IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received June 18, 2020, accepted July 19, 2020, date of publication July 28, 2020, date of current version August 7, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3012558

An Improved Faster R-CNN Pedestrian Detection Algorithm Based on Feature Fusion and Context Analysis

SHEPING ZHAI[1,2], SUSU DONG[1], DINGRONG SHANG[1], AND SHUHUAN WANG[1]

[1]School of Computer Science and Technology, Xi'an University of Posts and Telecommunications, Xi'an 710121, China
[2]Shaanxi Key Laboratory of Network Data Analysis and Intelligent Processing, Xi'an University of Posts and Telecommunications, Xi'an 710121, China

Corresponding authors: Susu Dong (dongsusu1996@163.com) and Dingrong Shang (sdr_person@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61373116; in part by the Ministry of Industry and Information Technology Communication Soft Science Project under Grant 2018-R-26; in part by the Shaanxi Provincial Social Science Foundation under Grant 2016N008; in part by the Shaanxi Provincial Department of Education Science Research Program Foundation under Grant 18JK0697; in part by the Xi'an Social Science Planning Foundation under Grant 17X63; and in part by the Xi'an University of Posts and Telecommunications Graduate Innovation Foundation under Grant CXJJLY2018064.

ABSTRACT Considering the multi-scale and occlusion problem of pedestrian detection in natural scenes, we propose an improved Faster R-CNN pedestrian detection algorithm based on feature fusion and context analysis (FCF R-CNN). We design a feature fusion method of progressive cascade on VGG16 network, and add LRN to speed up the convergence of the network. The improved feature extraction network enables our model to generate high-resolution feature maps containing rich, detailed and semantic information. We also adjust the RPN parameters to improve the proposal efficiency. In addition, we add a multi-layer iterative LSTM module to the detection model, which uses LSTM's memory ability to extract the global context information of the candidate boxes. This method only needs the feature map of the image itself as input, which highlights useful context information and enables the model to generate more accurate candidate boxes containing potential pedestrians. Our method performs better than existing methods in detecting small-size and occluded pedestrians, and has strong robustness in challenging scenes. Our method achieves competitive results in both accuracy and speed on Caltech pedestrian dataset, achieving a LAMR value of 36.75% and a runtime of 0.20 seconds per image. The validity of the algorithm has been proved.

INDEX TERMS Context analysis, faster R-CNN, feature fusion, pedestrian detection.

I. INTRODUCTION

Pedestrian detection has gradually become a research hotspot in the field of computer vision due to its wide application in many fields, such as intelligent video monitoring, vehicle assisted driving, intelligent robot and other fields [1]–[6]. Pedestrian detection is a technology used to judge whether there is a pedestrian in an input image or video frame by using computer vision technology, and quickly and accurately determine the location of the pedestrian. Although steady progress has been made in the past decade, there are still many challenges, such as multi-scale and occluded pedestrians in the scene.

The associate editor coordinating the review of this manuscript and approving it for publication was Fatos Xhafa.

Statistically, over 60% of the instances from the Caltech pedestrian dataset [7] have a height smaller than 100 pixels, so effective detection of these small-size pedestrian instances is crucial to improve the overall detection accuracy of pedestrian. In the convolution feature map, the features of small-size and large-size instances are greatly different. Most small-size instances have fuzzy boundaries and appearance, causing difficulties in distinguishing those instances from background clutters and other overlapping instances. Multi-scale pedestrian detection should be able to locate the pedestrian instances with different sizes in the image and accurately give the position and size of the pedestrian bounding boxes. Compared with the traditional object detection, the multi-scale pedestrian detection has a large difference in the object's size, resulting in great difficulties in the structural design and parameter selection of the detector.

We propose an improved algorithm called FCF R-CNN based on a Faster R-CNN [8] algorithm with better speed and accuracy. Faster R-CNN firstly extracts multi-layer convolution features through Convolutional Neural Network (CNN). Secondly, the Region Proposal Network (RPN) generates candidate regions that may contain objects according to convolution features, also known as Regions of Interest (RoI). Meanwhile, RoI is mapped back to the feature map generated by CNN. Then the classification regression sub-network Fast R-CNN [9] extracts the feature tensor with fixed length through Regions of Interest Pooling (RoI Pooling). Finally, it classifies and regresses the extracted feature tensor to obtain the final detection boxes.

Faster R-CNN only uses the last convolutional feature map. Although it has rich semantic information, the resolution is low and a lot of detail information will be lost after multi-layer pooling. Such feature map cannot meet the requirements of multi-scale pedestrian detection at all.

With the idea of feature fusion, we extract features from different layers for fusion on the basis of VGG16 [10], but instead of directly extracting features from each layer for fusion, we adopt a progressive cascade strategy. The shallow feature information is adequately integrated into the subsequent network layer to maximize the utilization rate of channel information, promote information transmission, and obtain better image features. And after each cascade, the Local Response Normalization (LRN) module is specially added to improve the convergence speed of the network.

Meanwhile, the fusion of context information has been proved to be an effective method to improve detection level [11], [12]. Context information is not only of great help to multi-scale pedestrian detection, but also can effectively solve the impact of occlusion and complex background on detection performance. Inspired by the excellent performance of Long Short-Term Memory networks (LSTM) [13] in natural language processing, we transform the LSTM model appropriately and design a multi-layer LSTM module. which uses the image's features as input, processes the feature information into the form of vector sequence, and inputs it into LSTM. After several iterations, the global context information in the image is extracted. The context of the final classification features is generated through the fully connected layers.

In summary, the main contributions of this paper are as follows:

1) We propose an improved pedestrian detection algorithm FCF R-CNN. We design a feature extraction network based on progressive cascade fusion that gives full play to the bottom-layer features. We integrate a global context information extraction module into the detector, the multi-layer iterative LSTM model is used to extract the context information of candidate boxes.
2) We handle the details of the model well. LRN is added after the connection of each layer feature map to accelerate the convergence speed of the network. In addition, the dense design of RPN and the Soft-NMS method are also used to improve the efficiency of candidate boxes proposal.
3) Our method achieves advanced performance on the most popular Caltech pedestrian dataset showing competitive speed, accuracy and robustness in challenging scene such as multi-scale and occlusion.

II. RELATED WORK

A. PEDESTRIAN DETECTION METHOD BASED ON DEEP LEARNING

Pedestrian detection is an important research direction of object detection. At present, object detection method based on deep learning can be divided into two categories according to the prediction process: one is the object detection algorithm based on regional proposal represented by Faster R-CNN, such as SPP-Net [14], Fast R-CNN, R-FCN [15], etc. The other is the object detection algorithm based on regression represented by YOLO [16], such as SSD [17], RetinaNet [18], etc. The algorithm based on regional proposal have two processes. First, the candidate regions that may contain targets are generated by the region recommendation algorithm. Then, the final detection box is obtained by the classification and position regression of the candidate regions through the convolutional neural network. The algorithm based on regression has no regional proposal section. For a given input image, it needs to be processed only once, and the target border and category of this position can be regressed in multiple positions of the image.

The regional proposal section provides a better quality detection box and makes the detection results more accurate. Most researchers apply the object detection algorithm based on the regional proposal to pedestrian detection because the pedestrian detection task requires high precision. Mao *et al.* [19] introduced other features (such as gradient, thermal information and optical flow) to Faster R-CNN to improve pedestrian detection performance, but at the same time, their method brought more computational consumption. Li *et al.* proposed an SA-Fast RCNN [20] method to solve the problem of excessive changes in pedestrian scale, and two sub-networks were designed to detect large-scale and small-scale pedestrians.

B. MULTI-SCALE FEATURE FUSION METHOD

Object detection model usually takes the last layer features to analyze, which directly leads to the partial loss of the bottom-layer details and seriously affects the detection of small objects. Feature fusion is a common method, which can solve the problem of feature loss well. It not only chooses the last convolutional feature map, but adopts the method of multi-layer feature fusion.

Bell *et al.* earlier proposed fusion convolution layer features, and proposed the Inside-Outside Net (ION) [21], The word Outside of the ION refers to context feature extraction using two Recurrent Neural Networks (RNN) outside the ROI region. The word Inside of the ION refers to extracting corresponding three scale features from three

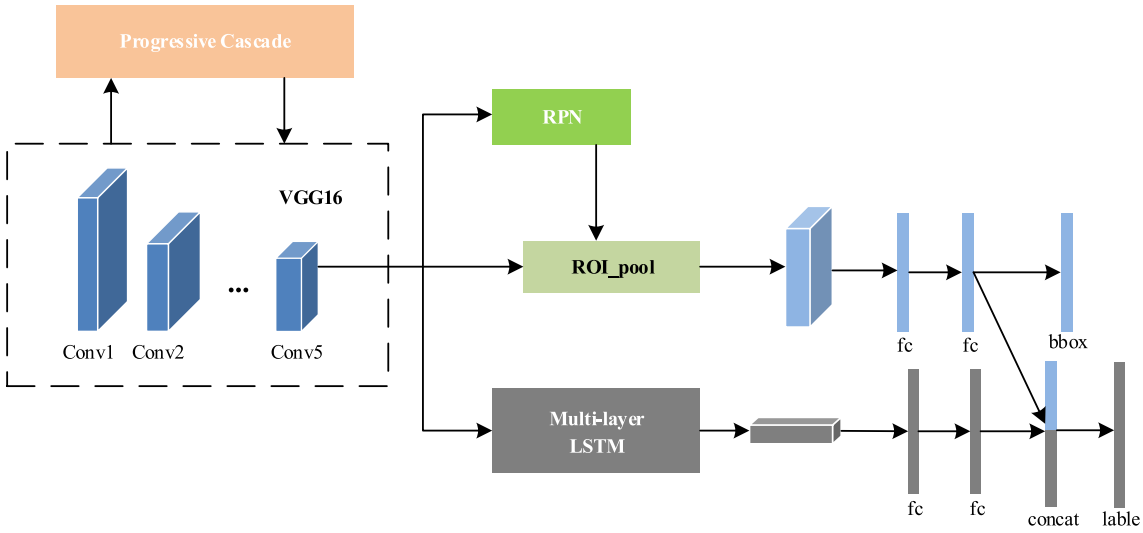

FIGURE 1. Network structure of FCF R-CNN. In the model, we improve the feature extraction network and add the extraction context information module. We propose a feature fusion method of progressive cascade to improve the feature extraction network. The fusion process is shown in Fig. 2. In addition, we use the multi-layer LSTM to extract the global context information.

convolution layers of conv3, conv4 and conv5 within the ROI region. Finally, the scale feature and context feature are fused to improve the accuracy of small target detection.

Some scholars have proposed different fusion methods to extract multi-scale features, Kong *et al.* proposed Hyper Net [22] and adopted the method of leap-through feature extraction for fusion. Their method uses max-pooling at the lower convolutional layers and adds a deconvolution operation at the deeper convolutional layers for up-sampling. Guo *et al.* proposed a multi-scale feature fusion convolutional neural network (MFF-CNN) [23] for pedestrian detection. Similar to ION, features from the latter three convolutional layers are pooled and then spliced and fused. However, unique scaling and standardization methods are added for pedestrian features. There are MS-CNN [24] and PVANet [25], both of which used the convolutional layers for multi-scale feature splicing and fusion. However, none of these feature fusion methods can make full use of the bottom-layer features. In this paper, a bottom-up and hierarchical multi-scale feature fusion method is proposed to extract more obvious feature information.

C. OBJECT DETECTION BASED ON CONTEXT ANALYSIS

Numerous studies have shown that if context information are around objects and backgrounds are introduced into the detection model, the information expression of features can be greatly enriched. Image context information can be divided into two types according to its functions. One is global context information describing the image as a whole. The other is local context information referring to the information of different areas in the image.

The SDN [26] model proposed by Luo *et al.* can learn the layered features of different parts of pedestrians, and then integrates these features into the object detection process to improve the ability of image representation. In [27], a Region Average Pooling method was proposed using contextual features from RoI in detection scene to enhance the semantics of candidate box features.

Some models directly design the network structure of context analysis and become part of the object detection process. Wang *et al.* proposed a PCN [28] in the field of pedestrian detection, used LSTM network to extract semantic information relating to body parts, and used local competition mechanism for adaptive context scale selection. Li *et al.* proposed AC-CNN [29], after the pooling layer of the detector, two sub-networks were introduced to effectively integrate global and local context information into the final detection process. These two sub-networks are trained and detected together with CNN. This method extracts rich context information and can effectively improve the target detection accuracy. However, the subnetwork structure for context analysis needs to be further optimized.

III. FCF R-CNN

Compared with Faster R-CNN, the improvement of our method is to design a multi-scale feature extraction network and a multi-layer LSTM module for global context extraction. The overall structure of the model is shown in Fig. 1.

FCF R-CNN uses VGG16 as the basic network. We propose a bottom-up and progressive cascade method to fuse the features of different layers. The specific fusion method is shown in Fig. 2. Different from Faster R-CNN, the final extracted feature maps are also input into the multi-layer LSTM network. Global context information is extracted through multiple stacked LSTM networks. In addition, according to the characteristics of pedestrian aspect ratio, we modify the anchor scale and proportion parameters to

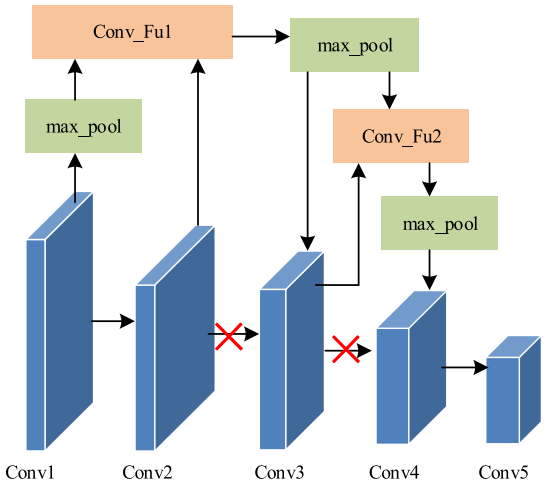

FIGURE 2. Feature extraction network based on progressive cascade.

make RPN network more suitable for pedestrian detection. There are described in detail below.

A. MULTI-SCALE FEATURE EXTRACTION NETWORK

Faster R-CNN only uses the last layer of convolution feature maps, which often results in small-scale pedestrian neglect. To solve this problem, we use the feature fusion method to improve the feature extraction capability of the backbone network for multi-scale pedestrian.

Fig. 2 shows the feature extraction network used by FCF R-CNN detector. We use a progressive cascading method to improve the VGG16 network. First of all, the convolution feature of conv1 is pooled and connected with the convolution feature of conv2 to generate the first fusion feature (Conv_Fu1). Then, we interrupt the connection between conv2 and conv3 of the original VGG16 network, and use the pooled Conv_Fu1 as the input of conv3 and connect with the output of conv3, thus generating the second fusion feature (Conv_Fu2). Finally, we interrupt the connection between conv3 and conv4 of the original VGG16 network, and Conv_Fu2 is pooled as the input to conv4. After that, feature information is transmitted to the conv5 to output the final feature maps.

This hierarchical feature fusion method promotes the flow of information in the image channel, and facilitates the transmission of more bottom-layer features. In other words, this fusion method better enriches the features of multi-scale pedestrians.

1) CONNECTION

In the feature extraction network of FCF R-CNN, the features of different layers are cascaded twice. Fig. 3 shows the first cascade process, which is also the generation process of the Conv_Fu1. The generation process of the Conv_Fu2 is similar too. After the conv1 feature map is output, the image size is compressed by the max-pooling operation. Max-pooling can avoid the fuzzy effect caused by average pooling. After

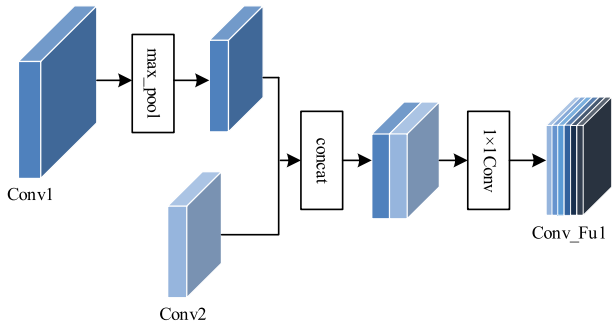

FIGURE 3. The first progressive cascade.

that, the pooled features are fused with the feature maps of conv2, generally there are two method of feature fusion: concat and element-wise add. Concat is an increase in the number of channels, That is, the number of feature channels of the image itself increases while the feature information of each layer remains unchanged. Element-wise add is the addition of feature maps, and the number of channels is unchanged. That is, the feature information of the image increases, but the dimension describing the image does not increase, only the amount of information in each dimension increases. Element-wise add requires that the number of feature channels must be the same, while concat does not. Moreover, element-wise add is carried out by adding each element of the feature maps, if the two features do not have the same feature information, the result of fusion is likely to have a negative impact. Therefore, we adopt concat method, and the calculation equation is as follows:

$$Z_{concat} = \sum_{i=1}^{c} X_i * K_i + \sum_{i=1}^{c} Y_i * K_{i+c} \tag{1}$$

where, X_i and Y_i respectively represent the feature of two paths to be fused, K_i represents the convolution kernel, and $*$ represents the convolution operation.

In the process of progressive cascade, it is possible to introduce unnecessary noise, which is not all beneficial to the detection of small-scale instances. Although the channel connection method will increase the information between the feature channels, the number of parameters will also greatly increase. Therefore, our method introduces 1×1 convolution operation to compress the fused features after feature fusion. The 1×1 convolution not only achieves cross-channel interaction and information integration, and play a role in reducing dimension and parameter quantity, but also can increase non-linearity in mapping. By learning weight allocation, the noise interference is reduced and the feature expression ability of the network is improved. The parameters of each layer of conv1 to conv5 of the improved feature extraction network are shown in TABLE 1.

2) NORMALIZATION MODULE

In the process of multi-scale feature extraction, we add a normalization module after the connection of each layer feature maps. Local response normalization (LRN) is

TABLE 1. Convolutional layer parameter.

Layers	Parameters	step
Conv1_1, Conv1_2	3×3×1×64	1
pool1	2×2	2
Conv2_1, Conv2_2	3×3×64×128	1
pool2	2×2	2
Conv3_1, Conv3_2, Conv3_3	3×3×128×256	1
pool3	2×2	2
Conv4_1, Conv4_2, Conv4_3	3×3×256×512	1
pool4	2×2	2
Conv5_1, Conv5_2, Conv5_3	3×3×512×512	1

introduced because of the different activation values and quantity differences between different layers. LRN can make the value with larger response become relatively larger, and inhibit other neurons with smaller feedback to highlight image features. In addition, LRN can solve the impact of data distribution changes on the CNN and accelerate the convergence speed of the network, thus improve the generalization ability of the model. The equation for LRN is as follows:

$$b^i_{x,y} = a^i_{x,y}/(k + \alpha \sum_{j=max(0,i-\frac{n}{2})}^{min(N-1,i+\frac{n}{2})} \left(a^j_{x,y}\right)^2)^{\beta} \qquad (2)$$

where, $a^i_{x,y}$ is the output of the ith kernel after the activation function ReLU is applied at (x, y), n is the number of adjacent nuclei at the same position, N is the total number of nuclei, and k, n, α, and β are all super parameters. We set $k = 2$, $n = 5, \alpha = 1 * e - 4, \beta = 0.75$.

B. CONTEXT INFORMATION EXTRACTION NETWORK BASED ON MULTI-LAYER LSTM

Generally, the pedestrian detection model based on regional proposal directly uses the convolution feature maps of the last layer for candidate boxes extraction and subsequent calculation, or uses the multi-scale features formed by the fusion of several layers of features. These operations can achieve good results on the premise of simple background environment and clear separation between pedestrians. Once there are problems such as dense, heavily occluded pedestrians and semantic confusion, the features extracted by CNN are only superficial representational information. Even if classification and regression can be carried out, the phenomenon of missing judgment will occur. In addition, whenever RPN recommends a set of candidate boxes, only the fragments corresponding to the candidate boxes in the feature are intercepted for pooling. The original RPN does not consider a lot of additional information, thus causing low detection accuracy.

Considering that context information can improve the performance of pedestrian detection, we use LSTM as the global context information extraction tool. We adopt a multi-layer iterative structure to extract the context features of candidate boxes. Context features can fully express the global context information of each pixel, and highlight the useful context location. This is why our method can produce more accurate candidate boxes containing potential pedestrians.

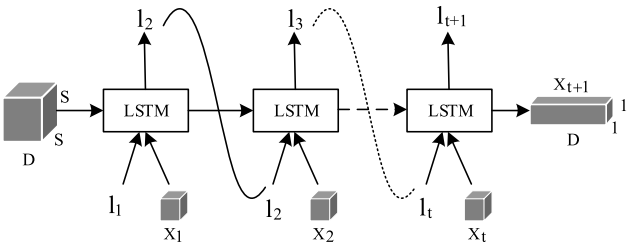

FIGURE 4. Multi-layer LSTM network structure.

The global context information extraction process is shown in Fig. 4. Due to the different sizes of the obtained fusion features, features need to be pooled to form a fixed size. Assuming that the feature size after pooling is $S \times S$ and the number of channels is D, this feature is input into a sub-network composed of a multi-layer LSTM structure. The way to change the feature into the sequence input sub-network is as follows. The feature after pooling is $F \in R^{S\times S\times D}$, it needs to be sliced to form $X = [x_i, \ldots, x_{i^2}]$, where x_i is the local description of the channel number D, and it is also a D-dimensional vector sequence. The input of LSTM is composed of these sequences. Then, one-way calculation is carried out in the multi-layer LSTM structure. After extracting the context information, it is scaled to the appropriate size and sent to the fully connected layer to generate the context part of the final classification features.

Using LSTM's memory function to perform multiple iterations can effectively extract context information. LSTM uses sequences as input, and then solves the hidden relationship between sequences. LSTM is widely used in natural language processing and other fields. We adjust the LSTM model appropriately for pedestrian detection. Fig. 5 shows the LSTM structure used in our model.

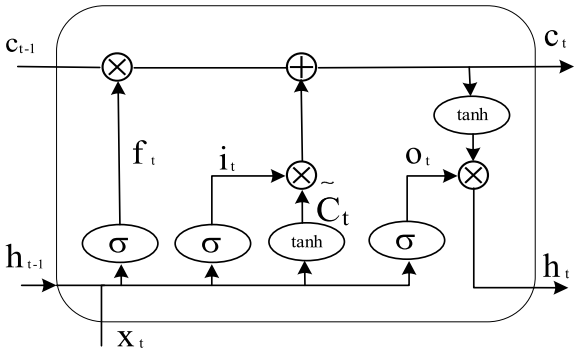

FIGURE 5. The LSTM structure used in our model.

In Fig. 5, c_t is the cell unit responsible for memory information. The input layer is composed of the output value h_{t-1} of the previous step and the new input x_t of this step. The h_t of this step is obtained through several operations in

the hidden layer. In addition to the memory unit c_t and input unit h_t at each step in the whole LSTM calculation process, there are many intermediate unit results, such as, forgotten gate unit f_t, input gate unit i_t, hidden memory unit $\widetilde{C}_t$, and output unit o_t, and so on. Two types of activation functions σ and *tanh* are used in the calculation of each intermediate unit. The following equation (3) - equation (8) is the calculation process of the whole LSTM.

$$f_t = \sigma\left(W_f\left[h_{t-1}, x_t\right] + b_f\right) \tag{3}$$
$$i_t = \sigma\left(W_i\left[h_{t-1}, x_t\right] + b_i\right) \tag{4}$$
$$\widetilde{c_t} = tanh\left(W_c\left[h_{t-1}, x_t\right] + b_c\right) \tag{5}$$
$$c_t = f_t \times c_{t-1} + i_t \times \widetilde{C}_t \tag{6}$$
$$o_t = \sigma\left(W_o\left[h_{t-1}, x_t\right] + b_o\right) \tag{7}$$
$$h_t = o_t \times tanh(c_t) \tag{8}$$

where, W_f, W_i, W_c and W_o in the equation represent the corresponding weights respectively, which can be obtained through dataset learning. Similarly, b_f, b_i, b_c and b_o represent the deviation values calculated accordingly. Activation function σ is the common Sigmoid function, and activation function *tanh* is the hyperbolic tangent function.

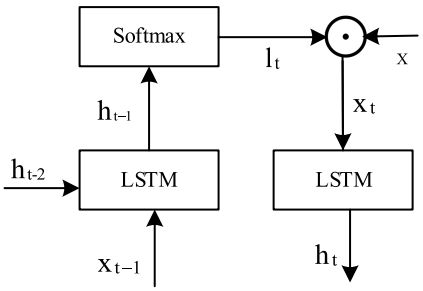

FIGURE 6. The iteration method between each LSTM structures.

The specific iterative mode among LSTM structures is shown in Fig. 6. In addition to the vector sequences sliced by pooling features, a kind of probability weight should be considered to judge the importance of each sequence. The weight of $S \times S$ mainly comes from softmax function calculation of h_t. The sequence X is weighted to form the input of each step. The weight l_t used for each step is derived from the softmax probability solution of h_{t-1} in the previous step. l_t is the probability that the corresponding candidate region of the input image can improve the global context information. The calculation is as follows:

$$l_{t,i} = p\left(L_i = i \mid h_{t-1}\right) = \frac{exp(W_i h_{t-1})}{\sum_{j=1}^{s^2} exp(W_j h_{t-1})} \tag{9}$$

where, $i = \{1, 2, \ldots, s^2$, W_i is the ith element of local softmax, which can be acquired through training and learning. L_i is a random variable. After solving these probabilities, we can find the input x_t of LSTM at each step:

$$x_t = \sum_{i=1}^{s^2} l_{t,i} x_{t,i} \tag{10}$$

In the experiment, we will take five steps of LSTM calculation, that is, five steps of iteration. Then, the iterative process can be performed according to the equation described above. But the initial values of the iteration boundaries h_t and c_t need to be set before the iteration can proceed. The model adopts the method of average evaluation according to the following equation:

$$c_o = \frac{1}{S^2} \sum_{i=1}^{s^2} X_i \tag{11}$$
$$h_o = \frac{1}{S^2} \sum_{i=1}^{s^2} X_i \tag{12}$$

Through c_o and h_o, l_1 can be obtained to calculate the input value d of the first step. After five steps of iteration, the context feature vectors whose output is a D-dimensional is obtained. The D-dimensional vector is then input into the fully connected network. The output of the two fully connected layers is 1024 dimensions. Finally, the context information part of the classification feature vector is formed to wait for splicing.

C. THE REGION PROPOSAL NETWORK MORE SUITABLE FOR PEDESTRIAN DETECTION

When using Faster R-CNN for object detection, in order to simultaneously detect different types of objects, three kinds of proportion (1:2, 1:1, 2:1) and three scales (128, 256, 512) are used for anchor. Each sliding window includes nine anchors. The primitive set based on the scale of the object and proportion assumes the assumption of uniform distribution. However, as a known object, pedestrians have a relatively fixed aspect ratio: 0.41. The original scale is more inclined to large-scale object detection, but in the pedestrian detection scenes, small-scale pedestrians account for the majority of pedestrians. Small-scale pedestrian detection directly determines the overall detection level. Therefore, we make a more elaborate design based on the original RPN. We fix an anchor ratio of 0.41, and set 9 scale parameters. The calculation equation is as follows:

$$scales = h \times 1.3^i \tag{13}$$

where, $i = \{0, 1, \ldots, 8\}$, The specific value of h is set according to the pixel height distribution of pedestrians in the dataset.

It is worth noting that both the modified network and the original network use nine different reference boxes, so the modification of the parameters will not affect the efficiency of the network. Different from the original design, we reduce the range of proportional parameters because pedestrians have a relatively fixed aspect ratio. On the other hand, our method achieves non-uniform processing of scale parameters, forming a reasonable distribution of small scale density and large scale sparseness, which is more in line with the characteristics of pedestrian detection scene. Moreover, since the loss function in RPN is composed of classification loss and regression loss used for position correction, the more aligned anchor makes the regression loss function smaller.

In the training process, the network will pay more attention to learning classification, thus making the final candidate region provided by RPN more accurate.

1) LOSS FUNCTION

In order to train RPN, we stipulate that the cof anchor and the mark box is a positive sample with the maximum value or more than 0.7, while the IoU less than 0.3 is a negative sample, and the anchor which neither belongs to the positive sample nor the negative sample does not participate in the training. The equation for calculating the IoU is as follows:

$$\mathrm{IoU} = \frac{Area\ of\ Overlap}{Area\ of\ Union} \tag{14}$$

The loss function in RPN is composed of classification loss and regression loss for position correction, as follows:

$$\mathrm{L}\left(p_i, t_i\right) = \frac{1}{N_{cls}} \sum\nolimits_i L_{cls}\left(p_i, p_i^*\right) + \lambda \frac{1}{N_{reg}} \sum\nolimits_i p_i^* L_{reg}\left(t_i, t_i^*\right) \tag{15}$$

where, p_i is $anchor[i]$ as the prediction probability of the foreground. If $anchor[i]$ is a positive sample, then the real tag $p_i^* = 1$. If $anchor[i]$ is a negative sample, $p_i^* = 0$. t_i represents the vector of the four parameterized coordinates of the predicted bounding boxes, while t_i^* is the vector of the real callout boxes associated with a positive anchor. λ is the weight balance parameter. The actual experimental results are not sensitive to the change of λ, if the value of λ changes from 1 to 100, the impact on the final results will be very small (within 1%).

$$L_{cls}\left(p_i, p_i^*\right) = -\log[p_i p_i^* + (1 - p_i)(1 - p_i^*)] \tag{16}$$

$$L_{reg}\left(t_i, t_i^*\right) = R\left(t_i - t_i^*\right) \tag{17}$$

$$Smooth_{L1}\ (\mathrm{x}) = \begin{cases} 0.5 * x^2 & |x| < 1 \\ |x| - 0.5 & otherwise \end{cases} \tag{18}$$

Classification loss L_{cls} is the logarithmic loss on two categories (foreground and background), as shown in equation (16). Regression loss L_{reg} is calculated as equation (17), where R is a function of $Smooth_{L1}$, as equation (18). Then $p_i^* L_{reg}(t_i, t_i^*)$ means that the regression loss is activated only for the positive sample anchor; otherwise, $p_i^* = 0$ is disabled.

IV. EXPERIMENTS

In our experiment, Caltech pedestrian detection dataset, the most popular and largest pedestrian detection dataset, is selected to verify the effectiveness of our algorithm FCF R-CNN. The Caltech pedestrian detection dataset is captured on a car driving through Los Angeles and contains 250k frames taken from 10 hours of 640 × 480.30Hz urban traffic video. The Caltech pedestrian detection dataset consists of 11 videos, among which set00-set05 is the training set and set06-set10 is the test set. The pedestrian annotation box is about 350,000, with about 2,300 pedestrians. In order to amplify the training set samples, 10Hz image data are collected, totaling 42,782 pieces. The test set is sampled by 1Hz, with a total of 4024 images. This experiment is trained and tested on the Caltech dataset and compared with other representative methods.

The module architecture of FCF R-CNN is similar to that of Faster R-CNN, so our experiment adopts the same multi-stage joint training method as that of Faster R-CNN. RPN proposal candidate box module is a stage. All the other modules constitute another end-to-end training stage. The context information extraction module is trained in the end-to-end overall network training stage, and the error from the regression layer and the classification layer is transmitted to the forward layer by the back propagation method. During the training process, the basic learning rate is set to 0.001 and the momentum coefficient is set to 0.9. The complete training and test code are built on Tensorflow. The entire network trains on an NVIDIA GeForce GTX TITAN X GPU with 12GB of RAM.

A. EVALUATION SETTINGS

In order to evaluate the detection performance of FCF R-CNN for pedestrians of different properties, we divide pedestrians into different evaluation settings according to the pixel height and visual range (0-1) for analysis. The evaluation settings are divided as shown in TABLE 2. We adopt the Log-Average Miss Rate (LAMR) [30], an evaluation method specially used for pedestrian detection, as the performance evaluation index. LAMR is computed evenly spaced in log-space in the range 10^{-2} to 10^{0} by averaging miss rate at the rate of nine false positives per image (FPPI). The lower the LAMR value, the better the detection effect.

B. IMPLEMENTATION DETAILS

In general, the number of candidate regions extracted by RPN network is often quite large and has a high degree of redundancy. When the original network is tested, the NMS algorithm (threshold value is 0.5) will be used to reduce the redundancy of network output and the number of candidate regions. At the same time, it can effectively reduce the computing burden of subsequent networks. The NMS algorithm sorts the detection boxes according to the degree of confidence, then retains the boxes with the highest degree of confidence, and removes the other boxes whose IoU is larger than a certain threshold. However, The NMS algorithm is not applicable to pedestrian detection where overlap is common. So the Soft-NMS [31] algorithm is adopted in the experiments. The Soft-NMS will not directly deleted overlaps confidence lower box, instead, the confidence of the detection box is reduced by linear weighting or Gaussian weighting. The higher the degree of overlap between the detection box and the box with the highest confidence, the faster the confidence decreases. In this way, the detection box which is mistakenly deleted due to overlap is largely retained.

C. RESULT ANALYSIS

1) ABLATIONS EXPERIMENTS

In order to verify the effectiveness of different components of FCF R-CNN, a simulation experiment is carried out on

TABLE 2. Evaluation.

	Overall		Scale			Occlusion		
	Resonable	All	Large	Medium	Small	None	Partial	Heavy
Height	[50,+∞)	[20,+∞)	[80,+∞)	[30,80]	(-∞,30]	(-∞,+∞)	(-∞,+∞)	(-∞,+∞)
Visibility	[0.65,1]	[0.2,1]	[0,1]	[0,1]	[0,1]	1	[0.65,1]	[0.20,0.65]

TABLE 3. Comparison of detection results of cascade information at different levels.

Network Structure	Context Module	Reasonable	Large	Medium	Small
1-2-3-4-5	×	18.44	0.89	42.43	82.24
1-(1,2)-3-(3,4)-5	×	14.30	1.17	41.26	79.07
1-(1,2)-(2,3)-4-5	×	11.28	0.92	37.69	77.33
1-(1,2)-(2,3)-(3,4)-5	×	12.04	1.10	40.64	81.04
1-(1,2)-(2,3)-4-5	√	**8.02**	**0.70**	**28.93**	**71.16**

TABLE 4. Performance comparison between different models.

Method	Reasonable	All	Large	Medium	Small	Occ.none	Occ.partial	Occ.heavy
SCF+AlexNet[34]	22.32	70.33	10.16	62.34	100	19.99	48.47	74.65
SAF R-CNN[20]	9.68	62.59	**0**	51.8	100	7.7	24.80	64.30
MS-CNN[24]	9.95	60.95	2.60	49.13	97.23	8.15	19.24	59.94
DeepParts[35]	11.89	64.78	4.78	56.42	100	10.64	19.93	60.42
ComepACT-Deep[36]	11.75	64.44	3.99	53.23	100	9.63	25.14	65.78
RPN+BF[33]	9.58	64.66	2.26	53.93	100	7.68	24.23	69.91
F-DNN+SS[32]	8.18	50.29	2.82	33.15	77.47	6.74	15.11	53.76
FCF R-CNN(ours)	**8.02**	**36.75**	0.70	**28.93**	**71.16**	**6.31**	**11.27**	**50.90**

the Caltech dataset in this section. After constant design and experimental verification of FCF R-CNN pedestrian detection network, the influence of different layers on detection accuracy is studied when selecting the feature extraction network. In the earlier version, FCF R-CNN does not introduce multi-layer LSTM to extract context information. Compared with the original Faster R-CNN, only RPN network is improved. In the experimental analysis, we also adopt the experimental results in the design process, and compared the effects of different layer features on pedestrian detection results in progressive fusion on the Reasonable, Large, Medium and Small evaluation setting of Caltech dataset.

As shown in the second to fourth rows of TABLE 3, the impacts of three different cascading modes on pedestrian detection performance are compared on VGG16 network. The numbers of 1∼5 respectively represent different convolutional layers. Parentheses indicate the cascading operation between layers, and cascading uses the last layer features of each stage (Conv1_2, Conv2_2, Conv3_3, Conv4_3, Conv5_3). It can be seen from the TABLE 3 that the more layers of feature fusion is not always better, On the contrary, excessive cascading sometimes reduces the accuracy of pedestrian detection. The network structure of the third row in the table shows better detection performance, with the LAMR value of 11.28% on the ''Reasonable'' evaluation setting, which is an improvement of 38.82% compared with the original network structure (rows 1). In particular, the network structure of the third row has a significant effect on the detection of small and medium-scale pedestrians. In ''Medium'' and ''Small'' evaluation Settings, the LAMR value reaches 37.69% and 77.33% respectively, 11.17% and 5.97% better than the original network structure. The validity of our feature extraction network is verified. The last row is the experimental result of using the network structure of the third row and adding the context information extraction module. It can be seen that the detection results are significantly improved after integrating the context information, reaching 8.02% on the ''Reasonable'' evaluation setting, 28.9% better than the detection result without adding the context module (rows 3).

We perform performance analysis on all evaluation settings of the Caltech dataset, as shown in TABLE 4. we compare with all the state-of-the-art methods reported on Caltech Pedestrian website. It can be seen that FCF R-CNN is significantly superior to other models in almost all evaluation setting. On the ''All'' evaluation setting, our method achieves 36.75%, a relative improvement of 26.92% from 50.29% by F-DNN + SS [32], and an improvement of 43.16% from 64.66% by RPN + BF [33]. In addition, significant improvements have been made in the evaluation settings at different scales. For example, when compared with

TABLE 5. Validation of improved RPN.

Method	Anchor	MAP(%)	Time(ms)
RPN	2000	69.21	231
Improved RPN	2000	75.26	231
RPN	300	67.74	224
Improved RPN	300	73.22	224

MS-CNN [24], there are 42.21% and 26.18% improvements in the pedestrian detection on the ''Medium'' and ''small'' evaluation settings.

Multi-layer LSTM module is added into the detection model to extract global context information. The context features can solve the pedestrian occlusion problems well. As shown in TABLE 4, the LAMR value on the evaluation setting of ''Occ.partial'' achieve a score of 11.27%, 25.41% better than the 15.11% of F-DNN + SS [32], F-DNN + SS is the best performance at present. On the ''Occ.heavy'' evaluation setting, the LAMR value also reach 50.90%, which is an improvement of 5.32% over F-DNN + SS at 53.76%. It shows that FCF R-CNN can effectively improve the detection accuracy of multi-scale pedestrians and solve the influence of pedestrian occlusion.

In addition, in order to verify the effectiveness of the improved RPN, we use different regional proposal networks for comparison. The classification and regression sub-networks all use the same original Fast R-CNN [9]. The results are shown in TABLE 5. While the improved RPN and the original RPN provide the same number of candidate regions, the accuracy of the improved RPN and the original RPN increased by 6.05% and 5.48% respectively, and there is no additional time consumption. These results show that the candidate regions provided by the improved RPN were of higher quality.

2) OVERALL MODEL PERFORMANCE

TABLE 6 compares the running time of some common algorithms and FCF R-CNN on the Caltech dataset with the result of miss rate when FPPI is 10^{-1}. It can be seen that our method is obviously superior to other methods, with an accuracy improvement of 5.52% and a relatively fast running time. Fig. 7 intuitively shows the running speed and miss rate distribution of different algorithms on the same hardware (TITAN X GPU). Compared with other algorithms, FCF R-CNN shows the fastest running time and the lowest miss rate, which verifies the effectiveness of our method.

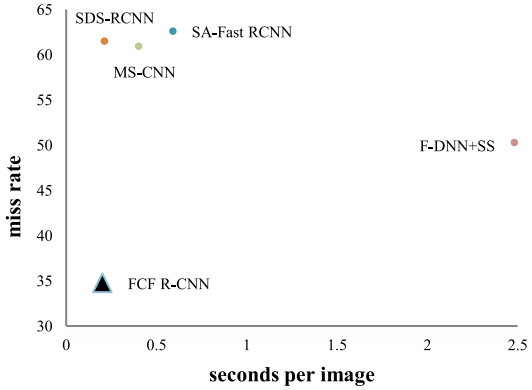

FIGURE 7. The running speed and miss rate distribution of different algorithms on the same hardware (TITAN X GPU).

3) VISUALIZATION

Fig. 8 shows four rows of four different sample detection images of different pedestrian detection networks in the Caltech dataset. In Fig. 8, (b) (c) (d) are the detection results of the current state-of-the-art methods, and are compared with the detection results of our method (a). The red bounding boxes show the detection results, and the green bounding boxes denote the ground truth. Small instances are also shown, and the height of these instance ground truth bounding boxes is less than 20 pixels. It can be observed that most pedestrians, including small instances, can be correctly detected. It can be analyzed from Fig. 8 that for some occlusion of the pedestrian instances, FCF R-CNN results are obviously better than other algorithms. In contrast, the state-of-the-art methods, such as ADM, F-DNN + SS and SDS-RCNN, generate more false positives as well as

TABLE 6. Comparison of running time and miss rate of different algorithms.

Method	Hardware	Runtime	Miss rate
ACF++[37]	Titan Z GPU	1.30	69.07
LDCF[38]	CPU	0.60	67.24
CompACT-Deep[36]	Tesla K40 GPU	0.50	64.44
RPN+BF[33]	Tesla K40 GPU	0.60	64.66
SAF R-CNN[20]	TITAN X GPU	0.59	62.59
SDS-RCNN[39]	TITAN X GPU	0.21	61.50
Ped-fused-mutiRPN[41]	GTX 1080	0.23	57.88
F-DNN+SS[32]	TITAN X GPU	2.48	50.29
MS-CNN[24]	TITAN X GPU	0.40	60.95
ADM[40]	-	0.58	42.27
FCF R-CNN(ours)	TITAN X GPU	**0.20**	**36.75**

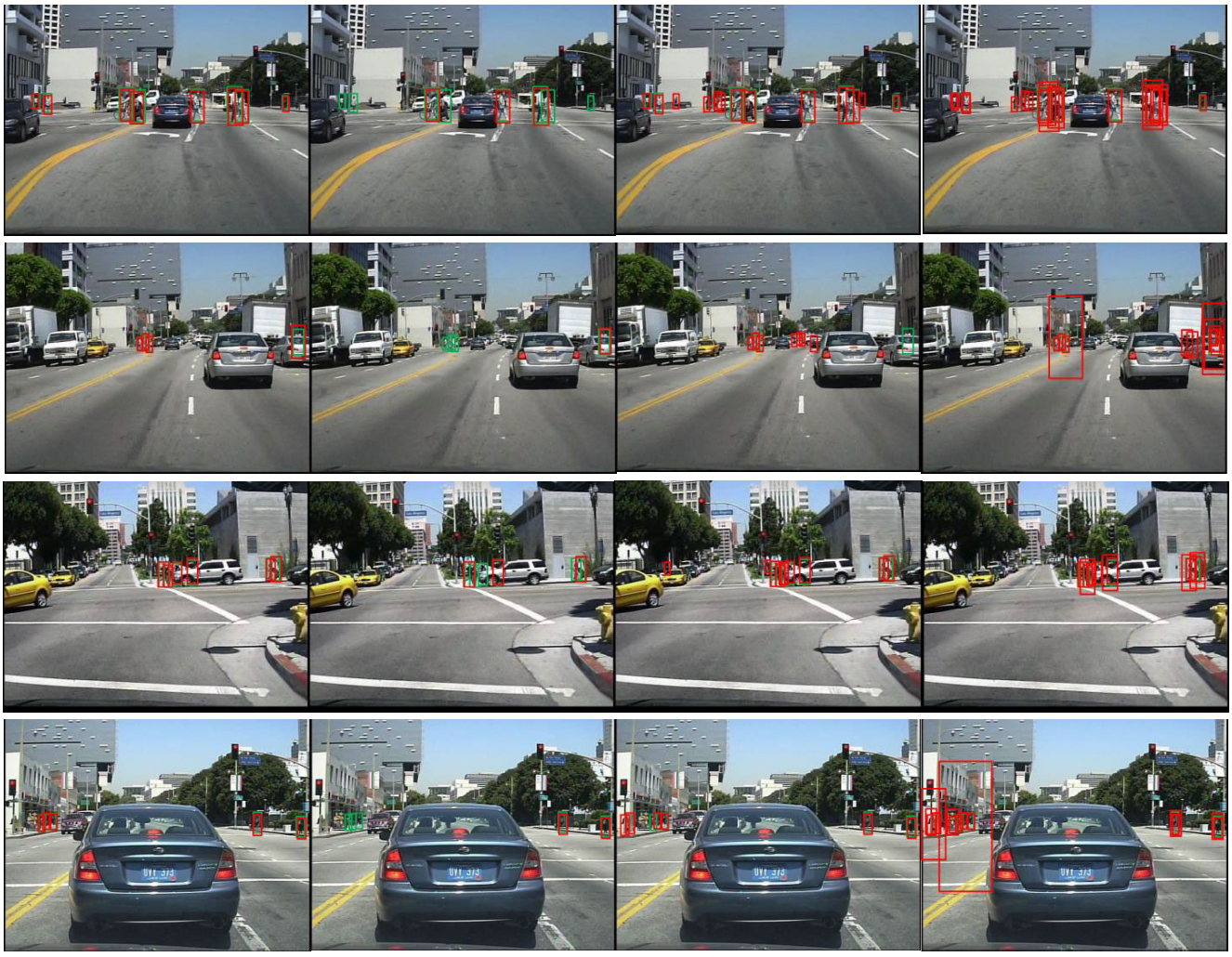

FIGURE 8. Visual comparison between our algorithm and the detection results of three state-of-the-art methods on the Caltech dataset.

more false negatives. Therefore, it can be proved that FCF R-CNN can accurately detect pedestrians of different scales more accurately, especially for small-scale pedestrians in complex backgrounds, and it is also robust in challenging scenes.

V. CONCLUSION AND FUTURE WORK

Based on Faster R-CNN, we propose a multi-scale feature fusion and context analysis of pedestrian detection algorithm (FCF R-CNN).We design a feature extraction network based on progressive cascade fusion, so as to promote the image channel flow of information, and greatly using the bottom-layer features. In addition, the local response normalization is added after the connection of each layer feature maps to accelerate the convergence speed of the network. Moreover, multi-layer LSTM modules are used to extract global context information. The context information extraction module can reduce the influence of environment such as occlusion or complex background on model detection results, and simplify the classification task. Finally, we design a regional proposal network more suitable for pedestrian detection, and use the Soft-NMS method to improve the efficiency of candidate boxes proposal. The improved RPN provides more accurate candidate boxes information. A large number of experiments on the Caltech dataset show that the FCF R-CNN not only has advantages in detecting small-scale pedestrian instances, but also performs well in complex scenes such as pedestrian occlusion. Our method achieves better performance in speed and accuracy than other state-of-the-art methods.

In the future work, we plan to simplify the network structure, optimize the detection model, and further improve the detection speed to meet the requirements of real scene detection. We will continue to explore better ways to extract context information to enhance the detection capability of the model, and continue to solve the existing problems

of pedestrian detection such as detecting heavily occluded pedestrian instances.

ACKNOWLEDGMENT

This work is supported by the special fund construction project of key disciplines of ordinary colleges and universities in Shaanxi Province.

REFERENCES

[1] W. Li, R. Zhao, T. Xiao, and X. Wang, "DeepReID: Deep filter pairing neural network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 152–159.

[2] N. K. Ragesh and R. Rajesh, "Pedestrian detection in automotive safety: Understanding State-of-the-Art," *IEEE Access*, vol. 7, pp. 47864–47890, 2019.

[3] X. Wang, M. Wang, and W. Li, "Scene-specific pedestrian detection for static video surveillance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 2, pp. 361–374, Feb. 2014.

[4] B. Nam, S.-I. Kang, and H. Hong, "Pedestrian detection system based on stereo vision for mobile robot," in *Proc. 17th Korea-Japan Joint Workshop Frontiers Comput. Vis. (FCV)*, Japan, South Korea, Feb. 2011, pp. 1–7.

[5] S. Alfasly, B. Liu, Y. Hu, Y. Wang, and C.-T. Li, "Auto-zooming CNN-based framework for real-time pedestrian detection in outdoor surveillance videos," *IEEE Access*, vol. 7, pp. 105816–105826, 2019.

[6] A. Mateus, D. Ribeiro, P. Miraldo, and J. C. Nascimento, "Efficient and robust pedestrian detection using deep learning for human-aware navigation," 2016, *arXiv:1607.04441*. [Online]. Available: http://arxiv.org/abs/1607.04441

[7] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: A benchmark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Miami, FL, USA, Jun. 2009, pp. 304–311.

[8] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[9] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, Dec. 2015, pp. 1440–1448.

[10] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[11] C. Zhou and J. Yuan, "Multi-label learning of part detectors for heavily occluded pedestrian detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 3506–3515.

[12] Z. Chen, S. Huang, and D. Tao, "Context refinement for object detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, 2018, pp. 74–89.

[13] K. Greff, R. K. Srivastava, J. Koutnik, B. R. Steunebrink, and J. Schmidhuber, "LSTM: A search space odyssey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 10, pp. 2222–2232, Oct. 2017.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.

[15] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, Barcelona, Spain, Dec. 2016, pp. 379–387.

[16] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 779–788.

[17] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2016, pp. 21–37.

[18] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020.

[19] J. Mao, T. Xiao, Y. Jiang, and Z. Cao, "What can help pedestrian detection?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 6034–6043.

[20] J. Li, X. Liang, S. Shen, T. Xu, J. Feng, and S. Yan, "Scale-aware fast R-CNN for pedestrian detection," *IEEE Trans. Multimedia*, vol. 20, no. 4, pp. 985–996, Apr. 2018.

[21] S. Bell, C. L. Zitnick, K. Bala, and R. Girshick, "Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 2874–2883.

[22] T. Kong, A. Yao, and Y. Chen, "Hypernet: Towards accurate region proposal generation and joint object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 172–178.

[23] A. Guo, B. Yin, J. Zhang, and J. Yao, "Pedestrian detection via multi-scale feature fusion convolutional neural network," in *Proc. Chin. Autom. Congr. (CAC)*, Jinan, China, Oct. 2017, pp. 1364–1368.

[24] Z. Cai, Q. Fan, and R. S. Feris, "A unified multi-scale deep convolutional neural network for fast object detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Amsterdam, Netherlands, Sep. 2016, pp. 354–370.

[25] S. Hong, B. Roh, and K. H. Kim, "PVANet: Lightweight deep neural networks for real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 412–425.

[26] P. Luo, Y. Tian, X. Wang, and X. Tang, "Switchable deep network for pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 899–906.

[27] K. Kuan, G. Manek, J. Lin, Y. Fang, and V. Chandrasekhar, "Region average pooling for context-aware object detection," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Beijing,China, Sep. 2017, pp. 1347–1351.

[28] S. Wang, J. Cheng, H. Liu, F. Wang, and H. Zhou, "Pedestrian detection via body part semantic and contextual information with DNN," *IEEE Trans. Multimedia*, vol. 20, no. 11, pp. 3148–3159, Nov. 2018.

[29] J. Li, Y. Wei, X. Liang, J. Dong, T. Xu, J. Feng, and S. Yan, "Attentive contexts for object detection," *IEEE Trans. Multimedia*, vol. 19, no. 5, pp. 944–954, May 2017.

[30] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 743–761, Apr. 2012.

[31] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, "Soft-NMS–improving object detection with one line of code," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 5562–5570.

[32] X. Du, M. El-Khamy, J. Lee, and L. Davis, "Fused DNN: A deep neural network fusion approach to fast and robust pedestrian detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Santa Rosa, CA, USA, Mar. 2017, pp. 953–961.

[33] L. Zhang, L. Lin, X. Liang, and K. He, "Is faster R-CNN doing well for pedestrian detection?" in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Amsterdam, Netherlands, Sep. 2016, pp. 443–457.

[34] J. Hosang, M. Omran, R. Benenson, and B. Schiele, "Taking a deeper look at pedestrians," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 4073–4082.

[35] Y. Tian, P. Luo, X. Wang, and X. Tang, "Deep learning strong parts for pedestrian detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, Dec. 2015, pp. 1904–1912.

[36] Z. Cai, M. J. Saberian, and N. Vasconcelos, "Learning complexity-aware cascades for pedestrian detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Apr. 11, 2020, doi: 10.1109/TPAMI.2019.2910514.

[37] P. Dollar, S. Belongie, and P. Perona, "The fastest pedestrian detector in the west," in *Proc. Brit. Mach. Vis. Conf.*, 2010, pp. 1–11.

[38] W. Nam, P. Dollár, and J. H. Han, "Local decorrelation for improved pedestrian detection," in *Proc. 27th Int. Conf. Neural Inf. Process. Syst.*, Montreal, QC, Canada, 2014, pp. 424–432.

[39] G. Brazil, X. Yin, and X. Liu, "Illuminating pedestrians via simultaneous detection and segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 4960–4969.

[40] X. Zhang, L. Cheng, B. Li, and H. Hu, "Too far to see? Not really!—Pedestrian detection with scale-aware localization policy," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3703–3715, Aug. 2018.

[41] J. X. Zeng, Q. Fang, X. Fu, and L. Leng, "Multi-scale pedestrian detection algorithm with multi-layer features," *J. Image Graph.*, pp. 1683–1691, May 2019.

SHEPING ZHAI was born in 1971. He received the B.S. degree in computer application from Beijing Jiaotong University, in 1995, and the M.S. degree in computer application and the Ph.D. degree in computer system architecture from Xi'an Jiaotong University, in 2004 and 2010, respectively. He is currently an Associate Professor and the Deputy Dean of the School of Computer Science, Xi'an University of Posts and Telecommunications, and the Director of the Shaanxi Computer Society. His main research interests include computer vision, artificial intelligence, and semantic web. In 2016, he became a member of the China Computer Federation.

SUSU DONG was born in Shaanxi, China, in 1996. He received the B.S. degree in network engineering from the Xi'an University of Posts and Telecommunications, in 2018, where he is currently pursuing the M.S. degree in computer system architecture. His research interests include machine learning and image processing.

DINGRONG SHANG was born in Shaanxi, China, in 1994. She received the B.S. degree in computer science and technology from the Baoji University of Arts and Sciences, in 2017. She is currently pursuing the M.S. degree in computer application technology with the Xi'an University of Posts and Telecommunications. Her research interests include machine learning and computer vision.

SHUHUAN WANG was born in Shanxi, China, in 1996. She received the B.S. degree in network engineering from Shanxi Agricultural University, in 2018. She is currently pursuing the M.S. degree in computer application technology with the Xi'an University of Posts and Telecommunications. Her research interests include object detection and video target tracking.