An IF-RCNN Algorithm for Pedestrian Detection in Pedestrian Tunnels

JIN REN 1,2, CHANGLIU NIU 1, AND JUN HAN 1

1 School of Information Science and Technology, North China University of Technology, Beijing 100144, China
2 Beijing Key Laboratory on Integration and Analysis of Large-Scale Stream Data, Beijing 100144, China

Corresponding author: Jin Ren (rj@ncut.edu.cn)

This work was supported in part by the 2017 Annual Outstanding Young Teacher Training Program Project of North China University of Technology under Grant XN019009, in part by the 2019 Education and Teaching Reform General Project of North China University of Technology under Grant 108051360019XN141/021, in part by the 2019 Fundamental Research Funds for Beijing Universities under Grant 110052971921/004, in part by the 2020 Beijing College Students Innovation and Entrepreneurship Training Program Project under Grant 218051360020XN214, in part by the Research Project on Teaching Reform and Curriculum Construction of North China University of Technology under Grant 18XN009-011, in part by the Beijing Urban Governance Research Center Project under Grant 20XN241, and in part by the Beijing Outstanding Young Backbone Individual Project under Grant 401053712002.

ABSTRACT In order to deal with the problems of background mixing, pedestrian blur and pedestrian multi-scale in pedestrian tunnels, we proposed an improved faster region based convolution neural network (IF-RCNN) pedestrian detection method, which uses deep CNN to automatically extract features from pictures instead of traditional manual design features. In this paper, an improved region proposal network (RPN) structure is proposed to solve the multi-scale problem of pedestrians in tunnel. The anchor size in RPN network is further improved in the face of pedestrian images in tunnels with small total pixels. Meanwhile, feature fusion technology is introduced to the algorithm to output the features of different convolution layers. The image is fused to enhance the detection performance of blurred and occluded pedestrians in tunnel. Experimental results show that IF-RCNN algorithm has better detection performance in tunnel data set and VOC2007 data set.

INDEX TERMS Pedestrian detection, deep learning, faster RCNN, feature fusion.

I. INTRODUCTION

Pedestrian detection is widely used in intelligent video monitoring, vehicle auxiliary automatic driving, target detection and other fields [1]. It is also a challenging problem in computer vision. Pedestrian tunnels have the characteristics of complex environment, dim light and large noise interference. The pedestrians monitored in video images have the problems of small size, low resolution, scale change and overlap of pedestrians. Because of its special environment, the tunnel image contains the common problems of target distortion, multi-scale, occlusion, illumination and so on in the problem of target detection and pedestrian detection. Pedestrian trampling accidents may be reduced by handling the increased pedestrian flow in pedestrian tunnels during rush hours in the mornings and evenings. Effective perception and monitoring of pedestrian in tunnel are of great significance in guiding pedestrians to pass orderly and ensuring the smooth operation of pedestrian tunnel.

Traditional pedestrian detection usually uses manual feature extraction, and then uses a classifier to achieve target detection. The pedestrian detection method of HOG+SVM proposed by Dalal et al [2]. This method generally uses a sliding window framework, which can be roughly divided into three steps: Firstly, sliding in the image using sliding windows of different scales, selecting a part of them as candidate target area. Secondly, extracting visual features of candidate target area, such as histogram of oriented gradient (HOG) features, commonly used for pedestrian target detection. Harr features [3] often used for face detection. Local binary pattern (LBP) feature [4], integral channel feature [5]. Lastly, a classifier is used for classification and recognition. This traditional method need to design specific features with good adaptability and poor generalization ability according to different detection tasks.

At present, CNN can replace traditional manual design features and extract features with advanced semantic
expression ability, strong feature expression ability and better robustness [6]–[8]. It has made great achievements in image classification, target detection and other computer vision fields [9]. Many detection algorithms based on deep learning have emerged in [7]–[11].

II. RELEVANT RESEARCH

There are a lots of research in the field of pedestrian detection. Deformable part model (DPM) algorithm [12] uses the traditional sliding window detection. By building scale pyramids to search candidate targets in a variety of scale spaces, DPM algorithm has strong robustness to target deformation. Dollar et al. [13] proposed a multi-channel feature that combines integral graph and original HOG features. The cascaded decision tree was introduced to construct adaboost classifier, which further improves the accuracy of pedestrian detection. Since 2014, great breakthrough have been made in target detection based on depth learning. The target detection framework represented by RCNN has achieved good results in VOC target detection data sets [14]. By combining candidate region extraction and CNN classification, the powerful learning feature ability of CNN has greatly improved the ability of feature extraction. Goirshick [9] proposed Faster RCNN universal target detection algorithm based on deep learning in 2015. It uses RPN network to generate candidate regions and sends them to Faster RCNN to achieve target detection, and achieves very high accuracy. Because of its excellent detection performance, it is widely used in various image classification tasks. Song Huansheng et al. [15] converted Faster RCNN to binary classification problem and applied it to vehicle detection in complex scenes. Sun et al. [16] improved faster RCNN through feature fusion, difficult case mining, and multi-scale training, and applied it to face detection tasks.

Currently, the main target detection algorithms depth learning based are mainly divided into two categories: one is region target detection algorithm based on Faster RCNN, which generates candidate target region and classifies them to achieve detection, such as Faster RCNN, Region based fully convolution network (R-FCN) [8]. The advantage of this kind of algorithm is that the detection accuracy is higher, but its detection speed is low. The other is to transform target detection into regression problem solving represented by you only look once (YOLO). In put the original picture and output the position and category of the object directly, such as YOLO [17], SSD [18]. The advantage of this method is that it can detect tens of frames per second, but the detection accuracy is relatively low, and it is insensitive to small target detection. The YOLO algorithm uses only one CNN network to directly predict the categories and locations of different targets. Although it can reduce the probability of background detection as an object, it also leads to a low recall rate. Its position accuracy is poor, and it is not good for small target objects and dense objects detection. The SSD algorithm in [18] is still poor in small size target recognition, and it can’t reach the level of Faster RCNN. Because SSD uses conv4_3

III. METHODS AND IMPLEMENTATION

Faster RCNN is the final version of RCNN series detection framework. RCNN [14] is the first CNN target detection framework based on candidate regions. It has undergone three improvements: SPP-Net, Fast RCNN [9] and Faster RCNN. The detection speed and accuracy have been greatly improved. The common point of these detection frameworks is that they are divided into three basic steps: Firstly, candidate regions are extracted from the images to be detected. Then, candidate regions are sent to the trained convolution network model to extract features. Finally, candidate regions are classified and edge box regression is used. Faster RCNN can generate candidate regions, and integrates the two into a complete network that can learn from end to end, which not only guarantees the accuracy of target detection, but also improves the speed.

In this paper, IF-RCNN algorithm is applied to pedestrian detection in tunnels. Because of the problems in the tunnel, such as mixed background, fuzzy pedestrian and multi-scale pedestrian. IF-RCNN improves the original Faster RCNN algorithm. The specific improvements are as follows:

1. In order to adapt to the pedestrian image in the tunnel scene with small overall pixels, the anchor size in the RPN network is further improved. The adjusted anchor size can be more suitable for the size of pedestrian image in tunnel data set, and pedestrian detection can be more accurate.
2. The original Faster RCNN uses a sliding window of 3 \times 3 to generate candidate regions on the last layer of feature graph. In this paper, three convolution kernels of different sizes are used to generate candidate regions on the last layer of feature maps.
3. In the original Faster RCNN, candidate regions are mapped to the last feature map of the feature extraction.
network, only using the features of Conv5_3, the deepest layer of the network, and the details of the shallow layer are not obvious. IF-RCNN uses feature fusion method to map candidate regions to feature maps generated by Conv5_3 and Conv4_3 layers of feature extraction network, which can obtain better deep semantic features and shallow detail features.

The pedestrian detection structure is designed in this paper is shown in Figure 1. The end-to-end training of the model is realized by using the approximate joint optimization mechanism in document [9]. Firstly, select the tunnel insider data set as training sample, input the data into the network, zoom the pictures to 600×1000 and send them to feature extraction network to generate feature maps. And then send the feature maps to RPN network to generate candidate regions, process the extracted features of candidate regions into fixed size feature vectors through region of interest (ROI) Pooling layer [17]–[23]. Lastly, input the full connection layer to achieve classification and regression. The whole scheme is an end-to-end structure, a network, four losses. Such a scheme design is a multi-tasks learning strategy, which helps to improve the accuracy of the model [24].

A. FEATURE EXTRACTION NETWORK

We choose VGG16 as feature extraction network. As shown in part A of Figure 1. The whole network is realized by stacking the same size convolution core (3×3) and pooling layer (1×1). VGG16 is selected to extract the features of the input image, and the full connection layer and the last pooling layer in the original network are removed. The specific network parameters are shown in Table 1.

| Type/ Layer number | Filters | Size/stride | Output |
|-------------------|---------|-------------|--------|
| Conv1_ x/2        | 64      | 3×3/1       | 600×1000 |
| Maxpool           | 2×2/1   |             |         |
| Conv2_ x/2        | 128     | 3×3/1       | 300×500 |
| Maxpool           | 2×2/2   |             |         |
| Conv3_ x/3        | 256     | 3×3/1       | 150×250 |
| Maxpool           | 2×2/2   |             |         |
| Conv4_ x/3        | 512     | 3×3/1       | 75×125  |
| Maxpool           | 2×2/2   |             |         |
| Conv5_ x/3        | 512     | 3×3/1       | 38×63   |

The convolution layers with the same output size in VGG16 network are grouped into five groups as shown in the first column in Table 1. Each group contains x layers, as shown in Conv5_ x/3, the fifth part contains three layers of convolution. From Table 1, we can see that the whole network convolution core size is 3×3. 3×3 convolution core is the smallest size to extract features. At the same time, stacking small convolution cores repeatedly can improve the learning ability of the network. Therefore, VGG16 is chosen as the feature extraction network for pedestrian detection in this paper. Columns 2, 3 and 4 in the table represent the number of convolution kernels, the size/stride of convolution kernels, and the output size of the feature map corresponding to each layer, respectively.

B. PROPOSED IMPROVEMENT TO RPN

The RPN input feature extracts the feature graph generated by the network and outputs the rectangular box set of target candidate regions. As shown in part B of Figure 1. The original RPN network structure directly extracts candidate regions by using sliding window and convolution of 3×3 on the output feature mapping, and then feeds them into the subsequent part of the network to further classify the foreground and background and to return the location box of candidate regions. In the last layer of feature extraction network output, each pixel is mapped back to the corresponding coordinate points of the original image after 3×3 convolution. Three scales, 128/256/512, and nine coarse-grained candidate regions of different sizes, namely “anchor”, are generated with the point as the center. As shown in Figure 2 and Figure 3.

The data of pedestrians in tunnels come from the surveillance cameras in tunnels. Pedestrians in surveillance videos are usually far away from the cameras, so the size of pedestrians in pictures is generally small. In order to make the model more sensitive to small targets, the scale of anchor is modified to 64/128/256, and the proportion remains unchanged, generating nine different candidate regions. In order to improve
the detection ability of multi-scale targets in the network, we propose an improved RPN structure. The original RPN network generates candidate regions by using the last layer feature map output by Conv5_3 of VGG16 convolution layer. The receptive field of each pixel is $228 \times 228$ after $3 \times 3$ sliding window. Candidate regions can’t be generated by only one receptive field, and targets of different scales can use different receptive fields to obtain better candidate regions. In this paper, three different sizes of sliding windows are used to generate candidate regions on the last layer of feature graph, which are realized by convolution of $1 \times 1$, $3 \times 3$ and $5 \times 5$, respectively. As shown in Figure 4, this RPN structure is named as an improved RPN.

The receptive field [24] is the area size of the original image mapped by the pixels on the feature mapping of each layer output of the CNN. Fisher and Valdlen [24] use Dilated Convolution to aggregate multi-scale context information to improve the accuracy of image segmentation. Expanded convolution is used to increase the receptive field without losing image information. CNN develops from 7-layer LeNet [6] in 2012 to 152-layer residual network [25] in 2015, which greatly improves the performance of image classification and detection. On the one hand, it benefits from the design of network structure and the extraction of more robust features from depth network, on the other hand, the deeper the network, the larger the field of receptivity. Inspired by the above work and the multi-scale problem in detection problems, RPN networks can locate targets of different scales by using different characteristic maps of receptive fields. Therefore, an improved RPN structure using three convolution kernels of different sizes is proposed. Such a kind of structure design can obtain better robustness and improve the detection ability of the model.

**C. PROPOSED FEATURE FUSION METHOD**

Part C of Figure 1 uses feature fusion method. The deep features of CNN have strong semantic features and large acceptance domain, namely global information and coarse-grained features. Shallow features have strong details, local information and fine-grained features. In the original faster RCNN, the candidate region generated by RPN network is mapped to the last feature map of feature extraction network by coordinates, and the feature map of candidate region is obtained. Fixed-size feature mapping is generated by ROI pool layer and sent to subsequent parts of the network to achieve target classification and regional frame regression. Only the deepest features of the network are utilized. Image semantics segmentation task is a prediction based on pixels. It needs to classify each object and separate it from the background image. FCN [26] used Skip Architecture structure to combine global information with local information, coarse-grained and fine-grained features to improve the accuracy of prediction. The specific operation is to add the pixels corresponding to the feature mapping of different layers of CNN. Pyramid scene parsing network (PSPNet) [27] uses feature maps from four different pooling layers to combine global, local and different levels of information, which improves the ability to distinguish different target categories. This method of combining feature information is called “feature fusion”, which combines the feature maps from different layers of CNN in different ways.

The pedestrian data set of tunnel pedestrians is vague, and there will be occlusion between pedestrians. Enlightened by the fusion of feature mapping from different layers in image segmentation [26], [27], Jie et al. [22], [28] used feature fusion to realize face recognition and face detection. In the task of pedestrian detection in tunnels, the feature mapping from different convolution layers can be fused to improve pedestrian detection performance. The feature fusion here uses the feature mapping concatenation, and the specific operation mode will be discussed in the following section. Pedestrians in tunnel pedestrian data sets are blurred and easy to overlap with the background. Shallow feature mapping contains some local information, which can help pedestrians to locate accurately. Only using depth feature can make the location of pedestrians that are seriously occluded from detection or detection inaccurate. The specific implementation process is as follows: Mapping candidate regions to feature mapping generated by Conv5_3 layer and Conv4_3 layer of feature extraction network, obtaining feature maps of candidate regions on these two layers. Fixed size feature vectors are obtained by ROI Pooling and L2 regularization, and then sent to the subsequent full connection layer to realize pedestrian detection and range box regression.

As shown in Figure 5, if L2 regularization is removed during the experiment, the network will be over fitted. The feature fusion method we use is splicing, which stacks the input feature map in the specified dimension. For example: input two groups of data with the size of $(n, C, h, w)$, and output data is $(n, 2C, h, w)$. Where $n$ represents the number of pictures, $C$ represents the number of channels, $h$ and $W$ represent the height and width of the feature map or picture respectively. It is worth noting that feature fusion will stack the feature maps output from different convolutions into the full connection layer, so the amount of partial operations connected with the full connection layer will increase. The ROI pooling layer pools the output characteristic map of convolution layer, and the output size is $7 \times 7$. It can be seen from Figure 5 that the ROI pooling 4 and 5 output $512 \times 7 \times 7$ characteristic images respectively. After feature fusion, $1024 \times 7 \times 7$ characteristic images are sent to the full connection layer, which contains 4096 neurons. Therefore, the full connection layer after feature fusion contains about...
Feature fusion structure for different levels.

20 million parameters. In the original fast RCNN, the candidate region is mapped to the last feature map of the feature extraction network, only using the features of the deepest conv5_3 of the network, while the details of the shallow layer are not obvious. If RCNN uses the feature fusion method to map the candidate region to the feature map generated by the conv5_3 and conv4_3 layers of the feature extraction network. We not only use the semantic features of the deep network, but also use the detailed features of the shallow network to achieve better pedestrian detection performance.

**D. DETECTION NETWORK AND TRAINING**
After the candidate regions are obtained by RPN, the region is acted on the feature map. As shown in part D of Figure 1. The detection network is used to pool the region of interest (ROI) and extract the corresponding regional features. The pedestrian classification and boundary prediction are carried out by using the border regression network and the border classification network. The detection network has two parallel output layers. The output of the classification layer is the probability distribution \( p = (p_0, p_1) \) of each border on pedestrian and non-pedestrian categories. The border regression network outputs border position parameters, \( t^i = (t^i_x, t^i_y, t^i_w, t^i_h) \), \( k \) represents category. Border Regression Network and Border Classification Network are trained by Joint Loss Function, that is:

\[
L(p, u, t^i, v) = L_{cls}(p, u) + \lambda [u \geq 1] \cdot L_{reg}(t^i, v) \quad (1)
\]

where \( L_{cls}(p, u) = -\log(p_u) \) is the logarithmic loss of real \( u \). \( L_{reg} \) is activated only when \( p^* = 1 \) occurs when the area to be detected is pedestrian. In order to get a precise rectangular box, two sets of parameters are defined: the real boundary \( v = (v_x, v_y, v_w, v_h) \) of category \( u \) and the predictive boundary \( t^u = (t_x, t_y, t_w, t_h) \) of category \( u \). The detailed procedure is as follows:

\[
t_x = \frac{x - x_a}{w_a}, \quad t_y = \frac{y - y_a}{w_a}, \quad t_w = \log \frac{w}{w_a}, \quad t_h = \log \frac{h}{h_a} \quad (2)
\]

In the formula \((x, y, w, h)\) is the center coordinate of the real target pedestrian and the width and height of the border, \((x_a, y_a, w_a, h_a)\) is the center coordinate of the candidate area and the width and height of the area. Border Regression Layer Loss is:

\[
L_{reg}(t^i, v) = \sum \text{smooth}_{L1}(t^i_y - v_i),
\]

\[
\text{smooth}_{L1}(x) = \begin{cases} 
0.5x^2, & |x| < 1 \\
|x| - 0.5, & \text{otherwise}
\end{cases}
\quad (3)
\]

**IV. EXPERIMENTAL RESULTS AND ANALYSIS**

**A. DATA SET**
The pedestrian tunnel data set in this paper comes from a pedestrian tunnel surveillance video. The whole data set contains 3210 pictures with the size of 1280 \( \times \) 720. 1605 pictures are selected as training set and 1605 pictures as test set. The number of pedestrians varies from 1 to 20 in each picture, including various scales.

**B. MODEL TRAINING**
In the part of network implementation, TensorFlow, the mainstream framework of deep learning, is chosen as the experimental platform. According to the current standard strategy of target detection based on depth learning [8]–[10], [28], the pre-trained model on ImageNet classification [29] task is selected to initialize the training network. The VGG16 convolution neural network obtained from the pre-training of ImageNet classification is used to initialize the weight of feature extraction network convolution layer. The whole network training process uses Stochastic Gradient Descent (SGD) back propagation to optimize the whole network model. The learning rate is 0.001, momentum is 0.9, weight_decay is 0.0001. The learning rate is attenuated once every 10,000 iterations, and the attenuation factor is 0.1. A total of 80,000 iterations are carried out. The equipment used in the experiment is Python 3.6, GeForce GTX 1060.

**C. TEST RESULTS AND ANALYSIS OF TUNNEL INSIDER DATA SET**
The IF-RCNN model can achieve an average detection accuracy of 90% on the pedestrian detection data set in tunnels. The training curve is shown in Figure 6, and the Precision Rate (P-R) curve of the network is shown in Figure 7. As can be seen from Figure 6, the model in this paper has converged after 80,000 iterations. The algorithm achieves an average detection accuracy of 90% in the tunnel pedestrian test data set. Through analysis, the IF-RCNN algorithm can obtain 90% detection accuracy in the tunnel data set. The main reasons are:

1. A large number of training data sets, training samples contain more than 30,000 pictures, and the whole training sample contains about 50,000 pedestrians with markers.
2. Faster RCNN generates pedestrian candidate areas by PRN, and then classifies the following areas by.
subsequent networks. Based on the excellent structure of area detection.

3. According to different scales of pedestrian images in tunnels, this paper proposes an improved RPN network structure to generate candidate areas for pedestrians of different scales.

4. Aiming at the problem of pedestrians blurring in images, the feature fusion method is used to classify and regress the range frame of the candidate areas. Finally, the improved algorithm achieves the above good detection results.

In order to prove the robustness of the model, we randomly find two images of tunnel pedestrians and use IF-RCNN and Faster RCNN to detect them. The experimental results show that the detection effect of IF-RCNN and Faster RCNN in Figure 8. The results of Faster RCNN detection show that there is a misjudgment, IF-RCNN can accurately detect pedestrians in the graph. It shows that IF-RCNN algorithm has better robustness.

**D. EFFECTIVENESS ON PUBLIC TEST SET VOC 2007**

In order to verify the effectiveness of the improved algorithm in this paper, it is further validated on the public measurement data set VOC 2007. The VOC 2007 data set includes 9963 pictures, 20 categories, and 24640 targets. The size of the VOC 2007 data set is basically 375 × 500 pictures in natural scenes. The parameter setting and experimental platform of the model training are the same as those of the previous section. The original Faster RCNN structure is reproduced in this paper, and the improved RPN structure and feature fusion described above are realized respectively. Some examples of detection effects are shown in Figure 9. The Mean Average Precision (MAP) of different models on VOC 2007 are shown in Table 2.

Observing the data in columns 2 and 3 of Table 2, the average detection performance of the improved RPN network is only 0.7% higher than that of the original algorithm on VOC 2007 data set, and the performance improvement is not obvious. It is noteworthy that the detection performance of bottles has been reduced by 5%, while the performance of chairs has been improved by 15%.
TABLE 2. Map comparison of different models.

| Model       | Faster RCNN | Faster RCNN+Improved RPN | Faster RCNN+Feature fusion | Faster RCNN+Improved RPN +Feature fusion (IF-RCNN) |
|-------------|-------------|--------------------------|---------------------------|--------------------------------------------------|
| Airplane    | 66.17       | 68.83                    | 69.43                     | 70.88                                            |
| Bicycle     | 78.13       | 78.41                    | 79.86                     | 79.69                                            |
| Bird        | 67.83       | 68.64                    | 69.97                     | 69.02                                            |
| Ship        | 56.25       | 56.75                    | 58.37                     | 57.43                                            |
| Bottle      | 50.54       | 45.71                    | 59.79                     | 55.38                                            |
| Bus         | 75.94       | 78.15                    | 76.63                     | 79.51                                            |
| Car         | 79.65       | 79.90                    | 80.04                     | 81.27                                            |
| Cat         | 86.96       | 83.74                    | 87.74                     | 86.30                                            |
| Chair       | 49.30       | 64.14                    | 52.07                     | 66.72                                            |
| Cow         | 75.77       | 74.16                    | 77.49                     | 77.07                                            |
| Table       | 64.36       | 64.16                    | 65.32                     | 66.74                                            |
| Dog         | 81.05       | 79.26                    | 82.57                     | 81.37                                            |
| Horse       | 80.18       | 80.58                    | 81.09                     | 82.26                                            |
| Bicycle     | 72.59       | 74.80                    | 76.55                     | 74.96                                            |
| **Person**  | **76.84**   | **77.85**                | **78.68**                 | **79.76**                                        |
| Botany      | 41.76       | 41.43                    | 41.63                     | 41.85                                            |
| Sheep       | 67.14       | 67.98                    | 69.25                     | 70.02                                            |
| Sofa        | 64.83       | 64.49                    | 65.03                     | 65.96                                            |
| Train       | 74.85       | 74.86                    | 76.98                     | 76.71                                            |
| Television  | 71.31       | 71.68                    | 73.65                     | 73.03                                            |
| Map         | 69.10       | 69.77                    | 71.16                     | 71.80                                            |

The target performance of other categories has basically remained unchanged or improved slightly. In view of this phenomenon, the data of bottle and chair in VOC 2007 data set are analyzed. The VOC 2007 data set contains 502 pictures of bottles, 244 of which are used for training, and a large number of smaller bottles are included in the labeled data. The data of VOC 2007 chair category contains 1117 pictures, of which 445 pictures contain 798 targets for training, and the size is generally larger and usually has different scales of occlusion. Because there are people sitting on the chair or other items in the data center. After analysis, it is concluded that “improved RPN structure” uses three convolution cores of different scales to locate targets.

The convolution core of $5 \times 5$ is too large in the original image and the size of the bottle is small. It can be divided into small target detection, and too large field will damage the detection performance of small targets; the chair of data set is larger and larger. Five-five convolution kernels in the improved RPN structure can help locate larger targets. The convolution kernels of three scales are more robust to targets of different scales. Therefore, the performance of the improved RPN chair detection algorithm is greatly improved compared with the original algorithm. In the fourth column of Table 2, feature fusion utilizes the shallow features and deep high-level semantic features of the image. Shallow features can use some detail information to help target localization, and improve the detection performance of each category on VOC 2007 data set in different ranges. Column 5 of Table 2 combines the “improved” and feature fusion results on VOC 2007 data sets.

This paper focuses on pedestrian detection in pedestrian tunnels, so it pays special attention to the results of IF-RCNN algorithm on VOC 2007 human target data. We plot the P-R curve of the improved algorithm on VOC 2007, as shown in Figure 10. At the same time, VOC 2007 human data is different from the data set in the tunnel. They are all full high definition color images, covering all kinds of gestures, including large-scale faces, pedestrians and so on, while the data in the tunnel is basically only pedestrians.

There are 4192 pictures of person in the whole VOC 2007 data set, of which 2096 are for training and 2096 are for testing. From Table 2 and Figure 10, we can see that the IF-RCNN algorithm is effective. In natural scenes, the IF-RCNN algorithm improves the performance of the original Faster RCNN by about 3%, which is better than the original algorithm. It also proves our algorithm is applicable to pedestrian detection in natural scenes.

At the same time, we compare our algorithm with Faster RCNN, SSD and Faster RCNN+Similarity measurement algorithm on the part of VOC2007 data set containing
human targets. From Table 3, it can be seen that the MAP of the proposed algorithm on the test set of VOC2007 data set is 2.92% higher than Faster RCNN algorithm, 8.79% higher than SSD algorithm and 0.56% higher than [19] algorithm. In terms of detection rate, the IF-RCNN proposed in this paper is 5.6 higher than the FPS (Frames Per Second) of the original Faster algorithm.

The main reasons for analyzing the results in Table 3 are as follows: SSD does not use the low-level characteristics of high resolution, which is very important for detecting small pedestrians. Faster RCNN+Similarity measurement method takes candidate regions with high classification score as feature templates, and candidate target positions with low confidence score are further analyzed and judged by comparing with the similarity of the model. Finally, the detection results are output. Although this method can reduce the error rate, it does not take advantage of the characteristics of the deep network, making it really difficult to detect some details. Our method uses different sizes of receptive fields to obtain better candidate regions for different scales of targets, and also uses the deep features of the network to obtain better detail features, and the detection effect is also better.

At the same time, we also compare our algorithm with other algorithms in common target detection data set coco. The experimental results are shown in Table 4. The results show that the map of IF-RCNN in coco data can reach 38.8%, which is 3.9% higher than the original fast RCNN and 0.6% higher than the mask RCNN. In terms of speed, it is 2.5fps faster than the original fast RCNN and 5.8fps faster than the mask RCNN.

**V. CONCLUSION**

This paper proposes an IF-RCNN based algorithm for Pedestrian detection in tunnels. It can better deal with the problem of pedestrian blur and multi-scale existing in the pedestrian image of tunnel. First of all, the size of anchor is adjusted for the small size of pedestrians in the tunnel. Secondly, an improved RPN network structure is proposed. Finally, the bottom features and high-level semantic features are fused to achieve the goal classification by using feature fusion technology. The results show that If-RCNN algorithm in tunnel pedestrian data set, the detection accuracy reaches 90%. In VOC 2007 pedestrian data, compared with the original faster RCNN, the MAP is improved by 2.92%, 0.56% higher than fast RCNN+similarity measurement [19], and 8.79% higher than SSD algorithm.

**REFERENCES**

[1] R. Benenson, M. Omran, J. Hosang, and B. Schiele, “Ten years of pedestrian detection, what have we learned,” in Proc. Eur. Conf. Comput. Vis. Zürich, Switzerland: CVRSUAD Workshop, 2014, pp. 613–627.

[2] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2005, pp. 886–893.

[3] T. Mita, T. Aneko, and O. Hori, “Joint Haar-like features for face detection,” in Proc. 10th IEEE Int. Conf. Comput. Vis., Oct. 2005, pp. 1619–1626.

[4] T. Ahonen, A. Hadid, and M. Pietikäinen, “Face recognition with local binary patterns,” in Proc. Eur. Conf. Comput. Vis. Berlin, Germany: Springer, 2004, pp. 469–481.

[5] P. Dollár, Z. Tu, P. Perona, and S. Belongie, “Integral channel features,” in Proc. Brit. Mach. Vis. Conf., London, U.K., Sep 2009, pp. 1–11.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in Proc. Int. Conf. Neural Inf. Process. Syst., 2012, pp. 1097–1105.

[7] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, “Backpropagation applied to handwritten zip code recognition,” Neural Comput., vol. 1, no. 4, pp. 541–551, Dec. 1989.

[8] Y. Li, Y. Li, K. He, and J. Sun, “R-FCN: Object detection via region-based fully convolutional networks,” in Proc. Adv. Neural Inf. Process. Syst., 2016, pp. 379–387.

[9] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” in Proc. Adv. Neural Inf. Process. Syst. Cambridge, MA, USA: MIT Press, 2015, pp. 91–99.

[10] L. Zhang, L. Liu, X. Liang, and K. He, “Is faster R-CNN doing well for pedestrian detection,” in Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer, 2016, pp. 443–457.

[11] 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.

[12] P. Dollár, 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. 746–761, Apr. 2012.
[13] P. Dollár, R. Appel, S. Belongie, and P. Perona, “Fast feature pyramids for object detection,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 36, no. 8, pp. 1532–1545, Aug. 2014.

[14] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Columbus, OH, USA, Sep. 2014, pp. 580–587.

[15] H. S. Song et al., “Vehicle target detection in complex scenes based on in-depth learning method,” (in Chinese), Appl. Res. Comput., vol. 35, no. 4, pp. 30–31, 2018.

[16] X. Sun, P. Wu, and S. C. H. Hoi, “Face detection using deep learning: An improved faster RCNN approach,” Neurocomputing, vol. 299, pp. 42–50, Jul. 2018.

[17] 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., Jun. 2016, pp. 779–788.

[18] 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. Cham, Switzerland: Springer, 2016, pp. 21–37.

[19] Z. M. Li, M. M. Xing, Y. J. Liu, and H. Li, “Pedestrian object detection based on faster R-CNN and similarity measurement,” J. Graph., vol. 39, no. 5, pp. 901–908, 2018.

[20] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), New York, NY, USA, Jul. 2017, pp. 2117–2125.

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

[22] X. Sun, P. Wu, and S. C. H. Hoi, “Face detection using deep learning: An improved faster RCNN approach,” Jan. 2017, arXiv:1701.08289.

[23] 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.

[24] F. Yu and V. Koltun, “Multi-scale context aggregation by dilated convolutions,” 2015, arXiv:1511.07122. [Online]. Available: http://arxiv.org/abs/1511.07122

[25] K. He, X. Zhang, S. Ren, and J. Sun, “Identity mappings in deep residual networks,” in Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer, 2016, pp. 630–645.

[26] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Boston, MA, USA, Jun. 2015, pp. 3431–3440.

[27] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, “Pyramid scene parsing network,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Honolulu, HI, USA, Jul. 2017, pp. 2881–2890.

[28] J. J. Wang, Z. Y. Liu, and H. Li, “Significant target detection based on feature fusion and objectness enhancement,” (in Chinese), J. Comput. Eng. Appl., vol. 53, no. 2, pp. 195–200, 2017.

[29] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, “ImageNet large scale visual recognition challenge,” Int. J. Comput. Vis., vol. 115, no. 3, pp. 211–252, Dec. 2015.

JIN REN received the Ph.D. degree in information and communications engineering from the School of Electronic and Information Engineering, Dankook University, South Korea, in 2011. He is currently an Associate Professor with the School of Information Science and Technology, North China University of Technology. His current research interests include intelligent information processing, wireless communication, and the Internet of Things (IoT) technology.

CHANGLIU NIU was born in 1967. He received the Ph.D. degree in field and wave electromagnetics from the Beijing University of Posts and Telecommunications, in 2006. He is currently a Communication Network Switching and Wireless Associate Professor with the School of Information Science and Technology, North China University of Technology. His research interests include, but not limited to, high-speed optical communication, communication network switching, and wireless sensor networks. The main courses he taught include field and wave electromagnetics, the principle of modern communication, and theoretical basis of communication networks.

JUN HAN was born in Qinghai, China, in 1994. He is currently the Graduate Student with the School of Information Science and Technology, North China University of Technology, and majoring in electronic and communication engineering. His research interests include artificial intelligence and image processing.