Journal of Computers Vol. 34 No. 3, June 2023, pp. 121-136
doi: 10.53106/199115992023063403009

Fault Diagnosis of Train Body Sign Abnormal Pattern with Deep Learning Based Target Detection

Yuanjiang Hu, Aisen Yang, Zonghong Zhang, Na Qin*

The Institute of Systems Science and Technology, Southwest Jiaotong University,
Chengdu, 611756, P. R. China

360832800@qq.com, aisen.yang@my.swjtu.edu.cn, zzh451045814@163.com,
qinna@swjtu.edu.cn

Received 5 April 2022; Revised 9 October 2022; Accepted 11 December 2022

Abstract. With the development of high-speed trains in recent years, security issues have received more attention. Automatic visual inspection of the train operation system for detecting abnormalities has become a fundamental element to guarantee the safety of the train operation. Train body sign patterns like the loss and fracture of signs and lock catch (SLC) on the electrical box cover (EBC) affect the regular operation of the train electrical system. In this paper, to ensure the safe operation of the train, a novel method combining a faster region-based convolutional neural network (Faster R-CNN) and similarity metrics is proposed to detect the abnormality of SLCs on train EBC. First, the positions of body train signs of multiple sizes are located by Faster R-CNN. Then, the regions of interest (ROI) are cut out and resized to the same size as the corresponding template images. Finally, by similarity measures, the status of the train body sign pattern is judged by comparing with the given threshold similarity value between ROIs and the template images. It is worth noting that the combination of Faster R-CNN and cosine similarity renders high accuracy in small target detection and strong robustness in image similarity comparison. The effectiveness of the proposed fault detection method and its superiority over the other types of combined methods are verified by actual experiments on the train of Guangzhou Metro Line 2.

Keywords: fault diagnosis, train operation system, faster R-CNN, image similarity

1 Introduction

The train system has been dramatically developed as universal and economical transportation. At the same time, with the increase in train speed, it is crucial to ensure a train's safe and stable operation. Due to the exposure of the train body to the external environment, it is easy to have problems such as aging, loosening, and loss of components. Under the influence of the driving and braking force, the running part, the suspension, the buffer connection device, the brake accessories, the bogie, and other critical visible parts may shake and loosen to various degrees. The failure of these components will increase the risk of train operation and even lead to grave danger. Therefore, timely detection of the failure of these critical components during the train operation plays a vital role in ensuring the safety of the train operation.

Train vehicle failure controls are often completed by train inspectors, with many drawbacks, such as low detection accuracy and low efficiency [1]. However, the complex situation of the train body will lead to high costs and inefficiency of the manual anomaly inspection process. Since common faults in trains are mainly caused by the breakage and position changes of small components, the difference between standard and faulty images is not apparent, which increases the difficulty of visual fault detection. The variety and complexity of faults make it difficult for conventional methods to achieve fast and accurate fault detection on train images. With the rapid development of artificial intelligence theory and the continuous increase in the number of trains, automatic fault detection systems based on machine learning have gradually replaced traditional manual methods [2]. It is necessary to apply computer vision technology to the fault detection of train components. By installing a series of cameras at the train maintenance place, distortion correction, registration, and detection are carried out on the images collected by the cameras to judge whether the critical components of the train have faults and the types and locations of faults to improve the efficiency and accuracy of train maintenance.

The electrical box is an essential component of the train body. As a part of the train components, train body

* Corresponding Author

signboards like the signs and lock catches (SLCs) on the electrical box cover (EBC) can prevent the door of the EBC from falling during the train running, endangering the safe operation of the train. Besides, the signs on the EBC remind the workers to take corresponding safety measures in repairing the train to prevent electric shock and other hazards. Hence, detecting whether the signs and lock catch on the EBC are faulty is essential to the train's safe operation. In addition to the normal working flow of the fault diagnosis technology mentioned above, there is still potential for improvement in the fault detection of SLCs of the train body, such as the various kinds of stains and marks under the influence of natural factors affecting the model accuracy. Besides, some researchers develop models with imbalanced training, and testing datasets may hamper the robustness of models [3].

This paper proposes a novel method to detect the fault of SLCs of the train body by combing target detection and template matching algorithms. Owing to the novel structure of the fault diagnosis model and application of transfer learning setup, high accuracy in small target detection and strong robustness in image similarity comparison can be achieved simultaneously in the fault diagnosis of SLCs. The superior combination of different types of model building blocks is demonstrated in the experimental test.

The rest of the paper is organized as follows. Section II outlines the primary methods used in the experimental model, including the object detection algorithm and related template-matching methods. Section III presents the experimental results and discussion. At last, Section IV concludes the work.

2 Related Work

Recently, computer-vision-based object detection technologies have been increasingly indispensable in industry. Much effort has been put into exploiting efficient fault diagnosis schemes for the components of a railway transportation system based on machine vision, including template matching, machine learning, and deep learning. In particular, the template-matching-based fault diagnosis methods calculate the matching degree between the image to be tested and the template and judge whether the image to be tested has a fault or not according to prespecified criteria [1]. For the sample images with high similarity to the template image, those template-matching-based methods would yield a high recognition rate; Otherwise, they will demonstrate relatively poor robustness. In principle, traditional machine-learning-based pattern recognition consists of two stages: feature extraction and classification, and applies to the case of a small sample set [4]. For the first stage, it is desirable to choose an appropriate feature extraction scheme, e.g., histogram of oriented gradients (HOG) and local binary patterns (LBP), according to the detection scene and artificial experience. The commonly used classifiers for machine-learning-based fault diagnosis include support vector machine (SVM) [5], k-nearest neighbor (KNN) [6], and decision tree (DT) [7]. Given statistical learning theory, SVM has been successfully used in various fault classification tasks. In [5], the fault diagnosis of the handle on the train gladiator was realized by applying the gradient-coded histogram and SVM to analyze the train angle cock image. Similarly, the brake shoe key of the train was located by combining the histogram of gradient coding and SVM in [8]. Further, a recognition algorithm was proposed in [9] using principal component analysis (PCA) and SVM classifiers, which were employed to identify the fault images of train key stations, pillow springs, and side pillars. Moreover, bolt faults were classified in [10] considering the combination of HOG features and SVM. All in all, it is challenging to design a general machine-learning-based method to detect and identify all kinds of faults at the same time.

In contrast to machine learning, deep learning can simultaneously complete feature extraction and classification and achieve high fault detection accuracy [11]. Deep learning learns features through a data-driven approach instead of establishing feature engineering, reduces the difficulty of deep learning technology in engineering applications, and makes it the first choice for fault diagnosis. In recent years, a series of CNN-based object detection algorithms have been proposed for fault diagnosis, among which Faster R-CNN [12], single-shot detector (SSD) [13], and You Only Look Once (YOLO) [14] are the mainstream. Compared with traditional image processing technologies that depend on many observable image features, deep-learning-based methods are still applicable when the features are difficult to extract manually. Deep-learning-based methods automatically extract image features through convolutional layers and pooling layers. In [15], CNN was utilized to recognize train side frame key loss, where the model accuracy was much higher than SVM. Meanwhile, the framework of Faster R-CNN was reconstructed in [16], yielding a higher detection accuracy and better real-time performance. In [17], a YOLO-based three-cascade neural network was built to detect the fasteners of high-speed railway catenary support devices, where the strong robustness of the model was verified in complex background environments. In [18], to evaluate the health status of turbine blades, a large number of fault samples were collected by UAV and classified by CNN, which achieved good results. Besides, in [19], a novel attention perception network APP-

UNet16 is proposed, which can detect the oil defects of the rolling shaft by segmenting the bearing oil. However, it is worth noting that the common point of the above cases is that there are many negative samples for network training to do multi-classification. However, in the actual situation faced by this paper, the number of negative samples is not enough for training to achieve data-driven end-to-end fault diagnosis.

After the target detection or localization, the similarity between the ROI and the corresponding template needs to be calculated to judge the faults in ROI by comparing it with a given threshold. What has attracted people's attention is some anomaly detection algorithms in recent years, including a series of anomaly detection algorithms based on generative adversarial networks (GAN) [20]. However, if an anomaly detection algorithm is used, each type of part (such as a lock) needs to build a model, which is more complicated in engineering applications. In addition, the anomaly detection algorithm re-quires some negative samples to fine-tune the neural network, while the number of negative samples in this experiment is limited, which is not enough to fine-tune the network. Therefore, it is considered to realize anomaly detection through image similarity comparison. The well-adopted methods for image similarity calculation include the histogram method, peak signal-to-noise ratio (PSNR) [21], structural similarity (SSIM) [22], and cosine similarity (CS) [23]. More detailedly, the histogram reflects the probability distribution of gray values of an image, and its performance in image similarity calculation is often degraded due to the lack of spatial position information. PSNR is calculated based on the error value of the corresponding pixels between the image to be tested and the template image. It does not take into account spatial frequency characteristics information [24]. Thus, the evaluation results obtained by PSNR are frequently inconsistent with the visual perception results.

3 Fault Diagnosis Method

The framework of the proposed method mainly includes two parts, as shown in Fig. 1. First, since CNN can extract images' features automatically, Faster R-CNN is used as the model for the recognition and localization of SLCs on EBC. Second, to realize the status judgment of SLCs, the similarity value of the template and the sample to be tested is calculated by the CS method. The details of Faster R-CNN and similarity calculation are described in the following.

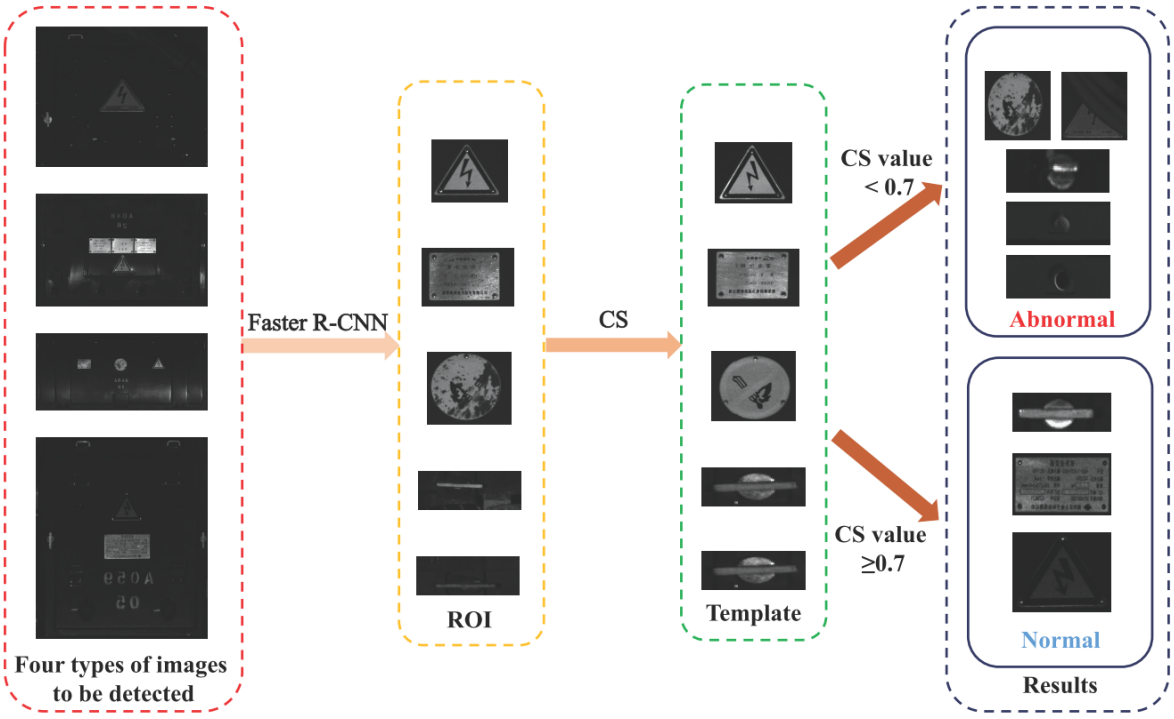


Fig. 1. The framework of the proposed method

3.1 The Architecture of Faster R-CNN

The object detection networks mainly consist of SSD and its derivative networks, the YOLO series, and the R-CNN series. YOLO and SSD are typical one-stage algorithms with the characteristics of fast detection speed, which combine the selection and classification of an anchor. However, in terms of accuracy, Faster R-CNN, as a two-stage algorithm, is higher than the one-stage algorithms. Meanwhile, the current research in object detection algorithms mainly focuses on lightweight object detection to facilitate the deployment of embedded devices. However, its research focuses on the embedded deployment, which can improve the inference speed based on reducing the inference accuracy [24]. If a light target detection network is selected, the positioning accuracy of components may not reach the expected level. The twice regression and classification make Faster R-CNN more suitable for small target detection. As a result of the small proportion of SLCs areas in EBC and the low requirement for real-time detection tasks, Faster R-CNN is selected to realize the location and identification of SLCs in this work.

As depicted in Fig. 2, the framework of Faster R-CNN contains three parts:

(1) Backbone, considered a network for extracting image feature information, usually consists of fully convolutional layers. The commonly used backbones are optical geometry group network 16 (VGG 16).

(2) Region Proposal Network (RPN) generates some proposal areas, judges whether these areas have attractive targets, and makes a bounding box regression for the regions with targets.

(3) ROI Pooling extracts the proposal feature maps that form the backbone network and RPN output and sends them to the subsequent complete connection layer for classification and regression.

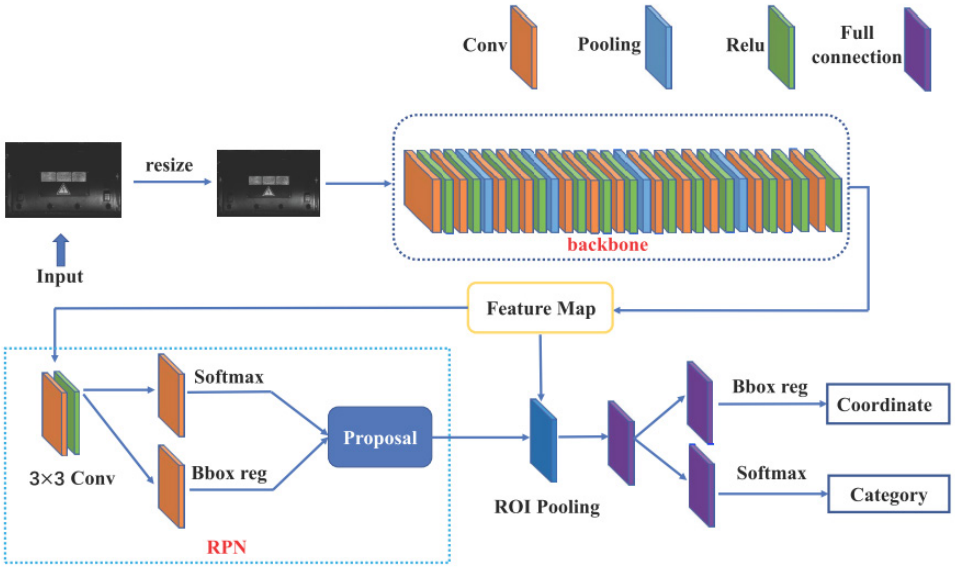


Fig. 2. Structure schematic diagram of Faster R-CNN

Region Proposal Network. RPN is a typical complete convolution network structure that cancels the entire connection operation widely used in the traditional convolution network, significantly saving the calculation cost. The anchor window mechanism plays an essential role in RPN. The different shapes and sizes of detection objects result in the need to use different boxes for frame selection when determining the location. The anchor window mechanism ensures that the detection objects of different sizes are in the receptive field.

First, RPN sets the anchor point on the convolution feature layer, which is responsible for prediction. It takes each pixel on the convolution feature layer as the center of all the anchors corresponding to the pixel. The anchor boxes are placed in three proportions, eight times, 16 times, and 32 times of the basic anchor box, respectively, and three aspect ratios, namely 1:1, 1:2, and 2:1, a total of nine sizes of anchor, as shown in Fig. 3. Second, the

3×3 convolution kernel convolutes the feature map containing anchor information, so that all the information in the feature map are fused. Then the processed feature maps are convoluted by 1×1 convolution kernels to complete the tasks of distinguishing foreground and background and regression of the bounding box. In the end, a series of object proposals are output through RPN. The target's position is determined in the training set, and the RPN network is optimized using the gradient descent method by comparing it with the target box position after regression.

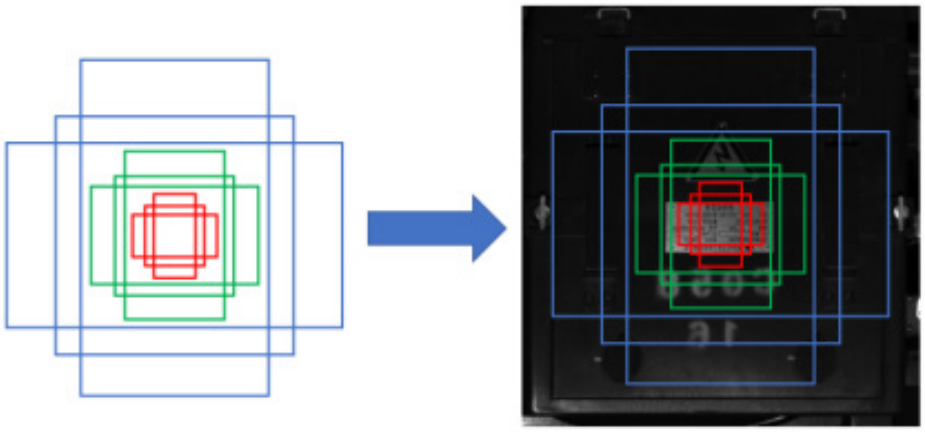


Fig. 3. Anchor generation schematic

In the loss function of RPN, classification and regression losses are considered simultaneously, or mathematically

$$L(\{p_i\},\{t_i\}) = \frac{\Sigma_i L_{cls}(p_i, p_i^*)}{N_{cls}} + \lambda \frac{p_i^* L_{reg}(t_i, t_i^*)}{N_{reg}} \ . \tag{1}$$

p_i and p_i^* represent the probability that the *i*-th anchor is the target and the label of the *i*-th anchor, respectively. Moreover, t_i and t_i^* are the coordinates of the *i*-th prediction box and standardization box, respectively, are the number of anchors, and N_{reg} is the size of the feature map. The classification loss function L_{cls} and the regression loss function L_{reg} are defined in the following

$$L_{cls}(p_i, p_i^*) = -\log[p_i^* p_i + (1-p_i^*)(1-p_i)] \ . \tag{2}$$

$$L_{reg}(t_i, t_i^*) = R(t_i - t_i^*) \ . \tag{3}$$

$$R(x) = \begin{cases} 0.5x^2 & \text{if } |x| < 1 \\ |x| - 0.5 & \text{otherwise} \end{cases} \ . \tag{4}$$

$$L_1(x) = |x| \ . \tag{5}$$

$$L_2(x) = x^2 \ . \tag{6}$$

R is a smooth $L1$ loss function. As shown in Eq. (4), smooth $L1$ loss function $R(x)$ is similar to the $L2$ loss function $L_2(x)$ when x belongs to (-1,1), which avoids the problem that $L_1(x)$ is not differentiable at x=0. When x is outside the range of (-1,1), $R(x)$ is similar to the $L1$ loss function $L_1(x)$, which avoids the gradient explosion of the $L2$ loss function. Thus, compared with $L1$ and $L2$ loss functions, the smooth $L1$ loss function has the following advantages: 1) faster convergence; 2) insensitive to outliers, and more stable during training.

ROI Pooling. Owing to the different sizes between the front-end network's feature maps output and the RPN's bounding boxes output, it is necessary to achieve the exact size of the proposal feature maps through the ROI pooling layer. First, the proposals generated by RPN are mapped to the corresponding positions of the feature map of the original image. Second, the feature map corresponding to each proposal is divided into 7×7 small areas. Finally, each small area after max pooling is transformed into a one-dimensional tensor sent to the subsequent complete connection layer.

3.2 Structural Similarity Index Method (SSIM)

When the detected object has specific faults, its overall structure will change. Therefore, whether the detected object is faulty or not could be judged by calculating the similarity between the template image and the image to be detected. The commonly used image similarity measurement methods are based on explicit numerical comparisons, such as the comparison of statistical parameters in a specific characteristic. SSIM considers image distortion by comparing the changes in image structure information to get an objective quality evaluation [22]. Inspired by such an idea, the fault state of the sample can be obtained by comparing the structure information between the standard sample and the sample to be tested. SSIM differs from other image similarity algorithms, e.g., PSNR and CS, since it evaluates image similarity by combining structural information, contrast information, and luminance information, as shown in Fig. 4.

Assume that and are image blocks with the size of H × W, located at the exact position of the template image and the image to be tested, respectively. The structural similarity information $s(x, y)$, the contrast similarity information $c(x, y)$, and the luminance similarity information $l(x, y)$ can be calculated as follows.

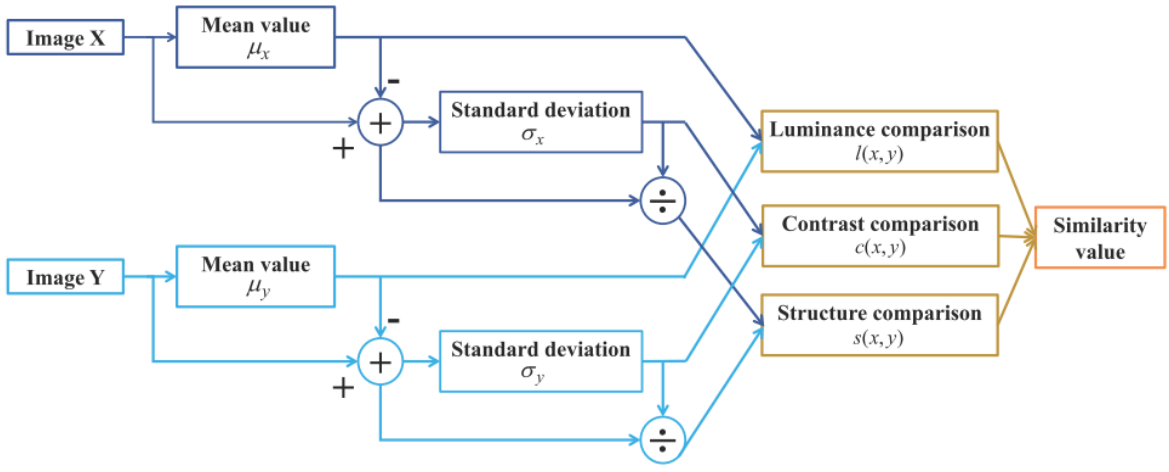


Fig. 4. The structure schematic diagram of SSIM

$$s(x,y)=\frac{\sigma_{xy}+C_1}{\sigma_x\sigma_y+C_1}\ . \tag{7}$$

$$c(x,y)=\frac{2\sigma_x\sigma_y+C_2}{\sigma_x^2+\sigma_y^2+C_2}\ . \tag{8}$$

$$l(x,y)=\frac{2\mu_x\mu_y+C_2}{\mu_x^2+\mu_y^2+C_3}\ . \tag{9}$$

σ_x, σ_y represent the standard deviation of image blocks *x* and *y* respectively, and σ_{xy} is the correlation of image blocks *x* and *y*. In addition, μ_x and μ_y are the average value of image blocks *x* and *y*, respectively. The rationality of (7)-(9) is maintained by the proper design of constants $C_i \geq 0, i=1,\cdots,3$, which are generally small.

The SSIM is defined as follows.

$$SSIM=\frac{1}{N}\Sigma_{i=1}^{N}[l(x_i,y_i)]^{\alpha}[c(x_i,y_i)]^{\beta}[s(x_i,y_i)]^{\gamma}\ . \tag{10}$$

x_i and y_i are the *i*-th image block of the reference image and the image to be tested, respectively, *N* is the number of image blocks, and the weights α,β,γ are all positive.

3.3 Cosine Similarity (CS)

For CS, the similarity between two vectors is measured by the cosine of their inner product. The larger the cosine value of eigenvectors of two images is, the greater the matching similarity is. Meanwhile, noticing that CS is suitable for vector comparison of any dimension, it has become increasingly popular in comparison to image similarity.

Analysis of image similarity is performed at the level of each image region. Let the image be divided into *M* regions, each with *N* pixels. After that, the regional characteristics of three channel image are calculated as follows:

$$f=\left((\Sigma_{i\in n}^{N}\frac{r_i}{N}),(\Sigma_{i\in n}^{N}\frac{g_i}{N}),(\Sigma_{i\in n}^{N}\frac{b_i}{N})\right)\ . \tag{11}$$

f represents the regional feature code, and *r*, *g* and *b* represent the red, green and blue component values of pixel *i*, respectively. Furthermore, the regional feature codes of the two images are transformed into one-dimensional feature vectors. Then, the degree of the *m*-th corresponding regions' similarity between images I_1 and I_2 is calculated using the CS between the feature vectors f_1^m and f_2^m [23]:

$$similarity(I_1^m,I_2^m)=CS(f_1^m,f_2^m)=\frac{f_1^m\bullet f_2^m}{|f_1^m|\cdot|f_1^m|}\ . \tag{12}$$

I_1^m and I_2^m represent the *m*-th region of the template image I_1 and the image to be tested I_2, respectively.

In particular, the CS between the regions I_1^m and I_2^m ranges from 0 to 1. In addition, $CS(f_1^m,f_2^m)=0$ implies that the regions I_1^m and I_2^m are completely different; $CS(f_1^m,f_2^m)=1$ represents that the regions I_1^m and I_2^m are completely same.

4 Experimental Setup and Results

4.1 Dataset

The signs of the train EBC could remind the workers to pay attention to the operation specifications and safety. In addition, the lock catch ensures that the EBC cannot fall off to avoid the threat to the operation of the train's electrical system. Therefore, when the SLCs of the electrical box cover fail, it is necessary to detect such an abnormality in time and complete the maintenance.

The images provided by Guangzhou Yunda Intelligent Technology Co., Ltd. are selected as the experimental objects. Specifically, the appearance images of the train in all directions are obtained through seven charge-coupled device (CCD) cameras installed in the train inspection station, as shown in Fig. 5. Then, the EBC images are cut out from the side image of the train using relative coordinate positioning, as shown in Fig. 6. Due to the changes of external environment and camera exposure parameters, the collected images might be significantly different. Thus, traditional image processing methods might not be applicable for fault diagnosis of the SLCs on the EBC of a train.

Fig. 5. Image acquisition system (The seven solid boxes refer to the CCD cameras, and the dotted box indicates the control box of the whole image acquisition system.)

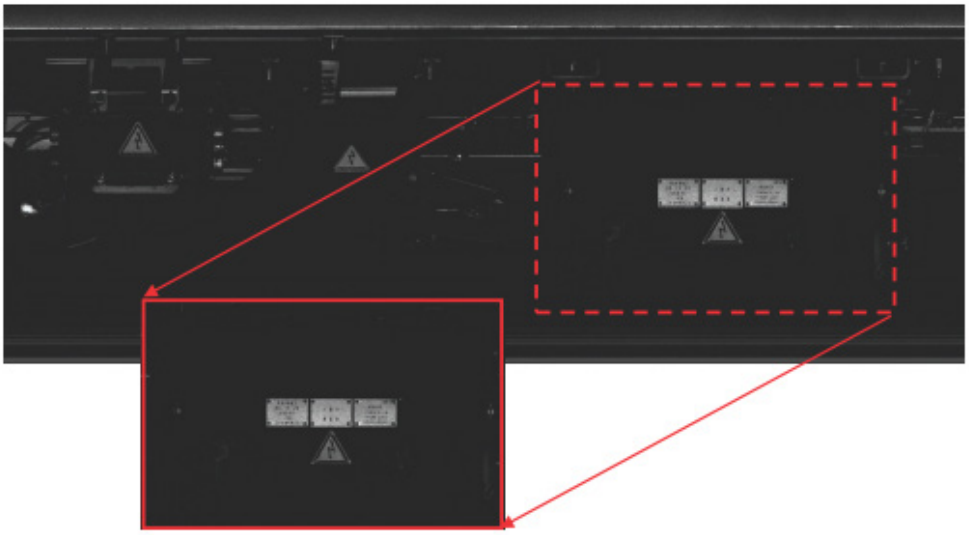


Fig. 6. The electrical box cover

The datasets used in this experiment are collected from the train of Guangzhou Metro Line 2. Considering that there are four types of EBC on the train, they are named type A, B, C, and D, respectively. Meanwhile, the experimental data are randomly reordered to avoid accidental experiments. The dataset, which contains 1468 images, is divided into a train set and a test set in a ratio of 9:1, and the statistics of four types of EBC used for training and testing are shown in Table 1. The objects to be detected include rectangular, triangular, circle signs, and lock catches, among which the detected objects in each image are different and contain multiple proportions of width and height. The statistics result of the number of each part in the four types of images is shown in Table 2. Significantly, the circle signs are marked as triangle signs during the model training because the number of circle signs is relatively tiny. Considering the small number of training samples, the idea of transfer learning is added to the Faster R-CNN model training to improve the feature extraction performance. The VGG16 pre-trained on the ImageNet dataset is used as the backbone network of Faster R-CNN.

Table 1. Statistics of four kinds of electrical box covers in the test set and train set

Type	Number of train set	Number of test set	Total nums
A	321	46	367
B	329	38	367
C	344	23	367
D	335	32	367

Table 2. Description of different types of electrical box covers

Type	Rectangle signs	Triangle signs (include circular signs)	Lock catch
A	0	1	1
B	3	1	2
C	1	2	2
D	1	1	2

4.2 Experimental Setup

To realize the fault diagnosis of SLCs on the EBC, Faster R-CNN and SSD512 are used to locate and cut out the detection target from the image. After the SLCs are located and intercepted by the object detection algorithms, the similarity values between the intercepted images and the corresponding template images are calculated with SSIM and CS, respectively. It should be pointed out that since the circle signs are marked as triangle signs in the positioning stage, the triangle and circle signs are put into the template library simultaneously to avoid the comparison error. Considering the complexity of the train inspection environment and the accuracy of fault diagnosis, the typical images of each type of object under sunny, rainy, firm, or weak light conditions are selected as reference templates. Moreover, selecting multiple templates for comparison also reduces accidental errors. Calculating the similarities between the images to be tested and the templates, the highest similarity value is output as the final value. If the final similarity value is less than 0.7, the object is considered to be faulty; otherwise, it is considered that there is no fault with the object. It is worth noting that the threshold used to distinguish between positive and negative samples in this work was obtained through engineering statistics.

Different methods are compared based on the same data set and experimental environment. The whole experimental process is completed on the TensorFlow platform, and the iterations of Faster R-CNN and additional SSD512 used for performance comparison are set as 20,000. The model's weight is updated based on the Adam optimizer, and its learning rate and weight decay rate are set to 0.001 and 0.0005, respectively. To objectively evaluate the performance of the proposed method, the test dataset is randomly divided three times. By calculating the evaluation indexes of each sub-dataset, the most representative average values of evaluation indexes are obtained.

Comparing the obtained similarity values with the preset threshold value, the fault state of the target to be tested is then judged, which can be divided into four scenarios: true positive (TP), false positive (FP), true negative (TN) and false negative (FN). As defined below, the metrics adopted for performance evaluation include the accuracy rate, false positive rate (FPR), and false negative rate (FNR).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad . \tag{13}$$

$$FPR = \frac{FP}{FP + TN} \quad . \tag{14}$$

$$FNR = \frac{FN}{TP + FN} \quad . \tag{15}$$

More specifically, Accuracy is the proportion of the image samples whose fault state is predicted correctly by the proposed model in all the testing samples; FPR is the proportion of all false positive samples predicted by the model in the total negative samples, and FNR refers to the proportion of all false negative samples predicted by the model in the positive samples.

4.3 Experimental Results

The four methods proposed for performance comparisons are Accuracy, FPR, FNR, average accuracy, average FPR, and average FNR. The statistical results in Table 3 show that the Faster R-CNN and CS combination is far superior to the other three combinations in terms of Accuracy and FNR. Specifically, the average accuracies of the other three methods are less than 90% on the three test subsets, and the average FNRs are pretty high, inducing low reliability in engineering applications. Consequently, combining Faster R-CNN and CS can fully meet the requirements of high accuracy and low FNR for detecting abnormal objects on the EBC. In addition, the loss function curves of the two algorithms are shown in Fig. 7, where the loss function of Faster R-CNN achieves a faster convergence and a smaller steady-state value. As a result, Faster R-CNN would have better detection performance than SSD512.

Table 3. The results of comparison experiments

Type		SSD512+SSIM	SSD512+CS	Faster-RCNN+SSIM	Faster-RCNN+CS
Sub dataset1	Accuracy	66.07%	89.29%	71.76%	94.91%
	FPR	0	0	0	0
	FNR	35.68%	15.02%	29.76%	5.37%
Sub dataset2	Accuracy	64.73%	88.84%	74.54%	97.22%
	FPR	0	0	0	0
	FNR	37.09%	11.74%	26.83%	2.93%
Sub dataset3	Accuracy	66.52%	87.05%	70.37%	96.30%
	FPR	0	0	0	0
	FNR	35.21%	13.62%	31.22%	3.91%
Average accuracy		65.77%	88.39%	72.22%	96.14%
Average FPR		0	0	0	0
Average FNR		35.99%	13.46%	29.27%	4.07%

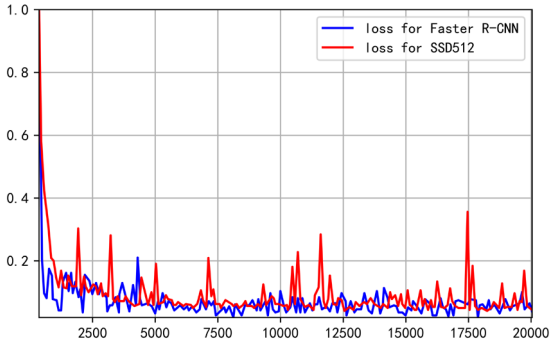
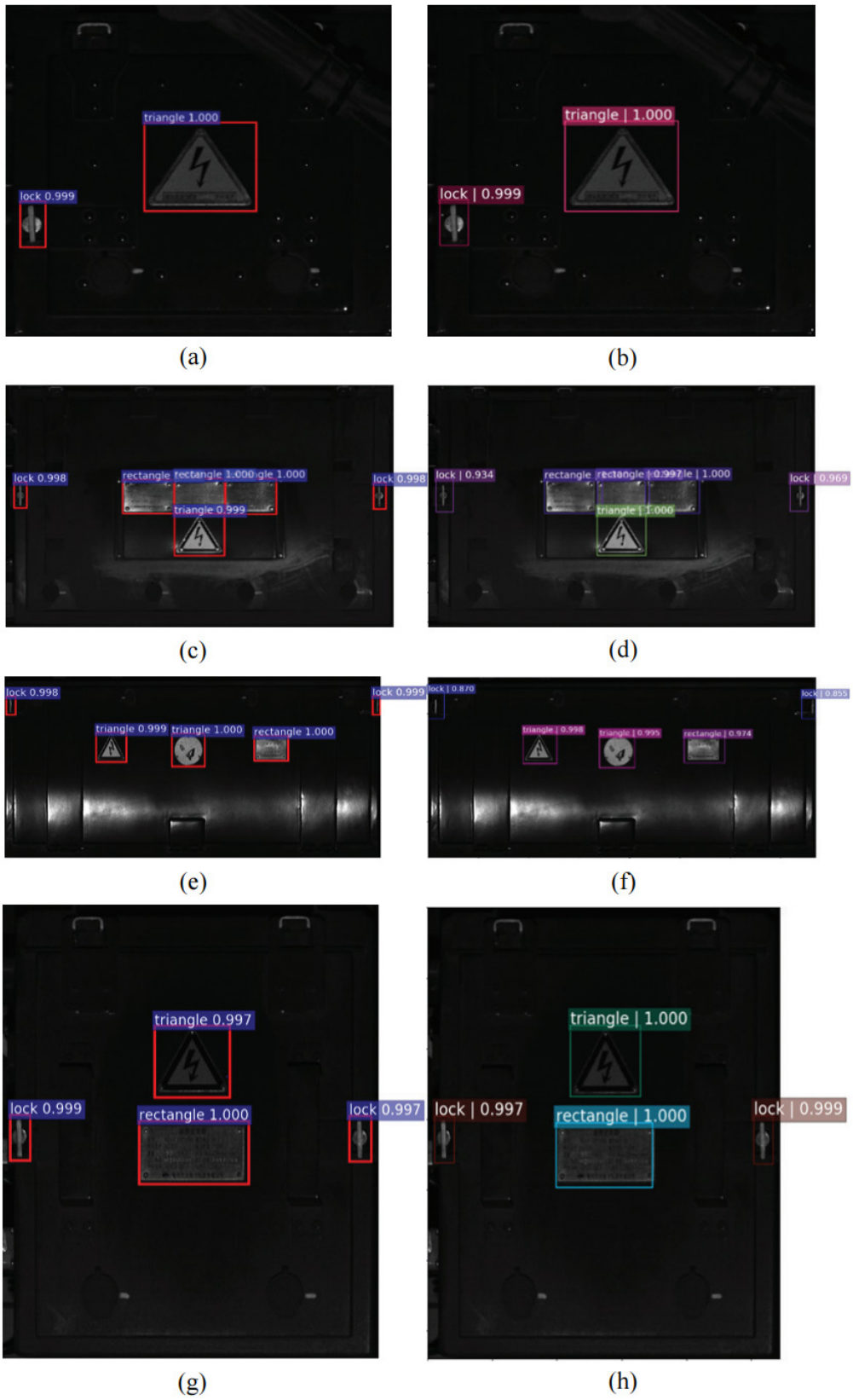


Fig. 7. The loss of Faster R-CNN and SSD512 for 20000 iterations on test sets



(a), (c), (e) and (g) on the left side are the position results of type 1, type 2, type3, and type 4 obtained from Faster R-CNN
(b), (d), (f) and (h) on the right side are the position results of type 1, type 2, type3, and type 4 obtained from SSD512

Fig. 8. The position results of different types of electrical box covers based on Faster R-CNN and SSD512

Fig. 8 shows the partial detection results on the same images after training 20,000 iterations using Faster R-CNN and SSD512, respectively. The position detection results of different electrical box covers are based on Faster R-CNN and SSD512. (a), (c), (e), and (g) on the left side are the position results of type 1, type 2, type 3, and type 4 obtained from Faster R-CNN, respectively; (b), (d), (f) and (h) on the right side are the position results of type 1, type 2, type3, and type 4 obtained from SSD512. The recognition results of both models are stable for large objects, such as triangles and quadrangles in the center of the image. In contrast to small objects, such as lock shapes on both sides of the image and figures (e) and (f) on the third row, the target detection results by the two models are quite different. The detection results of Faster R-CNN (figures on the left side) are more robust than those of SSD12 (the picture on the right) for small targets. When the confidence values of detection results of Faster R-CNN are greater than 0.99, many of the results of SSD512 have confidence values below 0.95.

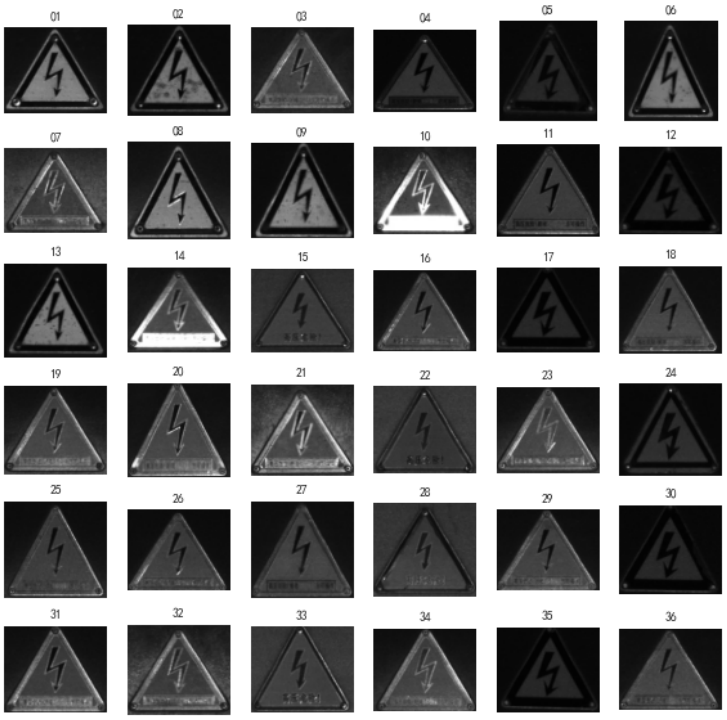
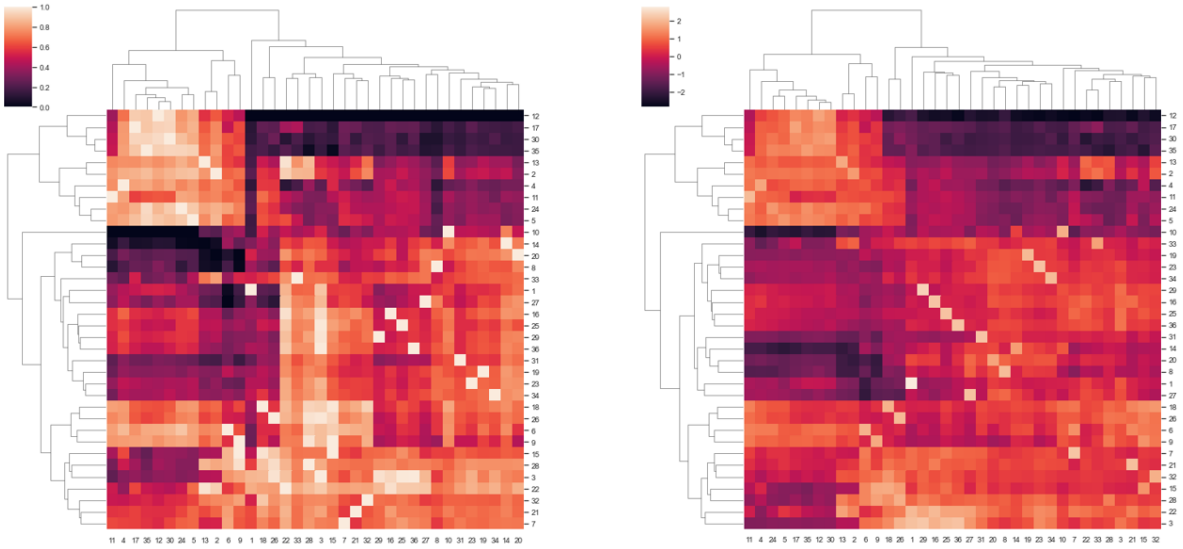


Fig. 9. The detection results of triangles signs in testing dataset obtained from Faster R-CNN

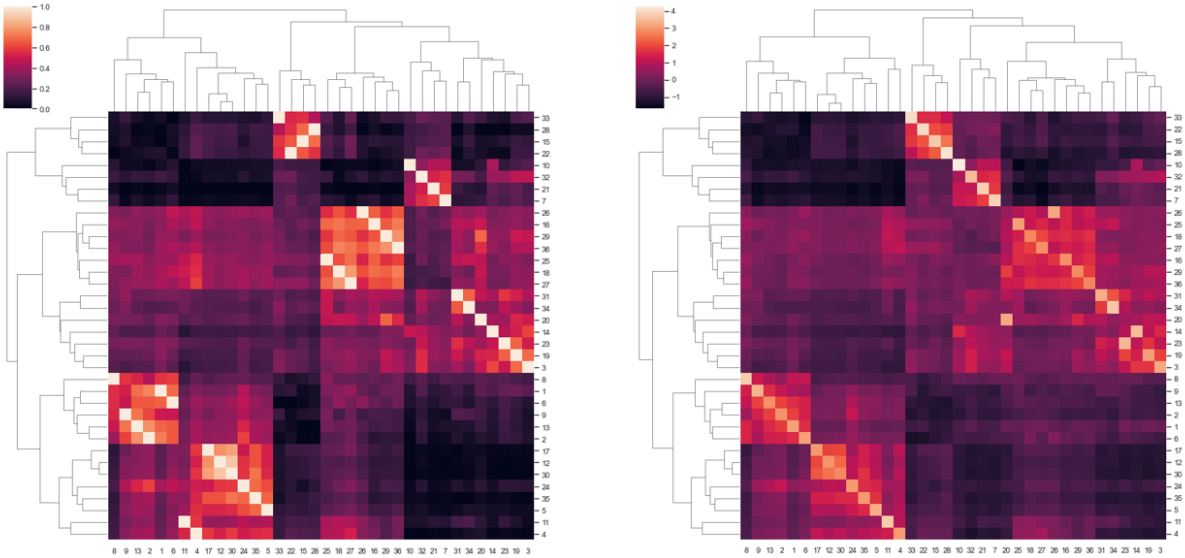
Fig. 9 shows the detection results of triangles signs in the testing dataset obtained from Faster R-CNN. The predicted results show accuracy and robustness to different perturbations like darkness, light exposure, and blurring. We apply cosine similarity and SSIM to the triangle sign sample pairs from the testing dataset in Fig. 9 and obtain similarity matrices between all samples. To identify groups of correlated triangle sign samples and analyze hierarchical clusters in the similarity matrices, we use cluster heatmaps to reveal the block structures along the diagonal of the triangle sign samples. Cluster heatmaps are commonly used to identify outliers and tissue subtypes. The results of the clustered heatmaps are shown in Fig. 10 and Fig. 11. Normalization results with Z-score and Min-Max scaling are calculated based on columns on the clustered heatmaps to enable a comparison between cosine similarities and SSIM of all triangle sign samples based on similar scaling. Z-score normalization rescales the columns values of similarity matrices to obtain the properties of a Gaussian distribution with a mean equal to 0 and standard deviation equal to 1, while Min-Max scaling shrinks the range of the data so that the range is fixed between 0 and 1. Fig. 10 shows the clustered heatmap results of cosine similarities. The picture on the left side shows the clustered heatmap result based on Z-score standardization, and the picture on the right shows the result based on Min-Max normalization. Similar to Fig. 10, Fig. 11 shows the clustered heatmap results of SSIM with the Z-score and Min-Max normalization.



(a) Clustered heatmap of standardized cosine similarities with Z-score between the predicted ROIs

(b) Clustered heatmap of Max/Min normalized cosine similarities between the predicted ROIs

Fig. 10. Clustered heatmap of cosine similarities with Z-score and Min-Max normalization



(a) Clustered heatmap of standardized SSIM with Z-score between the predicted ROIs

(b) Clustered heatmap of Max/Min normalized SSIM between the predicted ROIs

Fig. 11. Clustered heatmap of cosine similarities with Z-score and Min-Max normalization

Cosine similarity measures the cosine of the angle between two vectors in a multi-dimensional space. It determines the similarity between two image samples based on the orientation of the image samples, irrespective of their magnitude. Fig. 12 shows the clustered heatmaps results of cosine similarities with Min-Max (left) and Z-score (right) normalization for the testing images. Similar images show brighter colors on the heatmap. The triangle sign images clustered in the upper left corner show high similarities, mainly triangle sign images presenting dark backgrounds. From the perspective of image composition, SSIM defines structural information as an attribute independent of brightness and contrast, reflects the structure of objects in the scene, and models an im-

age as a combination of luminance, contrast, and structure. Fig. 13 shows the clustered heatmaps results of SSIM with Min-Max (left) and Z-score (right) normalization for the testing images. Several cluster blocks with brighter colors represent triangle sign images with similar textures as in the middle of Fig. 13. Based on the composite information from luminance, contrast, and structure, SSIM can better distinguish images sample with different types of perturbations like darkness, light exposure, and blurring. Comparison between the images on the left and right shows that Z-score provides more homogeneous similarity results than Min-Max normalization.

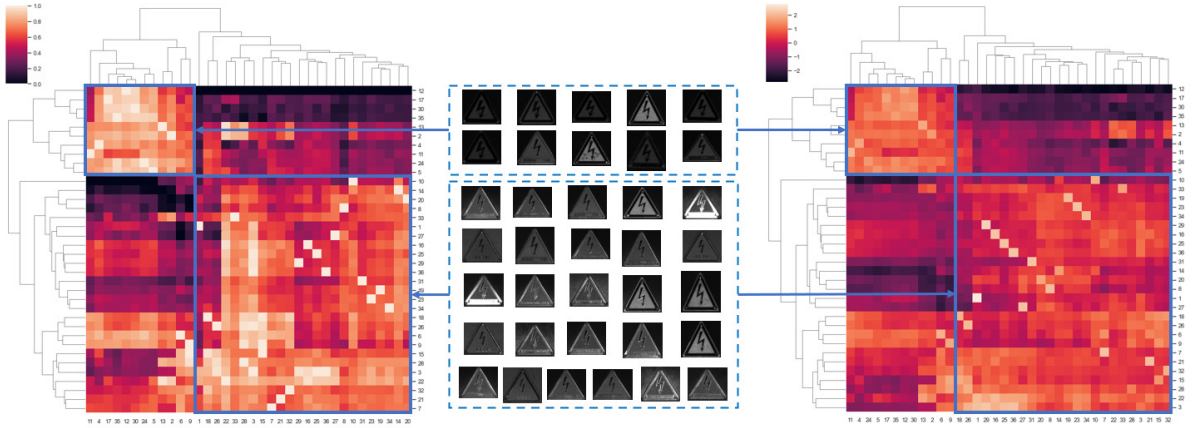


Fig. 12. Clustered heatmaps results of cosine similarities with Min-Max (left) and Z-score (right) normalization

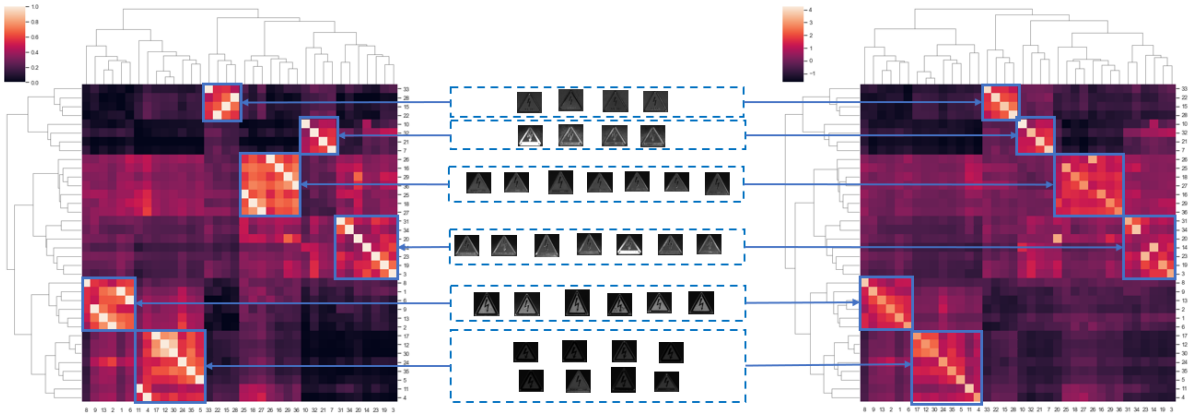


Fig. 13. Clustered heatmaps results of SSIM with Min-Max (left) and Z-score (right) normalization

5 Discussion

It is worth noting that under the same target detection algorithm (Faster R-CNN or SSD512), the performances of SSIM-based fault diagnosis methods are much worse on the three sub-datasets than those of CS-based methods. This is mainly because the images to be tested are greatly affected by ambient light, and SSIM is less robust than CS concerning the lighting factors. Reducing the influence of light in fault diagnosis would be an effective way to upgrade the diagnosis performance.

Fig. 8 shows that Faster R-CNN has higher average precision than SSD512 for detecting the ROI of different types of train body signs. This demonstrates that the two-stage algorithm has advantages in terms of detection accuracy compared to the one-stage algorithms. The Faster R-CNN first uses selective search or Region Proposal

Net (RPN) to generate the region of the target area, followed by classification and regression. This two-stage method can achieve higher accuracy compared to SSD512 but is limited to detection speed at the same time. Since SSD512 is an end-to-end object detection algorithm that uses a light network to predict the target boundary box and classification probability score with different ratios and scales in each feature map, the detection speed is improved. Faster R-CNN is a more suitable scenario when the highest priority is accuracy instead of real-time performance. Under acceptable model accuracy requirements, one way to increase the inference speed of Faster R-CNN is to use a smaller backbone network [25].

Cosine similarity can detect similar image samples based on a smaller cosine angle between them, even if the samples are far apart by the Euclidean distance because of their pixel magnitude. SSIM considers image transformation as a change in structural information in combination with luminance and contrast masking terms. Structural information is based on the idea that pixels in the image are strongly interdependent when spatially close. Luminance describes that image transformation and distortion are less visible for human eyes in bright regions in the image, and contrast masking describes the phenomenon that image distortions become less visible in regions with a significant texture pattern. In contrast to cosine similarity, SSIM detects more differences in texture in the image, which leads to a different appearance in the clustered heatmaps, as shown in Fig. 12 and Fig. 13.

In the future research phase, it is worth applying more deep-learning-based methods, e.g., DenseBox, to the abnormal detection of the EBC of the train body. For better model performance, the time required for single detection should be reduced as much as possible. Besides, in the subsequent optimization process of this method, the threshold value to distinguish between positive and negative samples can be selected more reasonably with hyperparameter tuning optimization methods to achieve more accurate and robust inferences.

6 Conclusion

This paper proposes a novel method of combining Faster R-CNN and CS for the fault detection of the train body signboard. First, the image's region of interest to be detected is extracted and clipped by multiple convolution kernels in Faster R-CNN. In consequence, the similarity between the ROI and the standard template image is calculated by CS and compared with the threshold value to determine the fault state of the detection target. The experimental results show that combining Faster R-CNN with cosine similarity (CS) is more effective than the other three types of combination methods for anomaly detection, namely, SSD512 and SSIM, SSD512 and CS, and Faster R-CNN and SSIM. Specifically, in the target detection process, the accuracy of Faster R-CNN is much higher than that of SSD512, where the average accuracy of Faster R-CNN is 0.96 while that of SSD512 is only 0.88. Compared with SSIM, the CS avoids the influence of light, thus yielding higher reliability in fault detection. Overall, the combination of Faster R-CNN and CS renders considerable reliability and applicability.

7 Acknowledgement

This work is supported by the Natural Science Foundation of China (62173279, U1934221, 61733015) and the Sichuan Science and Technology Program under Grant (2020YFQ0057, 2021JDJQ0012).

References

[1] G. Sun, W. Xu, Y. Liang, D. Zhao, Image recognition algorithm for side frame key of train based on shape context, Journal of Railway Science and Engineering 11(6)(2014) 127-131.

[2] J. Zhang, Y. Feng, S. Wang, Research on fault recognition algorithm of center plate bolts based on WLD-LPQ features, Journal of Railway Science and Engineering 15(9)(2018) 2349-2358.

[3] S. Jia, S. Jiang, Z. Lin, N. Li, M. Xu, S. Yu, A survey: Deep learning for hyperspectral image classification with few labeled samples, Neurocomputing 448(2021) 179-204.

[4] L. Nan, Z. Wei, Z. Cao, Automatic fault recognition for Brake-Shoe-Key losing of freight train, Optik - International Journal for Light and Electron Optics 126(23)(2015) 4735-4742.

[5] F. Zhou, R. Zou, Y. Qiu, H. Gao, Automated visual inspection of angle cocks during train operation, in: Proc. 2013 Institution of Mechanical Engineers Part F Journal of Rail & Rapid Transit, 2013.

[6] Y.S. Yang, A.B. Ming, Y.Y. Zhang, Y.S. Zhu, Discriminative non-negative matrix factorization (DNMF) and its application to the fault diagnosis of diesel engine, Mechanical Systems and Signal Processing 95(2017) 158-171.
[7] L. Kou, Y. Qin, X. Zhao, Y. Fu, Integrating synthetic minority oversampling and gradient boosting decision tree for bogie fault diagnosis in rail vehicles, in: Proc. 2019 the Institution of Mechanical Engineers, 2019.
[8] R. Zou, Z.Y. Xu, J.Y. Li, F.Q. Zhou, Real-time monitoring of brake shoe keys in freight cars, Frontiers of Information Technology & Electronic Engineering 16(3)(2015) 191-204.
[9] F.B. Cao, The fault automatic detection system based on PCA and SVM, [master dissertation] Beijing: Beijing Jiaotong University, 2011, <https://cdmd.cnki.com.cn/Article/CDMD-10004-1011102490.htm/>.
[10] J. Li, Key Locomotive Bolts Fault Detection Technique, [master dissertation] Chengdu: Southwest Jiaotong University, 2015, <https://cdmd.cnki.com.cn/Article/CDMD-10613-1015338720.htm/>.
[11] N. Qin, K. Liang, D. Huang, L. Ma, A.H. Kemp, Multiple convolutional recurrent neural networks for fault identification and performance degradation evaluation of high-speed train bogie, IEEE Transactions on Neural Networks and Learning Systems 31(12)(2020) 5363-5376.
[12] R. Girshick, Fast r-cnn, in: Proc. 2015 IEEE International Conference on Computer Vision, 2015.
[13] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.Y. Fu, A.C. Berg, Ssd: Single shot multibox detector, in: Proc. 2016 European Conference on Computer Vision, 2016.
[14] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: unified, real-time object detection, in: Proc. 2016 IEEE Conference on Computer Vision and Pattern Recognition, 2016.
[15] J. Sun, Z. Xiao, Y. Xie, Automatic multi-fault recognition in TFDS based on convolutional neural network, Neurocomputing 222(2017) 127-136.
[16] X.X. Liu, Research on rail surface defect recognition based on convolutional neural network, [master dissertation] Mianyang: Southwest University of Science and Technology, 2018, <https://cdmd.cnki.com.cn/Article/CDMD-10619-1018198980.htm/>.
[17] J.W. Chen, Study in the detection of defects of fasteners on high-speed railway catenary support devices, [master dissertation] Chengdu: Southwest Jiaotong University, 2018, <https://cdmd.cnki.com.cn/Article/CDMD-10613-1018709900.htm/>.
[18] A. Reddy, V. Indragandhi, L. Ravi, V. Subramaniyaswamy, Detection of cracks and damage in wind turbine blades using artificial intelligence-based image analytics, Measurement 147(2019) 106823.
[19] X. Fu, K. Li, J. Liu, K. Li, Z. Zeng, C. Chen, A two-stage attention aware method for train bearing shed oil inspection based on convolutional neural networks, Neurocomputing 380(2020) 212-224.
[20] S. Akcay, A. Atapour-Abarghouei, T.P. Breckon, Ganomaly: Semi-supervised anomaly detection via adversarial training, in: Proc. 2018 Asian Conference on Computer Vision, 2018.
[21] A. Hore, D. Ziou, Image quality metrics: PSNR vs. SSIM, in: Proc. 2010 International Conference on Pattern Recognition, 2010.
[22] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Transactions on Image Processing 13(4)(2004) 600-612.
[23] G. Pirlo, D. Impedovo, Cosine similarity for analysis and verification of static signatures, IET Biometrics 2(4)(2013) 151-158.
[24] Z. Qin, Z. Li, Z. Zhang, Y. Bao, G. Yu, Y. Peng, J. Sun, ThunderNet: Towards real-time generic object detection on mobile devices, in: Proc. 2019 IEEE/CVF International Conference on Computer Vision, 2019.
[25] J.A. Kim, J.Y. Sung, S.H. Park, Comparison of Faster-RCNN, YOLO, and SSD for real-time vehicle type recognition, in: Proc. 2020 IEEE International Conference on Consumer Electronics - Asia, 2020.