IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 16, 2020, accepted June 1, 2020, date of publication June 8, 2020, date of current version June 17, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3000506

An Improved Faster R-CNN for High-Speed Railway Dropper Detection

QIFAN GUO[1,2], LEI LIU[1,2], WENJUAN XU[1,2], YANSHENG GONG[3], XUEWU ZHANG[3], AND WENFENG JING[1,2]

[1]School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an 710049, China
[2]National Engineering Laboratory for Big Data Analytics, Xi'an Jiaotong University, Xi'an 710049, China
[3]China Railway First Survey and Design Institute Group Company, Ltd., Xi'an 710043, China

Corresponding author: Wenfeng Jing (wfjing@mail.xjtu.edu.cn)

This work was supported in part by the 2018 Major Science and Technology Special Project of China Railway Construction Corporation under Grant 18-A02, and in part by the Xi'an Scientific and Technological Innovation Project under Grant 201809164CX5JC6.

ABSTRACT Overhead contact systems (OCSs) are the power supply facility of high-speed trains and plays a vital role in the operation of high-speed trains. The dropper is an important guarantee for the suspension system of the OCS. Faults of the dropper, such as slack and breakage, can cause a certain threat to the power supply system. How to use artificial intelligence technologies to detect faults is an urgent technical problem to be solved. Because droppers are very small in whole images, a feasible solution to the problem is to identify and locate the droppers first, then segment them, and then identify the fault type of the segmented droppers. This paper proposes an improved Faster R-CNN algorithm that can accurately identify and locate droppers. The innovations of the method consist of two parts. First, a balanced attention feature pyramid network (BA-FPN) is used to predict the detection anchor. Based on the attention mechanism, BA-FPN performs feature fusion on feature maps of different levels of the feature pyramid network to balance the original features of each layer. After that, a center-point rectangle loss (CR Loss) is designed as the bounding box regression loss function of Faster R-CNN. Through a center-point rectangle penalty term, the anchor box quickly moves closer to the ground-truth box during the training process. We validate the improved Faster R-CNN through extensive experiments on the VOC 2012 and MSCOCO 2014 datasets. Experimental results prove the effectiveness of the proposed network combined with attention feature fusion and center-point rectangle loss. On the OCS dataset, the accuracy using the combination of the improved Faster R-CNN and ResNet-101 reached 86.8% mAP@0.5 and 83.9% mAP@0.7, which was the best performance among all results.

INDEX TERMS Dropper detection, feature fusion, improved Faster R-CNN, attention mechanism.

I. INTRODUCTION

In recent years, high-speed railway transport has developed rapidly worldwide. The overhead contact system (OCS) is the key equipment for powering electric locomotives. The continuous operation of the OCS ensures the high-speed running of the train. The dropper is one of the important components in the chain suspension of the OCS, and the carrier cable is suspended on the OCS through the dropper. Due to the open-air work all year round, the dropper is prone to breakdown. Once the dropper is loose or dropped, it will have a great impact on the power supply system of the high-speed railway, threatening the normal operation of trains and the safety of passengers. At present, the railway system still relies on manually viewing video images acquired through the 2C system to find dropper faults. Because of the influence of various human factors, omissions or misjudgments can easily occur. Image processing is a method for replacing manpower for fault diagnosis of droppers, the first step of which is to use an efficient detector to detect and locate the dropper in the high-definition image. With the development of artificial intelligence, it is an urgent problem to realize the dropper detection method based on deep learning.

The associate editor coordinating the review of this manuscript and approving it for publication was Vivek Kumar Sehgal.

Convolutional neural networks can learn the robustness and deep feature representation of an image and have good performance in computer vision. From LeNet [1], AlexNet [2] won the ImageNet [3] competition in 2012, and then to VGGNet [4] and ResNet [5], CNN has become deeper for better performance. With the development of CNNs, more powerful object detection algorithms have appeared one after another, such as the YOLO series [6]–[8] networks and Faster R-CNN [9], which are widely used in the engineering field. It is of great significance to use object detection networks to accurately locate and identify droppers for further research on dropper fault diagnosis. Therefore, the main purpose of this paper is to find a high-performance object detector.

However, the structure of the OCS components is complex and diverse, and the background is extremely complicated, which leads to poor feature representation of the dropper. There are many non-target parts that greatly affect the feature extraction of the dropper, such as wrist arms and wire rods. Therefore, using a deep learning network to achieve accurate dropper identification requires a more efficient object detection framework. With the introduction of Faster R-CNN [9], the accuracy of detection has been greatly improved. Faster R-CNN is widely used in some computer vision tasks in the engineering field and can solve the detection problem of small objects with different sizes. Due to the abundance of semantic information, the deep layer in feature extraction networks plays an important role in the classification stage, while the lower layer with more detailed information and content description is easy to ignore. Thus, the feature fusion of FPN [10] is of great significance to the performance improvement of object detection tasks. For example, the proposal of PANet [11] enables the feature pyramid to be enhanced through a bottom-up path, which can obtain more accurate positioning information from low-level features. In addition, the attention mechanism focuses information on key parts of the image and shows good performance in image classification and object detection tasks.

In this paper, to address the problem of dropper detection, we propose an improved Faster R-CNN with two innovative views. The first innovation is that a balanced attention feature pyramid network (BA-FPN) is proposed to obtain the fusion feature of multilevel feature maps. Specifically, by relying on an integrated semantic feature map to balance the original features of each layer of the pyramid, each resolution in the feature pyramid can obtain equal information from the other layers. The image information imbalance problem of FPN [10] can be solved by better fusion of shallow detailed information and deep semantic information. In addition, based on the attention mechanism, a new network module named the ''mixed attention block'' is designed to act on the integrated semantic feature map. By acquiring the channel and spatialwise attention, the mixed attention block reduces the information redundancy and extracts more useful image features. The second innovation is the proposal of a center-point rectangle loss (CR loss) to accelerate convergence and improve the accuracy of the model. In CR loss, we add a center-point rectangle penalty term to the coordinate regression loss function. The vertices of the center-point rectangle consist of the center points of the ground-truth box and the anchor box. By optimizing the area of the rectangle, the center distance between the anchor box and the ground-truth box is directly minimized, which provides a moving direction for the bounding box and accelerates convergence. In summary, the contributions of this paper are as follows:

1) We propose BA-FPN, a feature pyramid model based on an attention mechanism, which can better extract useful features.

2) We propose a center-point rectangle loss function, which uses a center-point rectangle penalty term to accelerate convergence.

3) We use the improved Faster R-CNN as the basic object detection network and validate the proposed method on VOC 2012 [12], MSCOCO 2014 [13] and our OCS datasets. Our method achieves state-of-the-art performance.

The remainder of this paper is organized as follows. Section II shows the recent research on engineering applications of OCSs and the development of detection tasks in the computer vision field. The dropper detection method proposed in this paper is described in Section III. Section IV presents the experimental datasets and parameter settings, and the experimental results are analyzed in detail. The relevant conclusions are given in Section V.

II. RELATED WORKS

A. THE OCS ANALYSIS AND DROPPER DETECTION

The OCS is an important part of the electrified railway system that is responsible for transferring the electric energy in the traction network to the electric locomotive. The specific structure of the OCS is shown in Figure 1. There are complex mechanical and electrical interactions between the pantograph and the catenary device. The vibration and impact generated by the long-term operation of the train will inevitably cause the failure of the catenary support device, such as the disappearance of the fasteners and breakage of the load-bearing cable, which can seriously affect train operation. In recent years, researchers have attempted to use image processing methods to detect the key components of the OCS. Karakose *et al.* [14] proposed a new approach using image processing-based tracking to diagnose faults in the pantograph-catenary system. Liu *et al.* [15] proposed a unified deep learning architecture for the detection of all catenary support components. Qu *et al.* [16] used a genetic optimization method based on an adadelta deep neural network to predict pantograph and catenary comprehensive monitor status. Zhong *et al.* [17] introduced a CNN-based defect inspection method to detect catenary split pins in high-speed railways.

This paper focuses on the dropper detection of the OCS. The dropper is one of the important components in the catenary suspension, which is of great significance to the normal operation of trains. Similar to the detection of other

FIGURE 1. The high-definition image of the OCS.

parts, dropper detection will also be interfered by the noise in the background of the complex OCS images. In addition, the main body of the dropper is filamentous and very small in the image, which creates some difficulties in feature extraction. Several years ago, Petitjean *et al.* [18], [19] introduced an original system for the automatic detection of droppers in the catenary, which used prior knowledge to obtain the location of the dropper. With the advancement of computer vision technology, Xu [20] used a Faster R-CNN to locate dropper images and then used the Hough transform to recognize dropper faults. Liu *et al.* [21] proposed a deep learning method based on depthwise separable convolution for dropper detection. In order to address the impact of image complexity, we propose an attention-based feature fusion method combined with a high-precision Faster R-CNN network to form an effective object detector and realize dropper detection in complex backgrounds.

B. OBJECT DETECTION NETWORK

With the development of CNN, image processing and object detection technology have achieved an improvement from traditional machine learning methods to deep learning. Girshick *et al.* [22] proposed R-CNN based on region proposal, which makes two-stage object detection a mainstream detection method. He *et al.* [23] used SPPNet to effectively solve the problem of computational redundancy of candidate regions. On the basis of R-CNN [22] and SPPNet [23], Fast R-CNN [24] realized a multitask learning method by simultaneously training object classification and bounding box regression. Immediately afterward, Ren *et al.* [9] proposed a region proposal network in Faster R-CNN to fuse the region proposal with CNN classification and realized a complete end-to-end CNN object detection model. After that, Cascade R-CNN [25] expanded Faster R-CNN [9] into a multistage detector through a powerful cascade structure. Lin *et al.* [10] proposed a feature pyramid network (FPN), which caused multiple detection ports from different levels in the network to detect objects of different scales. FPN [10] has now become a basic component in many detectors. In the path aggregation network proposed by Liu *et al.* [11], a bottom-up path augmentation structure was introduced to fuse FPN features and make full use of the features of the shallow layer.

A one-stage detection model can obtain the final detection result directly after a single detection and has a fast detection speed. YOLO [6] was the first proposed one-stage detection algorithm, which directly obtained the position of the bounding box and the classes of the object through only one convolutional neural network. Liu *et al.* [26] proposed the SSD algorithm, which absorbed the advantages of YOLO's fast speed and the precise positioning of RPN [9]. SSD [26] adopted multiwindow technology in RPN and detected multiple feature maps with different resolutions. To improve the detection accuracy of the one-stage method, Lin *et al.* [27] proposed "focal loss" to modify the traditional cross-entropy loss function and greatly improved the detection precision. The high-precision detectors of many algorithms rely on dense anchor strategies, resulting in a large number of redundant anchor boxes and a serious imbalance between positive and negative samples. To solve this problem, Wang *et al.* [28] proposed GA-RPN, which predicted the position and shape of the anchor to generate sparse and arbitrarily shaped anchors.

At present, object detection technology based on deep learning is also gradually used in various fields. Chen *et al.* [29] applied an attention mechanism to ship detection in satellite images. Cao *et al.* [30] designed an improved Faster R-CNN for small object detection. In the field of railway engineering, Wei *et al.* [31] used Faster-R-CNN to detect railway track fasteners. Juan *et al.* [32] proposed FB-NET detection based on a deep learning method for detecting the shape of railways and dangerous obstacles. In addition, He *et al.* [33] combined SSD and Faster-R-CNN to detect foreign matter in high-speed trains.

C. ATTENTION MECHANISM

The attention mechanism essentially imitates the way that humans observe objects. In recent years, most of the research work on the combination of deep learning and visual attention mechanisms has focused on the use of masks. By giving weight to the network layer to identify the key features of the image, an attention mechanism is formed. Wang *et al.* [34] introduced a residual attention network using a trunk-and-mask attention mechanism model. The trunk branch is similar to the traditional convolutional network, and features are extracted through multiple convolution operations. The mask branch is an encoder-decoder model with the output attention weight. Fu *et al.* [35] proposed RA-CNN, which combines area determination with fine-grained feature extraction. The region with a dense distribution of important features can be used as a key recognition region for further accurate judgment to promote feature extraction. Hu *et al.* [36] designed a squeeze-and-excitation block to explore the relationship between channels, which calculates the attention weight of each channel through a global pooling operation. Woo *et al.* [37] proposed the convolutional block attention module. In addition to considering the attention weight of

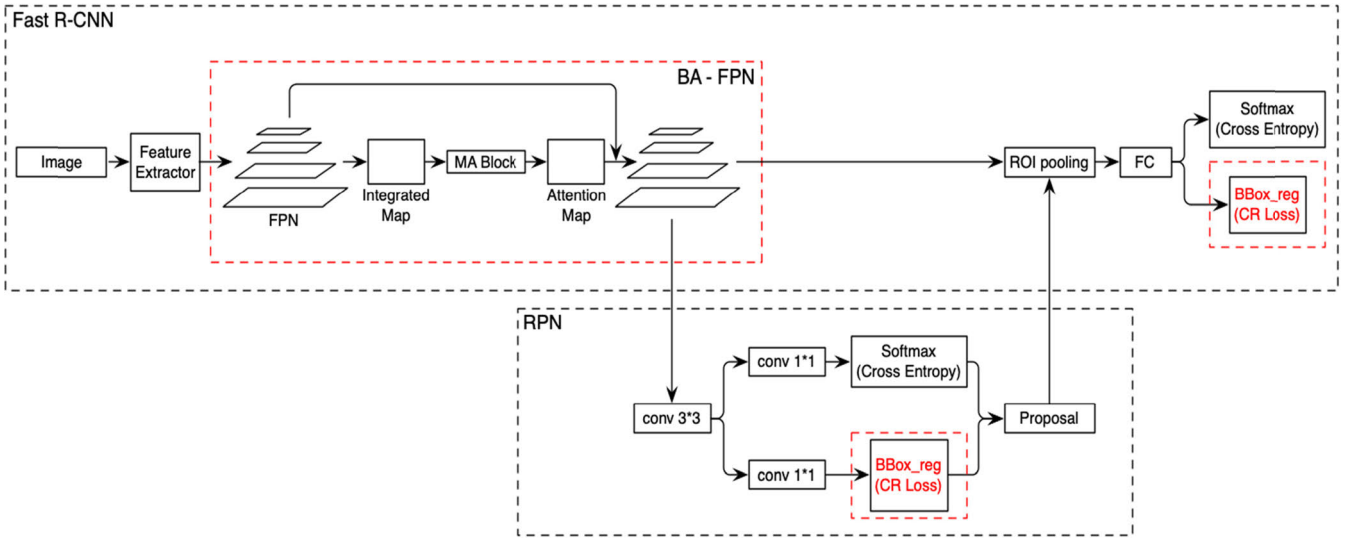


FIGURE 2. The overall algorithm framework of this paper.

the channels, a spatial attention branch was also added in the module.

In different visual tasks, the attention mechanism has also been applied accordingly. Ling *et al.* [38] proposed a self-residual attention network for deep face recognition. In the image translation task, a channel attention network was designed by Sun *et al.* [39], with which the original function in the encoder and the conversion function in the decoder can be better integrated. In addition, Liu *et al.* [40] proposed a spatiotemporal attention module for video action recognition. Gao *et al.* [41] introduced a residual attention mechanism to one convolutional layer object tracking network to avoid data imbalance.

III. OUR PROPOSED METHODS

To improve the performance of dropper detection, we develop an improved Faster R-CNN network. The architecture of the improved Faster R-CNN is shown in Figure 2. The proposed method contains two aspects: a balanced attention feature pyramid network (BA-FPN) and a center-point rectangle loss (CR loss).

The BA-FPN model balances the original feature of each layer by relying on an integrated semantic feature map. First, the feature maps of different levels of the feature pyramid are fused into an integrated semantic feature map. Then, we use the mixed attention block to extract the channel and spatial attention of the integrated feature map, which in turn acts on the integrated semantic feature map to generate an attention map. We combine the attention map with feature maps of the pyramid to balance the original feature. CR loss is an optimized bounding box regression loss function. Based on the regression of the prediction box vertex, we add a rectangular area penalty term to the function. The two diagonal vertices of the rectangle are composed of the center points of the predicted anchor box and the ground-truth box. By optimizing the rectangle penalty term, the convergence of loss is accelerated, and the accuracy is improved. In Section A, we introduce the feature extractor used in the proposed method. In Section B, we review the structure of the FPN and introduce the BA-FPN model in detail. In Section C, the proposed CR loss function is stated. Section D describes the generation process of the predicted bounding box.

A. FEATURE EXTRACTOR

It is important to select a high-performance convolutional neural network for the performance of the detection model. The depth and parameter settings of the feature extraction network directly affect the performance of the proposed method. A deep network can generate a feature map with rich semantic information, which is useful for achieving better feature pyramid fusion.

In this paper, we choose ResNet as the basic feature extractor of the proposed method. Instead of attempting to learn the mapping between the input and output directly as in VGGNet, ResNet can learn the representation of the input residual and output by using multiple residual blocks. The residual block is shown in Figure 3. It is much easier to learn residuals than to directly learn the mapping between the input and output, which is proven by a large number of experiments.

In the experiment, we used the models trained on ImageNet [3] as the basic pretrained parameter models of ResNet.

B. BALANCED ATTENTION FPN

There are objects of different sizes in the image, and different objects have different characteristics. Simple objects can be distinguished by shallow features, while complex objects can be distinguished by deep features. The emergence of the FPN can solve the above problem to some extent. FPN is a kind of enhancement of the image information expression

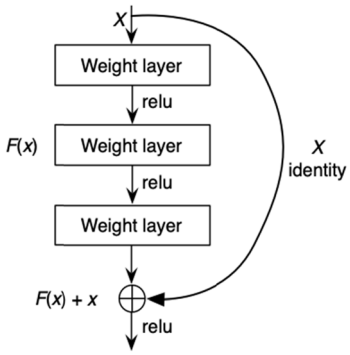


FIGURE 3. The residual block of ResNet.

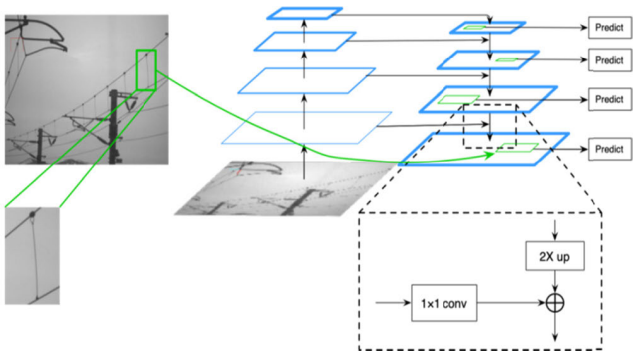


FIGURE 4. The FPN framework.

output of traditional CNN networks, which can be flexibly applied to different tasks. Figure 4 demonstrates the overall architecture of the FPN. First, FPN can efficiently calculate strong features through the hierarchical structure of the CNN network. By combining bottom-up and top-down methods, FPN obtains strong semantic features to improve the performance of object detection and semantic segmentation on multiple datasets. For small objects, FPN can utilize the high-level semantic information after the top-down model, which increases the resolution of the feature map and operates on a larger feature map to obtain more useful information of small objects.

However, in FPN, the semantic information contained in nonadjacent layers will be diluted in the information fusion process, resulting in information fusion imbalances of different scales. On the basis of FPN, BA-FPN fuses the feature maps of each level into an integrated semantic feature map, which in turn acts on the maps of the corresponding scales to balance the differences between the levels and enhance useful feature expression. The general framework of BA-FPN is shown in Figure 5.

Assuming the number of layers in the feature pyramid is L, the outputs of Conv2, Conv3, Conv4 and Conv5 are adopted here, denoted as $\{C_2, C_3, C_4, C_5\}$. To integrate features of different levels and retain their semantic information, the features of different levels $\{C_2, C_3, C_4, C_5\}$ were first reconstructed to the size of C_4 through interpolation or max-pooling, and then $\{F_2, F_3, F_4, F_5\}$ was obtained. After that, by calculating the mean value of $\{F_2, F_3, F_4, F_5\}$, the integrated semantic feature map F_b was obtained. The formula is defined as

$$F_b = \frac{1}{L} \sum_{l=l_{min}}^{l_{max}} F_l \tag{1}$$

To reduce the information redundancy of balanced semantic features and further enhance useful feature expression, we design a mixed attention block (MA block) based on an attention mechanism, including a channel attention branch and a spatial attention branch. The structure of the MA block is shown in Figure 6. The feature representation of the balanced semantic feature can be enhanced effectively by extracting the channel and spatialwise attention. Thus, the output of the MA block focuses on the most significant components of the information.

We took the integrated semantic feature map F_b as the input of the MA block, where $F_b \in R^{C\times H\times W}$. By calculating the channel attention branch and the spatial attention branch simultaneously, the corresponding attention maps were generated. In the channel attention branch, we aggregated the spatial information of F_b through an average-pooling operation to generate the spatial context descriptor: $F^c_{avg} \in R^{C\times 1\times 1}$, which generates a channel attention map $M_c \in R^{C\times 1\times 1}$ through a multilayer perceptron (MLP). The hidden layer size of the MLP was set to $R^{C/r\times 1\times 1}$, and r is the reduction ratio. Additionally, in the spatial attention branch, channel information is aggregated by averaging-pooling operation on the channel axis to generate a feature descriptor: $F^s_{avg} \in R^{1\times H\times W}$. Then, a convolutional layer was applied to F^s_{avg} to produce a spatial attention map $M_s \in R^{1\times H\times W}$. The overall attention process can be summarized as

$$\begin{aligned} M_c &= \sigma\left(MLP\left(AvgPool1\left(F_b\right)\right)\right) \\ &= \sigma\left(W_1\left(W_0\left(F^c_{avg}\right)\right)\right) \end{aligned} \tag{2}$$

$$\begin{aligned} M_s &= \sigma(f^{7\times 7}(AvgPool2(F_b))) \\ &= \sigma\left(f^{7\times 7}\left(F^s_{avg}\right)\right) \end{aligned} \tag{3}$$

where σ denotes the sigmoid function. $W_0 \in R^{C/r\times C}$, and $W_1 \in R^{C\times C/r}$ are the weight parameters of MLP in the channel attention branch. $f^{7\times 7}$ represents that the convolution kernel size of the convolution operation is 7*7 in the spatial attention branch. $AvgPool1$ and $AvgPool2$ are the channel and spatialwise global averaging-pooling, respectively.

After the above operation, we obtain the attention maps M_c and M_s acting on F_b. At the end of the MA block, the final refined attention feature map A is obtained.

$$A = (1 + M_c \otimes M_s) \otimes F_b \tag{4}$$

where $\otimes$ denotes elementwise multiplication. Considering that $M_c \otimes M_s$ belongs to [0, 1], if multiplied directly by F_b, it will lead to a weakened output response of the feature map.

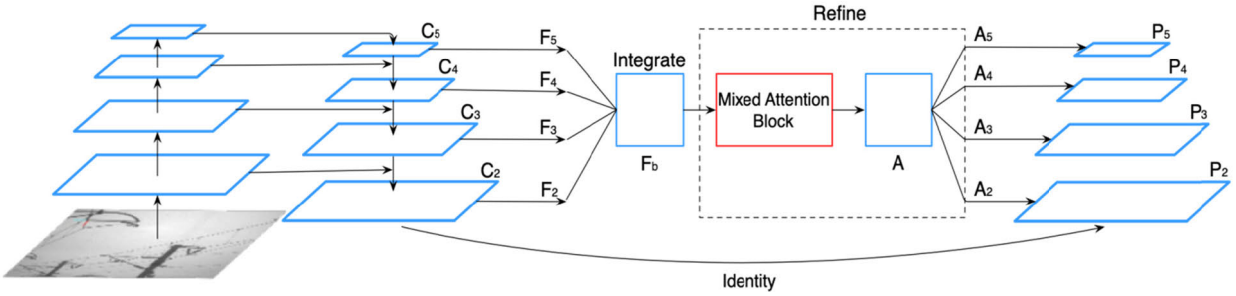


FIGURE 5. The framework of the balanced attention feature pyramid network.

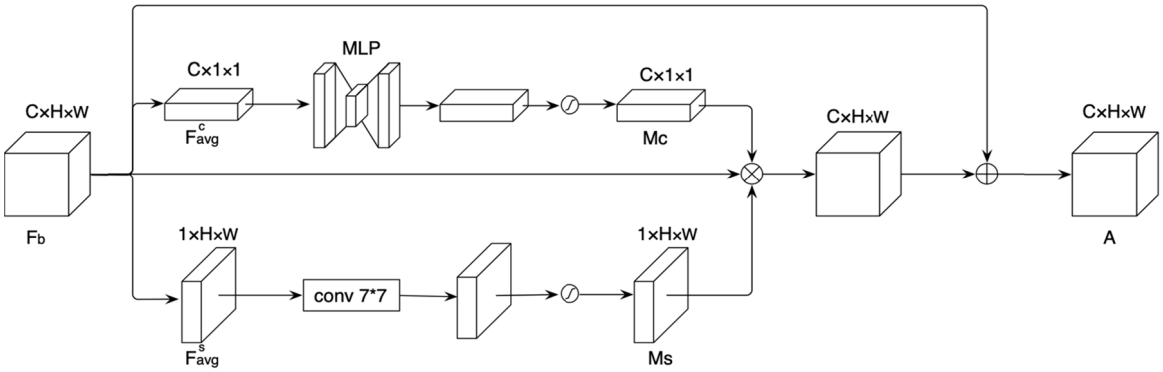


FIGURE 6. The structure of the mixed attention block.

Therefore, using $1 + M_c \otimes M_s$ can avoid the emergence of this problem.

To feed back the balanced semantic feature information to each level, the output A of the MA block is reconstructed to the same size corresponding to each level of $\{C_2, C_3, C_4, C_5\}$, and $\{A_2, A_3, A_4, A_5\}$ was obtained, which are then added with $\{C_2, C_3, C_4, C_5\}$ to obtain $\{P_2, P_3, P_4, P_5\}$. The process is expressed as follows:

$$P_i = A_i + C_i, \quad i = 2, 3, 4, 5 \tag{5}$$

Compared with $\{C_2, C_3, C_4, C_5\}$, $\{P_2, P_3, P_4, P_5\}$ balances the differences among the layers and enhances the original feature of each layer. For subsequent object detection, the following process of the model is the same as FPN.

C. CENTER-POINT RECTANGLE LOSS

From L1 loss and L2 loss to the proposal of smoothL1 loss, the optimization of regression loss makes the training process increasingly efficient. When the predicted value differs greatly from the target value, the gradient of L2 loss is (x-t), which is prone to gradient explosion, and the gradient of L1 loss is constant. At present, in the Faster R-CNN object detection network, smoothL1 loss is generally used as the loss function for bounding box regression. When the predicted value differs greatly from the target value, the gradient explosion can be prevented by changing from L2 Loss to L1 loss. The loss function of the original Faster R-CNN is expressed as follows:

$$L = \frac{1}{N_{cls}} \sum_i L_{cls}\left(p_i, p_i^*\right) + \lambda \frac{1}{N_{reg}} \sum_i p_i^* L_{reg}\left(t_i, t_i^*\right) \tag{6}$$

where i is the index of the predicted anchor box, and p_i represents the predicted probability of the i-th anchor box. p_i^* is the value of the i-th ground-truth box. If the anchor is a positive sample, the value of p_i^* is 1; otherwise, it is 0. t_i and t_i^* are the coordinate vectors of the predicted anchor box and ground-truth box, respectively. λ is the coefficient used to balance regression loss and classification loss, which was set to 1 in the experiment. N_{cls} and N_{reg} are the normalized and weighted parameters by λ. L_{reg} denotes the basis regression loss function (smooth L_1 loss).

$$L_{reg}\left(t_i, t_i^*\right) = S_{L1}\left(t_i - t_i^*\right) \tag{7}$$

where

$$S_{L1} = \begin{cases} 0.5x^2 & |x| < 1 \\ |x| - 0.5 & |x| \geq 1 \end{cases} \tag{8}$$

SmoothL1 has excellent performance in the Faster R-CNN network. This paper attempts to optimize the loss function by shortening the spatial distance between the predicted anchor box and the ground-truth box. In the DIoU loss function, Zheng *et al.* [42] rapidly reduced the distance between the predicted anchor box and the ground-truth box by adding a penalty term of center distance to the IOU loss. In this paper, center-point rectangle loss (CR loss) is designed based on

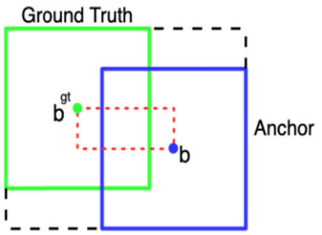


FIGURE 7. The CR loss for bounding box regression, where b_i and b_i^{gt} are the central points of the anchor box and the ground-truth box.

the smoothL1 loss function. We add a center-point rectangle term to L. The vertices of the center-point rectangle consist of the central points of the ground-truth box and the predicted anchor box. By optimizing the rectangular area, the distance between the two center points is directly minimized so that the anchor box quickly moves closer to the ground-truth box. As shown in Figure 7, our goal is to reduce the area of the rectangular box enclosed by the red dotted line. The formula of the CR loss function is defined as follows.

$$L_{CR}\left(t_i, t_i^*\right) = S_{L1}\left(t_i - t_i^*\right) + \frac{R(b_i, b_i^{gt})}{R_i'} \tag{9}$$

where b_i and b_i^{gt} are the center points of the anchor box and the ground-truth box. $R(b_i, b_i^{gt})$ is the center-point rectangle. R_i' represents the smallest rectangular box that can only contain both the anchor box and the ground-truth box. We replace $S_{L1}\left(t_i - t_i^*\right)$ with $L_{CR}\left(t_i, t_i^*\right)$ in the total loss function. In the experiment, the proposed loss function is proven to be effective.

D. DETECTION BOUNDING BOX GENERATION

Multilevel feature maps output by BA-FPN are used as the inputs of RPN, and the structure of RPN is shown in Figure 8. An n∗n sliding window is generated on the shared convolutional feature layer with the maximum number of k anchor boxes. After a 3 * 3 convolution operation, the feature map enters the regression layer and classification layer. Then, the regression layer and classification layer produce 4k and 2k outputs, which represent coordinate values of corresponding candidate regions and the probability of whether the area is the foreground.

The loss functions of the regression layer and classification layer are CR loss and cross-entropy loss, respectively. The total loss function is defined as follows:

$$L' = \frac{1}{N_{cls}} \sum_i L_{cls}\left(p_i, p_i^*\right) + \lambda \frac{1}{N_{reg}} \sum_i p_i^* L_{CR}\left(t_i, t_i^*\right) \tag{10}$$

Then, anchor boxes selected by NMS are output to train the Fast R-CNN. The position information output by RPN is mapped to the original feature map to obtain corresponding region proposals. These region proposals generate feature maps of size 7×7 through RoI pooling, which are then sent to the fully connected layer and softmax layer for the next classification operation. Additionally, the regression operation is used again to modify the region proposal to obtain a more accurate object anchor box.

IV. EXPERIMENTS

To validate the effectiveness of the proposed method, we first test the improved Faster R-CNN on VOC 2012 [12] and MSCOCO 2014 [13]. The results show that the proposed method has a significant performance improvement. Then, we apply the method to our OCS dataset and compare the performance with the experimental results of SSD [26] and RetinaNet [27]. In this section, we introduce the datasets used in the experiment and experimental implementation details. After that, the method is thoroughly tested on different datasets, and the results are presented. Finally, we conduct a detailed analysis of the experimental results.

A. DATASET

In the experiment, VOC 2012 and MSCOCO 2014 are used as validation datasets for the performance of the method. Specifically, VOC 2012 has 20 object categories, which contain 5,717 pictures for training and 5,823 images for validation. MSCOCO 2014 is another well-known object detection dataset with 80 object categories, which contains 5,717 pictures for training and 5,823 images for validation.

In this paper, 1,465 high-resolution OCS images are selected from the high-speed rail 2C system for engineering tests. Each OCS image contains several or dozens of dropper objects. We make them into the VOC dataset to perform dropper recognition experiments. The training set contains 1,172 images, and the test set contains 293 images.

B. IMPLEMENTATION DETAILS

1) TRAINING DETAILS

In the validation phase, we used Faster R-CNN as the basic detector and ResNet [5] as the feature extraction network to carry out experiments on the proposed method. On the VOC 2012 dataset, we trained the detector for 20 epochs with an initial learning rate of 0.01 and used stochastic gradient descent (SGD) with momentum 0.9 and a weight decay 0.0001. On the MSCOCO 2014 dataset, except that the epoch was set to 12, the other settings were the same as the VOC 2012 dataset.

In the test phase of dropper detection, we tested several detectors on the OCS dataset, including our improved Faster R-CNN, SSD512 and RetinaNet. Faster R-CNN and RetinaNet choose ResNet as the feature extraction network. We set the input size of training and testing to 1333×800 and 960×800 for Faster R-CNN. The other settings of Faster R-CNN were the same as the VOC 2012 dataset. We trained RetinaNet for 20 epochs with an input size of 960×800, an initial learning rate of 0.01 and a weight decay of 0.0005. SSD512 was trained for 24 epochs with an initial learning rate of 0.001 and a weight decay of 0.0005.

The entire experimental environment is described as follows: Deep learning framework Pytorch 1.1.0, centos7, and

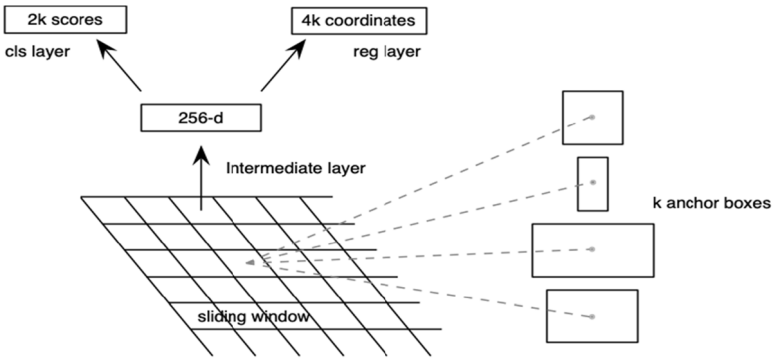


FIGURE 8. The structure of RPN.

the embedded artificial intelligence platform NVIDIA Tesla P100 GPU.

2) METRICS

The classification and location of the models in the object detection task need to be evaluated, and each image may have different objects in different categories. We use mAP (mean average precision) to evaluate the accuracy of the method. The formula is as follows:

$$R = \frac{TP}{TP + FN} \tag{11}$$

$$P = \frac{TP}{TP + FP} \tag{12}$$

$$mAP = \int_0^1 P(R)\, dR \tag{13}$$

where R is the recall rate and P is the accuracy rate. TP is the number of positive samples correctly divided into positive samples, FN is the number of positive samples incorrectly divided into negative samples, and FP is the number of negative samples incorrectly divided into positive samples. $TP + FN$ is the number of all actual positive samples, and $TP + FP$ is the total number of the samples divided into positive samples.

TP and FP were judged based on the IOU (intersection-over-union) threshold. The IOU calculation formula is as follows:

$$IOU(A, B) = \left| \frac{A \cap B}{A \cup B} \right| \tag{14}$$

where A represents the ground-truth box and B represents the anchor predicted by the detection model. The initial IOU threshold was set to 0.5. If IOU >0.5, the sample was TP; otherwise, FP.

C. EXPERIMENTAL RESULTS AND ANALYSIS

In the performance experiment of the VOC 2012 dataset, we used Faster R-CNN as the basic detector and ResNet as the feature extraction network to evaluate the proposed model. A total of 5,717 pictures were used to train the model, and 5,823 pictures were used for testing. First, to verify the effectiveness of BA-FPN, we conducted a comparative test on BA-FPN and FPN and set the IOU threshold to 0.5. Table 1 shows the experimental results of each combination. Compared to FPN, mAP@0.5 improved correspondingly on ResNet at all three depths, with ResNet50, ResNet101 and ResNet152 increasing by 1.3%, 0.9% and 0.7%, respectively. To learn more about mAP promotion details, we selected 5 small targets from 20 categories. The AP results of the

TABLE 1. The experimental results of BA-FPN on VOC 2012.

Backbone	mAP@0.5
ResNet50 + FPN	71.4
ResNet50 + BA-FPN	72.7
ResNet101 + FPN	73.2
ResNet101 + BA-FPN	74.1
ResNet152 + FPN	74.2
ResNet152 + BA-FPN	74.9

TABLE 2. The detection results of the small target on the VOC 2012 dataset.

Backbone	Bird	Bottle	Cat	Dog	Potted plant
ResNet50 + FPN	75.0	51.8	90.9	88.6	44.3
ResNet50 + BA-FPN	**76.7**	**52.7**	**92.0**	**90.5**	**45.4**

TABLE 3. The experimental results of the combination of BA-FPN and CR loss on the VOC 2012 dataset.

Backbone	Regression Loss	mAP@0.5
ResNet50+FPN	SmoothL1 Loss	71.4
ResNet50+FPN	CR Loss	71.7
ResNet50+BA-FPN	CR Loss	72.9
ResNet101+FPN	SmoothL1 Loss	73.2
ResNet101+FPN	CR Loss	73.6
ResNet101+BA-FPN	CR Loss	74.4

TABLE 4. The detection results on the MSCOCO 2014 dataset.

Backbone	Regression Loss	mAP@0.5	mAP@0.7	mAP@[0.5 , 0.95]
ResNet50	SmoothL1 Loss	47.0	27.4	26.8
ResNet50 + BA-FPN	CR Loss	48.5	28.8	27.9
ResNet101	SmoothL1 Loss	49.5	30.4	29.2
ResNet101 + BA-FPN	CR Loss	50.8	31.7	30.1

TABLE 5. The detection results on the OCS dataset based on different detection algorithms.

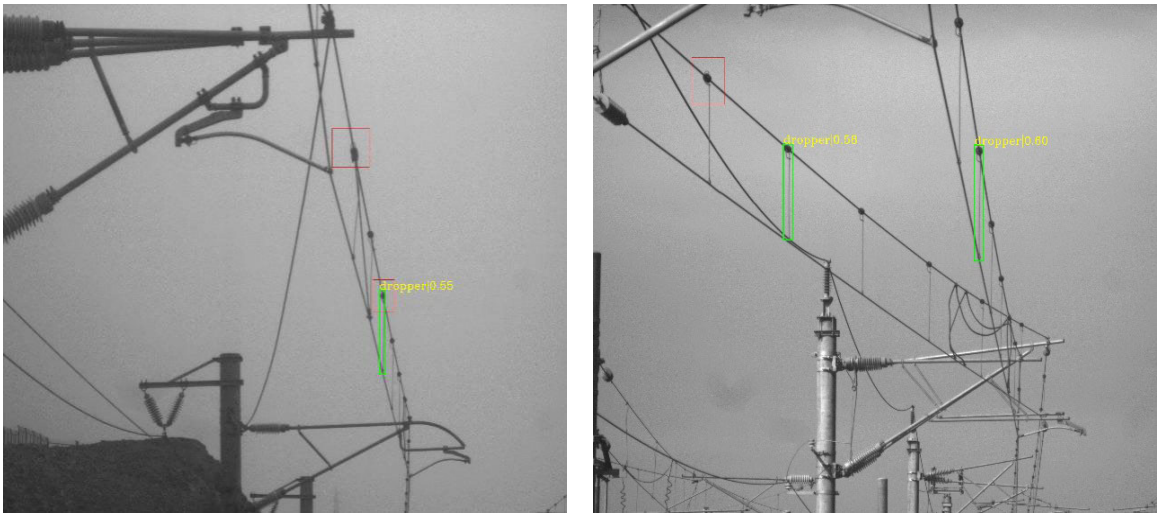
Train input	Test input	Detector	Backbone	Regression Loss	mAP@0.5	mAP@0.7	FPS
512×512	512×512	SSD	VGG16	—	67.6	—	22.3
960×800	960 × 800	RetinaNet	ResNet50 + FPN	—	77.6	71.9	12.3
			ResNet101 + FPN	—	78.8	72.7	10.9
1333×800	960 × 800	Faster R-CNN	ResNet50 + FPN	SmoothL1 Loss	83.6	81.4	8.2
			ResNet50 + BA-FPN	CR Loss	85.3	82.6	7.9
			ResNet101 + FPN	SmoothL1 Loss	85.2	82.8	6.8
			ResNet101+ BA-FPN	CR Loss	**86.8**	**83.9**	6.3

targets selected are shown in Table 2. The detection results of small targets improved considerably. Compared with FPN, the experimental results of BA-FPN showed a good performance improvement, indicating the effectiveness of the attention mechanism in FPN feature fusion.

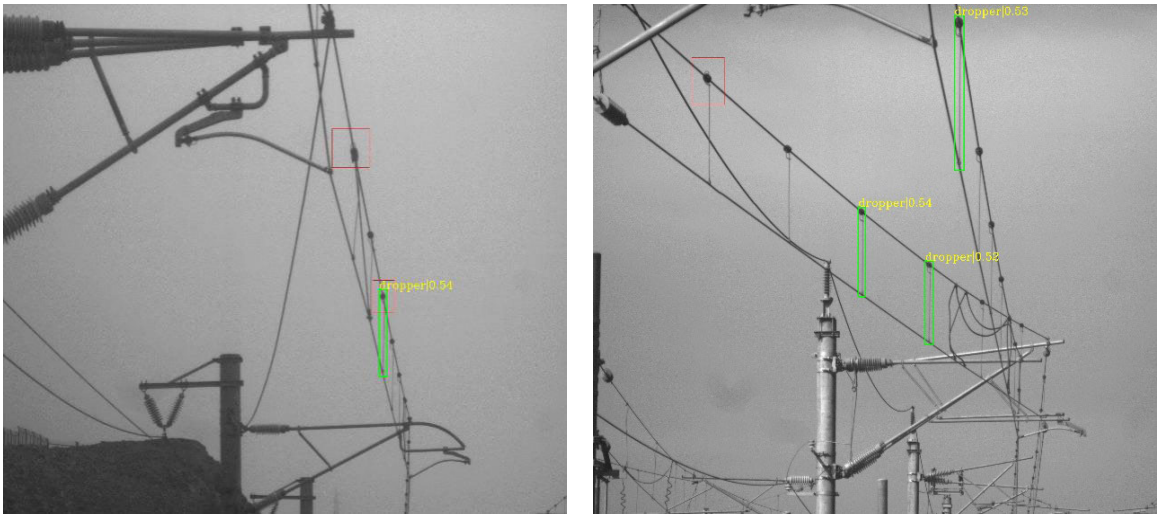
Table 3 shows the performance of CR loss on the VOC 2012 dataset. First, in the absence of BA-FPN, we compared the detection results of the original smoothL1 loss and CR loss. The mAP@0.5 of the model using CR loss was 0.3% and 0.4% higher than that of the model using smoothL1 loss on ResNet50 and ResNet101, respectively. Combining CR loss with BA-FPN, the performance of the detector was further improved. ResNet50 with BA-FPN and CR loss increased to 72.9% mAP@0.5 by 1.5% compared with ResNet50, and ResNet101 with BA-FPN and CR loss increased to 74.4% mAP@0.5 by 1.2% compared with ResNet101.

To further verify the performance of the proposed method, we tested the model on the MSCOCO 2014 dataset. The MSCOCO 2014 dataset contains 80 object categories and more than 80,000 pictures for training, which could test the performance of the detector better. In this paper, we used the training set for training and the val set for testing. The average mAP over different IOU thresholds from 0.5 to 0.95 was used for evaluation. The experiment used the Faster R-CNN detector and tested it on ResNet. The purpose of this experiment was to examine the effect of the combination of BA-FPN and CR loss on the whole detection network, so any performance improvement can prove its contribution to better performance. Table 4 describes the performance of the detector using ResNet50 and ResNet101 on the val set. ResNet50 with BA-FPN and CR loss achieved 48.5% mAP@0.5, 28.8% mAP@0.7 and 27.9% mAP@0.5:0.95 on the MSCOCO 2014 dataset, with 1.5 points higher mAP@0.5, 1.4 points higher mAP@0.7 and 1.1 points higher mAP@ [0.5, 0.95] compared to the original ResNet50. BA-FPN and CR loss also improved the performance of the model with the ResNet101 network. Compared with the basic ResNet101, the proposed model increased by 1.3% mAP@0.5 and 1.3% mAP@0.7. In general, the experiment on the MSCOCO 2014 dataset describes the promoting effect of the proposed method on the Faster R-CNN object detection network, showing significant performance improvement.

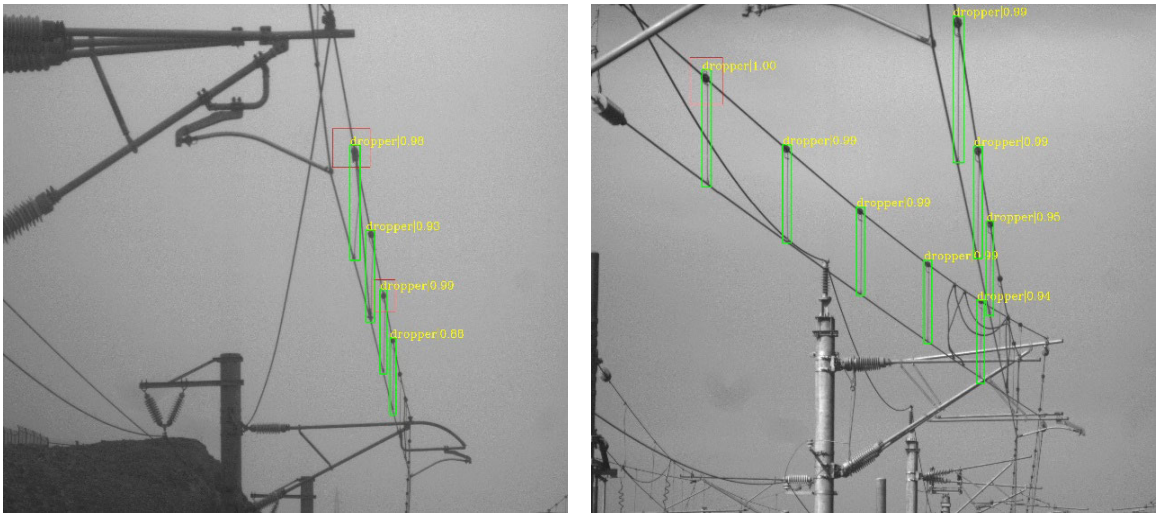
After testing on VOC 2012 and MSCOCO 2014, this paper carried out model testing on an engineering dataset of dropper detection. In this part, we chose three different detectors to conduct comparative experiments, including Faster R-CNN, RetinaNet and SSD. Considering that the pixel of the OCS dataset was high and the detection target was small, we used SSD512 instead of SSD300, which was faster. The experimental performance of different detectors is shown in Table 5. From Table 5, we learn that Faster R-CNN shows obvious advantages in test accuracy among the whole experiment, where resnet101 with BA-FPN and CR loss achieved 86.8% mAP@0.5 and 83.9% mAP@0.7, respectively, reaching the optimal performance. ResNet50 combined with BA-FPN and CR loss also improved compared to ResNet50. RetinaNet



(a)



(b)



(c)

FIGURE 9. The detection effect of different detectors. Figs (a) are the effect diagrams of SSD512. Figs (b) are the effect diagrams of RetinaNet. Figs (c) are the effect diagrams of our method.

performed best on resnet101, reaching 78.8% mAP@0.5 and 72.7% mAP@0.7. Compared with Faster R-CNN and RetinaNet, the input size of SSD is 512 × 512. SSD was faster than other detectors but performed poorly in accuracy, which only achieved 67.6% mAP@0.5.

To further describe the good performance of the proposed method in the dropper detection task, we trained different detection models on the OCS dataset and tested two input images from the dataset for performance verification. Figure 9 shows the detection effect of different detectors. The visualization results show that the Faster R-CNN with BA-FPN and CR loss had the best detection effect, significantly better than SSD512 and RetinaNet, and slightly better than that of the unimproved Faster R-CNN. The results also show the feasibility of the proposed method in the engineering testing task of droppers.

According to the comprehensive analysis, the OCS dataset used in this experiment for engineering detection of high-speed railways belongs to ultra HD images, and the detection object was too small, which required a more efficient and detailed object detection network. On the basis of the experimental results in Table 5 and Figure 9, Faster R-CNN shows great advantages in dropper recognition. On the premise that real-time detection is not required, Faster R-CNN becomes the preferred method in this project. BA-FPN and CR loss also further improved the performance of Faster R-CNN in dropper detection.

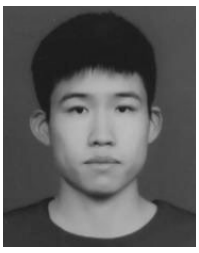
V. CONCLUSION

This paper proposes an improved Faster R-CNN for OCS dropper detection, including the balanced attention feature pyramid network (BA-FPN) and center-point rectangle loss (CR loss). First, we used an integrated semantic feature map to balance the original features of FPN and designed a mixed attention module to enhance the effective features by using an attention mechanism, making feature fusion of different scales more efficient. Second, CR loss accelerates the convergence of the regression function by optimizing the area of the rectangle, which is formed by the center points of the ground-truth box and the predicted anchor box. We carried out experiments on the VOC 2012 and MSCOCO 2014 datasets to verify the effectiveness of the proposed method and achieved great performance. In addition, compared with RetinaNet and SSD, the application experiment on the OCS dataset shows the effectiveness and feasibility of the proposed method in dropper detection, which lays a solid foundation for further dropper fault diagnosis.

REFERENCES

[1] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
[2] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, 2012, pp. 1097–1105.
[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[4] 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
[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[6] 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)*, Jun. 2016, pp. 779–788.
[7] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7263–7271.
[8] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*. [Online]. Available: http://arxiv.org/abs/1804.02767
[9] 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.
[10] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.
[11] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8759–8768.
[12] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. (2012). *The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Result*. [Online]. Available: http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html
[13] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. 13th Eur. Conf. Comput. Vis. (ECCV)*, vol. 8693. Zürich, Switzerland, 2014, pp. 740–755.
[14] E. Karakose, M. Gencoglu, M. Karaköse, I. Aydin, and E. Akin, "A new experimental approach using image processing-based tracking for an efficient fault diagnosis in pantograph–catenary systems," *IEEE Trans. Ind. Informat.*, vol. 13, no. 2, pp. 635–643, Apr. 2017.
[15] W. Liu, Z. Liu, A. Nunez, and Z. Han, "Unified deep learning architecture for the detection of all catenary support components," *IEEE Access*, vol. 8, pp. 17049–17059, 2020.
[16] Z. Qu, S. Yuan, R. Chi, L. Chang, and L. Zhao, "Genetic optimization method of pantograph and catenary comprehensive monitor status prediction model based on adadelta deep neural network," *IEEE Access*, vol. 7, pp. 23210–23221, 2019.
[17] J. Zhong, Z. Liu, Z. Han, Y. Han, and W. Zhang, "A CNN-based defect inspection method for catenary split pins in high-speed railway," *IEEE Trans. Instrum. Meas.*, vol. 68, no. 8, pp. 2849–2860, Aug. 2019.
[18] C. Petitjean, L. Heutte, R. Kouadio, and V. Delcourt, "Automatic extraction of droppers in catenary scenes," in *Proc. MVA*, 2009, pp. 497–500.
[19] C. Petitjean, L. Heutte, R. Kouadio, and V. Delcourt, "A top-down approach for automatic dropper extraction in catenary scenes," in *Proc. IbPRIA*, Póvoa de Varzim, Portugal, 2009, pp. 225–232.
[20] Y. Xu, "Application of image processing in detecting defects of catenary hanger," M.S. thesis, Dept. Elect. Eng., SWJT Univ., Chengdu, China, 2018.
[21] S. Liu, L. Yu, and D. Zhang, "An efficient method for high-speed railway dropper fault detection based on depthwise separable convolution," *IEEE Access*, vol. 7, pp. 135678–135688, 2019.
[22] 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.*, Jun. 2014, pp. 580–587.
[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] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.
[25] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6154–6162.
[26] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.
[27] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[28] J. Wang, K. Chen, S. Yang, C. C. Loy, and D. Lin, "Region proposal by guided anchoring," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2965–2974.
[29] C. Chen, C. He, C. Hu, H. Pei, and L. Jiao, "A deep neural network based on an attention mechanism for SAR ship detection in multiscale and complex scenarios," *IEEE Access*, vol. 7, pp. 104848–104863, 2019.
[30] C. Cao, B. Wang, W. Zhang, X. Zeng, X. Yan, Z. Feng, Y. Liu, and Z. Wu, "An improved faster R-CNN for small object detection," *IEEE Access*, vol. 7, pp. 106838–106846, 2019.
[31] X. Wei, Z. Yang, Y. Liu, D. Wei, L. Jia, and Y. Li, "Railway track fastener defect detection based on image processing and deep learning techniques: A comparative study," *Eng. Appl. Artif. Intell.*, vol. 80, pp. 66–81, Apr. 2019.
[32] J. Li, F. Zhou, and T. Ye, "Real-world railway traffic detection based on faster better network," *IEEE Access*, vol. 6, pp. 68730–68739, 2018.
[33] D. He, Z. Yao, Z. Jiang, Y. Chen, J. Deng, and W. Xiang, "Detection of foreign matter on high-speed train underbody based on deep learning," *IEEE Access*, vol. 7, pp. 183838–183846, 2019.
[34] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3156–3164.
[35] J. Fu, H. Zheng, and T. Mei, "Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4476–4484.
[36] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.
[37] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.
[38] H. Ling, J. Wu, L. Wu, J. Huang, J. Chen, and P. Li, "Self residual attention network for deep face recognition," *IEEE Access*, vol. 7, pp. 55159–55168, 2019.
[39] S. Sun, B. Zhao, X. Chen, M. Mateen, and J. Wen, "Channel attention networks for image translation," *IEEE Access*, vol. 7, pp. 95751–95761, 2019.
[40] Q. Liu, X. Che, and M. Bie, "R-STAN: Residual spatial-temporal attention network for action recognition," *IEEE Access*, vol. 7, pp. 82246–82255, 2019.
[41] L. Gao, Y. Li, and J. Ning, "Residual attention convolutional network for online visual tracking," *IEEE Access*, vol. 7, pp. 94097–94105, 2019.
[42] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-IoU loss: Faster and better learning for bounding box regression," 2019, *arXiv:1911.08287*. [Online]. Available: http://arxiv.org/abs/1911.08287



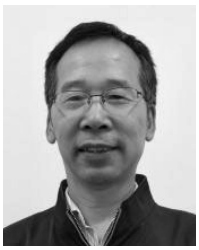
QIFAN GUO was born in Shijiazhuang, China, in 1996. He received the B.S. degree in statistics from Jilin University, Changchun, China, in 2018. He is currently pursuing the M.S. degree in applied statistics with Xi'an Jiaotong University. His current research interests include deep learning and data mining.

LEI LIU was born in Gansu, China, in 1995. He received the B.S degree in applied statistics from the Minzu University of China, Beijing, in 2018. He is currently pursuing the M.S. degree in applied statistics with Xi'an Jiaotong University. His current research interests include machine learning, and big data analysis and processing.

WENJUAN XU was born in Anhui, China, in 1995. She received the B.S. degree in information and computing science from the Anhui University of Finance and Economics, Anhui, in 2018. She is currently pursuing the M.S. degree in applied statistics with Xi'an Jiaotong University. Her current research interests include machine learning and big data analysis and processing.

YANSHENG GONG was born in Shandong, China, in 1965. He received the B.Eng. degree in design of electric locomotive from Southwest Jiaotong University, Sichuan, in 1987. His research interests include electrified railway traction power supply technology in complex and difficult environment and rail transportation intelligent power supply technology.

XUEWU ZHANG was born in Gansu, China, in 1984. He received the master's degree in electrical engineering and automation from Beijing Jiaotong University, in 2009. His research interests include the contact line technology of the electrified railways in complex and challenging environments and the intelligent monitoring technology of the contact line.



WENFENG JING was born in Xi'an, China, in 1963. He received the Ph.D. degree in applied mathematics from Xi'an Jiaotong University, China, in 2009. His current research interests include basic and core algorithms for big data, deep learning and AutoML methods, data analysis platforms, and applications of big data and deep learning.