Journal of Zhejiang University-SCIENCE A 2025 26(10):997-1009
www.jzus.zju.edu.cn; www.springer.com/journal/11582
E-mail: jzus_a@zju.edu.cn

Research Article
https://doi.org/10.1631/jzus.A2400549

Self-attention and convolutional feature fusion for real-time intelligent fault detection of high-speed railway pantographs

Xufeng LI[1], Jien MA[1]✉, Ping TAN[2]✉, Lanfen LIN[3], Lin QIU[1], Youtong FANG[1]

[1]*College of Electrical Engineering, Zhejiang University, Hangzhou 310027, China*
[2]*School of Automation and Electricity Engineering, Zhejiang University of Science and Technology, Hangzhou 310023, China*
[3]*College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China*

Abstract: Currently, most trains are equipped with dedicated cameras for capturing pantograph videos. Pantographs are core to the high-speed-railway pantograph-catenary system, and their failure directly affects the normal operation of high-speed trains. However, given the complex and variable real-world operational conditions of high-speed railways, there is no real-time and robust pantograph fault-detection method capable of handling large volumes of surveillance video. Hence, it is of paramount importance to maintain real-time monitoring and analysis of pantographs. Our study presents a real-time intelligent detection technology for identifying faults in high-speed railway pantographs, utilizing a fusion of self-attention and convolution features. We delved into lightweight multi-scale feature-extraction and fault-detection models based on deep learning to detect pantograph anomalies. Compared with traditional methods, this approach achieves high recall and accuracy in pantograph recognition, accurately pinpointing issues like discharge sparks, pantograph horns, and carbon pantograph-slide malfunctions. After experimentation and validation with actual surveillance videos of electric multiple-unit train, our algorithmic model demonstrates real-time, high-accuracy performance even under complex operational conditions.

Key words: High-speed railway pantograph; Self-attention; Convolutional neural network (CNN); Real-time; Feature fusion; Fault detection

1 Introduction

China's high-speed railway has gradually become a network. Since the introduction of the 'Mid-Long Term Railway Network Plan' in 2016, 70% of the total operational scale of the 'eight vertical and eight horizontal' high-speed railway network has been completed (Tan et al., 2016). By 2023, the total operating mileage of China's high-speed railway had exceeded 45000 km. The amount of new high-speed railway network is tremendous, which creates new requirements and challenges for the daily maintenance of railway safety operations (Tan et al., 2020, 2021, 2022; Fang and Ma, 2023; Wu et al., 2023).

✉ Jien MA, majien@zju.edu.cn
Ping TAN, 115011@zust.edu.cn
Xufeng LI, https://orcid.org/0000-0002-0506-4834
Jien MA, https://orcid.org/0000-0003-0775-2793
Ping TAN, https://orcid.org/0000-0001-8656-3514

Received Nov. 28, 2024; Revision accepted Jan. 3, 2025; Crosschecked Sept. 11, 2025

The kinetic energy of a high-speed train is provided by the pantograph-catenary system, as shown in Fig. 1. The pantograph obtains electric energy through sliding contact with the contact line and transmits it to the train to ensure normal operation (Zhou et al., 2011). Hence, high-quality current collection is a prerequisite for high-speed, safe, and reliable operation of trains. However, the contact position between the carbon slide and the contact line is the primary area of failure, and needs to be a focus of monitoring. In order to ensure the safe operation of high-speed trains, the railway operation department must evaluate the state of key components of the arch network in a timely way. Due to the pantograph structure and working characteristics, the non-contact monitoring method is safer. The most common method involves capturing pantograph surveillance video using a camera mounted on the train and transmitting it in real-time to the main system. At 350 km/h, any delay caused by pantograph failure can result in incalculable damage, with every second counting.

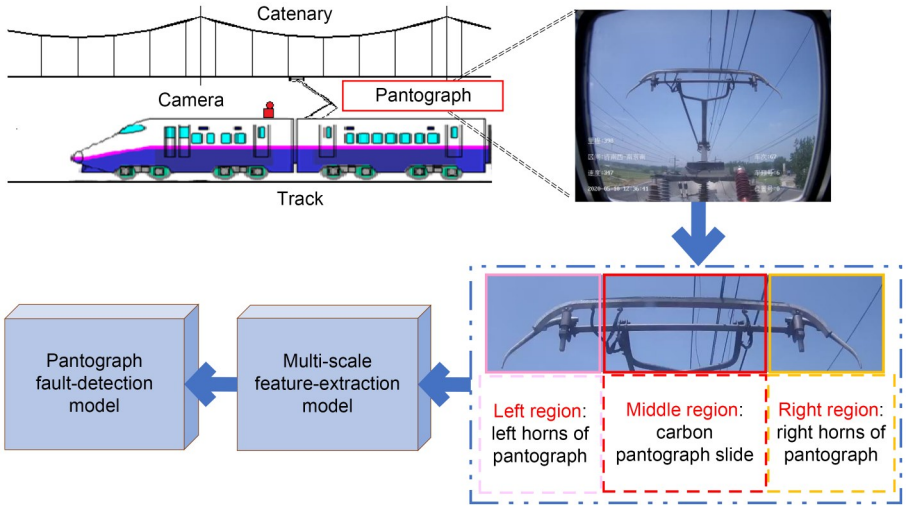

Fig. 1 Pantograph-catenary system and overall model

At present, the video monitoring system used for carriage pantographs only records and cannot perform real-time fault detection. In addition, the auto dropping device (ADD) does not effectively detect pantograph faults or provide fall protection, a shortcoming which can lead to serious accidents. The intelligent analysis function based on surveillance video attempts to address this issue by detecting faults in time and thus enabling railway engineers to take relevant measures to avoid further deterioration of accidents. Thus, many researchers use image-processing technology to detect pantograph-catenary faults. Deng et al. (2022) designed a pantograph slip-wear detection system using 3D structured light detection. Karaduman and Akin (2022) proposed a new predictive maintenance method using the fuzzy classifier in railway systems. Wei et al. (2020) proposed an innovative and intelligent method based on deep learning and image-processing technologies for online condition monitoring of the pantograph slide plate. Mo et al. (2022) used the YOLOv4 (You Only Look Once) model, along with edge extraction and other traditional image-processing algorithms, to detect pantograph slide defects. Liu et al. (2021) applied machine vision technology to measure the thickness of the slide, thereby reducing pantograph-catenary accidents caused by errors in manual measurement. Li et al. (2022) proposed an accurate real-time attitude-detection method for 3D monocular pantographs, which used the super ability of deep learning. Chen et al. (2022) proposed a depth-vision neural network detection method. Based on the YOLOv5 model, the depth pantograph-detection network (DPDN) was established to identify pantograph regions in different complex scenes, and then the image visual feature extraction (IVFE) algorithm was used to detect the uneven distribution of contact points between the pantograph and catenary in the pantograph region. Although various detection methods are available, significant challenges persist in complex real-world operational conditions, such as insufficient real-time capabilities and low fault-recognition rates.

As shown in Fig. 1, the pantograph is divided into three regions: left, middle, and right. Although the middle region is the main working region, the failure of the left or right region will also cause the pantograph to lose balance, so detection of the pantograph cannot focus only on its contact-point region. Furthermore, the actual operating conditions of trains are complex, for example, different bow types, different weather conditions, and trains passing over bridges and platforms, as shown in Fig. 2. Since the pantograph-monitoring scene is determined by actual operating conditions and the background is extremely complex, real-time processing is required in addition to efficient and robust image feature-extraction and classification methods. The integration of all these aspects is the main difficulty we hoped to solve in this study.

Compared with traditional image-processing methods, deep convolutional neural networks have shown relatively strong performance. Since AlexNet (Krizhevsky et al., 2017) was proposed, deep convolutional neural networks have gradually become a hot research topic in the field of computer vision. Simonyan and Zisserman (2015) proposed visual geometry group

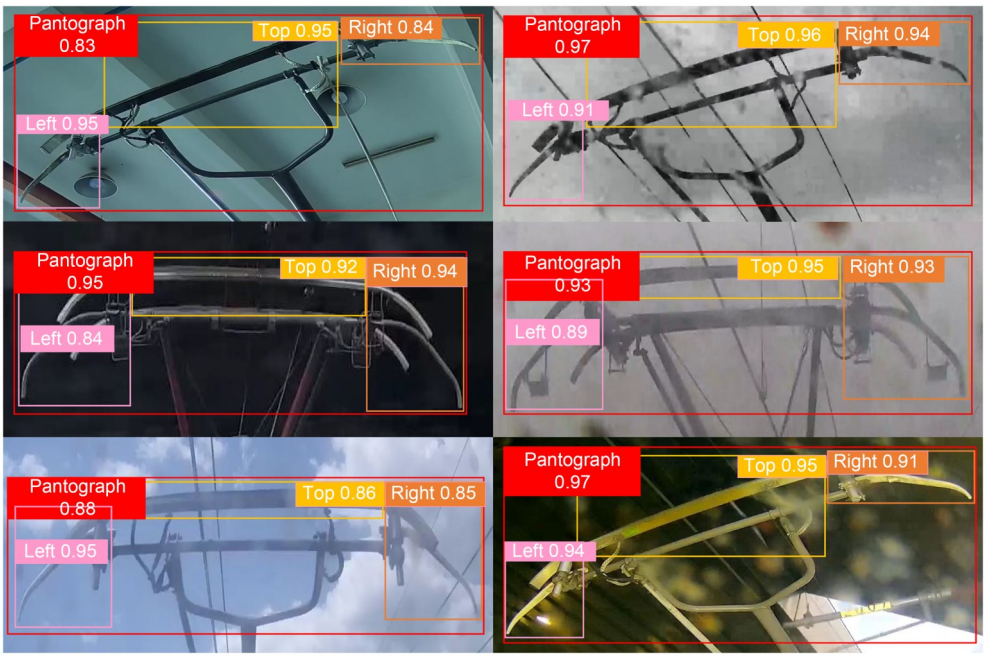

Fig. 2 Pantograph detection using the proposed method in complex conditions. The values in the figure refer to confidence coefficients

network (VGGNet), which has a deep network structure, a small convolutional kernel, and a pooled sampling domain, enabling it to offer good transfer learning ability in feature extraction. He et al. (2016) proposed the ResNet structure to solve the degradation problem caused by deep networks, and then proposed ResNetXt (Xie et al., 2017) and ResNet-D (He et al., 2019), which improved the feature-extraction performance. Ding et al. (2021) proposed RepVGG, which only consisted of recall, 3×3 convolution, and rectified linear unit (ReLU) activation functions, and further enhanced the feature-extraction performance of a VGGNet through a simple branch-free structure. Although the depth and structure of convolutional networks are constantly improved and optimized, the characteristics of convolutional kernels limit the capacity for long-distance dependence. As transformer (Vaswani et al., 2017) becomes a hot topic in natural language processing, more and more researchers are trying to transplant transformer methods to image processing with long-term dependent features, in order to break the inherent limitations of convolutional networks. Vision transformer (Dosovitskiy et al., 2021) realized the application of the transformer in images for the first time. It uses image block coding as the network input and replaces the convolutional operation with a self-attention module. It thus provides a new model in the field of computer vision, different from convolutional neural networks (CNNs), but its performance needs to be improved. Considering the advantages of convolution, some researchers began to try composite models with both CNN and self-attention. For example, the CMT (CNN meet transformer) (Guo et al., 2022) model greatly improved feature-extraction performance. This indicates that the composite model has definite research value. To tackle fault detection in complex scenarios, a feature-extraction method with higher performance is required. Therefore, we chose to design a detection model here by combining the transformer and convolutional architectures.

To overcome the challenges identified above, we propose an inter-block feature fusion method. The contribution of the study is threefold:

(1) The method integrates self-attention and convolutional features to improve the feature-extraction performance of convolutional networks and accurately identify pantographs in complex scenes.

(2) The lightweight multi-scale feature-extraction and fault-detection models are built to meet the requirements of real-time detection. Network parameters are reduced, and the model reasoning speed is improved.

(3) A complete and accurate fault-detection scheme for high-speed railway pantographs is established for the daily operation of trains.

2 Method

2.1 Overall model of pantograph detection

The overall model consists of two sub-models: a multi-scale feature-extraction model and a pantograph fault-detection model. In response to the limited number of samples of pantograph failures, we designed the method as follows: First, a feature-extraction network model is designed to learn the components of three regions using a large amount of normal pantograph image data. Then, a feature sample library of normal components is constructed. Finally, the presence of pantograph failures is detected by computing their confidence scores, as shown in Fig. 2. Compared to other types of faults, discharge spark faults have relatively simple shapes and colors, making them detectable through object recognition. As a result, these faults can also serve as subjects for learning. In this way, the problem of insufficient pantograph fault samples is avoided.

The multi-scale feature-extraction model is shown in Fig. 3, and the specific parameters are shown in Tables 1 and 2, in which the CBS (convolution, batch normalization, and sigmoid linear unit) module consists of three parts: convolution, batch normalization (BN), and sigmoid linear unit (SiLU) activation functions (Bochkovskiy et al., 2020). The function of BN is to make the network easier to converge. It also has a regularization function to prevent overfitting. Compared with ReLU, SiLU offers smoothness and non-monotonicity, which improves its effects. The calculation formula is given in Eq. (1):

$$f(x)=\frac{x}{1+\mathrm{e}^{-x}}. \tag{1}$$

To begin with, Table 1 shows a conventional feature-extraction network; this network is then connected with a feature pyramid network (FPN) structure (Lin et al., 2017; Hao et al., 2022) and pixel aggregation network (PAN) (Wang et al., 2019), as shown in Table 2. The FPN layer conveys strong semantic features from the top down (P1–P9), while the PAN conveys strong localization features from the bottom up (P10–P17). In this way, the parameters of different detection layers are aggregated from different backbone layers, further improving feature-extraction capability. The SACNNBlock (self attention and convolutional neural network block) module will be described

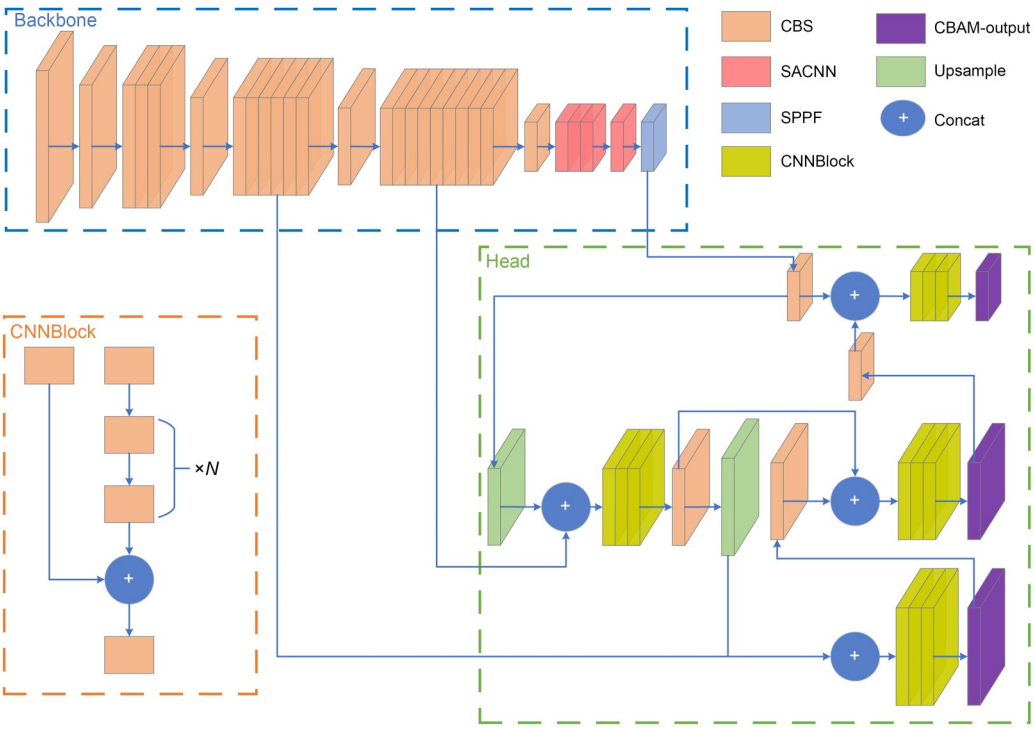

Fig. 3 Multi-scale feature-extraction model. CBAM: convolutional block attention module; SACNN: self attention and convolutional neural network; SPPF: spatial pyramid pooling-fast. References to color refer to the online version of this figure

Table 1 Backbone of the model

Layer number	Layer name	Output size	Built-in parameter
E1	CBS	320×320	6×6, 64, stride 2
E2	CBS	160×160	3×3, 128, stride 2
E3	CBSBlock	160×160	[1×1, 128]×3
E4	CBS	80×80	3×3, 256, stride 2
E5	CBSBlock	80×80	[1×1, 256]×6
E6	CBS	40×40	3×3, 512, stride 2
E7	CBSBlock	40×40	[1×1, 512]×9
E8	CBS	20×20	3×3, 1024, stride 2
E9	SACNNBlock	20×20	[1×1, 1024]×3
E10	SACNN	20×20	[1×1, 1024]×1
E11	SPPF	20×20	5×5, 1024, pool

Table 2 Head of the model

Layer number	Layer name	Output size	Built-in parameter
P1	CBS	20×20	1×1, 512
P2	Upsample	40×40	512
P3	Concat	40×40	1024, E7&P2
P4	CNNBlock	40×40	512
P5	CBS	40×40	1×1, 256
P6	Upsample	80×80	256
P7	Concat	80×80	512, E5&P6
P8	CNNBlock	80×80	256
P9	CBAM	80×80	256, output 1
P10	CBS	40×40	3×3, 256 stride 2
P11	Concat	40×40	512, P5&P10
P12	CNNBlock	40×40	512
P13	CBAM	40×40	512, output 2
P14	CBS	20×20	3×3, 512 stride 2
P15	Concat	20×20	1024, P1&P14
P16	CNNBlock	20×20	1024
P17	CBAM	20×20	1024, output 3

&: concat

in the next section. The CNNBlock module is similar to the C3 module in YOLO, as shown in Fig. 3. Similarly, the spatial pyramid pooling-fast (SPPF) (Bochkovskiy et al., 2020; Mu et al., 2023) module at the bottom of the left network processes parallel inputs through multiple max-pools of varying sizes, followed by further fusion. This approach can address the multi-scale target problem to some extent. Finally, the network outputs three feature maps of different sizes for prediction, which can improve the recognition accuracy of targets of different sizes.

2.2 SACNNBlock

Similar to CBS superimposed on a CBSBlock module, the SACNNBlock module is overlaid by SACNN, which is formed by the integration of self-attention and CNN, as shown in Fig. 4. It is an improved model based on ACmix (Pan et al., 2022). In the convolution path, it first uses a lightweight fully connected layer with a group number of N to aggregate the number of channels, and then restores the number of channels to C by a group convolution. By means of this, the amount of extra computation generated here

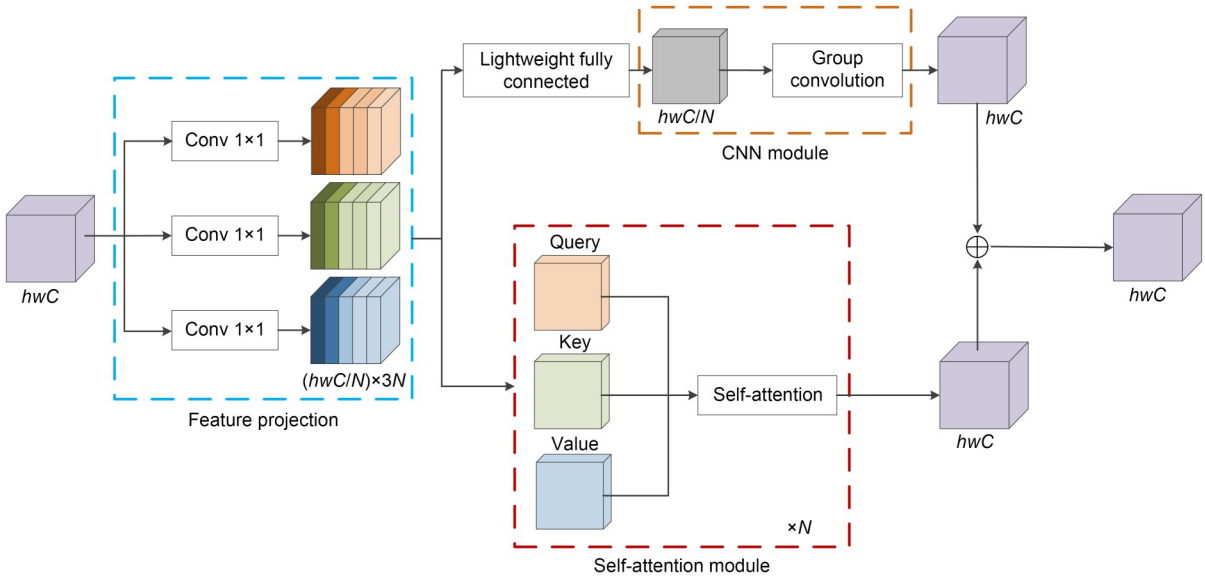

Fig. 4 A self-attention and convolutional neural network (SACNN)

is reduced from $(3k^2+k^4)C\times hw$ (Pan et al., 2022) to $(3+k^2)C\times hw$ compared to ACmix. k is the kernel size for convolution, C is the input or output channel, and h and w are the length and width of the feature map, respectively.

First, we used three 1×1 convolutional kernels for the convolutional operation of the input feature graphs, which are divided into N blocks to obtain $3\times N$ intermediate feature sets, and then we generated two calculation routes. The lower part of Fig. 4 shows the calculation route of self-attention, which generates three feature matrices, Query, Key, and Value, for the calculation of self-attention. The formula is given in Eq. (2) (Vaswani et al., 2017; Dosovitskiy et al., 2021; Hu et al., 2022):

$$\boldsymbol{Z}=\mathrm{Softmax}\left(\frac{\boldsymbol{QK}^{\mathrm{T}}}{\sqrt{d_{\boldsymbol{K}}}}+\boldsymbol{P}\right)\boldsymbol{V}, \tag{2}$$

where $\boldsymbol{Z}$ is the calculation result of self-attention, $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ are the three feature matrices of Query, Key, and Value, respectively, $d_{\boldsymbol{K}}$ is the feature dimension of $\boldsymbol{K}$, and $\boldsymbol{P}$ is the relative position coding.

The upper part of Fig. 3 shows the convolutional calculation route, where $3\times N$ intermediate feature sets are spliced. After that, a lightweight fully connected layer is adopted to generate C/N feature maps, and a 3×3 kernel is used for convolutional calculation. The obtained results are weighted and summed with the calculation result of self-attention. Finally, the calculation result of the fusion feature is output. It can be seen that the calculation process did not change the size of the feature map. The lightweight fully connected layer is composed of group convolution with the C/N group number, which is designed to reduce the number of parameters and the amount of computation.

2.3 Multi-scale feature output with convolutional block attention module (CBAM)

CBAM is a lightweight attention module that was created in 2018 and added to classical structures such as ResNet to improve the performance of convolutional networks (Woo et al., 2018; Wen et al., 2024). In this study, we added CBAM as a plug-in in front of three different output sizes for the network, as shown in Fig. 3.

CBAM consists of two parts, namely a channel attention module (CAM) and a spatial attention module (SAM) (Woo et al., 2018; Ni et al., 2022; Yao et al., 2022), and its main structure is shown in Fig. 5.

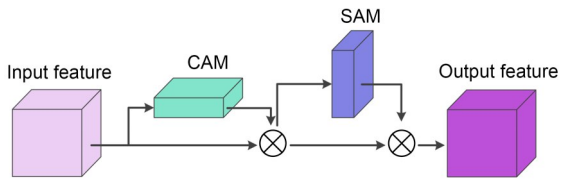

Fig. 5 CBAM

2.3.1 CAM

First, the input feature images are obtained by max pooling and average pooling based on width and height, respectively, to obtain two $1\times1\times C$ feature images. Then, they are respectively sent into a two-layer multilayer perceptron (MLP) (Tan et al., 2024). The number of neurons in the first layer is C/r (r is the reduction rate), the activation function is ReLU, and the number of neurons in the second layer is C. This two-layer neural network is shared. Then, the features output by MLP are added and activated by the sigmoid to generate the final channel-attention feature. Finally, the channel attention feature is multiplied by the input feature to generate the input features required by SAM.

2.3.2 SAM

The output features of CAM are used as the input features of this module. First, two $h\times w\times1$ features are obtained by max pooling and average pooling based on the channel, and then they are concatenated. After a 7×7 convolutional operation, the dimension is reduced to one channel, namely $h\times w\times1$. Next, a spatial attention feature is generated by the sigmoid. Finally, this feature is multiplied by the input feature of this module to get the final output feature.

2.4 Loss function

The loss function in this study includes classification loss (L_{cls}), bounding-box regression loss (L_{box}), and objectness loss (L_{obj}), in which the L_{cls} and L_{obj} are calculated by the binary cross-entropy loss function. The calculation formulas are given in Eqs. (3) and (4) (Bochkovskiy et al., 2020).

$$L_{\mathrm{cls}}=\frac{1}{n}\sum_{x}[y^{*}\times\ln y+(1-y^{*})\times\ln(1-y)], \tag{3}$$

$$L_{\text{obj}}=\frac{1}{n}\sum_{x}[z^{*}\times\ln z+(1-z^{*})\times\ln(1-z)], \tag{4}$$

where n is the total number of samples, x indicates samples, y^{*} and z^{*} are label values, and y and z are predictive values.

Complete intersection over union loss (CIoU loss) is used for L_{box}. Eqs. (5)–(8) give the calculation formulas (Zheng et al., 2020; Yao et al., 2022).

$$L_{\text{box}}=1-I_{\text{oU}}+\frac{\rho^{2}(\boldsymbol{A},\boldsymbol{B})}{c^{2}}+\alpha v, \tag{5}$$

$$I_{\text{oU}}=\frac{|\boldsymbol{A}\cap\boldsymbol{B}|}{|\boldsymbol{A}\cup\boldsymbol{B}|}, \tag{6}$$

$$v=\frac{4}{\pi^{2}}\left(\arctan\frac{w_{\boldsymbol{B}}}{h_{\boldsymbol{B}}}-\arctan\frac{w_{\boldsymbol{A}}}{h_{\boldsymbol{A}}}\right)^{2}, \tag{7}$$

$$\alpha=\frac{v}{1-I_{\text{oU}}+v}, \tag{8}$$

where $\boldsymbol{A}$ is the predictive box, $\boldsymbol{B}$ is the target box, ρ is the distance between the center points of $\boldsymbol{A}$ and $\boldsymbol{B}$, c is the diagonal length of the minimum surrounding of $\boldsymbol{A}$ rectangle and $\boldsymbol{B}$ rectangle, v is the similarity of the aspect ratios of $\boldsymbol{A}$ and $\boldsymbol{B}$, α is a positive trade-off parameter of v, and w and h are the width and height of the box, respectively. The larger the I_{oU}, the larger the overlap area between $\boldsymbol{A}$ and $\boldsymbol{B}$, resulting in a larger α and greater impact on the v. In this case, it focuses on optimizing the aspect ratios of $\boldsymbol{A}$ and $\boldsymbol{B}$. Otherwise, the smaller the I_{oU}, the smaller the overlap area between $\boldsymbol{A}$ and $\boldsymbol{B}$, resulting in a smaller α and smaller impact on the v. In this case, it focuses on optimizing the distance between $\boldsymbol{A}$ and $\boldsymbol{B}$.

Therefore, the total loss (L_{total}) is as follows: $L_{\text{total}}=\lambda_{\text{cls}}\times L_{\text{cls}}+\lambda_{\text{obj}}\times L_{\text{obj}}+\lambda_{\text{box}}\times L_{\text{box}}$, where λ is the gain coefficient of each loss function and the default value is 1.

2.5 Pantograph fault-detection model

The pantograph image is divided into four parts: the pantograph collector head, the left horns of the pantograph, the carbon pantograph slide, and the right horns of the pantograph. There is also an extra part, the discharge spark. Only abnormal discharge sparks with a large area are detected here. The detection process is shown in Fig. 6, and the pantograph fault types are shown in Fig. 7.

The blue dashed box in Fig. 6 illustrates the training process of the pantograph fault-detection model. It begins with the input of normal pantograph collector heads, extracted by the target-recognition network, which is subsequently divided into three regions and fed into a multi-scale feature-extraction model. Next, multiple independent logistic regression classifiers are utilized for training, resulting in the establishment of a feature sample library containing normal structures within these three regions.

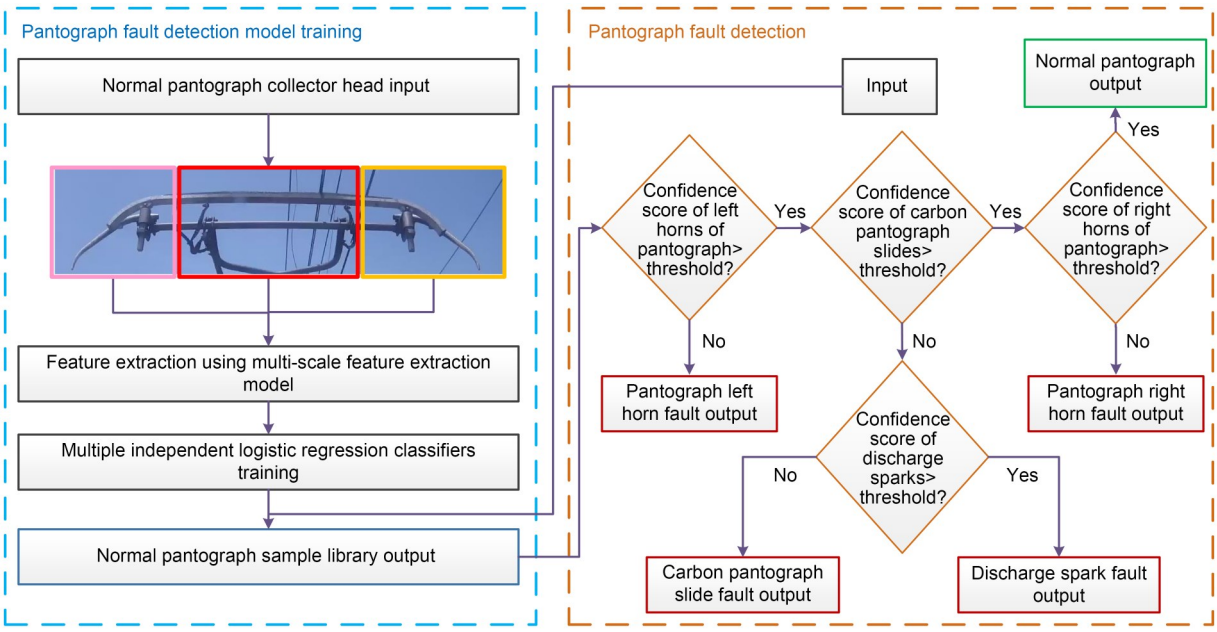

Fig. 6 Pantograph fault-detection model. References to color refer to the online version of this figure

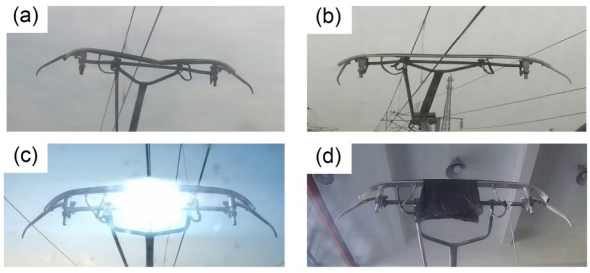

Fig. 7 (a) Carbon pantograph slide deformation; (b) horns of the pantograph fault; (c) discharge spark; (d) foreign body attachment

The orange dashed box in Fig. 6 depicts the detection process. Here, the pantograph image under examination is inputted into the model. Its features are then compared with those in the standard library, generating a confidence score. This score is determined by multiplying the objectness by the classification. Finally, these results are sequentially compared with the thresholds set within each region to ascertain the presence of a fault. If the score exceeds the threshold value, the object is considered normal; otherwise, it is deemed abnormal. The methodology for selecting these thresholds will be elucidated further in the subsequent chapter.

3 Experiments

3.1 Pantograph dataset

There are a total of 2954 images in the training set, and each image sample is marked according to the standards in Fig. 8. The test set consisted of 600 pantograph images, including 500 normal pantograph images and 100 pantograph fault images.

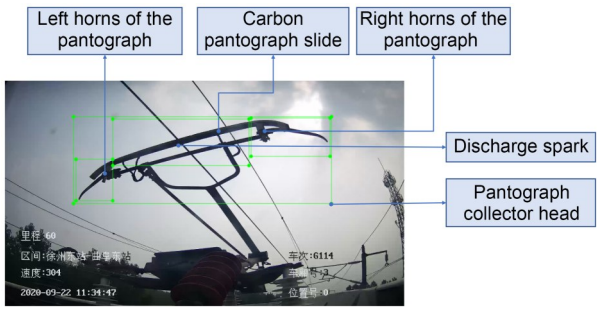

Fig. 8 Annotation of pantograph training images

Five categories of images are marked: pantograph collector head, left horns of the pantograph, carbon pantograph slide, right horns of the pantograph, and discharge spark. We used GTX 1050Ti for training and testing. The image size is set to 640 × 640 and the batch size is set to 2; the number of training epochs is set to 200.

3.2 Metrics

In this study, accurate fault detection is the most fundamental function, so the evaluation formula is as follows: accuracy $S_{Ac}=(N_{TP}+N_{TN})/(N_{TP}+N_{TN}+N_{FP}+N_{FN})$, recall $S_{Re}=N_{TP}/(N_{TP}+N_{FN})$, and precision $S_{Pr}=N_{TP}/(N_{TP}+N_{FP})$, where N is the number of samples, subscript TP means prediction is true and actual is true, TN means prediction is false and actual is false, FP means prediction is true and actual is false, and FN means prediction is false and actual is true. In addition, in order to better determine the confidence threshold, the measurement standard is introduced here, and the calculation formula is $S_{F1}=2S_{Pr}S_{Re}/(S_{Pr}+S_{Re})$, where S_{F1} is the harmonic mean of the precision and recall.

3.3 Analysis of the training process

For each of the three branches of the loss function, the analysis of the training process is shown in Fig. 9.

The CNN used represents the multi-scale feature-extraction model, whereas all modules in layers E1–E10 of Table 1 are convolutional modules. SACNN-A indicates that only the E10 module consists of the SACNN module. SACNN-ANC indicates that the CBAM is not used based on SACNN-A. SACNN-B indicates that only layer E9 consists of the SACNN-Block module and layer E10 consists of the SACNN module. SACNN-C indicates that only the layers E7 and E9 consist of the SACNNBlock module and layer E10 consists of the SACNN module. As can be seen from the diagram of the training process, in L_{cls}, there is little difference between CNN and SACNN-C, while the loss curves of SACNN-A and SACNN-B are all lower than the former, indicating lower loss (Fig. 9a). Moreover, on L_{obj} and L_{box}, when a SACNN module is added to E9 and E10 layers, the training effect is better than that of CNN, but when SACNN is added to the E7 layer, the training effect is worse. It can be concluded that when self-attention acts on the shallow layer of the network, the effect is worse than that of CNN, but when self-attention acts on the deep layer of the network, it can better improve the network performance of CNN.

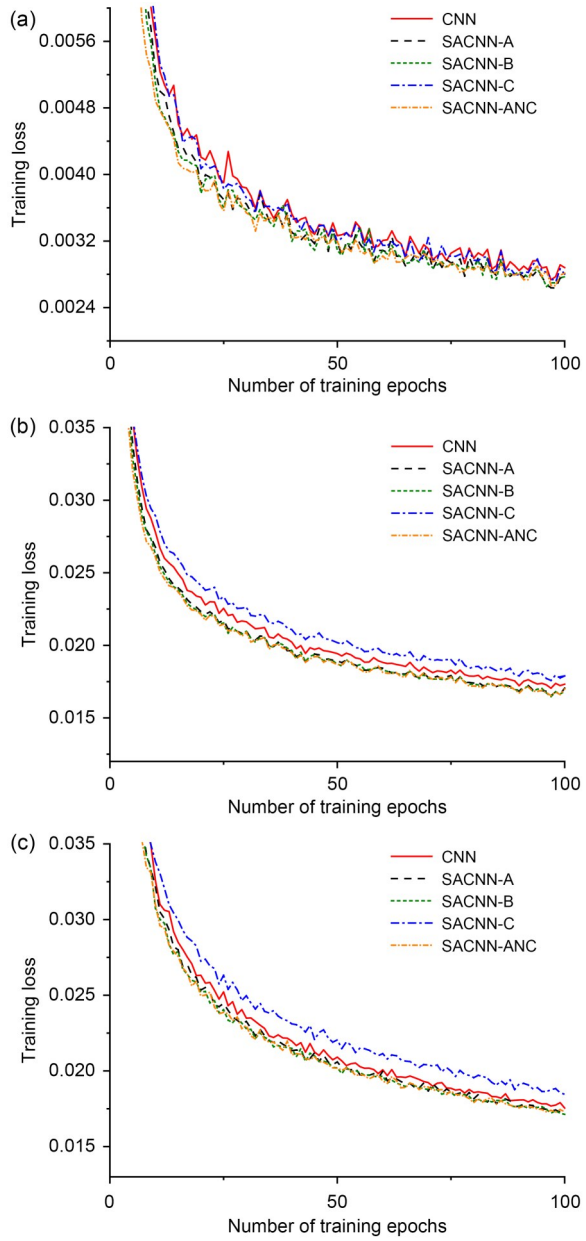

Fig. 9 (a) Training loss of classification; (b) training loss of objectness; (c) training loss of bounding-box regression

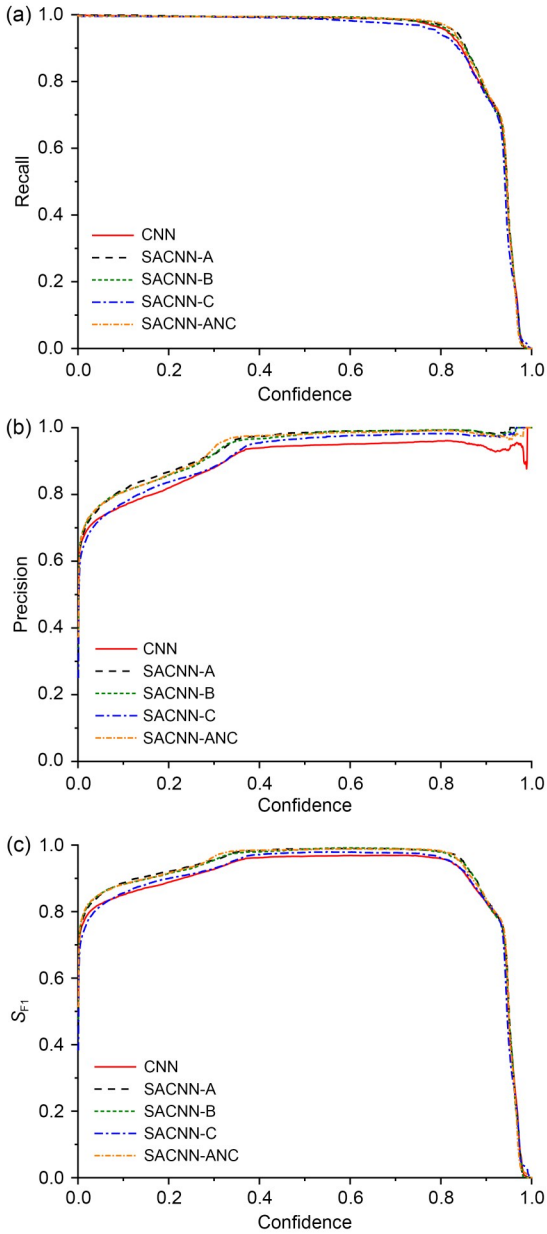

Fig. 10 Correlation curves in the training set for: (a) recall and confidence; (b) precision and confidence; (c) S_{F1} and confidence

3.4 Influence of the number of SACNN modules and the confidence-score threshold

The higher the confidence, the higher the inference accuracy of the object. As can be seen from Fig. 10, when the confidence level is 0.8, the recall and precision of all classification averages in the training set, as well as the confidence, are relatively high. Moreover, from the S_{F1} curve of integrated recall and precision, it is evident that when the confidence value exceeded 0.8, model performance began to decline. Therefore, the confidence of object detection is set at 0.8 as the threshold value in the fault-detection model.

Table 3 gives the test results of the pantograph fault detection. When the SACNN module is added, the accuracy of the models is better than that of CNN, but the reasoning speed of the models decreases as more modules are added, and the memory also increases. In the test of these four models, SACNN-A successfully detected all pantograph failures, and its overall performance was significantly better than that of the other

models; its speed also met the requirements of real-time detection. This proved that adding a CBAM module before network output is helpful in improving the accuracy of object recognition.

3.5 Pantograph-detection model test results

Fig. 11 shows the detection results for several types of pantograph faults. It is clear that when there is a fault in a certain region of the pantograph, the object in the region cannot be identified normally, so we determined that there is a fault in the region. Fig. 11c demonstrates that this method not only detected faults in the carbon slider but also identified the specific category of major spark faults. Fig. 11d shows foreign bodies attached to the carbon pantograph slide. Due to the wide variety of foreign bodies, direct identification is difficult to achieve, so the method proposed here instead aims to identify faults of the carbon pantograph slide in this region. Adopting the approach outlined by Chen et al. (2022) would be ineffective for identifying the two types of faults depicted in Figs. 11a and 11b. In Fig. 11a, despite deformation of the carbon slider, its contact point remains within the working area. Similarly, in Fig. 11b, the fracture at the left bow angle does not affect the contact point.

We then compared the model horizontally with other methods, and the results are shown in Table 4. In this test, the SACNN module of SACNN-A is replaced by an ACmix module to build an ACmix detection model. Based on the experimental results from Chen et al. (2022) in Table 4, it is evident that the comprehensive accuracy of a fault-detection method decreases significantly when the dataset is extended to encompass complex scenarios involving all-weather conditions and various shooting perspectives. Under the same conditions, the pantograph fault-detection model exhibited higher accuracy. Furthermore, compared to other object feature detection methods, it can be clearly seen that the performance of the proposed model is superior.

When the graphics card is configured as GTX 1050Ti, the number of frames per second (FPS) of the PFDM is 62.50, while the FPS of images captured by the surveillance camera on the high-speed railway is 25, which meets the requirements of real-time performance. Although the Faster-RCNN method can achieve

Table 3 Pantograph detection model test results with different numbers of SACNN modules

Model	Recall (%)	Precision (%)	Accuracy (%)	FPS (frame/s)	Memory (GB)
CNN+PFDM	93.0	63.7	90.0	66.67	**0.562**
SACNN-A+PFDM	**100.0**	85.2	**97.1**	62.50	0.686
SACNN-B+PFDM	89.0	77.3	68.8	32.25	0.797
SACNN-C+PFDM	93.0	68.8	91.8	16.13	1.660
SACNN-ANC+PFDM	93.0	**85.9**	96.3	**66.96**	0.682

PFDM: pantograph fault-detection model; FPS: number of frames per second

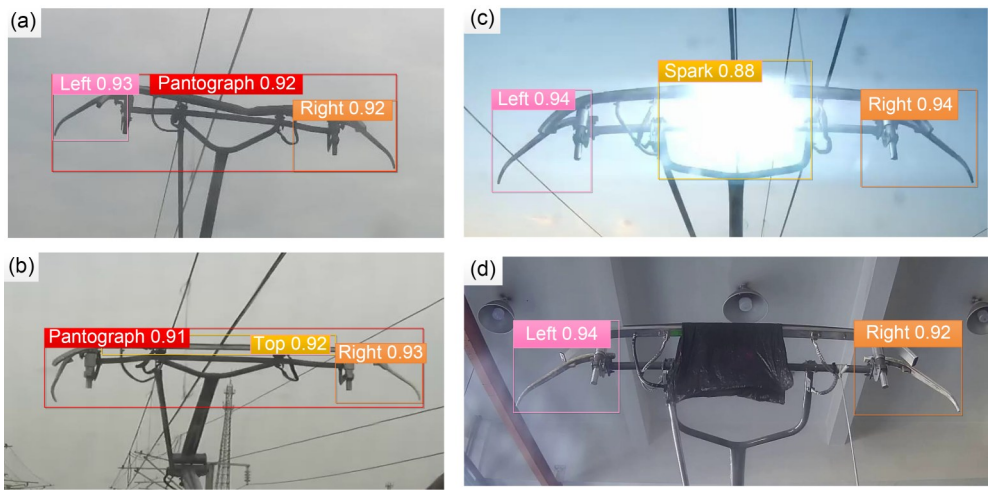

Fig. 11 Pantograph fault-detection results: (a) carbon pantograph slide deformation; (b) horns of the pantograph fault; (c) discharge spark; (d) foreign body attachment

Table 4 Comparison with other methods of pantograph fault detection

Model	Recall (%)	Precision (%)	Accuracy (%)	FPS (frame/s)	Memory (GB)
SSD+PFDM	82.0	65.2	89.7	5.88	2.581
Faster-RCNN+PFDM	89.0	**99.5**	**98.1**	5.55	2.581
YOLOv3+PFDM	82.0	90.1	95.5	14.08	2.330
YOLOv5+PFDM	88.0	88.0	96.0	40.00	1.740
YOLOv7+PFDM	**100.0**	34.6	68.5	8.11	1.980
ACmix+PFDM	94.0	81.7	95.5	56.38	0.686
DPDN+IVFE	43.0	11.3	34.5	22.22	1.788
Proposed method	**100.0**	85.2	97.1	**62.50**	**0.686**

SSD: single shot multibox detector; RCNN: region-based convolutional neural network

higher accuracy, its FPS is lower than the real-time requirement of 25. At the same time, the proposed method needs less memory and is lighter in weight, so it is convenient to use in high-speed trains for real-time detection of pantographs.

4 Conclusions

In order to solve the problems involved in obtaining real-time intelligence on high-speed railway pantograph faults, we propose a lightweight deep learning model based on the fusion of self-attention and convolutional features to achieve real-time and high-accuracy recognition of key components of pantographs. Our experiments show that the fusion module can effectively improve the performance of the original convolutional network. High recall and accuracy are achieved in training and testing sets, and the algorithm model has good performance. We designed a pantograph fault-detection model, which achieves fast and intelligent fault detection and accurately identifies all the faults in the test set. The detection model proposed here provides a set of real-time and accurate tools for intelligent fault detection in pantographs of high-speed trains, and provides an efficient auxiliary means for engineers on board to assess pantograph faults.

Acknowledgments

This work is supported by the National Key R&D Program of China (No. 2022YFB4301102).

Author contributions

Xufeng LI and Youtong FANG designed the research. Lin QIU processed the corresponding data. Xufeng LI wrote the first draft of the manuscript. Ping TAN and Lanfen LIN helped to organize the manuscript. Jien MA and Ping TAN revised and edited the final version.

Conflict of interest

Xufeng LI, Jien MA, Ping TAN, Lanfen LIN, Lin QIU, and Youtong FANG declare that they have no conflict of interest.

References

Bochkovskiy A, Wang CY, Liao HYM, 2020. YOLOV4: optimal speed and accuracy of object detection. arXiv: 2004.10934. https://doi.org/10.48550/arXiv.2004.10934

Chen RC, Lin YZ, Jin T, 2022. High-speed railway pantograph-catenary anomaly detection method based on depth vision neural network. *IEEE Transactions on Instrumentation and Measurement*, 71:1502710. https://doi.org/10.1109/tim.2022.3188042

Deng QH, Chen YG, Zhang YP, 2022. Research on wear detection of pantograph slide plate based on high speed 3D structured light detection. *Engineering and Technological Research*, 7(12):15-19 (in Chinese). https://doi.org/10.19537/j.cnki.2096-2789.2022.12.005

Ding XH, Zhang XY, Ma NN, et al., 2021. RepVGG: making VGG-style ConvNets great again. IEEE/CVF Conference on Computer Vision and Pattern Recognition, p.13728-13737. https://doi.org/10.1109/cvpr46437.2021.01352

Dosovitskiy A, Beyer L, Kolesnikov A, et al., 2021. An image is worth 16×16 words: transformers for image recognition at scale. International Conference on Learning Representations.

Fang YT, Ma JE, 2023. High-speed railway transport technology. *Journal of Zhejiang University-SCIENCE A*, 24(3): 173-176. https://doi.org/10.1631/jzus.A230HSRT

Guo JY, Han K, Wu H, et al., 2022. CMT: convolutional neural networks meet vision transformers. IEEE/CVF Conference on Computer Vision and Pattern Recognition, p.12165-12175. https://doi.org/10.1109/cvpr52688.2022.01186

Hao K, Chen GK, Zhao L, et al., 2022. An insulator defect detection model in aerial images based on multiscale feature pyramid network. *IEEE Transactions on Instrumentation and Measurement*, 71:3522412. https://doi.org/10.1109/tim.2022.3200861

He KM, Zhang XY, Ren SQ, et al., 2016. Deep residual learning for image recognition. IEEE Conference on Computer Vision and Pattern Recognition, p.770-778. https://doi.org/10.1109/cvpr.2016.90

He T, Zhang Z, Zhang H, et al., 2019. Bag of tricks for image classification with convolutional neural networks. IEEE/CVF Conference on Computer Vision and Pattern Recognition, p.558-567. https://doi.org/10.1109/cvpr.2019.00065

Hu TY, Ma HM, Liu H, et al., 2022. Self-attention-based machine theory of mind for electric vehicle charging demand forecast. *IEEE Transactions on Industrial Informatics*, 18(11):8191-8202. https://doi.org/10.1109/tii.2022.3180399

Karaduman G, Akin E, 2022. A new approach based on predictive maintenance using the fuzzy classifier in pantograph-catenary systems. *IEEE Transactions on Intelligent Transportation Systems*, 23(5):4236-4246. https://doi.org/10.1109/tits.2020.3042997

Krizhevsky A, Sutskever I, Hinton GE, 2017. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84-90. https://doi.org/10.1145/3065386

Li D, Pan X, Fu ZZ, et al., 2022. Real-time accurate deep learning-based edge detection for 3-D pantograph pose status inspection. *IEEE Transactions on Instrumentation and Measurement*, 71:5001012. https://doi.org/10.1109/tim.2021.3137558

Lin TY, Dollár P, Girshick R, et al., 2017. Feature pyramid networks for object detection. IEEE Conference on Computer Vision and Pattern Recognition, p.936-944. https://doi.org/10.1109/cvpr.2017.106

Liu WJ, Zhao J, Wang SF, 2021. Pantograph slide thickness detection method research based on machine vision. *Electronic Measurement Technology*, 44(24):128-133 (in Chinese). https://doi.org/10.19651/j.cnki.emt.2107928

Mo XF, Wang KL, Pan CQ, et al., 2022. Intelligent defect-detection technology of pantograph pan based on the image from railway 5C device. *China Railway*, (2):148-155 (in Chinese). https://doi.org/10.19549/j.issn.1001-683x.2021.07.14.002

Mu ZH, Qin Y, Yu CC, et al., 2023. Adaptive cropping shallow attention network for defect detection of bridge girder steel using unmanned aerial vehicle images. *Journal of Zhejiang University-SCIENCE A*, 24(3):243-256. https://doi.org/10.1631/jzus.A2200175

Ni XF, Ma ZJ, Liu JW, et al., 2022. Attention network for rail surface defect detection via consistency of intersection-over-union (IoU)-guided center-point estimation. *IEEE Transactions on Industrial Informatics*, 18(3):1694-1705. https://doi.org/10.1109/tii.2021.3085848

Pan XR, Ge CJ, Lu R, et al., 2022. On the integration of self-attention and convolution. IEEE/CVF Conference on Computer Vision and Pattern Recognition, p.805-815. https://doi.org/10.1109/cvpr52688.2022.00089

Simonyan K, Zisserman A, 2015. Very deep convolutional networks for large-scale image recognition. International Conference on Learning Representations, p.1-14.

Tan BK, Wang D, Shi JL, et al., 2024. Temperature field prediction of steel-concrete composite decks using TVFEMD-stacking ensemble algorithm. *Journal of Zhejiang University-SCIENCE A*, 25(9):732-748. https://doi.org/10.1631/jzus.A2300441

Tan P, Ma JE, Zhou J, et al., 2016. Sustainability development strategy of China's high speed rail. *Journal of Zhejiang University-SCIENCE A*, 17(12):923-932. https://doi.org/10.1631/jzus.A1600747

Tan P, Li XF, Xu JM, et al., 2020. Catenary insulator defect detection based on contour features and gray similarity matching. *Journal of Zhejiang University-SCIENCE A*, 21(1):64-73. https://doi.org/10.1631/jzus.A1900341

Tan P, Li XF, Wu ZG, et al., 2021. Multialgorithm fusion image processing for high speed railway dropper failure-defect detection. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 51(7):4466-4478. https://doi.org/10.1109/tsmc.2019.2938684

Tan P, Li XF, Ding J, et al., 2022. Mask R-CNN and multifeature clustering model for catenary insulator recognition and defect detection. *Journal of Zhejiang University-SCIENCE A*, 23(9):745-756. https://doi.org/10.1631/jzus.A2100494

Vaswani A, Shazeer N, Parmar N, et al., 2017. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems, p.6000-6010.

Wang WH, Xie EZ, Song XG, et al., 2019. Efficient and accurate arbitrary-shaped text detection with pixel aggregation network. IEEE/CVF International Conference on Computer Vision, p.8439-8448. https://doi.org/10.1109/iccv.2019.00853

Wei XK, Jiang SY, Li Y, et al., 2020. Defect detection of pantograph slide based on deep learning and image processing technology. *IEEE Transactions on Intelligent Transportation Systems*, 21(3):947-958. https://doi.org/10.1109/tits.2019.2900385

Wen YT, Cheng JX, Ren YX, et al., 2024. Complex defects detection of 3-D-printed lattice structures: accuracy and scale improvement in YOLO V7. *IEEE Transactions on Instrumentation and Measurement*, 73:5013209. https://doi.org/10.1109/tim.2024.3370765

Woo S, Park J, Lee JY, et al., 2018. CBAM: convolutional block attention module. The 15th European Conference on Computer Vision, p.3-19. https://doi.org/10.1007/978-3-030-01234-2_1

Wu Y, Fu HR, Bian XC, et al., 2023. Impact of extreme climate and train traffic loads on the performance of high-speed railway geotechnical infrastructures. *Journal of Zhejiang University-SCIENCE A*, 24(3):189-205. https://doi.org/10.1631/jzus.A2200341

Xie SN, Girshick R, Dollár P, et al., 2017. Aggregated residual transformations for deep neural networks. IEEE Conference on Computer Vision and Pattern Recognition, p.5987-5995.
https://doi.org/10.1109/cvpr.2017.634

Yao H, Liu YH, Li X, et al., 2022. A detection method for pavement cracks combining object detection and attention mechanism. *IEEE Transactions on Intelligent Transportation Systems*, 23(11):22179-22189.
https://doi.org/10.1109/tits.2022.3177210

Zheng ZH, Wang P, Liu W, et al., 2020. Distance-IoU loss: faster and better learning for bounding box regression. Proceedings of the AAAI Conference on Artificial Intelligence, p.12993-13000.
https://doi.org/10.1609/aaai.v34i07.6999

Zhou N, Zhang WH, Li RP, 2011. Dynamic performance of a pantograph-catenary system with the consideration of the appearance characteristics of contact surfaces. *Journal of Zhejiang University-SCIENCE A*, 12(12):913-920.
https://doi.org/10.1631/jzus.A11GT015