

## RESEARCH ARTICLE

# A Dense Attention Railway Foreign Object Detection Algorithm Based on Mask R-CNN

SHUANG GAO 

Railway Department, Hohhot Vocational College, Hohhot 010010, China

e-mail: gshuang1207@163.com


**ABSTRACT** In railway safety monitoring, foreign object detection is a key task, especially in infrared and low illumination conditions. In order to improve the detection accuracy of railway foreign bodies, an optimized railway foreign bodies detection algorithm named O-Mask R-CNN is proposed in this study. Firstly, by integrating the densely connected feature pyramid network (FPN) and the convolutional attention mechanism (CBAM), the recognition ability of low-contrast objects and the detection accuracy of small-size foreign objects are significantly improved. In addition, O-Mask R-CNN uses an improved Zone Suggestion Network (RPN) and ROIAlign layer to ensure feature alignment, thereby optimizing the precision of the bounding box and the quality of the instance segmentation. Finally, a method of adjusting the size of anchor frame based on cluster analysis is introduced to adapt to the characteristics of different scale railway foreign bodies. After only 46 iterations, the developed railway foreign item identification algorithm achieved a constant total loss value and has strong iterative performance. Mean square error and mean absolute error are the smallest, 1.35 and 1.21 respectively, and the detection error performance is also very good. Finally, the real average detection accuracy of the algorithm on infrared and common railway foreign body photos is tested, which is up to 98.26 and 98.85, respectively. Therefore, the detection method created in this paper not only has high performance, but also provides technical support for foreign object recognition in practical application environment.

**INDEX TERMS** Attention mechanism, convolutional neural networks, mask, foreign object detection, railway, night

## I. INTRODUCTION

The continuous increase in the operating mileage of High-speed Railways (HSR) in China has raised concerns about the safety of HSR operations [1]. Due to the fast operation speed of HSRs, once personnel or other foreign objects invade the railway line, it will pose a huge threat to the safety of railway trains [2]. Therefore, real-time monitoring of railways and using Object Detection (OD) algorithms to determine whether there has been foreign object intrusion can eliminate safety hazards, reduce the incidence of dangerous accidents, and ensure the safe operation of trains. In recent years, with the increasing number of monitoring

facilities along the railway and the rapid growth of video monitoring, computer vision technology has been combined with monitoring images in railway scenes to achieve foreign OD and real-time monitoring of the railway operating environment. Region-based Convolutional Neural Network (Mask R-CNN) is a deep learning algorithm mainly used for instance segmentation problems, which can accurately detect the target object and perform the corresponding classification operations. Mask R-CNN can not only identify the position of different objects in the image in the form of a border, but also generate an accurate pixel-level mask for each object, so as to distinguish the specific outline of the object [3]. SSD is a fast object detection algorithm, mainly used in real-time processing scenarios. SSDs eliminate the candidate region extraction step in the traditional target detection

The associate editor coordinating the review of this manuscript and approving it for publication was Antonio J. R. Neves .

process by directly predicting bounding box and category probabilities in a single forward pass, and they can also make predictions on multi-scale feature maps to improve detection performance for objects of various sizes [4]. MSSIF-Net is an innovative object detection framework designed to process multi-scale image features. It effectively utilizes image information from different levels through independent and shared feature extraction modules, which is designed to enhance the detection ability of the model for objects with large size changes [5]. Currently, many experts have studied the Railway Foreign Object Detection (RFOD), but most of the detection models designed by scholars can only accurately detect foreign objects in good daylight conditions, so they do not have universality. In response to the problems of insufficient target feature extraction and low detection accuracy in RFOD in Infrared Weak Light Environments (IWLE), this study proposes a dense attention RFOD method suitable for this environment based on Mask R-CNN.

The main innovation of the research is to combine the improved Feature Pyramid Network (FPN) with the Convolutional Block Attention Module (CBAM). The detection ability of infrared low illumination railway foreign object image in Mask R-CNN is optimized. Specific research innovation points are the following three points. Firstly, the traditional FPN is improved. By introducing dense connection mode, the ability of transmitting and integrating characteristic information in different scales is enhanced. This densely connected FPN not only optimizes the flow of information between the feature layers, but also improves the detection of small-scale targets, which is particularly critical for the small foreign objects commonly seen in infrared images. The second is the integration of CBAM in Mask R-CNN, an attention mechanism focused on enhancing the discriminability of the model by learning the importance of features. CBAM weights the channel and spatial dimensions of the feature map sequentially, effectively enhancing the model's response to key features, especially under low illumination conditions with more background noise. Finally, the improved K-means clustering algorithm is used to adjust the anchor frame size of the Region Proposal Network (RPN) to make it more suitable for the feature distribution of infrared railway foreign body images. This strategy reduces the errors in the conventional anchor frame presets and improves the accuracy of the model for locating foreign bodies.

The main contributions of the research are as follows. First, by integrating densely connected FPN and CBAM into Mask R-CNN, the final detection algorithm improves the target detection performance of the network under low illumination conditions. This structural improvement enhances the feature hierarchy and detail capture ability, especially for the recognition of small and low-contrast targets. Secondly, the size and proportion of the anchor frame are redesigned using the method based on cluster analysis to better adapt to the feature distribution of infrared railway foreign object image. This strategy reduces the false detection rate and improves the accuracy of the model to locate the foreign

object boundary. Thirdly, according to the requirement of real-time monitoring, the computational efficiency of the algorithm is optimized to ensure the fast response even in the environment with limited computing resources. In addition, extensive experiments have been carried out on several infrared railway foreign body detection data sets to verify the effectiveness of the proposed method.

## II. RELATED WORK

Regarding the difficulties in instance segmentation tasks in computer vision, X. Bi et al. proposed an information enhanced Mask R-CNN. This model effectively solved the problems of losing important channel information and inaccurate instance classification in high-level feature maps by enhancing useful channel and global information in FPN, as well as enhancing the processing of local global information in mask branches. This model outperformed traditional Mask R-CNN in benchmark testing on different datasets [6]. Y. Liu et al. proposed a method grounded on Mask R-CNN for detecting and segmenting apples in complex orchard environments. The squeeze excitation module was introduced into the ResNet-50 to effectively allocate computing resources until the most informative channel feature map is obtained. Then, the aspect ratio parameter in bounding box regression loss was introduced to adapt to the shape of the apple by changing the shape of the bounding box, thereby improving the effectiveness of bounding box regression. This method surpassed multiple advanced technologies in apple detection and segmentation [7]. Regarding the issue of Railway Surface Defect Detection (RSDD), F. Guo et al. proposed a instance segmentation built on computer vision. Firstly, a railway surface database containing 1040 images was constructed in the framework, and then the Mask R-CNN model was retrained and fine tuned using this database, aiming to complete the RSDD task through the trained Mask R-CNN. At the 0.005 learning rate, the retrained Mask R-CNN performed the greatest in RSDD [8].

At present, many experts have conducted target detection research on railway objects using various intelligent algorithms. However, there are still huge challenges in accurately detecting railway objects under the interference of complex railway backgrounds, adverse weather conditions, and low-quality images. T. Ye et al. proposed a multi-modal feature enhanced CNN, aimed at accurately detecting railway traffic targets in changeable railway scenes using this network. The network consists of three parts, namely the improved Darknet53 basic network, spatial feature extraction, and attention fusion enhancement. The constructed network model achieved an average accuracy of 0.9439 on the railway transportation dataset. In addition, it has also been proven that the model has certain feasibility in the practical application of railway object monitoring [9]. In the safe operation of HSRs, the intrusion of foreign objects into the airspace is a big threat. In response to this current situation, R. Tian et al. suggested a multi-scale RFOD algorithm suitable for HSRs, and named the algorithm CenterNet and variable focal length multi-scale

enhancement algorithm. This algorithm generated confidence scores for each region in the image by rendering true values with a 2D Gaussian distribution. Moreover, this algorithm outperformed other comparative algorithms in both detection accuracy and speed [10].

In summary, many experts have conducted research on Mask R-CNN and RFOD problems, but most of the research is limited to detecting railway foreign objects with good lighting conditions, so it does not have universality. To further improve the detection performance of Infrared Railway Foreign Object Images (IRFOI), this study optimizes Mask R-CNN. By combining improved FPN and CBAM structures, Mask R-CNN accurately detects foreign object images under normal lighting conditions and IRFOI.

### III. RFOD UNDER IWLE

This study proposes an RFOD algorithm applicable to the detection environment for RFOD tasks in nighttime IWLE. Firstly, the main network structure of the detection algorithm, Mask R-CNN, is optimized, including the introduction of dense connection FPN structure and CBAM structure. Then, an optimized K-means is taken to adjust the detection anchor box's size for the Region Proposal Network (RPN) in Mask R-CNN, aiming to further prompt foreign object localization in the network. Finally, the optimized parts are combined to design the RFOD algorithm.

#### A. OPTIMIZATION DESIGN OF MASK R-CNN FEATURE EXTRACTION

Mask R-CNN is a type of deep learning model mainly used for computer vision tasks, and currently has good performance in the fields of monitoring object and instance segmentation. This model was designed by Kaiming He et al. in 2017. As an extended network model of Faster R-CNN, Mask R-CNN adds a branch to it and uses this branch to generate masks for targets [11], [12]. The generated mask area enables Mask R-CNN to achieve both OD and instance segmentation simultaneously. Figure 1 shows the network structure of Mask R-CNN.

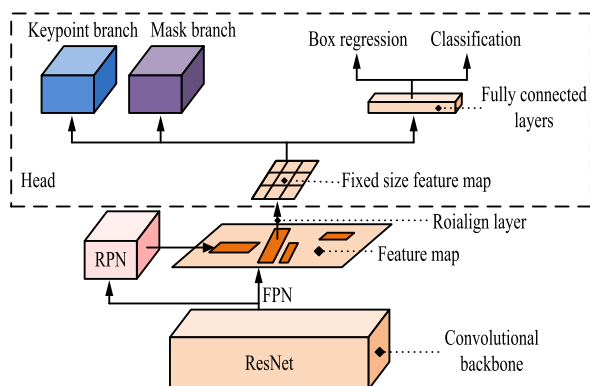


FIGURE 1. Structure of mask R-CNN.

The Mask R-CNN mainly contains four parts: basic network, RPN, ROIAlign, and head network [13]. As the core part of feature extraction, the basic network adopts ResNet50 architecture to extract complex feature information from input images. The output of this network is a series of feature maps at different scales, providing basic data support for subsequent RPN and ROIAlign. The RPN runs directly on the feature map of the underlying network to generate a bounding box of candidate objects, and by scoring these proposed regions, the RPN determines which regions are most likely to contain detection targets. ROIAlign is used to precisely extract a fixed-size feature map from each candidate region proposed by the RPN. Compared with traditional ROI Pooling technology, ROIAlign improves feature spatial accuracy by avoiding quantization error and can achieve better feature segmentation. The header network includes a classification header and a mask dock. The classification header is responsible for determining the category of each ROI, while the mask dock generates the corresponding pixel-level mask to achieve accurate instance segmentation. Overall, the hierarchical structures in Mask R-CNN efficiently collaborate to process OD tasks in images, achieving precise recognition and pixel level segmentation of detected targets. Analyzing ResNet in Mask R-CNN and its structural diagram is obtained as shown in Figure 2.

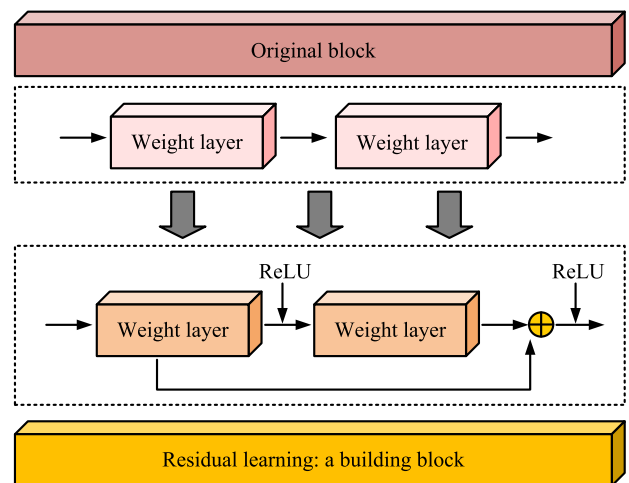


FIGURE 2. Composition of residual network and its structure.

Figure 2 shows the residuals unit structure composed of ResNet, which mainly consists of four modules: residuals block, skip connection, global average pooling and output layer. In a residual block, each residual block contains two or three convolution layers, each of which is followed by the Batch Normalization layer and the ReLU activation function. These convolutional layers are designed with smaller filter sizes to extract local features of the image. Skip joins are a prominent feature of residual blocks, which directly add input to the output of the convolution layer. This connection helps the network maintain a steady propagation of gradients during training, allowing deeper network models to be built.

Global average pooling means that the global average pooling layer is used to replace the traditional fully connected layer at the end of the network, which can greatly reduce the number of parameters of the model, reduce the risk of overfitting, and improve the adaptability of the model to the input image size. In the output layer, a fully connected layer and Softmax activation function are used to output the final classification results [14], [15].

FPN is an effective feature extraction method commonly used in OD tasks. To lift the detection accuracy of Mask R-CNN for target objects, this study adds an optimized FPN to the ResNet basic network, aiming to enhance multi-scale feature representation. Figure 3 shows the structures of traditional FPN and improved dense connected FPN.

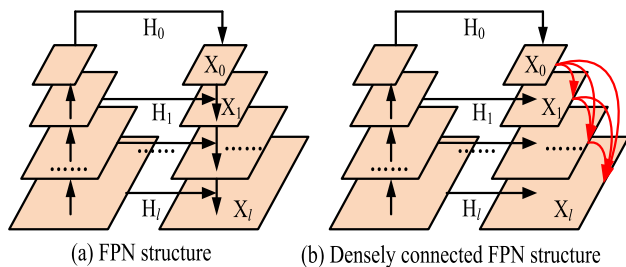


FIGURE 3. Structure of FPN with densely connected FPNs.

Figures 3 (a) and (b) show the traditional FPN and the improved dense connected FPN, respectively. In 3 (a), the conventional FPN has a simple structure and mainly uses multiple single-layer pyramid structures to complete OD and segmentation tasks. Although FPN effectively processes targets of different scales, it has certain limitations in processing inter layer semantic information. The limitations of FPN result in sub-optimal performance in handling small OD. To compensate for the shortcomings of the original FPN in semantic information transmission, the traditional FPN has been improved by drawing on the dense connection concept of DenseNet. In Figure 3 (b), based on the traditional FPN, the concept of dense connections from low to highdimensional features is introduced, which effectively utilizes the information of each feature layer in the improved FPN. The fusion formula for feature maps in traditional FPN is equation (1) [16], [17].

$$X_l = H_l + X_{l-1} \tag{1}$$

In equation (1),  $l$  means the amount of layers in the FPN network.  $X_l$  and  $X_{l-1}$  are the feature maps of the  $l$  and  $l - 1$  layers.  $H_l$  represents the horizontal connection feature map of the layer  $l$ . Introducing the idea of dense connections to improve the FPN network, the improved expression of equation (1) is obtained as shown in equation (2).

$$X_l = H_l + (X_0 + X_1 + \dots + X_{l-1}) \tag{2}$$

In equation (2), the FPN feature map that integrates the idea of dense connections is no longer solely determined by the horizontally connected feature map and the previous feature

map, but is fused at multiple scales using dense connections and horizontally connected features. Thus, it ensures that the multi-scale information of the image can be fully extracted.

In the field of deep learning, attention mechanisms have become an important technique, especially when dealing with images and language tasks. The attention mechanism mimics the function of human visual attention, which can make the model focus on the important part of the input data, thereby improving the performance and efficiency of the model. Specifically, the attention mechanism strengthens the model's processing ability of key information by assigning different weights to different parts of the input. In computer vision tasks, CBAM, as an improved model of attention mechanism, can implement attention in both spatial dimension and channel dimension, so as to enhance the ability of feature expression. When detecting foreign objects in the railway environment at night, the small difference in infrared radiation rates between different foreign objects can lead to blurry imaging details of these foreign objects under low light conditions at night. To improve the efficiency of target detection in low light conditions and enable Mask R-CNN to quickly recognize target areas and suppress background noise in complex low light infrared images, this study incorporates CBAM into the dense connected FPN of Mask R-CNN. Figure 4 shows the structure of CBAM.

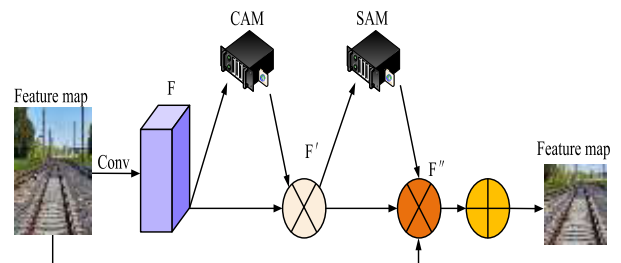


FIGURE 4. Structure diagram of CBAM.

The CBAM in Figure 4 is an efficient attention mechanism fusion module, which mainly has two main parts: Channel Attention Mechanism (CAM) and Spatial Attention Mechanism (SAM) [18], [19]. In the CAM section, it learns the features of different channels and weights them, allowing the model to focus more on information rich channels. In the SAM section, it concentrates on selecting key features from feature images, guiding the network to focus on these key regions and extract relevant feature information. This dual attention mechanism that combines channel and space enables CBAM to further improve the efficiency of network processing of image data while refining feature extraction. The calculation formula of CBAM is equation (3) [20].

$$F' = M_c(F) \otimes F \tag{3}$$

In equation (3),  $F$  is the feature map obtained through the previous convolution operation.  $\otimes$  is the multiplication symbol between features.  $M_c$  represents CAM.  $F'$  means the

new feature map handled by CAM, while that processed by SAM is equation (4) [21], [22].

$$F'' = M_s (F') \otimes F' \quad (4)$$

In equation (4),  $M_s$  represents SAM.  $F''$  is the attention weight value obtained by processing  $F'$  through SAM. The final attention weight value is weighted and summed with the original feature map, and then output as a new one. The processing flow of CAM and SAM is Figure 5.

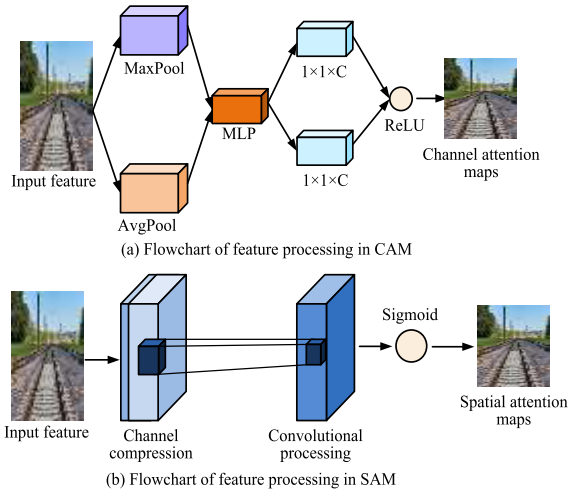


FIGURE 5. Feature processing flowchart for CAM and SAM.

Figure 5 (a) shows the feature processing flow of CAM, which is mainly divided into three steps: global pooling, multi-layer perceptron, and activation of applications [23]. In CAM, the input feature map is first subjected to global Average Pooling (AP) and global Maximum Pooling (MP) to generate two different channel description information. The information described in these two channels is processed through a multi-layer perceptron with shared weights [24]. Finally, the output information of the two multi-layer perceptrons is processed by adding feature elements and generating a Channel Attention map (CAM) through a sigmoid function. Figure 5 (b) shows the feature processing flow of SAM, which is mainly divided into three steps: channel compression, convolution processing, and activation application. Firstly, to compress the feature map on the channel axis, and then to generate two 2D feature maps through AP and MP. Next, to merge these two feature maps and process them through a  $3 \times 3$  convolutional layer to generate a single channel Spatial Attention map (SAM). Finally, the attention map of this single channel is processed using Sigmoid and the final SAM is generated [25]. The formula for generating CAMs is equation (5).

$$MC(F) = \sigma (MLP(AvgPool(F)) + MLP(MaxPool(F))) \quad (5)$$

In equation (5),  $MC(F)$  is the CAM.  $\sigma$  represents the activation function Relu.  $MLP$  means a multi-layer perceptron.

$AvgPool(F)$  is AP.  $MaxPool(F)$  represents MP. The formula for generating SAMs is equation (6).

$$MS(F) = \theta \left( f^{3 \times 3} [AvgPool(F); MaxPool(F)] \right) \quad (6)$$

In equation (6),  $MS(F)$  represents the SAM.  $\theta$  is the function Sigmoid.  $f^{3 \times 3}$  denotes a  $3 \times 3$  filter size's convolution operation.

## B. DESIGN OF DENSE ATTENTION RFOD ALGORITHM BASED ON OPTIMIZED MASK R-CNN

After optimizing each part of Mask R-CNN, a brand new backbone network can be obtained as the basic network part of Mask R-CNN. Recording this new basic network as the Convolutional Block Attention Mechanisms-Residual Network-Density Connected FPN (CBAM-ResNet-DCFPN) that integrates dense connections between FPN and CBAM. Figure 6 shows the composition of CBAM-ResNet-DCFPN.

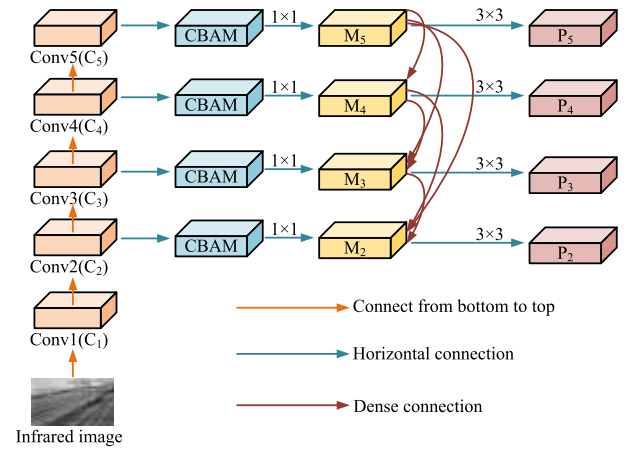


FIGURE 6. Architecture of CBAM-ResNet-DCFPN.

The CBAM-ResNet-DCFPN structure in Figure 6 adopts a bottom-up approach to extract hierarchical features from the image, thereby enabling better capture of detailed features in the image. In CBAM-ResNet-DCFPN, ResNet is the backbone network containing five layers of residual modules. The residual module outputs from the first layer to the fifth layer are  $C_1$ ,  $C_2$ ,  $C_3$ ,  $C_4$ , and  $C_5$ . The resolution of the first to fifth layers is set to half of the original. The output of the  $C_2 \sim C_5$  layers is selected as the feature map of DCFPN, and then the CBAM is utilized for subsequent processing. After using the CBAM to enhance the effective feature information of the  $C_2 \sim C_5$  layers, to perform  $1 \times 1$  convolution processing on the information of these four layers separately. The reduced feature maps  $C_2$ ,  $C_3$ ,  $C_4$ , and  $C_5$  are obtained in sequence, denoted as  $M_2$ ,  $M_3$ ,  $M_4$ , and  $M_5$ . Introducing the concept of dense connections between  $M_2$  and  $M_5$ , that is, the dimensionality reduction feature map of this layer is the fusion of feature information from the upper layers. Finally, a  $3 \times 3$  convolution operation is applied to the fusion results of each layer, and the final features are obtained, denoted as  $P_2$ ,  $P_3$ ,  $P_4$ , and  $P_5$ . By implementing dense

connections, high-level semantic information can be more accurately transmitted to lower levels, thereby reducing the loss of feature information. This feature processing method helps to capture more detailed and comprehensive feature information. Starting from the dimensionality reduction feature map at the top layer, the calculation formula for  $M_5$  is equation (7) [26], [27].

$$M_5 = f_{CBAM}(C_5) \tag{7}$$

In equation (7),  $f_{CBAM}$  represents the use of CBAM for feature processing. The expression for  $M_4$  obtained from  $M_5$  is equation (8).

$$M_4 = f_{1*1}(C_4) \oplus f_{2up}(M_5) \tag{8}$$

In equation (8),  $f_{1*1}$  represents  $1*1$  convolution processing.  $f_{2up}$  represents a 2-fold upsampling operation.  $\oplus$  represents the addition symbol between features. According to  $M_5$  and  $M_4$ , the expression for  $M_3$  is equation (9).

$$M_3 = f_{1*1}(C_3) \oplus [f_{2up}(M_4) \oplus f_{4up}(M_5)] \tag{9}$$

In equation (9),  $f_{4up}$  represents a 4-fold upsampling operation. According to  $M_5$ ,  $M_4$ , and  $M_3$ , the expression for  $M_2$  is equation (10).

$$M_2 = f_{1*1}(C_2) \oplus [f_{2up}(M_3) \oplus f_{4up}(M_4) \oplus f_{8up}(M_5)] \tag{10}$$

In equation (10),  $f_{8up}$  represents an 8-fold upsampling operation.

In Mask R-CNN, RPN is responsible for quickly and effectively generating a series of region proposals from the image, which annotate feature regions that may contain targets. Subsequently, RPN predicts the presence and bounding box positions of targets by sliding windows on the image and combining anchor frames, laying the foundation for subsequent OD and segmentation tasks [28], [29]. In addition to optimizing the basic network of Mask R-CNN and improving its ability to extract feature information, it is also necessary to optimize the localization function of RPN to ensure that the entire Mask R-CNN can accurately detect defect positions and locate them accurately. The traditional Mask R-CNN uses nine anchor boxes that are paired with three sizes and three aspect ratios as the prediction boxes for the target bounding box. This design is reasonable in traditional Railway Foreign Object Image (RFOI) detection. However, in the IRFOI detection task, due to the fact that the targets are mostly slender orbital shapes, the original anchor boxes are not always effective, leading to false positioning. Based on this, this study optimizes the anchor box positioning position using the K-means algorithm [5]. The traditional K-means uses Euclidean distance as a measure of similarity between anchor boxes and cluster center boxes, but this measure not only requires a large amount of computation but also cannot effectively segment targets. Therefore, this study introduces the Intersection of Union (IoU) to measure the similarity between anchor boxes and cluster center boxes. The improved K-means is referred to as the Intersection of Union K-means

Clustering Algorithm (IoU-K-means) that integrates IoU. IoU more effectively reflects the overlap degree between anchor and cluster boxes, enabling the selection of preset anchor box positions that are more fit to IRFOI data. The calculation of the distance between the anchor box and the cluster center  $d$  is equation (11).

$$d(b, o) = 1 - IOU(b, o) \tag{11}$$

In equation (11),  $d$  has a value range of  $[1, 0]$ .  $b$  means anchor box and  $o$  is cluster center point. According to the IOU index, the clustering formula for distance measurement is further obtained as shown in equation (12).

$$IOU(b_{pred}, b_{truth}) = \frac{b_{pred} \cap b_{truth}}{b_{pred} \cup b_{truth}} \tag{12}$$

In equation (12),  $IOU(b_{pred}, b_{truth})$  represents the IOU metric clustering value between the anchor box and the actual box.  $b_{pred}$  and  $b_{truth}$  represent anchor boxes and actual boxes, respectively. The various optimization parts in Mask R-CNN are combined to obtain the optimized RFOD algorithm under Mask R-CNN. The complete detection RFOD algorithm is referred to as the OptimizedMask R-CNN (i.e. O-Mask R-CNN). Figure 7 shows the flow of O-Mask R-CNN.

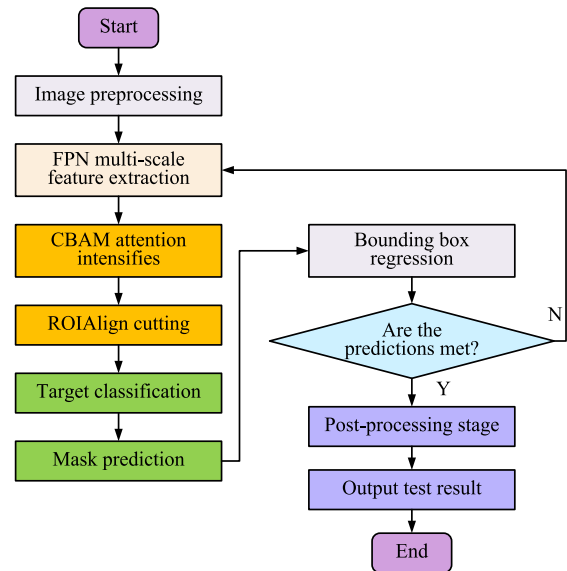


FIGURE 7. Flow of O-Mask R-CNN.

In Figure 7, the image is first preprocessed. Input images are first dimensioned and normalized, and image enhancement techniques are applied to infrared and low-light images to improve quality. Secondly, multi-scale feature maps are extracted by improved FPN network and processed by CBAM module to enhance the model's focus on key features. Then, the feature map is fed into the RPN, and the redundancy of overlapping proposal regions is reduced by non-maximum suppression processing. The proposed area obtained from the RPN is then precisely adjusted through the ROIAlign layer to ensure feature alignment. Classification and bounding box

regression are processed through a header network to output category probabilities and position adjustments for each region. Finally, in the post-processing phase, thresholds and NMS are applied to remove the low confidence results and integrate the detection results of the output content.

#### IV. PERFORMANCE TESTING AND APPLICATION EFFECT ANALYSIS OF RFOD ALGORITHM UNDER IWLE

To demonstrate the effectiveness of the O-Mask R-CNN algorithm in RFOD, this paper first conducted ablation experiments on the CBAM-ResNet-DCFPN network in O-Mask R-CNN. After ensuring excellent benchmark performance of the basic network, this study further introduced other detection algorithms for comparative experiments.

##### A. PERFORMANCE TESTING OF OBJECT DETECTION ALGORITHMS

To verify the O-Mask R-CNN effectiveness, this study develops a dataset for RFOD. RFOD uses night traffic infrared data provided by railway transportation companies, and directly captures night images on railway sites using infrared cameras, which are annotated through image processing software. There are a total of 5000 railway infrared images in the dataset, which are separated into training and testing datasets in a 9:1 ratio. To avoid experimental errors, this study conducts various experiments using the same equipment. Table 1 shows the experimental environment and basic network parameters.

**TABLE 1. Configuration of experimental environment and network parameter settings.**

Experimental equipment	Value
CPU	Intel(R) Core i7-9700K @ 3.6GHz
Epoch	100
Graphics Card	NVIDIA GeForce GTX 1660
Learning Rate	0.001
RAM	64.0GB RAM
Batch-size	32

In Table 1, the main network part of O-Mask R-CNN, namely CBAM-ResNet-DCFPN, is composed of multiple different networks combined. In this paper, the Mask R-CNN algorithm and its components are parameterized in detail to ensure the optimal performance of railway foreign object detection. ResNet50 was chosen first as the base network because it provides a deep enough network to capture complex features while keeping the computational complexity modest. Next, using the Adam optimizer, set its initial learning rate to 0.001, attenuation factor to 0.1, and decay every 10 epoches. Then He weight initialization method is used to optimize the ReLU activation function to prevent gradient disappearance or explosion in the initial training of the network. In the RPN, first of all, according to the statistics of the target size, the anchor box of  $\{32 \times 32, 64 \times 64, 128 \times 128\}$  three sizes is selected, and these sizes cover various sizes

of foreign bodies. Then set the anchor frame ratio to  $\{0.5, 1, 2\}$  to accommodate the different shapes of the foreign body. Finally, 9 anchor boxes are set at each position, and the non-maximum suppression threshold is set to 0.7, which is used to filter candidate boxes with large overlap in the detection phase to reduce redundancy. For FPN and CBAM, four levels of FPN are first used, and the size of the feature map at each level is reduced to correspond to 1/4, 1/8, 1/16, and 1/32 of the original map respectively. Secondly, CBAM modules are added after each FPN output layer, multi-layer perceptrons are used for channel attention, and  $7 \times 7$  convolution nuclei are used for spatial attention. The selection of these parameters is based on extensive experimental testing and literature review to ensure the high performance and robustness of the model in practical applications.

To test the O-Mask R-CNN, this study first set up ablation experiments to test the performance of each part of the network structure. Detection accuracy, recall rate, F1 value, and detection time are used as detection indicators for ablation experiments. Table 2 shows the result data obtained from the ablation experiment.

**TABLE 2. Results of ablation experiments.**

Network structure	Accuracy value	Recall value	F1 value	Network response time/s
ResNet+FPN (Model 1)	0.80	0.76	0.78	1.23s
VGG+FPN (Model 2)	0.74	0.71	0.73	1.36s
ResNet+DCFPN+CBAM (Model 3)	0.88	0.85	0.87	0.41s
VGG+DCFPN+CBAM (Model 4)	0.87	0.82	0.84	0.55s
ResNet+DCFPN+CBAM+Dense (Model 5)	0.97	0.98	0.98	0.02s
VGG+DCFPN+CBAM+Dense (Model 6)	0.90	0.93	0.92	0.05s

Table 2 lists a total of six different network combinations, denoted as Model 1~6. Model 5 is the main network combination method used in this study. Model 2 shows the worst performance in all aspects of ablation experiments, with accuracy values, recall values, F1 values, and network response times of 0.80, 0.76, 0.78, and 1.23 seconds. Model 5 performs the best in all aspects of ablation experiments, with values of 0.97, 0.98, 0.98, and 0.02 seconds, respectively. The benchmark test results of other models under ablation experiments are between Model 2 and Model 5.

After verifying the performance of the CBAM-ResNet-DCFPN, this study selects Single Shot Multi-Box Detector (SSD), Mask R-CNN, Multi-Scale Shared and Independent Feature Network (MSSIF-Net) as comparative algorithms for O-Mask R-CNN. Further testing is conducted on the loss values of four detection algorithms in the training dataset, as shown in Figure 8.

Figures 8 (a) to (c) show the total loss curve, segmentation loss curve, and regression loss curve of SSD, Mask R-CNN, MSSIF-Net, and O-Mask R-CNN detection algorithms. In Figure 8 (a), as the iteration period increases,

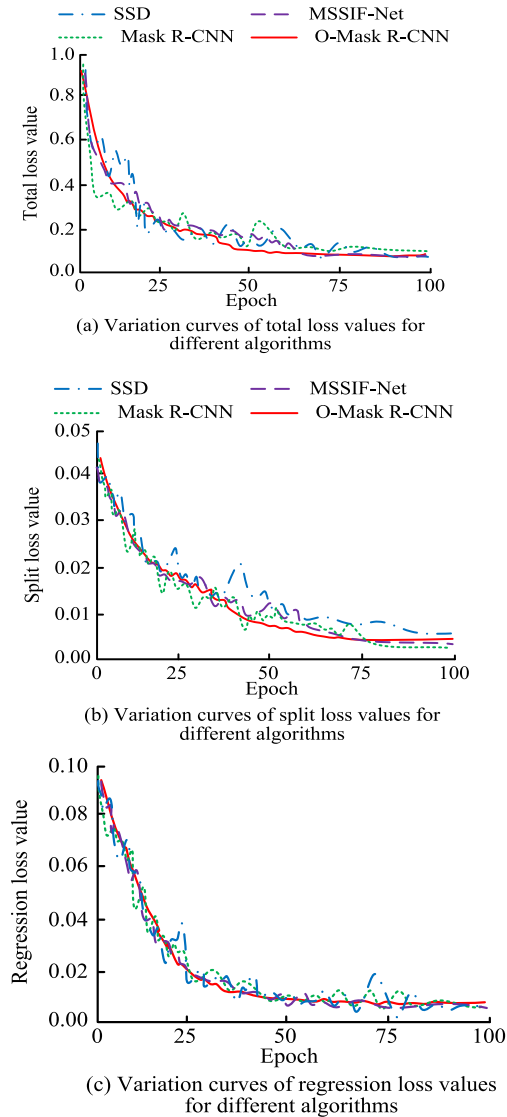


FIGURE 8. Loss value curves for different detection algorithms.

the total loss value curves of the four algorithms will gradually stabilize. Among the four algorithms, the total loss curve of O-Mask R-CNN can reach a stable state first, requiring only 46 iterations to reach stability. In Figures 8 (b) and(c), O-Mask R-CNN can achieve stable segmentation loss and regression loss values after 59 and 41 iterations, respectively. Its oscillation amplitude during the iteration process is much smaller than the other three comparison algorithms. Comparing the error performance of four algorithms during training, the Mean Squared Error (MSE) and Mean Absolute Error (MAE) curves shown in Figure 9 are obtained.

Figures 9 (a) and (b) show the MSE and MAE of the four detection algorithms, respectively. In Figure 9 (a), as the samples increase, the MSE of SSD, Mask R-CNN, MSSIF-Net, and O-Mask R-CNN will continue to increase and eventually stabilize at 3.13, 2.97, 2.71, and 1.35, respectively. In Figure 9 (b), as the sample increases, the MAE of each

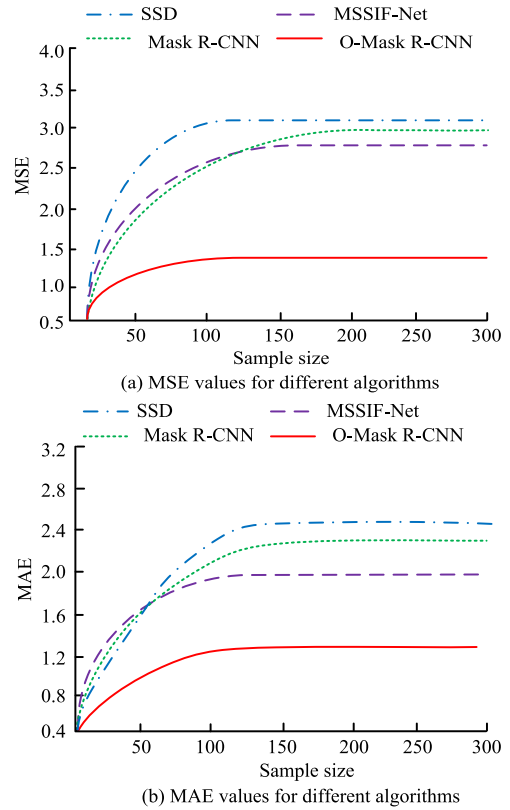


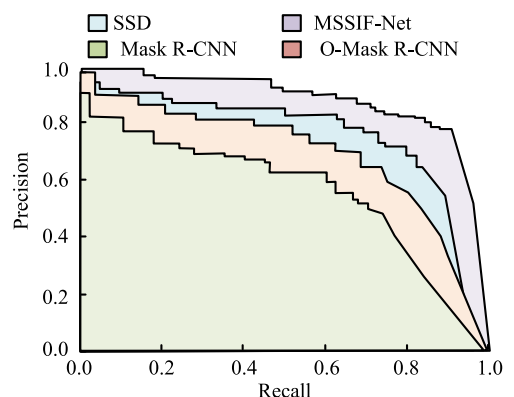
FIGURE 9. Error performance of different detection algorithms.

algorithm also increase. When the four algorithms approach a stable state, the stable MAE values are 2.48, 2.32, 1.90, and 1.21, respectively. As a tool for evaluating the classification performance of four types, the area enclosed below the PR curve can usually demonstrate the overall classification performance. Figure 10 shows the PR curves of four algorithms in different datasets.

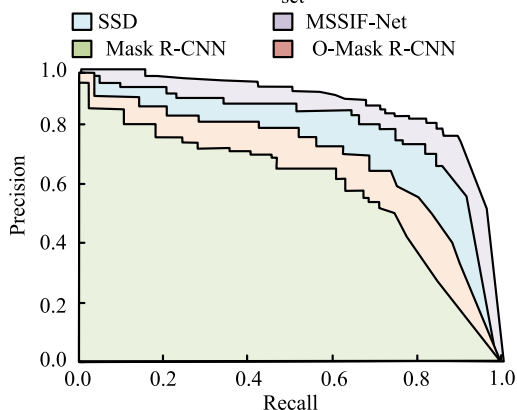
Figures 10 (a) and(b) show the PR curves of the four detection algorithms on the training and testing datasets, respectively. The O-Mask R-CNN has the highest Area Under the Curve (AUC) in both the training and testing datasets, showing that the algorithm has performed well and can accurately distinguish between defective and normal samples in the dataset.

Table 3 shows the average detection accuracy and false detection rates of the four algorithms in the training set and the test set. As can be seen from Table 3, the average detection accuracy rates of SSD, Mask R-CNN, MSSIF-Net and O-Mask R-CNN in the training set are 86.71%, 89.52%, 91.34% and 98.26%, respectively. The average false detection rates were 0.75%, 0.62%, 0.48% and 0.25%, respectively, while the average detection accuracy rates of the four algorithms in the test set were 86.24%, 89.39%, 91.06% and 98.15%, respectively, and the average false detection rates were 0.73%, 0.59%, 0.45% and 0.21%. In summary, O-Mask R-CNN has higher average detection accuracy and lower average detection false detection rate in training set and test set.





(a) PR curves of different algorithms in the training set



(b) PR curves of different algorithms in the test set

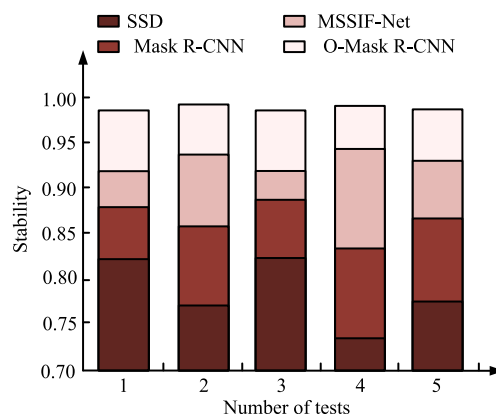
FIGURE 10. PR curves for different detection algorithms.

TABLE 3. Average detection accuracy and average false detection rate of the four algorithms.

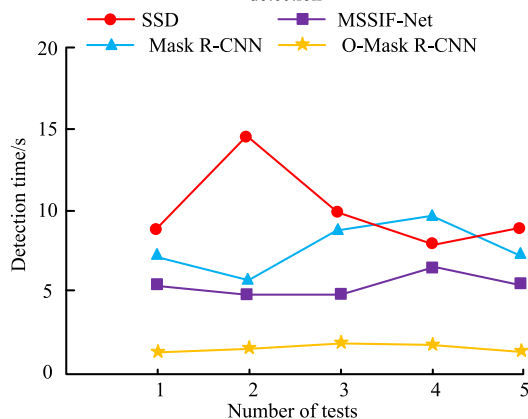
Network structure	The average accuracy of the training set	The average false detection rate of the training set	The average accuracy of the test set	The average false detection rate of the test set
SSD	86.71%	0.75%	86.24%	0.73%
Mask R-CNN	89.52%	0.62%	89.39%	0.59%
MSSIF-Net	91.34%	0.48%	91.06%	0.45%
O-Mask R-CNN	98.26%	0.25%	98.15%	0.21%

**B. ANALYSIS OF RFOD APPLICATION EFFECTIVENESS**

To further demonstrate the application effect of O-Mask R-CNN in practical problems, this study selects two different types of RFOIs as detection targets. One type is a normal RFOI, which is taken during the day and can accurately see various types of foreign objects on the railway. Another type is nighttime IRFOI, which displays the location of foreign objects in the form of infrared images. Firstly, four algorithms are used to detect two types of foreign object images. After a total of five tests, the stability and detection time of the four algorithms in the actual detection process are shown in Figure 11.



(a) Stability of different algorithms under multiple detection



(b) Detection time of different algorithms in multiple detections

FIGURE 11. Stability and time of different algorithms for multiple detection of foreign object images.

Figures 11 (a) and (b) show the stability and detection time of four methods in five tests. In Figure 11 (a), the stability of the four algorithms will change in all five detections. O-Mask R-CNN has the best stability and the smallest fluctuation amplitude among the five tests. The highest stability of each algorithm in five detections can reach 0.83, 0.88, 0.94, and 0.98, respectively. In Figure 11 (b), the detection time of O-Mask R-CNN in all five detections is within 5 seconds, while the detection time of SSD, Mask R-CNN, and MSSIF-Net fluctuates significantly. The shortest detection time for the four algorithms in five detections is 0.82s, 0.67s, 0.48s, and 0.25s, respectively. The missed detection rate and Mean Average Precision (mAP) are used as practical performance indicators for the four models. Table 4 shows the detection performance of four models in two image detection tasks.

Table 4 shows the missed detection rates and mAP of four algorithms for detecting two types of images. The detection performance of the four algorithms in normal RFOIs is better than that in IRFOI. Among them, the missed detection rate and mAP value of O-Mask R-CNN in normal RFOIs are 0.09% and 98.85%, respectively, and the missed detection rate and mAP value in IRFOI are 0.14% and 98.26%, respectively. Overall, O-Mask R-CNN performs better in both types

TABLE 4. Performance results of different algorithms in real foreign OD.

Detection Images	Network structure	Leakage/ %	mAP/ %
Normal railway foreign body images	SSD	0.69	85.37
	Mask R-CNN	0.45	90.36
	MSSIF-Net	0.18	94.21
	O-Mask R-CNN	0.09	98.85
Infrared railway foreign body image	SSD	0.73	82.21
	Mask R-CNN	0.58	87.14
	O-Mask R-CNN	0.22	90.05
	O-Mask R-CNN	0.14	98.26

of foreign object image detection. A set of physical images is selected to test the actual detection performance of four models on railway track images, as shown in Figure 12.

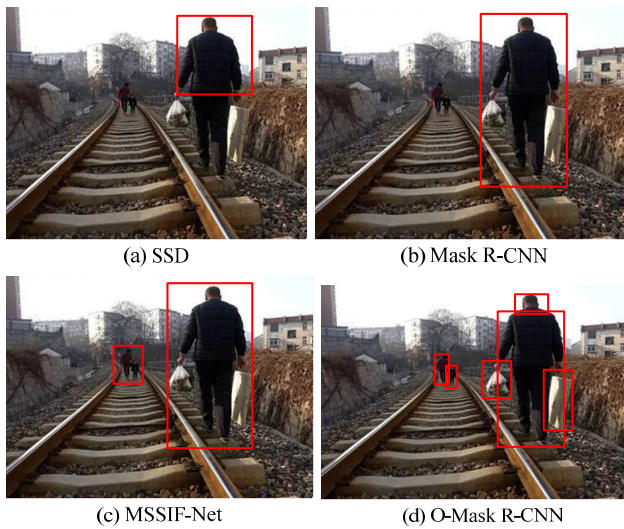


FIGURE 12. Effectiveness of different algorithms for detecting normal railway foreign body images.

Figures 12 (a) - (d) show the detection performance of SSD, Mask R-CNN, MSSIF-Net, and O-Mask R-CNN in normal RFOIs, respectively. The last one has the best detection performance, which can not only fully detect all foreign objects on the railway, but also perform differential detection on different foreign objects. As the detection target, IRFOI obtains the detection results of four models as shown in Figure 13.

Figures 13 (a) - (d) show the detection performance of SSD, Mask R-CNN, MSSIF-Net, and O-Mask R-CNN in IRFOI, respectively. The first three detection methods can only detect a rough framework of foreign objects, and cannot detect various foreign objects in a targeted and classified manner. O-Mask R-CNN can monitor foreign objects in all infrared images, and classify various foreign objects on railway tracks, making the detection model have higher detection accuracy and better detection performance.

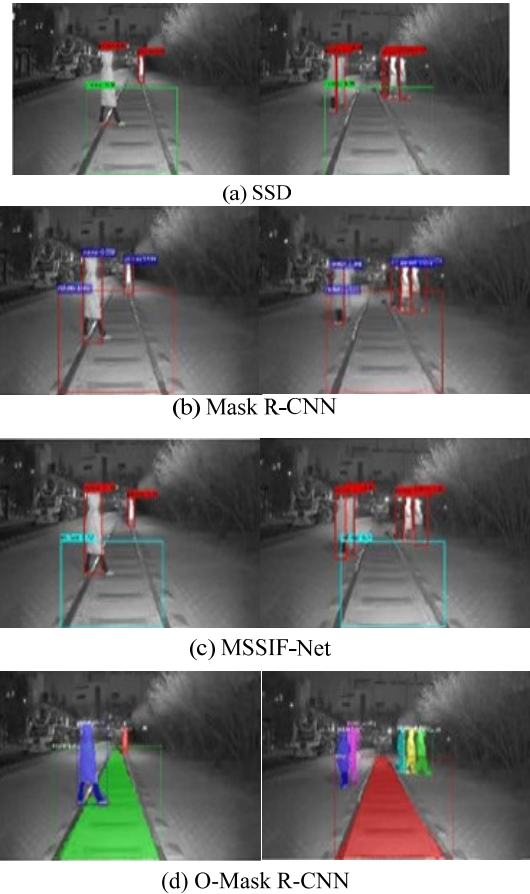


FIGURE 13. Effectiveness of different algorithms in detecting IRFOI.

TABLE 5. The advantages and limitations of the four algorithms.

Network structure	Advantage	limitation
SSD	It has faster processing speed and lower resource consumption	Poor performance in small target detection
Mask R-CNN	It has high precision case segmentation effect	The processing power in complex background is limited and the speed is slow.
MSSIF-Net	Multi-scale feature sharing enhances the adaptability and robustness of the network	The complexity of the model is high and the calculation is slow.
O-Mask R-CNN	The detection effect is better in infrared and low light environment, and the detection precision is higher for small targets	The training time of the model is longer

The advantages and limitations of the four comparison algorithms are described in Table 5. As can be seen from Table 5, although Mask R-CNN performs well in instance segmentation, it has limitations when dealing with complex backgrounds or rapid application scenarios. O-Mask R-CNN specifically optimizes the processing of low-light and infrared images, significantly improving detection

accuracy and small target recognition under these conditions. At the same time, SSD is more suitable for applications requiring real-time detection due to its fast processing ability, while MSIF-NET enhances the model's adaptability to scale changes by using multi-scale feature sharing technology, especially for railway monitoring in dynamic environments. In general, O-Mask R-CNN has better detection performance, and its detection effect is best in small targets and special environments.

## V. CONCLUSION

To improve the detection accuracy of traditional neural networks for RFOIs under infrared weak light conditions, this study optimized Mask R-CNN and proposed an O-Mask R-CNN detection algorithm. The data showed that Model 5 performed better than other network structures in ablation experiments on the main network part of O-Mask R-CNN. The accuracy value, recall value, F1 value, and network response time were 0.97, 0.98, 0.98, and 0.02 seconds, respectively. When testing the performance of O-Mask R-CNN, it was found that it can achieve stable total loss, segmentation loss, and regression loss values after 46, 59, and 41 iterations, respectively. It used far fewer iterations than SSD, Mask R-CNN, and MSSIF-Net. In addition, O-Mask R-CNN also exhibited good error performance, with MSE and MAE values as low as 1.35 and 1.21, respectively, far lower than Mask R-CNN's 2.97 and 2.32. Applying this algorithm to practical problems, the missed detection rate and mAP value of O-Mask R-CNN in normal RFOIs were 0.09% and 98.85%, respectively. Its missed detection rate and mAP value in IRFOI were 0.14% and 98.26%, respectively, and all indicators were better than the other three comparative algorithms. Finally, the actual performance of four algorithms in detecting normal RFOIs and IRFOI was tested. O-Mask R-CNN had the most detection boxes, indicating that its detection effect was the best.

Although this study has made significant progress in the detection of railway foreign bodies under infrared low illumination conditions, there are still some limitations and challenges. First, the computational efficiency of the model in processing large-scale data sets still needs to be improved, and more efficient network structures and optimization algorithms need to be combined to shorten the reasoning time of the model, achieve faster detection speed and meet the needs of real-time monitoring. Second, there is room for improvement in the stability and accuracy of the current model under changeable environmental conditions, such as rain and fog weather and strong light irradiation. Third, considering the complexity of railway foreign object detection, visual information alone is not enough to solve all the challenges at present, and future research will explore how to effectively integrate multiple sensor data, such as sound, temperature, etc., to provide more comprehensive monitoring results and enhance the comprehensive performance of the detection system.

## VI. ABBREVIATIONS

Mask R-CNN:	Region-based Convolutional Neural Network.
FPN:	Feature Pyramid Network.
CBAM:	Convolutional Block Attention Module.
RPN:	Region Proposal Network.
Faster R-CNN:	Faster Region-based Convolutional Neural Network.
CAM:	Channel Attention Mechanism.
SAM:	Spatial Attention Mechanism.
CBAM-ResNet-DCFPN:	Convolutional Block Attention Mechanisms-Residual Network-Density Connected FPN.
K-means:	K-means clustering algorithm.
IoU:	Intersection of Union.
IoU-K-means:	Intersection of Union-K-means clustering algorithm.
O-Mask R-CNN:	Optimized-Mask Region-based Convolutional Neural Network.
SSD:	Single Shot MultiBox Detector.
MSSIF-Net:	Multi-Scale Shared and Independent Feature Network.
MSE:	Mean Squared Error.
MAE:	Mean Absolute Error.
AUC:	Area Under the Curve.
mAP:	Mean Average Precision.

## REFERENCES

- [1] T. Mahmood and Z. Ali, "Analysis of Maclaurin symmetric mean operators for managing complex interval-valued q-rung orthopair fuzzy setting and their applications," *J. Comput. Cognit. Eng.*, vol. 2, no. 2, pp. 98–115, Apr. 2022, doi: [10.47852/bonviewjccce2202164](https://doi.org/10.47852/bonviewjccce2202164).
- [2] P. M. Blok, F. K. van Evert, A. P. M. Tielen, E. J. van Henten, and G. Kootstra, "The effect of data augmentation and network simplification on the image-based detection of broccoli heads with mask R-CNN," *J. Field Robot.*, vol. 38, no. 1, pp. 85–104, Jan. 2021, doi: [10.1002/rob.21975](https://doi.org/10.1002/rob.21975).
- [3] A. Podorozhniak, N. Liubchenko, M. Sobol, and D. Onishchenko, "Usage of mask R-CNN for automatic license plate recognition," *Adv. Inf. Syst.*, vol. 7, no. 1, pp. 54–58, Mar. 2023, doi: [10.20998/2522-9052.2023.1.09](https://doi.org/10.20998/2522-9052.2023.1.09).
- [4] J. Zhao, Y. Pan, H. Zhang, M. Lin, X. Luo, and Z. Xu, "InPlaceKV: In-place update scheme for SSD-based KV storage systems under update-intensive workloads," *Cluster Comput.*, vol. 27, no. 2, pp. 1527–1540, Apr. 2024, doi: [10.1007/s10586-023-04031-9](https://doi.org/10.1007/s10586-023-04031-9).
- [5] L. Zhang, Y. Hu, J. Chen, C. Li, and K. Li, "MSSIF-Net: An efficient CNN automatic detection method for freight train images," *Neural Comput. Appl.*, vol. 35, no. 9, pp. 6767–6785, Mar. 2023, doi: [10.1007/s00521-022-08035-1](https://doi.org/10.1007/s00521-022-08035-1).
- [6] X. Bi, J. Hu, B. Xiao, W. Li, and X. Gao, "IEMask R-CNN: Information-enhanced mask R-CNN," *IEEE Trans. Big Data*, vol. 9, no. 2, pp. 688–700, Apr. 2023, doi: [10.1109/TBDATA.2022.3187413](https://doi.org/10.1109/TBDATA.2022.3187413).
- [7] Y. Liu, G. Yang, Y. Huang, and Y. Yin, "SE-mask R-CNN: An improved mask R-CNN for apple detection and segmentation," *J. Intell. Fuzzy Syst.*, vol. 41, no. 6, pp. 6715–6725, Dec. 2021, doi: [10.3233/jifs-210597](https://doi.org/10.3233/jifs-210597).
- [8] F. Guo, Y. Qian, D. Rizos, Z. Suo, and X. Chen, "Automatic rail surface defects inspection based on mask R-CNN," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2675, no. 11, pp. 655–668, Jul. 2021, doi: [10.1177/03611981211019034](https://doi.org/10.1177/03611981211019034).

- [9] T. Ye, J. Zhang, Z. Zhao, and F. Zhou, "Foreign body detection in rail transit based on a multi-mode feature-enhanced convolutional neural network," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 10, pp. 18051–18063, Oct. 2022, doi: [10.1109/TITS.2022.3154751](https://doi.org/10.1109/TITS.2022.3154751).
- [10] R. Tian, H. Shi, B. Guo, and L. Zhu, "Multi-scale object detection for high-speed railway clearance intrusion," *Int. J. Speech Technol.*, vol. 52, no. 4, pp. 3511–3526, Mar. 2022, doi: [10.1007/s10489-021-02534-9](https://doi.org/10.1007/s10489-021-02534-9).
- [11] J. Zheng, L. Wang, J. Liu, H. Wang, S. Wang, L. Wang, and J. Zhang, "An inspection method of rail head surface defect via bimodal structured light sensors," *Int. J. Mach. Learn. Cybern.*, vol. 14, no. 5, pp. 1903–1920, May 2023, doi: [10.1007/s13042-022-01736-y](https://doi.org/10.1007/s13042-022-01736-y).
- [12] K.-M. Na, K. Lee, and H. Kim, "Condition monitoring of railway pantograph using R-CNN and image processing," *J. Electr. Eng. Technol.*, vol. 18, no. 3, pp. 2407–2416, May 2023, doi: [10.1007/s42835-022-01229-6](https://doi.org/10.1007/s42835-022-01229-6).
- [13] L. Xiao, W. Li, N. Deng, B. Yuan, Y. Bi, Y. Cui, and X. Cui, "Automatic pavement crack identification based on an improved C-mask region-based convolutional neural network model," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2677, no. 3, pp. 1194–1216, Mar. 2023, doi: [10.1177/03611981221122778](https://doi.org/10.1177/03611981221122778).
- [14] G. Hu, T. Wang, M. Wan, W. Bao, and W. Zeng, "UAV remote sensing monitoring of pine forest diseases based on improved mask R-CNN," *Int. J. Remote Sens.*, vol. 43, no. 4, pp. 1274–1305, Mar. 2022, doi: [10.1080/01431161.2022.2032455](https://doi.org/10.1080/01431161.2022.2032455).
- [15] T. Vu, T. Bao, Q. V. Hoang, C. Drenbenstetd, P. V. Hoa, and H. H. Thang, "Measuring blast fragmentation at nui phao open-pit mine, Vietnam using the mask R-CNN deep learning model," *Mining Technol.*, vol. 130, no. 4, pp. 232–243, Jun. 2021, doi: [10.1080/25726668.2021.1944458](https://doi.org/10.1080/25726668.2021.1944458).
- [16] J. Jiang, Y. Bie, J. Li, X. Yang, G. Ma, Y. Lu, and C. Zhang, "Fault diagnosis of the bushing infrared images based on mask R-CNN and improved PCNN joint algorithm," *High Voltage*, vol. 6, no. 1, pp. 116–124, Feb. 2021, doi: [10.1049/hve.2019.0249](https://doi.org/10.1049/hve.2019.0249).
- [17] M. Frei and F. E. Kruis, "Image-based analysis of dense particle mixtures via mask R-CNN," *Eng.*, vol. 3, no. 1, pp. 78–98, Jan. 2022, doi: [10.3390/eng3010007](https://doi.org/10.3390/eng3010007).
- [18] D. Tiede, G. Schwendemann, A. Alobaidi, L. Wendt, and S. Lang, "Mask R-CNN-based building extraction from VHR satellite data in operational humanitarian action: An example related to COVID-19 response in Khartoum, Sudan," *Trans. GIS*, vol. 25, no. 3, pp. 1213–1227, May 2021, doi: [10.1111/tgis.12766](https://doi.org/10.1111/tgis.12766).
- [19] S. Moccia, M. C. Fiorentino, and E. Frontoni, "Mask-R<sup>2</sup>CNN: A distance-field regression version of mask-RCNN for fetal-head delineation in ultrasound images," *Int. J. Comput. Assist. Radiol. Surgery*, vol. 16, no. 10, pp. 1711–1718, Jun. 2021, doi: [10.1007/s11548-021-02430-0](https://doi.org/10.1007/s11548-021-02430-0).
- [20] M.-C. Chiu and T.-M. Chen, "Applying data augmentation and mask R-CNN-based instance segmentation method for mixed-type wafer maps defect patterns classification," *IEEE Trans. Semicond. Manuf.*, vol. 34, no. 4, pp. 455–463, Nov. 2021, doi: [10.1109/TSM.2021.3118922](https://doi.org/10.1109/TSM.2021.3118922).
- [21] A. Saveliev, E. Aksamentov, and E. Karasev, "Automated terrain mapping based on mask R-CNN neural network," *Int. J. Intell. Unmanned Syst.*, vol. 10, no. 2/3, pp. 267–277, Mar. 2022, doi: [10.1108/ijius-11-2019-0066](https://doi.org/10.1108/ijius-11-2019-0066).
- [22] F. Zhang, D. Zhao, Z. Xiao, J. Wu, L. Geng, W. Wang, and Y. Liu, "Rodlike nanoparticle parameter measurement method based on improved mask R-CNN segmentation," *Signal, Image Video Process.*, vol. 15, no. 3, pp. 579–587, Apr. 2021, doi: [10.1007/s11760-020-01779-0](https://doi.org/10.1007/s11760-020-01779-0).
- [23] U. Gawande, K. Hajari, and Y. Golhar, "SIRA: Scale illumination rotation affine invariant mask R-CNN for pedestrian detection," *Appl. Intell.*, vol. 52, no. 9, pp. 10398–10416, Jan. 2022, doi: [10.1007/s10489-021-03073-z](https://doi.org/10.1007/s10489-021-03073-z).
- [24] L. Petrucci, F. Ricci, R. Martinelli, and F. Mariani, "Detecting the flame front evolution in spark-ignition engine under lean condition using the mask R-CNN approach," *Vehicles*, vol. 4, no. 4, pp. 978–995, Sep. 2022, doi: [10.3390/vehicles4040053](https://doi.org/10.3390/vehicles4040053).
- [25] A. Droby, B. Kurar Barakat, R. Alaasam, B. Madi, I. Rabaev, and J. El-Sana, "Text line extraction in historical documents using mask R-CNN," *Signals*, vol. 3, no. 3, pp. 535–549, Aug. 2022, doi: [10.3390/signals3030032](https://doi.org/10.3390/signals3030032).
- [26] X. Xu, C. Li, X. Fan, X. Lan, X. Lu, X. Ye, and T. Wu, "Attention mask R-CNN with edge refinement algorithm for identifying circulating genetically abnormal cells," *Cytometry A*, vol. 103, no. 3, pp. 227–239, Mar. 2023, doi: [10.1002/cyto.a.24682](https://doi.org/10.1002/cyto.a.24682).
- [27] F. Bagheri, M. J. Tarokh, and M. Ziaratban, "Skin lesion segmentation based on mask RCNN, multi atrous full-CNN, and a geodesic method," *Int. J. Imag. Syst. Technol.*, vol. 31, no. 3, pp. 1609–1624, Mar. 2021, doi: [10.1002/ima.22561](https://doi.org/10.1002/ima.22561).
- [28] Q. Wang, T. Gao, Q. He, Y. Liu, J. Wu, and P. Wang, "Severe rail wear detection with rail running band images," *Comput.-Aided Civil Infrastruct. Eng.*, vol. 38, no. 9, pp. 1162–1180, Jun. 2023, doi: [10.1111/mice.12948](https://doi.org/10.1111/mice.12948).
- [29] Z. Ma, Y. Li, M. Huang, Q. Huang, J. Cheng, and S. Tang, "Automated real-time detection of surface defects in manufacturing processes of aluminum alloy strip using a lightweight network architecture," *J. Intell. Manuf.*, vol. 34, no. 5, pp. 2431–2447, Jun. 2023, doi: [10.1007/s10845-022-01930-3](https://doi.org/10.1007/s10845-022-01930-3).



**SHUANG GAO** was born in Heze, Shandong, Han Nationality, in 1987. She received the B.E. degree in transportation from Shandong Jiaotong University, in 2011, and the M.E. degree in application engineering of transport vehicles from Shijiazhuang Tiedao University, in 2014. She is currently a full-time Teacher in the Railway Department, Hohhot Vocational College. She has published nine academic papers, three textbooks, six research projects, and two patents.

...