



An improved one-stage pedestrian detection method based on multi-scale attention feature extraction

Jun Ma¹ · Honglin Wan¹ · Junxia Wang² · Hao Xia¹ · Chengjie Bai¹

Received: 18 August 2020 / Accepted: 8 January 2021 / Published online: 22 January 2021
© The Author(s), under exclusive licence to Springer-Verlag GmbH, DE part of Springer Nature 2021

Abstract

In recent years, the performance of the convolutional neural network-based pedestrian detection method has improved significantly. However, an imbalance remains between detection accuracy and speed. In this paper, we employ a one-stage object detection framework and propose a pedestrian detection method based on the multi-scale attention mechanism of a convolutional neural network to improve the imbalance between accuracy and speed. First, a multi-scale convolution module is designed to extract corresponding features at different scales. Second, using the attention module, association information between features is mined from space and channel perspectives to strengthen the original features. Then, the enhanced features are passed through a classification and regression module to perform object positioning and bounding box regression. Finally, to learn more pedestrian location information, we improve the loss function to realise better network training. The proposed method achieved considerable results on the challenging CityPersons and Caltech pedestrian detection datasets.

Keywords Pedestrian detection · Multi-scale · Attention mechanism · Classification and regression module

1 Introduction

In the computer vision field, object detection remains a relatively active research area, and pedestrian detection is a specific object detection task in which pedestrians are detected in pictures or video sequences. Compared to general object detection, pedestrian detection has its own differentiated characteristics, and from practical value and landing space perspectives, pedestrian detection is worthy of study. For example, it has various useful applications, e.g. driverless car systems, intelligent robots, intelligent video surveillance and intelligent transportation. Pedestrian detection has strong application value and has been the focus of many researchers. Despite the rapid development of pedestrian detection technology, it must consider influences from many factors in practical applications, e.g. pedestrian posture, wear and viewing angle and occlusion between pedestrians, illumination and other factors. These issues make pedestrian

detection a challenging topic in the computer vision field. For example, achieving quick and accurate pedestrian detection in various scenes remains a hot topic.

Traditional pedestrian detection methods primarily use machine learning methods, and such methods mainly focus on manual feature extraction, feature classifier learning and post-processing, which have low hardware requirements. Machine learning-based pedestrian detection methods also provide fast detection speed. However, such methods generally only achieve good detection results under specific conditions, and the detection effect is greatly reduced in the present of occlusion or poor lighting conditions. As a result, such methods are frequently impractical because they cannot satisfy real-world requirements.

Convolutional neural networks have been used in pedestrian detection, and such methods screen out all the area frames where pedestrians may appear by traversing the entire picture. Then, these methods classify the extracted features of the area frames and finally suppress the output results by non-maximum values [1]. Pedestrian detection methods based on convolutional neural networks have achieved significant improvements in speed and accuracy and have become the current mainstream methods.

In recent years, the pedestrian detection method based on deep learning [2–5] has entered a stage of rapid

✉ Chengjie Bai
790110977@qq.com

¹ School of Physics and Electronics, Shandong Normal University, Jinan 250358, China

² School of Information Science and Engineering, Shandong Normal University, Jinan 250358, China

development; however, some problems remain. For example, the main problem is that the performance and speed have not yet reached a reasonable trade-off. In addition, research and development of autonomous driving technology is in full swing, and such technologies must be able to detect pedestrians quickly and accurately to ensure the safety of pedestrians. We adopt a one-stage object detection framework and propose a convolutional neural network pedestrian detection method based on a multi-scale attention mechanism, which we refer to as the multi-scale convolution module-attention network (MSCM-ANet), to improve the trade-off between detection accuracy and speed in pedestrian detection methods.

To improve the detection accuracy of the MSCM-ANet method, we designed a multi-scale convolution module and an attention module. We also employ a pyramid feature fusion method to generate rich feature maps. Simultaneously, to maintain faster detection speed, a one-stage object detection framework is adopted. Through the classification and regression module, the object category probability and position coordinate value is regressed directly on the feature map. In addition, we primarily use small 1×1 and 3×3 convolution kernels in the network structure and compress feature maps channels to 256 when output, which reduces the number of parameter calculations. Our primary contributions are summarised as follows.

1. We propose a pedestrian detection method for multi-scale attention feature extraction named MSCM-ANet. The proposed method includes a multi-scale convolution module and attention module for feature extraction and enhancement, a classification and regression module (CRM) for object classification and regression and an improved loss function to optimise network training. By joint action of these modules, the detection accuracy and speed of the detector are optimised.
2. The proposed method performs multi-scale extraction of features and fully mines the associated information of the feature channel and space dimensions via the attention module, which enriches feature information and ensures effective improvement of accuracy.
3. MSCM-ANet achieves satisfactory results in experiments on the challenging CityPersons [6] and Caltech [7] pedestrian detection datasets.

2 Related work

2.1 Object detection

Object detection is a basic component of computer vision research. Its purpose is to locate and classify object instances in a large number of predefined categories of natural images.

Since 2006, many deep neural networks have emerged in the object detection field. At the 2012 ImageNet image recognition competition, Hinton et al. won the championship by building the AlexNet [1] CNN. Since then, CNN has received significant attention and is widely used in object detection research.

In 2014, Girshick et al. proposed RCNN [8], which integrates AlexNet and region proposal selective search to form a prototype two-stage detector, which effectively improves general object detection accuracy. Although the RCNN shows improvement in detection accuracy, it suffers from slow detection speed and high resource consumption. Inspired by RCNN, in 2015 and later, many improved detection frameworks have been proposed, e.g. Fast-RCNN [9] and Faster-RCNN [10]. The above two-stage detectors abandon SVM, Softmax and bounding box regression training separately, instead of using Softmax and bounding box regression training simultaneously, which greatly reduces training time. These studies have promoted the development of the two-stage detection framework.

However, the abovementioned detection methods based on region proposal are computationally expensive. Therefore, researchers began to develop a unified detection strategy, i.e. a one-stage detector prototype, to reduce computational costs. In 2016, Redmon et al. designed YOLO [11], which treats the object detection problem as a regression problem. It uses a convolutional neural network structure to directly predict the bounding box and category probability from the input image, which improves detection speed and maintains good accuracy. Also in 2016, Liu et al. proposed a one-stage detector SSD [12], which uses a CNN to extract feature maps of different scales for direct detection. This method effectively reduces detection time and provides better detection accuracy than YOLO. Subsequently, many high-performance one-stage methods emerged, e.g. DSSD [13], RefineDet [14], YOLOV2 [15] and YOLOV3 [16].

Since 2020, object detection methods have developed rapidly. For example, Han et al. designed GhostNet [17], which generates a small amount of internal feature maps through a few convolution kernels, and then efficiently generates ghost feature maps through simple linear changes. This reduces the overall number of parameters and calculation costs of the network. Fan et al. introduced a few-shot object detection method, and they proposed the Attention-RPN [18] multi-relation detector and a comparative training strategy. By constructing the few-shot detection FSOD dataset, which contains 1000 categories, a model trained on FSOD can be directly migrated to the detection of new categories without transfer learning. In another study, Wang et al. proposed SEPC [19], which used deformable convolution to adapt to the corresponding irregularities between actual features and maintain a balanced scale to improve detection accuracy. In addition, Bochkovskiy et al. presented YOLOV4 [20],

which summarised and improved a large number of previous object detectors to make the detection model faster and more accurate, where only a single GPU is required to complete model training.

2.2 Pedestrian detection

Pedestrian detection is one branch of object detection. The traditional pedestrian detection method is based on motion detection and machine learning methods. This type of method is fast; however, accuracy is not very high, and it is more prone to false detections. In 2014, researchers applied RCNN [8] to pedestrian detection, which greatly improved detection accuracy. With the successful application of RCNN in pedestrian detection, an increasing number of studies have used deep learning methods to handle pedestrian detection tasks.

In 2015, Tian et al. proposed TA-CNN [2], which uses an ACF detector to generate proposals, and then combines pedestrian detection with some specific tasks, e.g. pedestrian wear and scenes, to extract the required information optimise. Different from using traditional detectors to generate proposals, Zhang et al. designed RPN + BF [3], which is based on Faster-RCNN. Here, the original RPN was adjusted to generate proposals, and then the generated region proposals, confidence scores and extracted features are used to generate a cascaded boosted forest classifier, which effectively improves detection accuracy. In 2016, to solve the retrieval problem when multiple scales exist simultaneously, Cai et al. developed MSCNN [21], which employs different detectors for different layers and uses feature up-sampling rather than input image up-sampling to improve detection speed.

In recent years, the use of deep learning has developed rapidly in pedestrian detection. The direct use of convolutional networks for sliding window detection incurs huge computational costs; thus, Cai et al. developed a network that uses cascaded convolution, which they refer to as a cascade CNN [22]. This method can optimise the pedestrian detection scheme and quickly determine whether the detection object is a pedestrian. Targeting the occlusion problem in pedestrian detection, Wang et al. proposed RepLoss [23] and Zhang et al. designed OR-CNN [24] with two new regression loss functions, which effectively alleviated the pedestrian occlusion problem. In another study, Liu et al. presented CSP [4], which abandoned the traditional window detection method and predicted the centre position and dimension of pedestrians directly through a convolution operation, which effectively improved detection accuracy and speed.

Since 2020, many good pedestrian detection methods have emerged. For example, Wang et al. designed DR-CNN [25], which employs a new coulomb loss function to regress the bounding box. In addition, it introduced an effective

semantic-driven strategy to select the anchor point position and effectively improve the geometric and appearance interference problems in pedestrian detection. In another study, Chu [26] et al. developed a simple and effective proposal-based pedestrian detector, which allows each detection scheme to predict a set of related instances rather than the traditional single instance. This detector adopts new technologies, e.g. EMD loss and NMS, which effectively solves the problem of highly overlapping pedestrian detection in crowded scenes. In addition, Ma et al. proposed the DDFE pedestrian detection scheme [5], which is based on deep feature extraction. This method improves the receptive field and feature information of the deep feature map through a dilate convolution layer, which effectively improves imbalance between pedestrian detection accuracy and speed.

2.3 Attention mechanism

The attention mechanism originated from the study of human vision. In cognitive science, due to the bottleneck of information processing, humans will selectively focus on part of the known information while ignoring other visible information. We call this method the attention mechanism. The attention mechanism mainly focuses on two aspects: one is to decide which part of the input needs to be paid attention to, and the other is to allocate limited information processing resources to the concerned part. Through the attention mechanism, limited brain resources can be used to screen out valuable information from a large amount of information.

In computer vision, the attention mechanism is used to process visual information, by assigning different weights to different parts of the input, to find the most useful image information for detection.

In recent years, an increasing number of object detection methods have used attention mechanisms to improve the performance of algorithm networks. For example, in 2015, Jaderberg et al. proposed STN [27], which enables a neural network to actively transform the feature map into space by learning the deformation of the input to complete the pre-processing operation suitable for the task and achieved good test results. With the success of STN, the use of attention mechanisms has increased gradually in the object detection field. In 2017, Hu et al. considered the relationship between feature channels and proposed SENet [28], which automatically obtains the importance of each feature channel through learning. Based on this, network performance was improved by improving important features and suppressing unimportant features. Fu et al. designed DANet [29], which is based on the self-attention mechanism, to capture feature dependencies in the spatial and channel dimensions separately, which significantly improves network performance.

Since 2020, the use of attention mechanisms in object detection has been developed continuously. For example, Zhu et al. introduced an end-to-end attention mechanism network named RSSAN [30], which enhances the required feature information using a spatial attention mechanism and suppresses the amount of information through weighting fewer feature information to achieve effective feature learning. In another study, Ji et al. proposed ACNet [31], which uses a binary neural tree combined with attention convolution to perform weakly supervised fine-grained classification and combines attention convolution operations on the edges of the tree structure. The method uses a routing function at each node to define the calculation path from the root node to the leaf node and combines the predicted values of all leaf nodes to generate the final prediction with improved accuracy. In addition, Li et al. introduced the MAF [32] network, which uses a new channel feature attention mechanism to pass messages layer by layer in the global view and uses semantic cues in higher convolution blocks to indicate the lower feature selection in the block, which effectively improves the accuracy of saliency maps. Thus, the application of attention mechanisms in general object detection has been successful, and the same method can be attempted in pedestrian detection tasks.

3 Method

The proposed MSCM-ANet method is an algorithm based on the one-stage detection framework. The backbone includes building a multi-scale attention mechanism on the Resnet-50 [33] network, and CRM is built using the stack predictor scheme. Simultaneously, the detector automatically generates a score for each instance category, and then determines

whether it is a pedestrian instance according to an established threshold.

3.1 Overall architecture

The proposed MSCM-ANet method uses Resnet-50 pre-trained on ImageNet as its backbone. Resnet-50 [33] is commonly used in the object detection field. It is divided into five stages, and each stage contains multiple convolutional layers. After an image is input to the network, a feature map is extracted through each stage, and then down-sampling is performed. Here, the down-sampling factors are 2, 4, 8, 16 and 32, corresponding to the outputs of stages 1–5. The overall structure of the proposed MSCM-ANet method is shown in Fig. 1.

For convenience, Fig. 1 intercepts the last three stages of Resnet-50 [33]. To achieve better perform deep feature extraction, we added stage 6 (green block in Fig. 1), which comprises a series of small convolution layers at the end of the backbone network. The entire network can be divided into three parts, i.e. a multi-scale convolution module, an attention module and a CRM.

Due to different objects and different shooting angles, the object size in the image will differ. If there is only one scale in the detection method, loss of feature information will occur, which will result in blurred object feature maps. Thus, we designed the multi-scale convolution module to address the problem and obtain better features. Note that single feature detection cannot greatly mine the correlation between features, resulting in not all the detection content is we need. Therefore, we have added the attention module to the network to make the detector focus on the characteristics of pedestrians in the image, which can achieve better detection results.

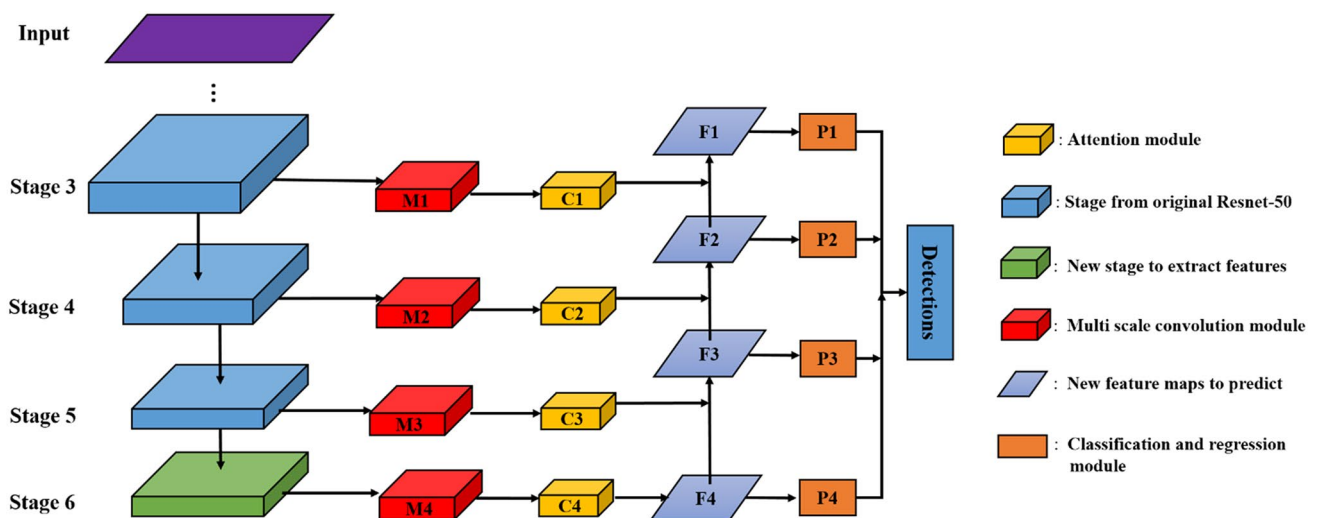


Fig. 1 Structure of proposed MSCM-ANet

The feature map extracted from the attention module enters a pyramid fusion mechanism. Here, the features are sequentially up-sampled from the deepest part of the network and merged with sub-deep feature maps such that the features contain both deep semantic information and shallow location information, which is conducive to detection. Then, the fused features are sent to the CRM to achieve bounding box regression and object classification. Next, the network predicts the feature maps extracted in stages 3–6 in order. The proposed anchor frame sizes are (16, 24), (32, 48), (64, 80) and (128, 160) pixels (aspect ratio: 0.41). These different sizes of anchor boxes are allocated to the feature map of the corresponding stage.

3.2 Multi-scale convolution module

To optimise the performance of the network, the architecture of the multi-scale convolution module (MSCM) is based on Hebb theory and multi-scale processing intuition. We refer to the structural characteristics of the inception network, and the MSCM primarily uses 1×1 and 3×3 small convolution kernels to reduce computational costs [34]. Among them, the 1×1 convolution kernel can naturally combine features with high correlation in the same spatial location but in different channels. Convolution kernels of other sizes, e.g. 3×3 , can guarantee the diversity of characteristics. The specific structure of the multi-scale feature detection block is shown in Fig. 2.

The MSCM comprises four different scales corresponding to the four levels (from top to bottom) in Fig. 2. Except for the fourth scale in Fig. 2, each scale passes a 1×1 convolution kernel to reduce the number of parameters before detection. In Fig. 2, the first scale uses a 1×1 small convolution to detect small-scale objects. The second scale uses two 3×3 convolutions to detect medium-scale objects,

while reducing the amount of parameters, expanding the receptive field and enriching the nonlinearity of the network. In the third scale, to reduce the number of calculations incurred by 7×7 convolution, we use 1×7 and 7×1 asymmetric convolutions to detect large-scale objects. The fourth scale is the average pooling layer, which is used to change the input feature arrangement and reduce the thickness of the feature map. Finally, the feature information of each scale is merged through a concatenation layer, and then passes a normalisation layer and activated layer to forms a new feature map.

3.3 Structure of attention module

According to the deep learning research in recent years [27–29], here, we summarise three important factors that affect the network under normal circumstances, i.e. depth, width and cardinality. Attention must highlight key points and increase the representativeness of the points of interest. Therefore, we employ the attention mechanism to increase expressiveness and consider important features, e.g. pedestrians, while suppressing unnecessary features, e.g. backgrounds, statues and posters. Here we combine channel attention and spatial attention in the attention module such that the detector focuses more on the characteristics of pedestrians in the image and improves the detection effect. The overall framework is shown in Fig. 3.

The convolution operation extracts features by mixing multiple channels and spatial information; thus, we design the attention module along the features in the channel dimension feature and spatial dimension feature. As a result, each branch can learn classification and positioning information on the channel and spatial dimensions, respectively. The equations for the attention network are given as follows:

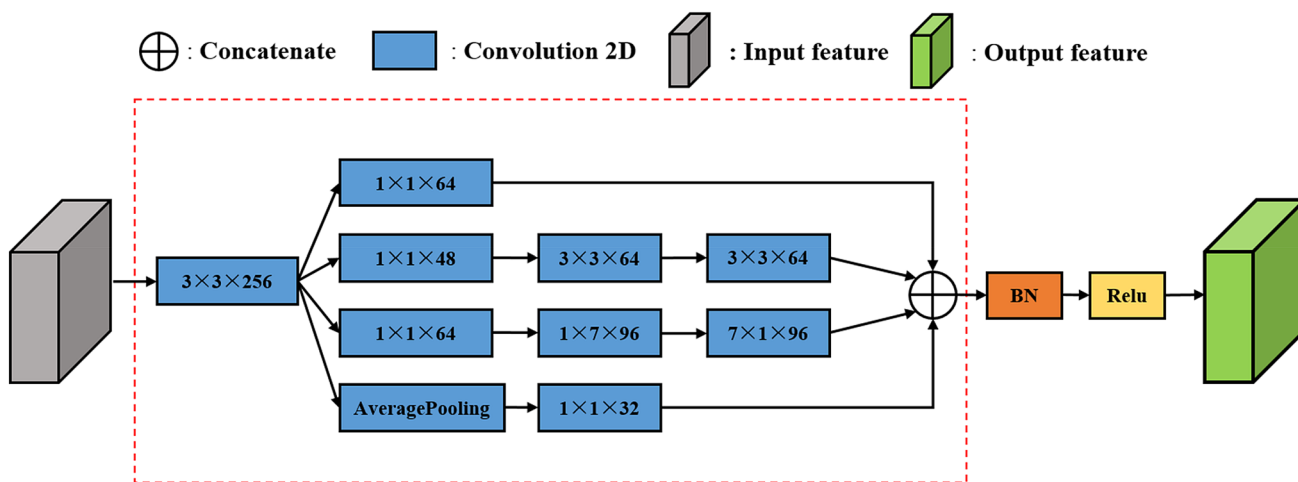


Fig. 2 Structure of multi-scale convolution module

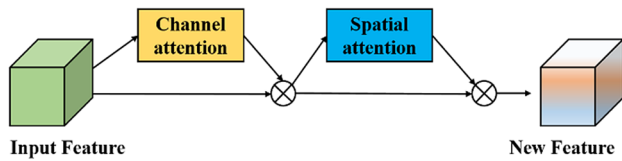


Fig. 3 Structure of attention module. The green box represents the feature extracted from the previous layer. The multi-coloured box represents the new feature handled by attention module

$$x_{Mid} = CH(x) \otimes x, \tag{1}$$

$$x_{New} = S(x_{Mid}) \otimes x_{Mid}. \tag{2}$$

Here, x is the original input feature, and C , H and W represent the channel number, height and width of the feature, respectively. x_{New} is the new output feature map that incorporates attention weights. $CH(\bullet)$ represents the attention extraction operation in the channel dimension of the feature, where $CH(\bullet) \in R^{C \times 1 \times 1}$. $S(\bullet)$ represents the attention extraction operation in the spatial dimension, where $S(\bullet) \in R^{1 \times H \times W}$. \otimes represents element-wise multiplication.

The attention network in the channel dimension is called channel attention. Its overall structure is shown in Fig. 4.

To gather and calculate channel information, we first compress the feature map in the spatial dimension to obtain a set of one-dimensional vectors for subsequent extraction. When performing dimensional compression on the input feature map, we use average pooling and maximum pooling. Average pooling can effectively learn the object, and maximum pooling can collect another important clue about the characteristics of unique objects to infer attention in the channel. In addition, the average pooling layer has feedback for each pixel in the feature map, while the maximum pooling layer performs gradient backpropagation calculations, only the place with the largest response in the feature map has gradient feedback. After obtaining two one-dimensional vectors, the vectors are input to a shared network comprising a hidden layer and multilayer perceptron (MLP). Here, to reduce computational costs, we set the activation size of the hidden layer to $R^{C/r \times 1 \times 1}$, where r is the channel compression rate. After applying the shared network to the vector, we perform

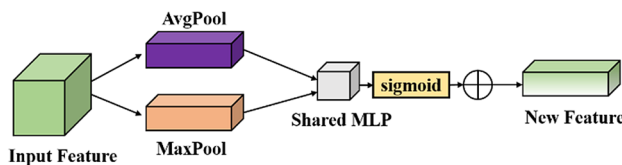


Fig. 4 Structure of channel attention module

element-wise summation, then go through the activation function and finally merge the output feature vectors to obtain new feature weights. The equations for the channel attention is given as follows.

$$CH(x) = \sigma(\text{MLP}(\text{Avgpool}(x))) + \sigma(\text{MLP}(\text{Maxpool}(x))) \\ = \sigma(W_1(W_0(x_{avg}))) + \sigma(W_1(W_0(x_{max}))). \tag{3}$$

Here, $\sigma(\bullet)$ is the activation function sigmoid, and W_0 and W_1 are the weights of the multilayer perceptron, where $W_0 \in R^{C/r \times C}$ and $W_1 \in R^{C \times C/r}$, respectively. $\text{AvgPool}(\bullet)$ and $\text{MaxPool}(\bullet)$ represent the average pooling and maximum pooling layers, respectively.

After the channel attention network, new feature weights are generated and sent to the spatial attention network. The overall structure is shown in Fig. 5.

In addition to generating the attention model in the channel dimension, we consider that the network must also detect which parts of the feature map should have a higher response at the spatial level. Therefore, we also designed a spatial attention network.

The first input is the characteristics of the channel attention module, and the average pooling layer and the maximum pooling layer are used to compress the input feature map. Here, the compression is performed at the channel level, and the input features are respectively in the channel dimension do mean and max operations. Finally, two two-dimensional features are obtained, and we splice them together according to the channel dimensions to obtain a feature map with two channels. We then use a hidden layer containing a single convolution kernel to ensure that the final feature is consistent with the input feature map in the spatial dimension, where we use a 3×3 convolution operation. Next, after the scale operation, a feature map adjusted by double attention is obtained. The spatial attention is calculated as follows.

$$S(x) = \sigma(h^{3 \times 3}([\text{AvgPool}(x); \text{MaxPool}(x)])) \\ = \sigma(h^{3 \times 3}([x_{avg}^s; x_{max}^s])). \tag{4}$$

Here, $h^{3 \times 3}$ is the 3×3 convolution operation, $[\bullet; \bullet]$ is the feature fusion operation, $\text{AvgPool}(\bullet)$ and $\text{MaxPool}(\bullet)$ represent the average pooling and maximum pooling layers, respectively, and $\sigma(\bullet)$ is the sigmoid activation function.

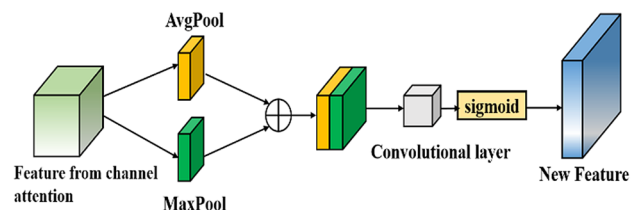


Fig. 5 Structure of spatial attention module

After the feature map generates new features through channel and spatial attention, the next feature fusion operation is performed. Finally, the fused features will be sent to the CRM for detection to obtain the results.

3.4 Classification and regression module

To improve detection accuracy while ensuring sufficient detection speed, we draw on the experience of ALFNet [35]. Under the one-stage framework, we cascade two identical predictors to form CRM (Fig. 6).

The new feature maps generated from the attention network enter the first predictor. After being processed through the convolution kernel (3×3 convolutional layer with 256 channels), the classification and regression information is generated, and the score and regressed proposals are calculated. In Fig. 6, k represents the number of anchor boxes. Then, the feature enters a second equal predictor and performs classification and regression again. Finally, the results of each scale are combined to obtain the final detection result. Here, each predictor uses the regressed anchor for optimisation. With multiple classification and regression of the detector, the anchor is refined gradually, and the number of obtained positive samples increases gradually, which is conducive to generating accurate positioning information. As a result, the one-stage detector can achieve results that are equivalent relative to accuracy of a two-stage detector while maintaining detection speed.

3.5 Loss function

In pedestrian detection, the category distribution in a single picture is typically not balanced. For example, there may only be a single person in an image, and the rest of the image is background, which will lead to extremely unbalanced numbers of positive and negative samples. Therefore, the design of the loss function should be considered carefully. Inspired by mitigating imbalance in the number of positive and negative samples by assigning different weights to such samples in focal loss [36], we designed the

loss function to keep positive and negative samples in relative balance. Here, the loss function comprises both classification loss and regression loss. During training, when the IOU values of the detected anchor box and ground truth box are greater than threshold I_h , we treat this anchor box as a positive sample and use e_+ as a representation. When the IOU value is less than threshold I_l , anchor boxes are considered to be negative samples denoted e_- . Note that other anchor boxes between I_l and I_h are ignored and not calculated.

For the classification loss function, we made improvements based on focal loss. The classification loss function is expressed as follows:

$$L_{\text{classification}} = -\mu \sum_{b \in e_+} (1 - T_b)^\gamma \log T_b - (1 - \mu) \sum_{b \in e_-} T_b^\gamma \log (1 - T_b). \tag{5}$$

Here, T_b is the probability that sample b is a positive sample, μ is the weight parameter and γ is the focusing parameter. In our experiment, we set $\mu = 0.25$ and $\gamma = 2$. With this adjustment, the weight of easier to classify samples is lower, and the model is more focused on training more difficult samples.

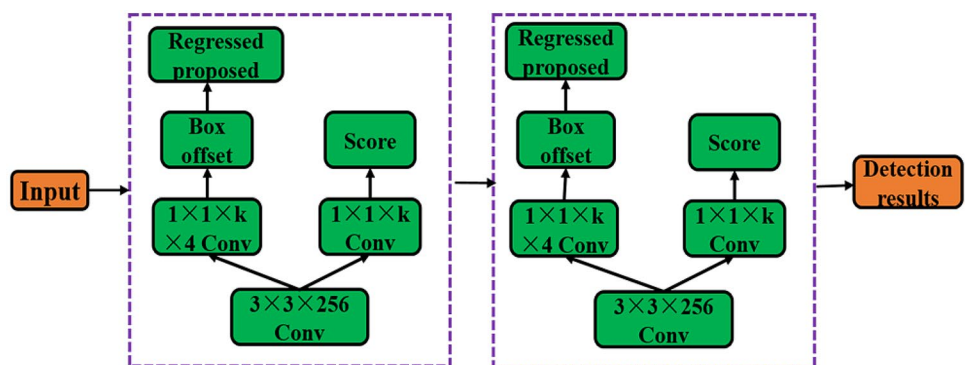
For the regression loss function, we employ the smooth L1 loss [10] function:

$$L_{\text{location}} = \lambda \frac{1}{N_{\text{location}}} \sum_b T_b^* L_{\text{regression}}(f_b - f_b^*), \tag{6}$$

$$L_{\text{regression}}(f_b - f_b^*) = \begin{cases} 0.5 |f_b - f_b^*|^2, & |f_b - f_b^*| < 1 \\ |f_b - f_b^*| - 0.5, & \text{otherwise} \end{cases}. \tag{7}$$

Here, λ is the weight balance parameter (set to 1 in our experiments). N_{location} is the number of anchor locations, b is a sample and T_b is the probability that sample b is a positive sample. When the anchor box is a positive sample,

Fig. 6 Structure of classification and regression module



the value of the ground truth label T_b^* is 1, and when the anchor box is a negative sample, the value of the ground truth label T_b^* is 0. $L_{\text{regression}}$ is the smooth L1 function, where f_b represents the bounding box parameterised coordinates generated during anchor box prediction, and f_b^* is the bounding box parameterised coordinates of the anchor box ground truth.

3.6 Inference

The MSCM-ANet's network comprises forward propagation and back propagation. The input image is propagated forward through the backbone network. Here, stages 3–6 generate corresponding feature maps. Then, the feature maps are sent to the MSCM and attention module to extract features. These features are passed through the feature pyramid network with sequential up-sampling from the deep network and fusion of the shallow features to generate a new feature map. Finally, the bounding boxes and confidence scores of the feature map are regressed through CRM. We filter all boxes with a confidence score below 0.01, and we employ a non-maximum suppression method to retain boxes with the highest scores, merge the remaining boxes and delete redundant anchor boxes, thereby improving the detection efficiency. Note that the threshold of ground truth boxes and generated boxes was set to 0.5 during testing.

4 Experiments

To verify the effectiveness of the proposed MSCM-ANet, experiments and evaluations were performed on two commonly used pedestrian detection datasets, i.e. the CityPersons [6] and Caltech [7] datasets. The system used in the experiment is Ubuntu 16.04, two RTX2080ti GPUs and 11 GB of main memory. The network construction, training and testing were all performed using in Keras environment.

4.1 Datasets

The CityPersons [6] dataset is a high-resolution and large-scale public pedestrian detection dataset with various occlusion levels. This dataset has a total of 2975 training set images, 500 validation set images and 1525 test set images, with a total of 5000 1024×2048 pixel values in PNG format images. These images are from 27 different cities, and each instance is labelled accurately. The CityPersons dataset has more labelled pedestrian instances. Its training set contains 19,654 pedestrian labels, and the pedestrian poses are diverse. There are more pedestrians in a single picture.

The Caltech [7] dataset is currently the most commonly used and largest pedestrian database. It was capturing using a vehicle-mounted camera for approximately 10 h.

The video resolution is 640×480 at 30 frames per second. This dataset is labelled with 350,000 rectangular boxes and 2300 pedestrians. In addition, the time correspondence between the rectangular boxes and occlusion conditions are also labelled. The dataset is divided into set00–set10. According to official usage, the training set has a total of 42,788 frames (set00–set05), and the test set has a total of 4024 frames (set06–set10).

The CityPersons [6] and Caltech [7] datasets use the log-average miss rate over the false positive per image as the evaluation metric, and we fixed the range to $[10^{-2}, 10^0]$, which we denote MR^{-2} .

4.2 Training settings

The Adam optimiser was used in training to optimise the weight parameters. To make the training weights data smoother, we also employed the moving average weights method in the experiments. We set the attenuation rate to 0.999 during training. In addition, the min-batch method was used in training, and the batch size of a single GPU was set to 10.

To reduce computational costs during training, we cropped the CityPersons [6] dataset into 640×1280 images, and the Caltech [7] dataset was cropped into 336×448 images. For testing and evaluation, images in the two datasets were maintained at their original size. The pretrained Resnet-50 [33] model was used to initialise the weight parameters, and the remaining network layer used the Xavier method for random parameter initialisation. The number of iterations of the entire network is 160 k and the learning rate was set to 10^{-4} during the initial 120 k iterations. Then, the learning rate was changed to 10^{-5} for the remaining 40 k training iterations (at which time training was terminated).

4.3 Data augment

Due to the limited number of samples in the original datasets, we adopted a data augment strategy in the training phase to increase sample diversity and improve the training effect of the pedestrian dataset.

In this experiment, we used online data augmentation strategies to reduce computer resource consumption. First, we randomly flipped the image horizontally, adjusted its brightness randomly and set the probability to 0.5. Then, the image was randomly scaled to $[0.3, 1.0]$ of the original size. Finally, we randomly cropped or filled the image which in the CityPersons [6] training set with a width and height of 640×1280 pixels. Note that the images in the Caltech [7] dataset were randomly cropped to 336×448 pixels.

Table 1 Results of ablation experiment

Backbone	Method	MR ⁻²
Resnet-50 + MSCM	Without attention	13.07
	Channel attention	12.36
	Spatial attention	12.50
	Channel attention first and spatial attention	11.95
	Spatial attention first and channel attention	12.17

4.4 Ablation study

The proposed MSCM-ANet uses the attention module to mine the correlation between features, thereby improving the detection effect. The attention module uses a combination of channel attention and spatial attention mechanisms to make the generated features targeted at detecting pedestrians. However, effectively combining these attention mechanisms for joint detection and obtaining better results was a problem we need to address. Thus, we conducted ablation experiments on the CityPersons [6] dataset.

We conducted multiple sets of experiments to verify the effectiveness of adding an attention network by whether to add an attention mechanism and where the two attention mechanisms are located. Table 1 shows the detection results obtained when adding different attention mechanisms.

As can be seen, when no attention mechanism was used, the MR⁻² value was 13.07, and the effect is general. When a single attention network mechanism was used, the network considered the correlation in the channel and spatial dimensions and focused on generating features for pedestrians to improve the detection effect (the MR⁻² value were 12.36 and 12.50, respectively). When two attention mechanism networks were used simultaneously, the generated features obtained relevant information in the channel dimension and combined information in the spatial dimension, thereby making the content of the feature richer. When entering the channel attention network first and then the spatial attention network, the MR⁻² value was 11.95. However, when features first entered the spatial attention network and then the channel attention network, the MR⁻² value was 12.17. As shown in Table 1, when entering the channel attention network first and then entering the spatial attention network, the correlation information between features can be fully utilised to achieve better feature detection results.

In addition, we performed statistics analysis and comparisons of the number and distribution of anchor boxes generated by the CRM with different threshold intervals during training. Here, $[a, b)$ represents the threshold between a and b . The specific data are shown in Fig. 7.

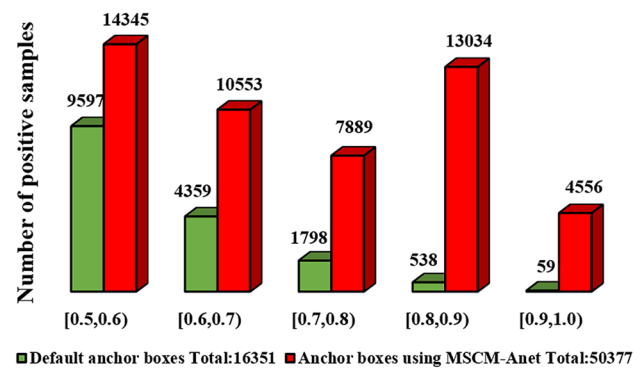


Fig. 7 Comparison of anchor boxes with and without proposed MSCM-ANet. Green blocks are the number of anchor boxes using the default network. Red blocks are the number of defined anchor boxes using MSCM-ANet. The threshold range was between 0.5 and 1

As shown in Fig. 7, a total of 16,351 positive samples were generated when only using Resnet50 [33], which is a small number of samples. When using the proposed MSCM-ANet method, the anchor frame can be refined, and more positive samples can be generated in each threshold interval by the combined effect of multi-scale modules, attention network and pyramid fusion. Compared to the default network, the total number of samples increased to 50,377 positive samples, which verifies the effectiveness of the proposed method.

4.5 Experimental results and analysis

In this section, we verify the effectiveness of the proposed MSCM-ANet method experimentally and compare it to the current advanced detection methods. All experiments were performed using the CityPersons [6] validation set and the Caltech [7] test set.

Figure 8 is a feature map extracted from different layers of a picture in the CityPersons validation set using the MSCM-ANet method. For simplicity, we only extract a feature map from the first channel in each network layer. The feature maps are extracted from the shallow network to the deeper network, i.e. from left to right and top to bottom.

It is evident from Fig. 8 that the extracted shallow features cover the location information of pedestrians and objects in the picture. When the deepest point of the multilayer network layer is reached, the extracted deep features contain the image's semantic information. The proposed network enriches the content of feature maps by combining multi-scale feature information and a fusion mechanism, which is conducive to prediction.

The experimental results for the CityPersons [6] validation set are shown in Table 2.

According to the experimental results in Table 2, without the MSCM-ANet method, the MR⁻² value with the

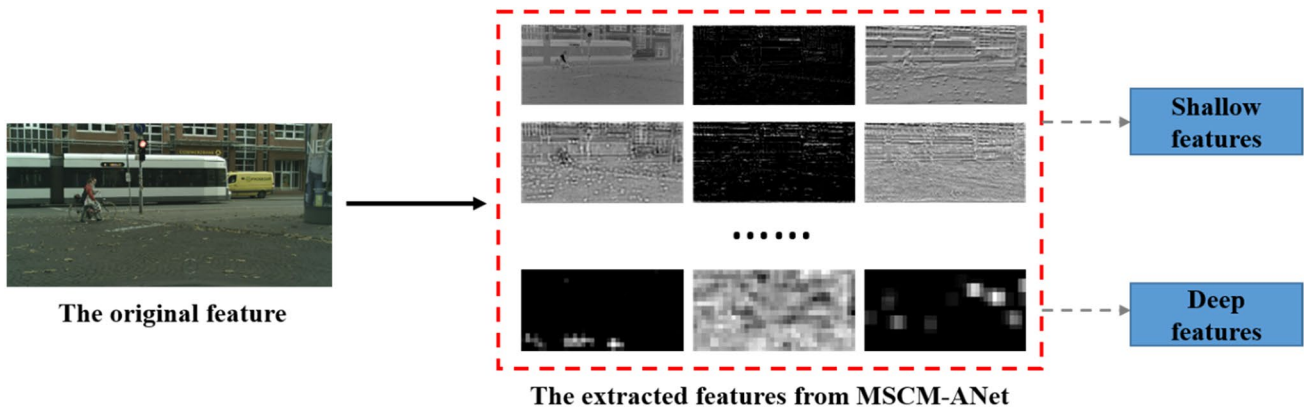


Fig. 8 Extracted features from MSCM-ANet

Table 2 Comparison of whether the backbone uses the MSCM-ANet

Backbone	Multi-scale convolution module-attention	MR ⁻²	
		IoU=0.5	IoU=0.75
Resnet-50		16.01	48.94
Resnet-50		11.95	38.04

Resnet-50 [33] network was 16.01 with the CityPersons benchmark. When the MSCM-ANet method is used, MR⁻² is 11.95, which proves that the proposed method is effective.

The result for the proposed method and existing mainstream methods on the CityPersons [6] validation dataset are listed in Table 3. We choose FRCNN [10] (the early classic pedestrian detection method), RepLoss [23] (the typical pedestrian detection method for pedestrian occlusion), CSANet [39] (the pedestrian detection method based on key points), and PedJointNet [37], AMS-Net [38], DDFE [5], TLL + MRF [40], ALFNet [35] (the

high-performance pedestrian detection method in recent years) to compare with the method we designed. During testing, the detection method usually up-samples the image to achieve good detection accuracy. However, to maintain the detection speed of the pedestrian algorithm, we choose to test using the original image. In the table, the reasonable column represents the detection value in an ideal state, and heavy, partial and bare represent three different degrees of occlusion. Test time is the time required to test each picture.

As can be seen from Table 3, MSCM-ANet achieved considerable results on the CityPersons dataset. Under reasonable conditions, the proposed method achieved a value of 11.95, which is slightly lower than the value achieved by CSANet, the suboptimal method. However, when comparing pedestrian occlusion levels, the proposed algorithm has obvious advantages under heavy and partial occlusion conditions. With heavy occlusion, our algorithm achieved a value of 50.1, which is lower than both CSANet (51.3) and the suboptimal DDFE method (50.5). With partial occlusion, the

Table 3 Comparison of the performance between MSCM-ANet and other mainstream methods on CityPersons

Method	Backbone	Reasonable	Heavy	Partial	Bare	Parameters (MB)	Test time (s/img)
FRCNN [10]	VGG-16	15.4	–	–	–	–	–
FRCNN + Seg [10]	VGG-16	14.8	–	–	–	–	–
RepLoss [23]	ResNet-50	14.6	60.6	18.6	7.9	188.5	0.35
PedJointNet [37]	ResNet-50	13.5	52.1	–	–	–	–
AMS-Net [38]	ResNet-50	13.9	–	–	–	–	–
CSANet [39]	ResNet-50	12.0	51.3	11.9	–	–	–
DDFE [5]	ResNet-50	12.9	50.5	12.1	8.2	168.8	0.22
TLL + MRF [40]	ResNet-50	14.4	52.0	15.9	9.2	230.5	0.41
ALFNet [35]	ResNet-50	13.1	51.9	11.4	8.4	191.1	0.27
MSCM-ANet	ResNet-50	11.95	50.1	11.1	7.9	154.1	0.15

The best result is shown in bold

proposed algorithm achieved a value of 11.1, which is 0.8 lower than CSANet. The proposed method is able to achieve these results because it extracts feature information at different scales, mines the correlation between features, and pays more attention to required features. Consequently, the proposed algorithm can achieve better detection result even when it is not designed for pedestrian occlusion problems. Under bare conditions, the proposed algorithm achieved a value of 7.9, which is equivalent to the results obtained by RepLoss result, a localisation technique that is designed for pedestrian occlusion. In addition, the method proposed in this paper is based on one-stage framework that performs classification and regression on the feature map directly during detection. In addition, we also use multi-scale convolution modules to reduce the number of parameters, which effectively reduces test time. The test time for the proposed method was 0.15 s/img, and the weight parameters of model were as low as 154.1 MB (Table 3).

To verify the generalisation of our algorithm network, we conducted experiments on Caltech dataset. Here, we used the same network configuration and compared the proposed

algorithm to various mainstream algorithms, such as DeepParts [41] (the early application of deep learning pedestrian detection algorithm), FasterRCNN + ATT [10], RPN + BF [3], MSCNN [21], FDNN [43] (the early recognised pedestrian detection method), RTPD [42], TLL-TFA [40] and DM-PPP [44] (the pedestrian detection method with good performance in the near future). Figure 9 shows the L-AMR vs. FPPI plot of the state-of-the-art methods under four different evaluation settings on Caltech test set.

As can be seen from Fig. 9, the proposed method achieves a MR^{-2} value of 7.10 under reasonable evaluation settings on Caltech test sets, which is 0.18 lower than the suboptimal result. Under no occlusion and partial occlusion, evaluation settings, MSCM-ANet gets the lowest MR^{-2} values which are 5.82 and 11.25. Under the evaluation setting of heavy occlusion, our method achieves suboptimal result which MR^{-2} value is 30.05 slightly higher than the best results (28.66 TTL-TFA). Based on the experimental performance on two different datasets, it can be said that the MSCM-ANet has certain advantages. Table 4 shows the comparison results of different methods on the Caltech test subset.

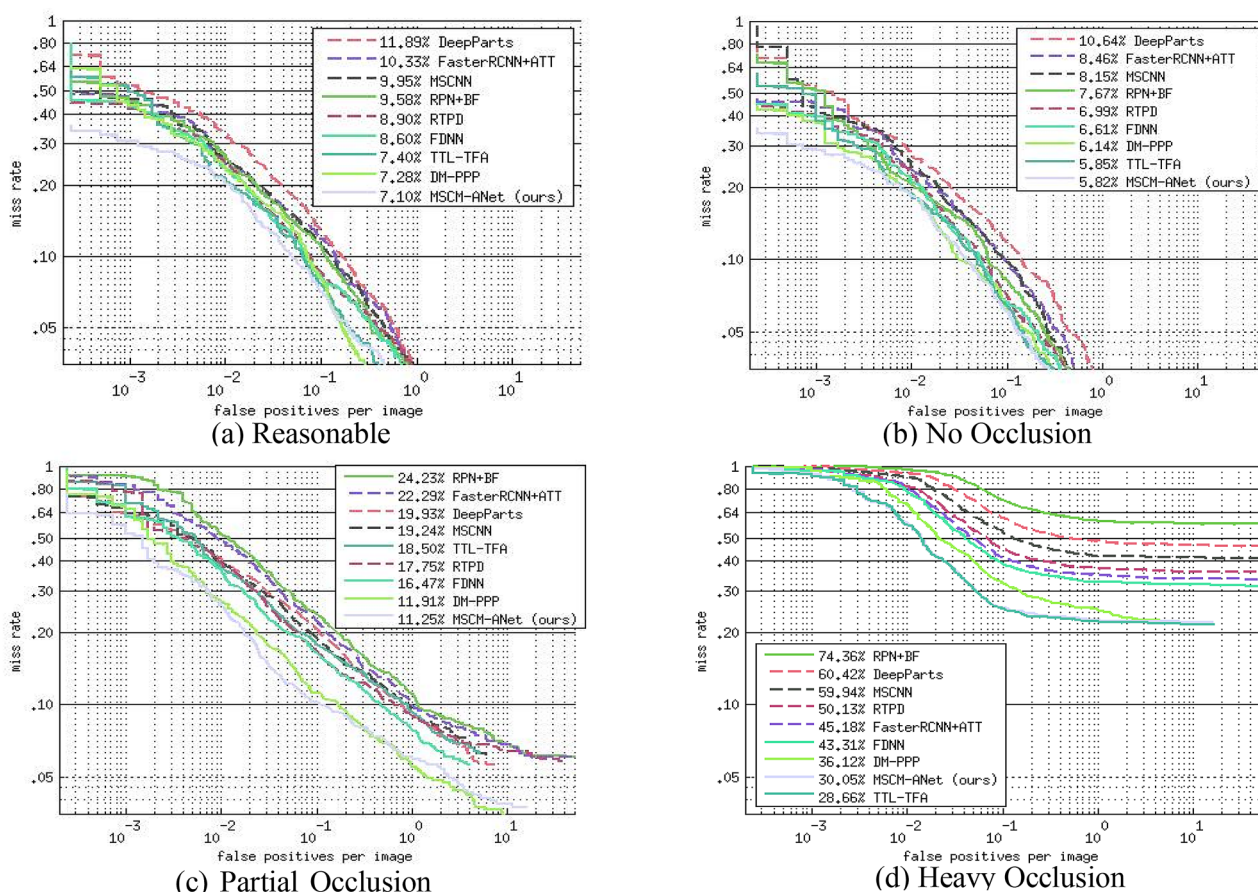


Fig. 9 L-AMR vs. FPPI plot of the state-of-the-art methods and the proposed MSCM-ANet under four different evaluation settings on Caltech test sets. **a-d** "Reasonable", "no occlusion", "partial occlusion", and "heavy occlusion" evaluation settings respectively

Table 4 Comparisons of the performance among the state-of-the-art methods on Caltech test subset

Method	Parameters (MB)	Test time (s/ Reasonable img)	
RPN + BF [3]	–	0.50	9.58
FDNN [43]	–	0.16	8.60
TLL-TFA [40]	218.5	0.25	7.40
RTPD [42]	160.5	0.13	8.90
MSCM-ANet	154.1	0.06	7.10

The best result is shown in bold

From Table 4, MSCM-ANet achieves the fewer amount of parameters, the faster test speed, and the lower value than other methods we listed on Caltech. And it is worth mentioning that our method achieves detect speed of 16.67 FPS.

Figure 10 is a partial display of the detection results using the CityPersons dataset. Here, the display includes various situations, such as pedestrian occlusion, multi-pedestrian detection, detection under different lighting conditions,

detection of highly similar objects, for example, star posters in the figure.

From the point of view of detection effect, the proposed method rarely returns false or missed detections and can more accurately mark pedestrians in the picture. The quantitative and qualitative experimental results demonstrate that the proposed MSCM-ANet achieves satisfactory results for both performance and speed.

5 Conclusions

In this paper, we have proposed a one-stage convolutional neural network pedestrian detection method based on a multi-scale attention mechanism, which we referred to as MSCM-ANet. By fully mining multi-scale feature information combined with the one-stage detection framework, the proposed method aimed to improve the imbalance problem of detection accuracy and speed in pedestrian detection. We identified the limitations of existing methods and conducted



Fig. 10 Detection examples in different environments using MSCM-ANet

experiments to prove the feasibility and effectiveness of the proposed method.

With the MSCM-ANet method, we first extract input features at multiple scales, and then add an attention mechanism network at each scale to mine important feature information that is beneficial for pedestrian detection. Second, multi-scale features are fused through the pyramid mechanism and sent to the CRM predictor to obtain more positive samples. The proposed method achieved satisfactory results on the CityPersons and Caltech datasets. Experimental results demonstrated that better feature location and classification can be obtained through the proposed MSCM-ANet method while maintaining acceptable detection speed.

Multi-scale feature extraction has always been of interest to us. In future, we will focus on the feature up-sampling operation. We believe that feature up-sampling can be further optimised and given new weight to improve the performance of pedestrian detection algorithms.

Acknowledgements This study is sponsored by the China Shandong Key R&D Plan (2018GGX106008), and is supported by the China Shandong Key Laboratory of Medical Physical Image Processing Technology.

References

- Krizhevsky, A., Sutskever, I., Hinton, G. E.: ImageNet classification with deep convolutional neural networks. In: *Neural Information Processing Systems*, pp. 1097–1105 (2012)
- Tian, Y., Luo, P., Wang, X., Tang, X.: Pedestrian detection aided by deep learning semantic tasks. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5079–5087 (2015)
- Zhang, L., Lin, L., Liang, X., He, K.: Is faster r-cnn doing well for pedestrian detection? In: *European Conference on Computer Vision (ECCV)*, pp. 443–457 (2016)
- Liu, W., Liao, S., Ren, W., Hu, W., Yu, Y.: High-level semantic feature detection: a new perspective for pedestrian detection. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5187–5196 (2019)
- Ma, J., Wan, H., Wang, J., Xia, H., Bai, C.: An improved scheme of deep dilated feature extraction on pedestrian detection. *SIVIP* (2020). <https://doi.org/10.1007/s11760-020-01742-z>
- Zhang, S., Benenson, R., & Schiele, B.: CityPersons: a diverse dataset for pedestrian detection. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3213–3221 (2017)
- Dollar, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: An evaluation of the state of the art. *IEEE Trans. Pattern Anal. Mach. Intell.* **34**(4), 743–761 (2012)
- Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 580–587 (2014)
- Girshick R.: Fast R-CNN. In: *IEEE International Conference on Computer Vision (ICCV)*, pp. 1440–1448 (2015)
- Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**(6), 1137–1149 (2017)
- Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 779–788 (2016)
- Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., Berg, A. C.: Ssd: single shot multibox detector. In: *European Conference on Computer Vision (ECCV)*, pp. 21–37 (2016)
- Fu, C., Liu, W., Ranga, A., Tyagi, A., Berg, A. C.: DSSD: deconvolutional single shot detector. [arXiv:1701.06659](https://arxiv.org/abs/1701.06659) (2017)
- Zhang, S., Wen, L., Bian, X., Lei, Z., Li, S.Z.: Single-shot refinement neural network for object detection. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4203–4212 (2018)
- Redmon, J., Farhadi, A.: YOLO9000: better, faster, stronger. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6517–6525 (2017)
- Redmon, J., Farhadi, A.: YOLOv3: an incremental improvement. [arXiv:1804.02767](https://arxiv.org/abs/1804.02767) (2018)
- Han, K., Wang, Y., Tian, Q., Guo, J., Xu, C., Xu, C.: GhostNet: more features from cheap operations. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2020)
- Fan, Q., Zhuo, W., Tang, C., Tai, Y.: Few-shot object detection with attention-RPN and multi-relation detector. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2020)
- Wang, X., Zhang, S., Yu, Z., Feng, L., Zhang, W.: Scale-equalizing pyramid convolution for object detection. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [arXiv:2005.03101](https://arxiv.org/abs/2005.03101) (2020)
- Bochkovskiy, A., Wang, C., Liao, H. M.: YOLOv4: optimal speed and accuracy of object detection. [arXiv:2004.10934](https://arxiv.org/abs/2004.10934) (2020).
- Cai, Z., Fan, Q., Feris, R. S., Vasconcelos, N.: A unified multi-scale deep convolutional neural network for fast object detection. In: *European Conference on Computer Vision (ECCV)*, pp. 354–370 (2016)
- Cai, Z., Vasconcelos, N.: Cascade r-cnn: delving into high quality object detection. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6154–6162 (2018)
- Wang, X., Xiao, T., Jiang, Y., Shao, S., Sun, J., Shen, C.: Repulsion loss: detecting pedestrians in a crowd. In: *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7774–7783 (2018)
- Zhang, S., Wen, L., Bian, X., Lei, Z., Li, S. Z.: Occlusion-aware R-CNN: detecting pedestrians in a crowd. In: *European Conference on Computer Vision (ECCV)*, pp. 637–653 (2018)
- Wang, Z., Wang, J., Yang, Y.: Resisting the distracting-factors in pedestrian detection. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [arXiv:2005.07344](https://arxiv.org/abs/2005.07344) (2020)
- Chu, X., Zheng, A., Zhang, X., Sun, J.: Detection in crowded scenes: one proposal, multiple predictions. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [arXiv:2003.09163](https://arxiv.org/abs/2003.09163) (2020)
- Jaderberg, M., Simonyan, K., Zisserman, A., Kavukcuoglu, K.: Spatial transformer networks. In: *NIPS*, pp. 2017–2025 (2015)
- Hu, J., Shen, L., Albanie, S., Sun, G., Wu, E.: Squeeze-and-excitation networks. In: *IEEE Trans. Pattern Anal. Mach. Intell.*, p. 1 (2019)
- Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H.: Dual attention network for scene segmentation. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3146–3154 (2019)
- Zhu, M., Jiao, L., Liu, F., Yang, S., Wang, J.: Residual spectral-spatial attention network for hyperspectral image classification. In: *IEEE Trans. Geosci. Remote Sensing*, pp. 1–14 (2020)
- Ji, R., Wen, L., Zhang, L., Du, D., Wu, Y., Zhao, C., Liu, X., Huang, F.: Attention convolutional binary neural tree for fine-grained visual categorization. In: *IEEE Conference on Computer*

- Vision and Pattern Recognition (CVPR). [arXiv:1909.11378](https://arxiv.org/abs/1909.11378) (2020)
32. Li, A., Qi, J., Lu, H.: Multi-attention guided feature fusion network for salient object detection. *Neurocomputing* 416–427 (2020)
 33. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778 (2016)
 34. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.: Inception-v4, inception-ResNet and the impact of residual connections on learning. In: *AAAI* (2017)
 35. Liu, W., Liao, S., Hu, W., Liang, X., Chen, X.: Learning efficient single-stage pedestrian detectors by asymptotic localization fitting. In: *2018 European Conference on Computer Vision (ECCV)*, pp. 618–634 (2018)
 36. Lin, T. Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2980–2988 (2017)
 37. Lin, C.Y., Xie, H.X., Zheng, H.: PedJointNet: joint head-shoulder and full body deep network for pedestrian detection. *IEEE Access* 7, 47687–47697 (2019)
 38. Zhang, S., Yang, X., Liu, Y., Xu, C.: Asymmetric multi-stage CNNs for small-scale pedestrian detection. *Neurocomputing* 12–26 (2020)
 39. Zhang, Y., Yi, P., Zhou, D., Yang, X., Zhang, Q., Wei, P.: CSANet: channel and spatial mixed attention CNN for pedestrian detection. *IEEE Access* 8, 76243–76252 (2020)
 40. Song, T., Sun, L., Xie, D., Sun, H., Pu, S.: Small-scale pedestrian detection based on topological line localization and temporal feature aggregation. In: *2018 European Conference on Computer Vision (ECCV)*, pp. 536–551 (2018)
 41. Tian, Y., Luo, P., Wang, X., Tang, X.: Deep learning strong parts for pedestrian detection. In: *2015 IEEE international conference on computer vision*, pp. 1904–1912 (2015)
 42. Li, Z., Chen, Z., Wu, Q.J., Liu, C.: Real-time pedestrian detection with deep supervision in the wild. *SIVIP* 13(4), 761–769 (2019)
 43. Du, X., El-Khamy, M., Morariu, V., Lee, J., Davis, L.: Fused deep neural networks for efficient pedestrian detection. [arXiv:1805.08688](https://arxiv.org/abs/1805.08688) (2016)
 44. Saeidi, M., Ahmadi, A.: High-performance and deep pedestrian detection based on estimation of different parts. *J Supercomput* (2020). <https://doi.org/10.1007/s11227-020-03345-4>

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



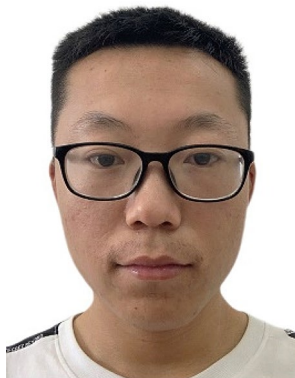
Jun Ma received his B.S. degree in Electronic Information Engineering from Shandong Youth University of political Science, China, in 2018. He is currently pursuing the master's degree in Electronic Science and Technology with Shandong Normal University. His research interests include pedestrian detection and crowd counting.



Hong-Lin Wa obtained his Ph.D. degree in Signal and Information Processing from Shandong University, China, in 2008. He is currently an Associate Professor in School of Physics and Electronics, Shandong Normal University, Jinan. His research interests include artificial intelligence, computer vision and medical image processing.



Jun-Xia Wang received her B.S. degree in Computer Science and Technology from Shandong Youth University of political Science, China, in 2019. She is currently a master student in Computer Science and Technology, Shandong Normal University. Her research interest include object detection and semantic segmentation based on deep learning.



Hao Xia became a master at the School of Physics and Electronic Science at Shandong Normal University in 2019. He is currently a master student in electronic science and technology. His research interest include object detection.



Cheng-Jie Bai received his M.S. degree in Educational Technology from Shandong Normal University, China, in 1998. He is currently a Professor in School of Physics and Electronics, Shandong Normal University, Jinan. His research interests include image processing and information fusion, pedestrian detection and medical image processing.