IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 12, 2021, accepted June 3, 2021, date of publication June 9, 2021, date of current version June 23, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3087705

Modeling and Optimization of Semantic Segmentation for Track Bed Foreign Object Based on Attention Mechanism

HAORAN SONG[1], SHENGCHUN WANG[2], ZICHEN GU[3], PENG DAI[2], XINYU DU[2], AND YU CHENG[2]

[1]China Academy of Railway Sciences, Beijing 100081, China
[2]Infrastructure Inspection Research Institute, China Academy of Railway Sciences Corporation Ltd., Beijing 100081, China
[3]INMAI Railway Technology Company Ltd., Beijing 100081, China

Corresponding author: Peng Dai (daipeng_iic@qq.com)

This work was supported by the Scientific Research Projects of China Academy of Railway Sciences under Grant 2019YJ158.

ABSTRACT The problem of foreign object intrusion onto the track bed often occurs in the actual operation process of high-speed railways. To solve the problem, we propose an anomaly detection method for the ballastless track bed, which is based on semantic segmentation. Firstly, we put forward the RFODLab semantic segmentation network according to the randomness of foreign objects distribution, and a small proportion of target pixels in the track image. The segmentation results of track image obtained through this model can be used to obtain the accurate pixel information of foreign objects. To further improve the recall and precision, the channel attention mechanism is introduced for the backbone network of the model to aggregate the context information of images, which achieves the weighted constraints of the model on the area to be recognized. Furthermore, to improve the model performance affected by unbalanced sample category distribution during the anomaly detection, we modify the loss function by balancing distribution of each category. The experimental results show that our proposed method can effectively segment various types of anomalies on the ballastless track bed including broken elastic strips, animal carcasses, and fallen pieces. The precision of anomaly detection on the test set can reach 90% while the recall can be maintained at more than 95%. The anomaly detection results on actual lines also verify the effectiveness of the method.

INDEX TERMS Foreign object, railway safety, deep learning, anomaly detection, semantic segmentation, attention mechanism, loss function.

I. INTRODUCTION

China's railway lines feature the long mileage, large spatial span, and complex and changeable conditions, thus raising high requirements for the efficient operation and maintenance of railway infrastructures. Due to the strong airflow generated by the train running at a high speed, the broken parts or foreign abnormal objects that are very likely to appear beside the ballastless track bed might collide with the train body. If so, these foreign objectswill cause structural damage and serious safety hazards to the train. Thus, detection and removal of foreign objects on the track bed in an accurate and efficient manner is required to ensure the safe train operation. At present, track bed anomalies are detected mainly through manual inspection. However, this method has such disadvantages as low detection efficiency, large influence by human factors, and massive missed detection. Therefore, more effective technical means are needed for detection of foreign objects on the ballastless track bed.

We design to continuously collect image information of rails and their vicinity through specialized inspection trains, so as to identify foreign objects in a timely manner according to such information. The ballastless track bed images collected through the vehicle-mounted dynamic detection method are specific in shape and pattern, but foreign objects on the ballastless track bed usually appear randomly in small quantities. In other words, few foreign objects on the ballastless track bed need to be identified according to a large number of normal track images. The key point is to distinguish foreign object that does not belong to the inherent track facilities from a large number of normal track images. Therefore, this problem belongs to the category of anomaly

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

detection. To be specific, it is necessary to identify the location and shape of any foreign object (if any) shown in the images. Thus, higher requirements are raised for final anomaly detection.

However, it is really hard to find out foreign objects on the ballastless track bed because some track bed facilities really resemble certain foreign objects. Thus, a false alarm might be given during anomaly detection. In addition, foreign objects usually occupy a low proportion in the collected track image samples, and there exists the problem of unbalanced distribution in the category of image samples.

In order to solve the above problems, we mainly study the semantic segmentation network for anomaly detection of the ballastless track bed, and optimize the network according to the image characteristics of ballastless track. Specifically, we propose to develop the semantic segmentation network RFODLab (Railway Foreign Object Detection Lab) through which the pixel-level information of foreign objects on the track bed can be extracted out of track images. In addition, the mask generated by the network can be filtered according to the contour area to avoid false alarms, and then the pixel information of foreign objects can be obtained. Moreover, the attention mechanism is introduced for the backbone network of the model, to aggregate the context information of images. The loss function combining Focal Loss and Dice Loss is also developed to achieve a balance in sample category.

Our innovative contributions are summarized as follows:

1. We put forward a semantic segmentation network for anomaly detection of the ballastless track bed through the semantic segmentation network RFODLab. This method makes it possible to accurately distinguish the foreign objects from the image background, and enables the anomaly segmentation of the ballastless track bed at the pixel level.

2. The attention mechanism is introduced for the backbone part of the RFODLab network, to aggregate the context information of images through the semantic segmentation network. The model imposes the weighted constraints on the images to be identified, and further improves the segmentation fineness. Most importantly, the model makes it possible to solve the problem of a large number of false alarms caused by the great similarity between some foreign objects and certain existing facilities of the ballastless track bed.

3. The loss function combining Focal Loss and Dice Loss is adopted for the RFODLab network. Through this loss function, a balance is achieved in the category proportion of foreign objects and backgrounds, and the problem of unbalanced sample category distribution caused by the excessive proportion of backgrounds in track images is solved to a certain extent.

II. RELATED WORK

Foreign objects on the track bed bring great potential dangers to the safe operation of railways. To detect foreign objects, multiple non-contact technologies have been developed and applied, including ultrasound, radar, and infrared technologies. For example, Fernando J. Alvarez *et al.* used the ultrasonic detector to detect foreign objects [1]. A. Mroue' *et al.* put forward a method for detecting the platform track area through the radar UWB (Ultra-Wideband) transmission technology [2]. Dhiraj Sinha *et al.* used the track accelerometer to detect any vibrations caused by falling foreign objects at specific locations [3]. Juan Jesús García *et al.* proposed to detect the foreign objects on the track by using multiple sensors such as infrared and ultrasonic sensors [4].

The aforesaid methods are all effective to detect the foreign objects fallen on the track bed in specific areas, with the desired results achieved. However, such fixed-point monitoring methods are costly and hard to practice. What's worse, they are limited in the detection range, and only applicable to the long-term fixed-point monitoring of certain key areas of railway. Compared with the fixed-point monitoring methods, the method of vehicle-mounted dynamic detection based on the computer vision technology (hereinafter referred to as ''the computer vision method'') is characterized by large range and low cost of detection, and enables the large-scale detection of track bed state. Therefore, the computer vision method is more ideal for the detection of foreign objects on the track bed along the whole railway line [5]. In recent years, it as an effective method for anomaly detection has been widely applied in many fields, with remarkable results achieved. Liupeng Jiang *et al.* used the improved Canny operator edge detection (image edge detection) method to identify the foreign objects in the port transportation channels [6]. Haoyu Xu *et al.* proposed to identify the possible foreign objects on the airport runway by making use of the deep convolutional network [7]. Jinguo Zhu suggested detecting the foreign objects along the electric transmission lines by using the deep learning method based on the regression strategy [8]. Deqiang He put forward a object detection method combining SSD and MobileNet, which is favorable in the detection of foreign objects under the high-speed trains [9]. Hongxia Niu *et al.* suggested the rapid detection of foreign objects by using the method of background modeling and pixel differencing [10]. Baoqing Guo *et al.* came up with a method for the detection of intrusive objects according to the YOLO v3 object detection network and the images of pedestrians and animals in railway areas generated by using GAN [11]. Wang Shengchun *et al.* proposed to detect foreign objects on the track bed in the high-speed railway running scenario by using a variety of object detection techniques based on deep learning [12].

At present, the methods of anomaly detection based on the computer vision technology have been developing rapidly, and been gradually applied for industrial detection. The methods of object detection, image segmentation, generative adversarial network, meta-learning and others have been put forward in succession [13].Dong Gong *et al.* designed the autoencoder named MemAE which is augmented with a memory module, to strengthen anomaly detection by amplifying the error between the abnormal reconstruction sample

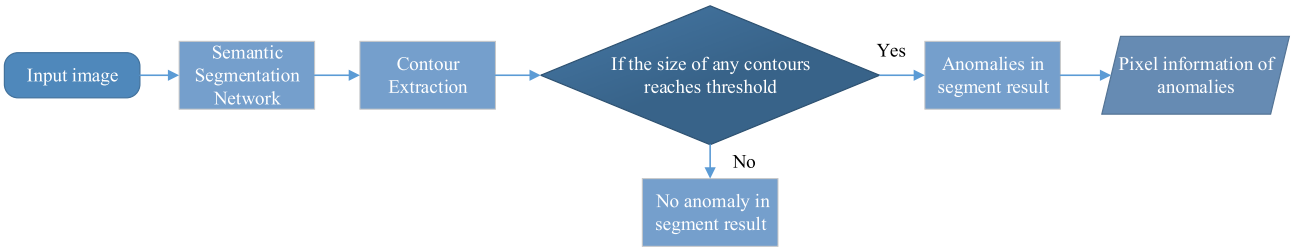


FIGURE 1. Anomaly detection algorithm of the ballastless track bed based on semantic segmentation.

and the original image [14].Hyunjong Park *et al.* proposed to introduce the memory module with a new update scheme into the convolutional neural network for anomaly detection, and train the memory according to new feature compactness and separateness losses [15]. Muhammad Zaigham Zaheer *et al.* presented a method of anomaly detection based on the images reconstructed by GAN. The approach uses the discriminator of GAN to distinguish the quality of the reconstructed images, so the discriminator can detect subtle distortions that often appear during the reconstruction of the anomaly inputs [16]. Jihun Yi *et al.* put forward a method of anomaly detection and anomaly segmentation through the deep learning variant based on support vector data description (SVDD) at the patch level [17]. In addition, Shashanka Venkataramana et al design a Convolutional Adversarial Variational Autoencoder with Guided Attention (CAVGA) to locate anomalies and preserve spatial information through the convolution latent variable [18]. Yingda Xia *et al.* came up with a method of anomaly detection through comparing the original image with the image reconstructed according to the results of semantic segmentation [19].

Traditional techniques of image processing and object detection are usually adopted for anomaly detection of the ballastless track beds of high-speed railways. However, anomaly detection results are mostly at the bounding box level (relatively rough), so that it is hard to distinguish the foreign objects accurately, and is likely to cause false alarms. In contrast, the method of anomaly detection based on the semantic segmentation technology is favorable and characterized by image classification in pixel level. The semantic segmentation technology makes it possible to distinguish the foreign objects on the ballastless track bed from the surrounding background and locate the anomalies by pixel, thus significantly improving the fineness of anomaly detection. Compared to the method of anomaly detection based on semi-supervised / unsupervised semantic segmentation, our method based on fully-supervised semantic segmentation enables more effective and precise detection because a certain amount of foreign object data has been accumulated during previous researches, and clear data label information is available.

III. METHODOLOGY

A. PRELIMINARY

The algorithm for anomaly detection of track bed based on semantic segmentation is detailed as follows. Foreign objects on the ballastless track bed are segmented and extracted through the semantic segmentation network, and then accurately identified and located according to the segmentation results. The detection process is shown in Figure 1.

As shown in Figure 1, the algorithm of anomaly detection of track bed based on the semantic segmentation according to the track image $I(x, y)$ is described as follows:

1) Obtain the mask image from the input image by using the semantic segmentation network, which is formulated as:

$$seg\left(I\left(x,y\right)\right) \rightarrow I_m\left(x,y\right), \tag{1}$$

where: $seg()$ represents the semantic segmentation network for the detection of foreign objects on the ballastless track bed; $I_m(x, y)$ denotes the mask image obtained after segmentation.

2) Conduct contour extraction of the mask image, which is formulated as:

$$ce\left(I_m\left(x,y\right)\right) \rightarrow \bigcup_{n=1}^{N_{ini}} \bigcup_{x=I_b}^{I_s} \bigcup_{y=J_b}^{J_s} I'_m\left(n,x,y\right), \tag{2}$$

where: $ce()$ represents the contour extraction algorithm; N_{ini} denotes the number of contour areas initially extracted; $I'_m(n, x, y)$ refers to the mask image for the n^{th} contour area of segmentation, which satisfies the condition of $I_m{}'(n, x, y) \subseteq I_m(x, y)$; $[I_b, I_s]$ and $[J_b, J_s]$ indicates the pixel coordinate range of the mask contour area.

3) Compare the extracted contour size with the contour threshold, which is formulated as

$$filt\left(\bigcup_{n=1}^{N_{ini}} \bigcup_{x=I_b}^{I_s} \bigcup_{y=J_b}^{J_s} I'_m\left(n,x,y\right)\right) \xrightarrow{threshold} \bigcup_{n=1}^{N_{large}} \bigcup_{x=I'_b}^{I'_s} \bigcup_{y=J'_b}^{J'_s} I''_m\left(n,x,y\right), \tag{3}$$

where: $filt()$ represents the contour area filtering algorithm; $threshold$ refers to the contour area threshold; N_{large} denotes the number of contour areas higher than the threshold; $I_m{}''(n, x, y)$ indicates the mask image for the nth contour area higher than the threshold, which satisfies the condition of $I_m{}''(n, x, y) \subseteq I_m{}'(x, y)$; $[I_b{}', I_s{}']and[J_b{}', J_s{}']$ represents the pixel coordinate range of the mask contour area greater than the threshold, which satisfies the condition of $[I_b{}', I_s{}'] \subseteq [I_b, I_s]$ and $[J_b{}', J_s{}'] \subseteq [J_b, J_s]$.

4) Obtain the contour pixel information after confirming the contour size, which is formulated as:

$$loc\left(\cup_{n=1}^{N_{large}} \cup_{x=I_b}^{I_s} U_{y=J_b}^{J_s} I'_m(n,x,y)\right) \rightarrow \cup_{n=1}^{N_{large}} \cup_{x=I_b}^{I_s} \cup_{y=J_b}^{J_s} pix(n,x,y), \tag{4}$$

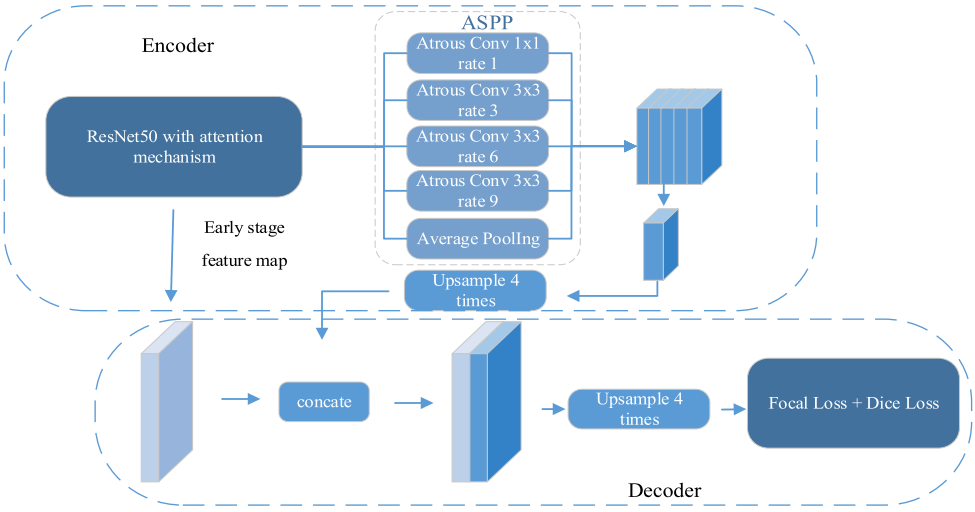


FIGURE 2. Overall framework of RFODLab.

where: $loc()$ represents the algorithm of contour pixel information localization; $pix(n, x, y)$ denotes the pixel information of the nth contour area higher than the threshold.

Whereas the foreign objects on the ballastless track appear randomly, and foreign object images occupy a small proportion, the ballastless track images should be firstly introduced to the semantic segmentation network to obtain mask images. The mask contour will be searched and extracted to identify any possible foreign objects. In order to prevent the almost negligible sporadic pixel interference in the anomaly detection due to the similarity between foreign objects and track facility components, the area greater than the contour area threshold should be delimited by comparing the extracted contour size with the contour area threshold. If the contour area is greater than the threshold, it is deemed that there exist foreign objects, and the pixel-level information of foreign objects will be returned. Otherwise, it is judged that no anomaly is detected.

B. RFODLab SEMANTIC SEGMENTATION NETWORK

For more effective anomaly detection, higher requirements are raised for the efficiency of segmenting and extracting foreign objects through the semantic segmentation network. The segmentation effect of the model directly affects the possibility of finding out foreign objects. We propose to establish the semantic segmentation network named RFODLab according to the characteristics of images on foreign objects on the ballastless track beds of high-speed railways.

RFODLab adopts the encoder-decoder structure similar to that of DeepLab v3+ [20] network. The encoder is composed of the backbone network and ASPP module. For the backbone network, the features are extracted through the ResNet 50[21] network with the channel attention mechanism, and the ASPP module behind the backbone network structure is used to extract multi-scale features of the feature map outputted by the backbone network. In addition, up-sampling and feature fusion operations are performed for the decoder. The loss function combining Focal Loss [22] and Dice Loss [23] is adopted. The structure of the RFODLab network is shown in Figure 2.

After the feature information of the encoder is extracted from the given pictures, the feature map will be resized to fit the original image through the decoder to obtain segmentation results. The encoder-decoder structure avoids the loss of accuracy caused by the massive image upsampling, and improves the edge accuracy of semantic segmentation, so that foreign objects can be distinguished from the surrounding environment to the greatest extent, thus reducing false alarms.

The foreign object images inputted in the encoding process are processed through the backbone network with the channel attention mechanism to obtain corresponding feature maps, some of which (one layer) will be inputted in the subsequent decoding process to realize the effective fusion of low-layer feature location information and high-layer feature semantic information. The feature maps obtained through the convolutional layer are processed via the ASPP module to obtain the features of different scales, so as to combine a wide range of contextual semantic information. Because foreign objects occupy a small proportion in the image area, and the excessively large receptive field plays a limited role in improving the effect of feature extraction, the atrous rates of atrous convolution for the ASPP module are set as 1, 3, 6, and 9.

In the decoding process, the final feature map obtained through encoding process is subject to channel dimension concatenation firstly, and then upsampling by 4 times. Final feature maps are connected with the previous feature maps outputted by the convolutional network from the channel dimension, and then subject to upsampling by 4 times, followed by the calculation through the loss function combining Focal Loss and Dice Loss.

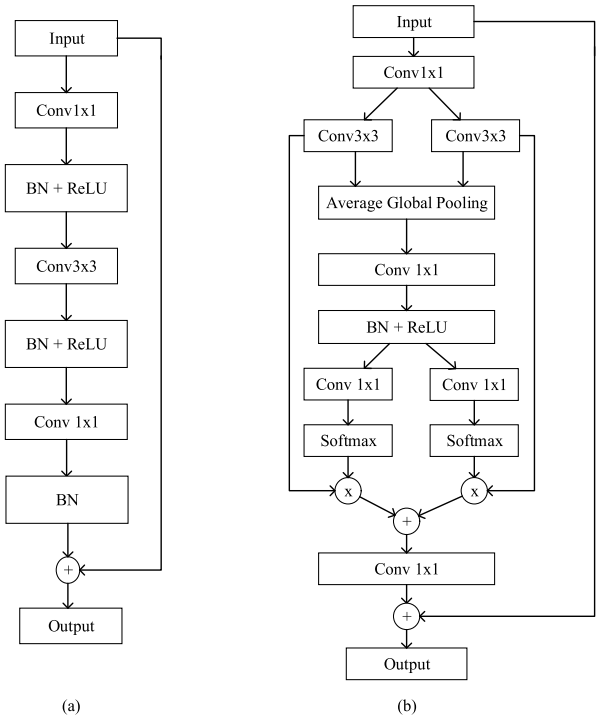


FIGURE 3. RFOD backbone network block. (a) Original ResNet block; (b) ResNet block with channel attention mechanism.

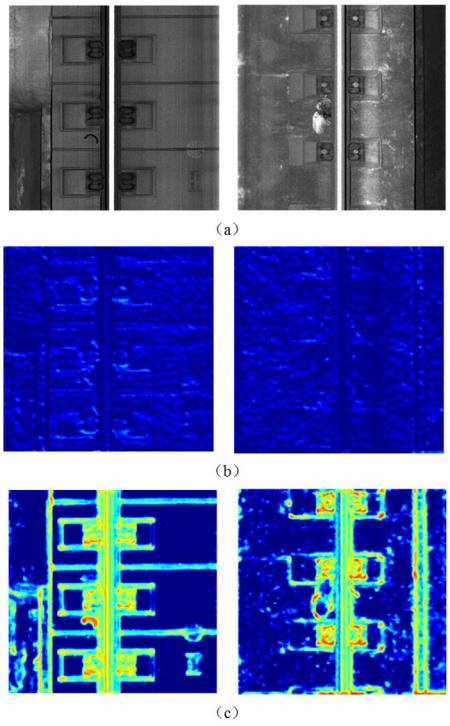


FIGURE 4. Activation heat map of output feature map processed by Grad-CAM (a) input image; (b) Activation heat map of output feature map by original ResNet 50 backbone network; (c) Activation heat of output feature map by ResNet 50 backbone network with attention mechanism.

C. CHANNEL ATTENTION MECHANISM

In order to further improve the semantic segmentation network's capacity to extract the features of foreign objects on the track bed, the channel attention mechanism is introduced for the backbone network (ResNet 50) of the model. The foreign objects on the track bed look very small in the corresponding images collected. To be specific, the size of a foreign object is only about 0.045-0.08 of the entire image size in length or width, with a relatively small pixel occupancy. Because foreign objects appear randomly, very high requirements are raised for the model's capacity to extract effective features of images. The attention mechanism based on the characteristics of human attention can be deemed as a kind of adaptive pooling of the model. The greater weight can be allocated to the specific location in an image according to the contextual semantic information captured. After the introduction of the attention mechanism, more weight can be allocated to the area whose features are similar to those of foreign objects, to detect foreign objects more effectively.

The two-branch channel attention mechanism is introduced for the original backbone network of ResNet [24]. The ResNet module structure before and after the improvement is shown in Figure 3.

The feature map obtained after the feature fusion through the RFODLab's ASPP module is visualized via the Grad-CAM [25] operation, so as to analyze the area more concerned by the RFODLab network. The obtained activation heat map is shown in Figure 4. The redder the area is, the more sensitive the model is.

When extracting features, the backbone network with the introduced attention mechanism can effectively capture the contextual semantic information of track images, adaptively allocate larger weights for the pixels that may be recognized as foreign objects, and inhibit the areas around the image pixels whose features are more similar to those of the background. As shown in the activation heat map, the model with the attention mechanism effectively integrates the relevant features of channel dimension mapping at the time of feature fusion, and adaptively allocates the weights according to the context information, allocating more weight for the areas around the pixels whose features are similar to those of foreign objects. The results of subsequent anomaly detection show that, the model can effectively distinguish the inherent infrastructure of the track bed that is allocated with a higher weight according to the extracted context information of images, thus avoiding the false alarms about foreign objects.

D. LOSS FUNCTION

The semantic segmentation network classifies each pixel usually through the cross-entropy loss function, and then averages all pixels. In essence, it is a process of learning of the same importance for each image pixel, which, however, might give rise to a problem. Specifically, if multiple categories of image pixels are unbalanced in representation, the training will be dominated by the category of pixels with the largest proportion.

At present, we have summarized foreign object datasets manually labeled by pixel category. There are 9,580,666 pixels labeled as foreign objects, accounting for only 0.27% of the total 3,599,373,056 pixels. In addition, foreign object image samples are significantly imbalanced in category, and there are far more background pixels than foreign object pixels. Due to the cumulative effect, background losses take up a large proportion, while foreign object losses occupy a small proportion. In the training process, the model is more inclined to learn the background, which reduces the network's capacity to extract foreground objects, and affects the model's performance in detecting foreign objects on the track bed. In this context, we adjust the loss function for the model, adopting the loss function combining Focal Loss and Dice Loss.

This loss function has the characteristics of both Focal loss and Dice loss. Due to the reduction of loss function weights for easy-to-classify samples through the Focal loss, the model can focus more on difficult-to-classify samples during training, which can effectively lower the large proportion of easy-to-classify background pixels in the loss function. For the Dice loss, all the pixels of the same category can be regarded as a whole, and the proportion of the intersection will be calculated. Not subject to the influence by a large number of background pixels of the ballastless track bed, the Dice loss is applicable to anomaly detection of the ballastless track bed in case of serious imbalance between foreground and background samples. In addition, desired results are achieved in the segmentation of small targets through the combination of Focal Loss and Dice Loss [26]. This means the loss function is an ideal choice for detection of small and medium-sized foreign objects on the ballastless track bed.

The formula of the loss function combining Focal Loss and Dice Loss is as follows:

$$TP_p(c) = \sum_{n=1}^{N} p_n(c)\, g_n(c), \tag{5}$$

$$FN_p(c) = \sum_{n=1}^{N} (1 - p_n(c)) g_n(c), \tag{6}$$

$$FP_p(c) = \sum_{n=1}^{N} p_n(c)\,(1 - g_n(c)), \tag{7}$$

$$\begin{aligned} L &= L_{Dice} + \lambda L_{Focal} \\ &= C - \sum_{c=0}^{C-1} \frac{TP_p(c)}{TP_p(c) + \alpha FN_p(c) + \beta FP_p(c)} \\ &\quad -\lambda \frac{1}{N} \sum_{c=0}^{C-1} \sum_{n=1}^{N} g_n(c)\,(1 - p_n(c))^2 \\ &\quad \times log((p_n(c)), \end{aligned} \tag{8}$$

where: $TP_p(c)$, $FN_p(c)$, $FP_p(c)$ respectively represent the true positivity, false negativity, and false positivity of category c; $p_n(c)$ denotes the prediction probability of pixel n; $g_n(c)$ refers to the possibility of pixel n as that of category c (0 indicates that pixel n belongs to the category c, while 1 means that pixel n does not belong to the category c); λ is used to balance the proportions of Dice Loss and Focal Loss; α and β are used to balance the proportions of false negativity and false positivity.

IV. EXPERIMENTS

A. DATASET AND PREPROCESSING

The datasets used herein are collected from the images of track components of high-speed railways collected by the integrated inspection trains through the non-contact dynamic detection method. Because there are very few foreign objects along the actual railway lines, the accumulated data of foreign objects are relatively limited. We select 1,217 ballastless track images in which partial foreign objects and complex track bed backgrounds are included, mainly including broken elastic strips, fallen rocks, animal carcasses, fallen wastes, scattered debris, etc. In fact, the situation reflected in such ballastless track images is in line with the distribution of foreign objects along the actual railway lines. To be specific, there are 675, 63, 84, 156, 167, 21 and 51 images about broken elastic strips, fallen rocks, animal carcasses, fallen wastes, scattered debris, other foreign objects, and complex backgrounds, respectively. Among all foreign objects, broken elastic strips occupy a large proportion, mainly because a large number of rail fasteners are used, some of which might be broken. In addition, there are 906 training images and 311 testing images (including 212 foreign objects). At present, foreign objects shown in the datasets are limited in category, so all foreign objects are classified as those of the same category. The data are labeled through the Labelme labeling toolbox, and stored at the VOC 2012 format. The dataset images and labeled images of foreign objects on the ballastless track bed are shown in Figure 5.

The hardware configuration of the experimental computing platform is as follows: Intel Xeon@2.4GHz CPU, Nvidia Geforce TitanX GPU × 4 and 256GB memory. The software configuration is Ubuntu 16.04, Pytorch 1.3.1, Python 3.7, and CUDA 10.1.

Because the railway line environments are complex and diverse, and the illumination conditions at the time of image collection are different, the model needs to overcome the interference of image data by different conditions, having strong robustness. In order to extend the datasets and enhance the robustness of the anomaly detection method, some preprocessing operations are performed in the training process of the semantic segmentation network, including random image resizing, random cropping, random flipping, and image measurement distortion. All these preprocessing operations aim to improve the model's anomaly detection capacity under different conditions in the training process.

B. MODEL TRAINING

In order to accelerate the convergence in the training process, the backbone network model pre-trained on the ImageNet [27] dataset is adopted. 64 images are set for each training batch, and iterative training is performed for 40,000 times in total. The Online Hard Example Mining (OHEM) [28] method and the pixels with a confidence coefficient of less than 0.7 (particularly the difficult-to-classify

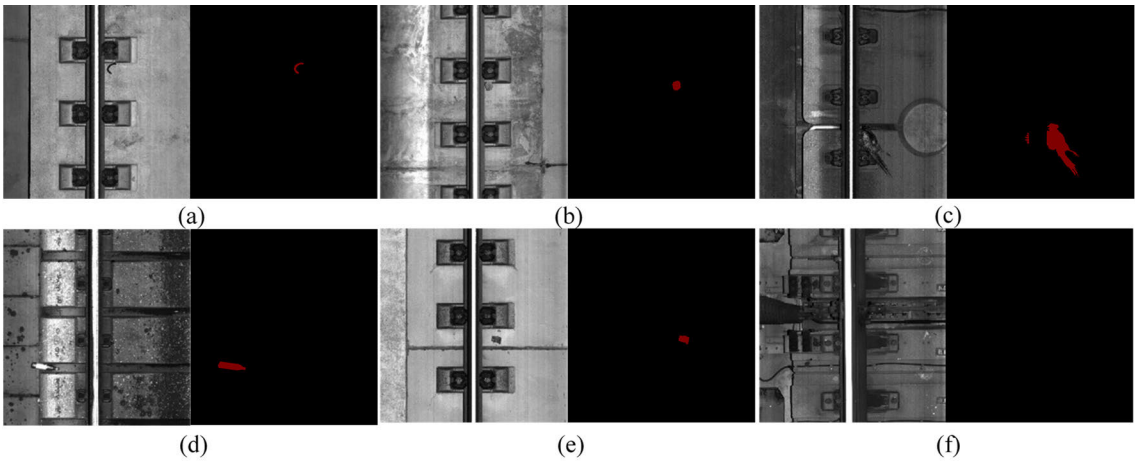


FIGURE 5. Image and annotation of foreign object data set on ballastless track bed (a) Broken elastic bar (b) Falling rocks (c) Animal carcasses (d) Falling garbage (e) Scattered debris (f) Complex background.

pixels) are adopted for the model training, to improve the semantic segmentation effect.

C. EVALUATION INDICATOR

For the anomaly detection model, such indicators as recall and precision are used to evaluate the effect of detecting foreign objects on the ballastless track bed. At the same time, *mIOU* indicator is introduced to assess the segmentation effect of the semantic segmentation network, so as to compare the model's capacity to extract the features of foreign objects. The recall, precision, and *mIOU* calculation formulas are shown below.

$$Recall = \frac{TP}{TP + FN} \times 100\% \tag{9}$$

$$Precision = \frac{TP}{TP + FP} \times 100\% \tag{10}$$

where: *TP* is the number of true foreign objects correctly identified, *FN* represents the number of true foreign objects unidentified, and *FP* indicates the number of articles wrongly identified as foreign objects.

$$mIOU = \frac{1}{2} \sum\nolimits_{i=0}^{1} \frac{p_{ii}}{\sum_{j=0}^{1} p_{ij} + \sum_{j=0}^{1} p_{ji} - p_{ii}} \tag{11}$$

where: p_{ij} represents the number of pixels whose true value is i and is predicted as j.

D. RESULTS AND ANALYSIS

311 images (including 212 foreign objects) extracted from the datasets are selected to evaluate the performance of the semantic segmentation network in anomaly detection. Firstly, the confidence thresholds of pixels for the RFODLab semantic segmentation network are set as 0.3, 0.5 and 0.7, respectively. The experimental results are detailed in Table 1.

When the confidence threshold is set to 0.5 (rather than 0.3 or 0.7), the RFODLab semantic segmentation network achieves a good balance in terms of recall and precision. The confidence threshold is ultimately set as 0.5.

TABLE 1. Performance of RFODLab Network with different confidence thresholds on test set.

Confidence threshold	Foreign objects number	TP	FP	Recall	Precision
0.3		207	52	97.64%	79.92%
0.5	212	205	19	96.70%	91.52%
0.7		173	11	81.60%	94.02%

The segmentation results of images on foreign objects on the track bed shown when the threshold is 0.5 are shown in Figure 6.

As shown in the results, the RFODLab network with the introduced attention mechanism can better capture the contextual semantic information of images, and allocate more weights to the image areas where foreign objects really exist. The mask segmentation results show that the RFODLab network enables the accurate segmentation and extraction of foreign objects on the ballastless track bed, while preventing the misclassification of some inherent infrastructures of track bed as foreign objects.

The ablation experiment was conducted for the RFODLab network is composed of the attention mechanism for the backbone network, and the loss function combining Focal Loss and Dice Loss. The backbone network refers to the ResNet50 network without/with the attention mechanism. Five loss functions are adopted, including the cross-entropy loss, weighted cross-entropy, Dice loss, Focal loss, and Dice loss + Focal loss. The performance of the RFODLab network is assessed, and the results of the ablation experiment are shown in Table 2 and Figure 7. The performance parameters of the backbone network ResNet 50 are not bold in Table 2, and the performance parameters of the backbone network ResNet 50 with the attention mechanism are bold in Table 2.

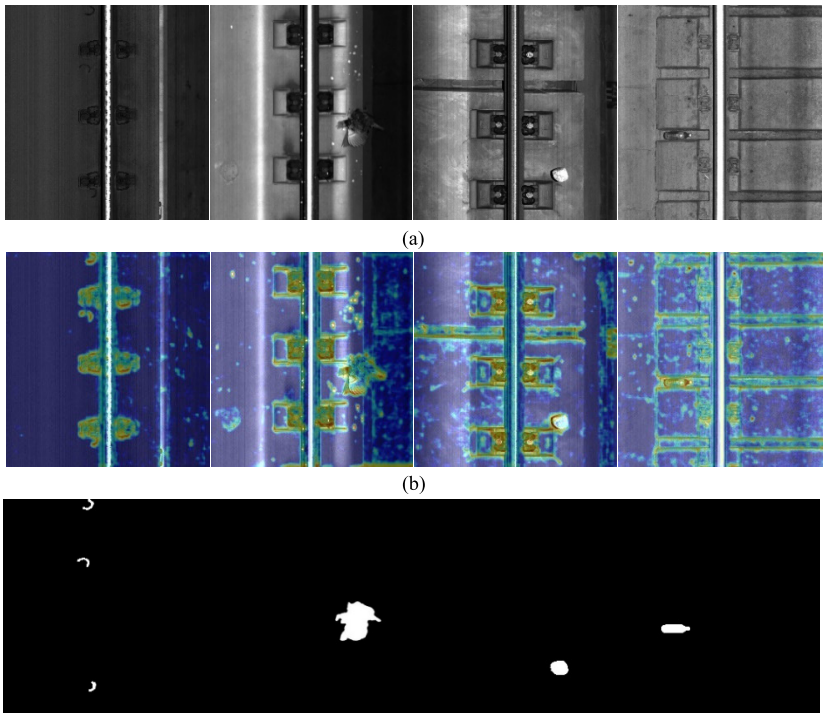


FIGURE 6. Segmentation result of foreign object image through RFODLab (a) Input image; (b)Heat map of output feature map; (c) Segmentation mask.

TABLE 2. Performance of RFODLab network in ablation experiment on test set.

Backbone network type	Loss function type	TP	FP	Recall	Precision	mIOU
ResNet50 /ResNet50 with attention mechanism	Cross Entropy Loss	193/**197**	26/**19**	91.04%/**92.92%**	88.13%/**91.20%**	77.25%/**80.09%**
	Weighted Cross Entropy Loss	194/**198**	24/**17**	91.51%/**93.40%**	88.99%/**92.09%**	77.43%/**81.39%**
	Dice Loss	195/**203**	21/**13**	91.04%/**95.75%**	90.19%/**93.98%**	78.57%/**83.26%**
	Focal Loss	195/**200**	23/**17**	91.98%/**94.34%**	89.45%/**92.17%**	78.45%/**81.96%**
	Dice Loss + Focal Loss	197/**205**	23/**19**	92.92%/**96.70%**	89.55%/**91.52%**	79.06%/**83.78%**

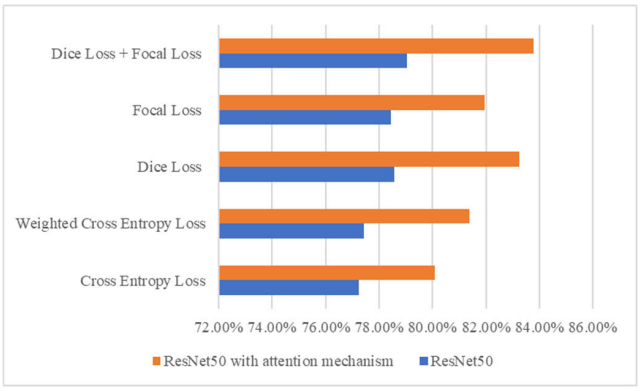


FIGURE 7. mIOU of RFODLAB network in ablation experiment on test set.

TABLE 3. Anomaly detection result of multiple semantic segmentation networks on test set.

Method	Foreign objects number	TP	FP	Recall	Precision	mIU
FCN	212	191	25	90.09%	88.43%	78.35%
DeepLab v3+		194	18	91.51%	91.51%	80.09%
CCNet		193	20	91.04%	90.61%	79.56%
DANet		194	23	91.51%	89.40%	79.43%
GCNet		196	25	92.45%	88.69%	80.41%
DNLNet		198	22	93.40%	90.00%	81.32%
RFODLab		205	19	96.70%	91.52%	83.78%

Experimental results show that the RFODLab network with the attention mechanism introduced in the backbone network enhances the capacity to obtain the contextual semantic information, which significantly improves the

TABLE 4. Performance of anomaly detection algorithm for ballastless track bed based on semantic segmentation in actual line test.

Line type	Picture number	Foreign objects number	TP	FP	Recall	Precision	Accuracy
Line 1	173702	81	80	234	98.77%	25.48%	99.86%
Line 2	149054	56	56	157	100.00%	26.29%	99.89%
Line 3	104250	106	103	379	97.17%	21.37%	99.63%

weighted constraints on the area to be recognized. As a result, the recall and precision of anomaly detection, as well as the fineness of anomaly segmentation increase greatly. Experimental results also show that the RFODLab with adjusted loss function combining Dice Loss and Focal loss has a very outstanding performance in the recall and the mIOU. It proves that the improved loss function reduces the excessive proportion of the background pixels, which, to a certain extent, can ease the negative impact from the imbalance in the sample category. The precision of the model decreased slightly means that the valid feature of foreign object has been extracted effectively by the model, but may cause inherent facilities with features close to foreign objects have a greater chance to be recognized as false alarm.

Despite the loss function combining Dice Loss and Focal loss slightly reduces the anomaly detection precision to 91.52%, it significantly improves the recall. In view of railway safety foreign objects should be detected to the greatest extent. Therefore, this loss function is most ideal choice for the RFODLab network, which is conducive to ensuring the safe operation of railways and the effective anomaly detection.

From the perspectives of anomaly detection and semantic segmentation performance, the RFODLab segmentation network is compared with the recent state-of-the-art semantic segmentation networks such as FCN[29], DeepLab v3+, CCNet[30], DANet[31], GCNet[32], DNLNet[33]. All the confidence thresholds of the semantic segmentation network pixels are set as 0.5, with the experimental results shown in Table 3 and Figure 8.

As shown in the Table 3, compared to the state-of-the-art semantic segmentation network, the RFODLab semantic segmentation network shows the optimal performance in the recall, precision, and segmentation fineness mIOU (these indicators are used to evaluate the performance of anomaly detection on the ballastless track bed) on test set. In contrast with other semantic segmentation networks, the RFODLab semantic segmentation network proposed herein is more applicable to the anomaly detection of the ballastless track bed.

E. ACTUAL LINE TEST

The RFODLab semantic segmentation network for anomaly detection of the ballastless track bed is tested along three

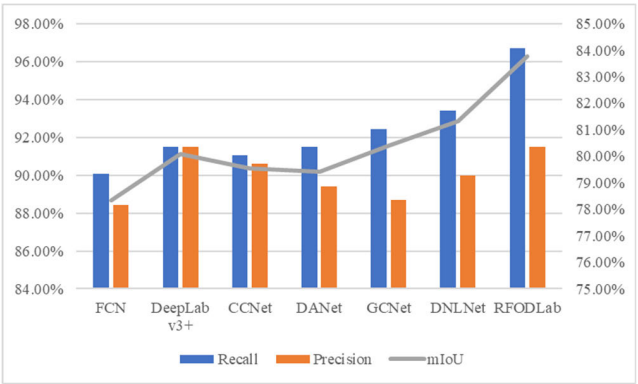


FIGURE 8. Performance of multiple semantic segmentation networks on test set.

sections of high-speed railway lines. The original detection results are obtained by manually checking and reviewing the images continuously captured by the integrated inspection trains, and compared with the results of anomaly detection through the RFODLab semantic segmentation network. The indicator *Accuracy* is added to verify the actual overall performance of the model.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \times 100\% \qquad (12)$$

where: *TP* is the number of true foreign objects correctly identified, *TN* denotes the number of backgrounds correctly identified, *FN* refers to the number of true foreign objects unidentified, and *FP* indicates the number of foreign objects wrongly identified.

The experimental results of RFODLab semantic segmentation network in actual line test are shown in Table 4.

To ensure the safe operation of high-speed railways, foreign objects should be detected in the actual scenario as far as possible. In case of ensuring the recall as much as possible, the original limitation conditions of model detection, and the effect of distinguishing the foreign objects from the objects with similar features will be weakened (in other words, various shapes of paint or normal facilities, cables and other objects may be identified as foreign objects). Therefore, certain false alarm is allowed to some extent.

In reality, the results of anomaly detection through the detection model need to be manually reviewed. Experimental results show that the recall and accuracy of detection in

actual scenes of three existing lines all exceed 97% and 99%, respectively. Compared to the original method of manually reviewing the foreign object images, the new method combining the semantic segmentation network detection and manual image reviewing is more applicable to actual scenarios, which can significantly reduce the manpower.

V. CONCLUSION

We propose the RFODLab semantic segmentation network for detection of foreign objects on the ballastless track bed. In order to improve the model's capacity to extract the features of foreign objects on the ballastless track bed, the channel attention mechanism is introduced for the backbone network, which can allocate more weights to the image areas in which foreign objects to be identified. In consideration of the extreme imbalance exist in the category distribution of samples for anomaly detection of the ballastless track bed, the loss function combining Dice Loss and Focal loss is adopted for the RFODLab semantic segmentation network, to improve the recall. The RFODLab semantic segmentation network shows a good performance in the anomaly detection of the ballastless track bed. To be specific, the precision and recall on test set reach up to 90% and 96.7%. In multiple actual scenarios, the recall and accuracy exceed 97% and 99%, respectively. All these prove that this model is effective to detect foreign objects on the ballastless track bed.

The semantic segmentation network for the anomaly detection of the ballastless track bed is a method of supervised anomaly detection based on sample data. In the future, we will further explore the methods of unsupervised anomaly detection based on a small number of samples or even positive samples only. In addition, with the continuous accumulation of data on foreign object samples, it is necessary to further classify foreign objects on the ballastless track bed and establish a system for effective assessment of safety risks from foreign objects of railways.

REFERENCES

[1] F. J. Alvarez, J. Urefia, M. Mazo, A. Hernandez, J. J. Garcia, and P. Donato, "Ultrasonic sensor system for detecting falling objects on railways," in *Proc. IEEE Intell. Vehicles Symp.*, Jun. 2004, pp. 866–871.

[2] A. Mroue, M. Heddebaut, F. Elbahhar, A. Rivenq, and J. M. Rouvaen, "UWB radar and leaky waveguide for fall on track object identification," in *Proc. IEEE Radar Conf.*, May 2010, pp. 573–577.

[3] D. Sinha and F. Feroz, "Obstacle detection on railway tracks using vibration sensors and signal filtering using Bayesian analysis," *IEEE Sensors J.*, vol. 16, no. 3, pp. 642–649, Feb. 2016.

[4] J. J. Garcia, A. Hernandez, J. Urena, and E. Garcia, "FPGA-based architecture for a multisensory barrier to enhance railway safety," *IEEE Trans. Instrum. Meas.*, vol. 65, no. 6, pp. 1352–1363, Jun. 2016.

[5] G. Xu, T. Shi, S. Ren, Q. Han, and D. Wang, "Development of on-board track inspection system based on computer vision," *China Railway Sci.*, vol. 34, no. 1, pp. 139–144, 2013.

[6] L. Jiang, G. Peng, B. Xu, Y. Lu, and W. Wang, "Foreign object recognition technology for port transportation channel based on automatic image recognition," *EURASIP J. Image Video Process.*, vol. 2018, no. 1, p. 147, Dec. 2018.

[7] H. Xu, Z. Han, S. Feng, H. Zhou, and Y. Fang, "Foreign object debris material recognition based on convolutional neural networks," *EURASIP J. Image Video Process.*, vol. 2018, no. 1, p. 21, Dec. 2018.

[8] J. Zhu, Y. Guo, F. Yue, H. Yuan, A. Yang, X. Wang, and M. Rong, "A deep learning method to detect foreign objects for inspecting power transmission lines," *IEEE Access*, vol. 8, pp. 94065–94075, 2020.

[9] 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.

[10] H. X. Niu and T. Hou, "Fast detection study of foreign object intrusion on railway track," *Arch. Transp.*, vol. 47, no. 3, pp. 79–89, Sep. 2018.

[11] B. Guo, G. Geng, L. Zhu, H. Shi, and Z. Yu, "High-speed railway intruding object image generating with generative adversarial networks," *Sensors*, vol. 19, no. 14, p. 3075, Jul. 2019.

[12] S. Wang, Z. Gu, Q. Han, P. Dai, and X. Du, "Intelligent recognition of foreign object invasion in high-speed railway movement scene," in *Proc. 13th Nat. Conf. Vib. Theory Appl.*, 2019, pp. 2–7.

[13] R. Chalapathy and S. Chawla, "Deep learning for anomaly detection: A survey," 2019, *arXiv:1901.03407*. [Online]. Available: https://arxiv.org/abs/1901.03407

[14] D. Gong, L. Liu, V. Le, B. Saha, M. R. Mansour, S. Venkatesh, and A. Van Den Hengel, "Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1705–1714.

[15] H. Park, J. Noh, and B. Ham, "Learning memory-guided normality for anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 14372–14381.

[16] M. Z. Zaheer, J.-H. Lee, M. Astrid, and S.-I. Lee, "Old is gold: Redefining the adversarially learned one-class classifier training paradigm," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 14183–14193.

[17] J. Yi and S. Yoon, "Patch SVDD: Patch-level SVDD for anomaly detection and segmentation," in *Proc. Asian Conf. Comput. Vis.*, Nov. 2020, pp. 1–16.

[18] S. Venkataramanan, K. C. Peng, R. V. Singh, and A. Mahalanobis, "Attention guided anomaly localization in images," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 485–503.

[19] Y. Xia, Y. Zhang, F. Liu, W. Shen, and A. L. Yuille, "Synthesize then compare: Detecting failures and anomalies for semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 145–161.

[20] L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 801–818.

[21] 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.

[22] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[23] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. J. Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Cham, Switzerland: Springer, 2017, pp. 240–248.

[24] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, H. Lin, Z. Zhang, Y. Sun, T. He, J. Mueller, R. Manmatha, M. Li, and A. Smola, "ResNeSt: Split-attention networks," 2020, *arXiv:2004.08955*. [Online]. Available: https://arxiv.org/abs/2004.08955

[25] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[26] W. Zhu, Y. Huang, L. Zeng, X. Chen, Y. Liu, Z. Qian, N. Du, W. Fan, and X. Xie, "AnatomyNet: Deep learning for fast and fully automated whole-volume segmentation of head and neck anatomy," *Med. Phys.*, vol. 46, no. 2, pp. 576–589, Feb. 2019.

[27] 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.

[28] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 761–769.

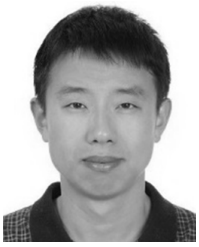
[29] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.
[30] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "CCNet: Criss-cross attention for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 603–612.
[31] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3146–3154.
[32] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "GCNet: Non-local networks meet squeeze-excitation networks and beyond," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1–10.
[33] M. Yin, Z. Yao, Y. Cao, X. Li, Z. Zhang, S. Lin, and H. Hu, "Disentangled non-local neural networks," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 191–207.

HAORAN SONG was born in Henan, China, in 1996. He received the B.S. degree from the Harbin Institute of Technology, China, in 2019. He is currently pursuing the M.S. degree in traffic information engineering with the China Academy of Railway Sciences, Beijing, China. His research interests include railway infrastructure inspection, object detection, and semantic segmentation.

SHENGCHUN WANG was born in 1985. He received the B.S. and Ph.D. degrees from Beijing Jiaotong University, China, in 2008 and 2015, respectively. He worked as an Associate Researcher with the Railway Infrastructure Inspection Institute, China Academy of Railway Science, Beijing. His research interests include scene representation for railway environment, high resolution reconstruction, and object detection.

ZICHEN GU received the M.S. degree from the University of Bristol, U.K., in 2012. He is currently an Engineer with INMAI Railway Technology Company Ltd., Beijing, China. His research interests include image recognition and object detection.

PENG DAI received the B.S., M.S., and Ph.D. degrees from the Department of Control Science and Engineering, Harbin Institute of Technology. He is currently a Researcher with the Railway Infrastructure Inspection Institute, China Academy of Railway Science, Beijing. His research interests include statistical machine learning, visual detection, and its application to high-speed railway.



XINYU DU received the Ph.D. degree from the University of Electronic Science and Technology of China, Chengdu, China, in 2012. He is currently a Researcher with the Infrastructure Inspection Institute, China Academy of Railway Sciences, Beijing, China. His research interests include image processing and measurement.

YU CHENG received the M.S. degree from Beijing Jiaotong University, Beijing, China, in 2014. He is currently a Research Associate with the Infrastructure Inspection Research Institute, China Academy of Railway Sciences Corporation Ltd., Beijing. His current research interests include track inspection technology, embedded data acquisition and processing technology, and system safety verification technology in railway systems.