et al (2020) Effective object proposals: size prediction for pedestrian detection in surveillance videos. *Electron Lett* 56(14):706–709
31. Qiu M, Huang L, Tang BH (2022) ASFF-YOLOv5: Multielement Detection Method for Road Traffic in UAV Images Based on Multiscale Feature Fusion. *Remote Sens* 14(14):3498
32. Ramírez I, Cuesta-Infante A, Pantrigo JJ et al (2020) Convolutional neural networks for computer vision-based detection and recognition of dumpsters. *Neural Comput Appl* 32(17):13203–13211
33. Ren J, Han J (2021) A new multi-scale pedestrian detection algorithm in traffic environment. *J Electri Eng Technol* 16:1151–1161
34. Shao X, Wang Q, Yang W et al (2021, 1820) Multi-scale feature pyramid network: A heavily occluded pedestrian detection network based on ResNet. *Sensors* 21(5)
35. Sun C, Ai Y, Qi X et al (2022) A single-shot model for traffic-related pedestrian detection. *Pattern Anal Appl* 25(4):853–865
36. Szegedy C, Vanhoucke V, Ioffe S et al (2016) Rethinking the inception architecture for computer vision. In Proceedings of the IEEE conference on computer vision and pattern recognition:2818–2826
37. Tian Y, Luo P, Wang X, et al (2015) Pedestrian detection aided by deep learning semantic tasks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5079–5087
38. Wang M, Chen H, Li Y et al (2021) Multi-scale pedestrian detection based on self-attention and adaptively spatial feature fusion. *IET Int Trans Syst* 15(6):837–849
39. Wang Z, Feng J, Zhang Y (2022) Pedestrian detection in infrared image based on depth transfer learning. *Multimed Tools Appl* 81(27):39655–39674
40. Woo S, Park J, Lee JY et al (2018) Cbam: Convolutional block attention module. In Proceedings of the European conference on computer vision (ECCV):3–19
41. Xiao F, Liu B, Li R (2020) Pedestrian object detection with fusion of visual attention mechanism and semantic computation. *Multimed Tools Appl* 79(21):14593–14607
42. Xiao Y, Zhou K, Cui G et al (2021) Deep learning for occluded and multi-scale pedestrian detection: A review. *IET Image Proc* 15(2):286–301
43. Xie H, Chen Y (2019) Shin H (2019) Context-aware pedestrian detection especially for small-sized instances with Deconvolution Integrated Faster RCNN (DIF R-CNN). *Appl Int* 49(3):1200–1211

44. Xue Y, Ju Z, Li Y et al (2021) MAF-YOLO: Multi-modal attention fusion based YOLO for pedestrian detection. *Infrared Phys Technol* 118:103906
45. Xue P, Chen H, Li Y et al (2023) Multi-scale pedestrian detection with global-local attention and multi-scale receptive field context. *IET Comput Vis* 17(1):13–25
46. Yi Z, Yongliang S, Jun Z (2019) An improved tiny-yolov3 pedestrian detection algorithm. *Optik* 183:17–23
47. Zhang C, Kim J (2019) Multi-scale pedestrian detection using skip pooling and recurrent convolution. *Multimed Tools Appl* 78:1719–1736
48. Zhang L, Lin L, Liang X, et al (2016) Is faster R-CNN doing well for pedestrian detection? In *Computer Vision–ECCV*, pp 443–457
49. Zhang X, Cao S, Chen C (2020) Scale-aware hierarchical detection network for pedestrian detection. *IEEE Access* 8:94429–94439
50. Zhang Q, Ren J, Liang H et al (2022) BFE-Net: Bidirectional Multi-Scale Feature Enhancement for Small Object Detection. *Appl Sci* 12(7):3587
51. Zhou H, Yu G (2021) Research on pedestrian detection technology based on the SVM classifier trained by HOG and LTP features. *Future Gen Comput Syst* 125:604–615

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.