



Multiview deep learning based on tensor decomposition and its application in fault detection of overhead contact systems

Xuewu Zhang¹ · Yansheng Gong¹ · Chen Qiao² · Wenfeng Jing²

Accepted: 3 February 2021 / Published online: 19 February 2021
© The Author(s) 2021

Abstract

This article mainly focuses on the most common types of high-speed railways malfunctions in overhead contact systems, namely, unstressed droppers, foreign-body invasions, and pole number-plate malfunctions, to establish a deep-network detection model. By fusing the feature maps of the shallow and deep layers in the pretraining network, global and local features of the malfunction area are combined to enhance the network's ability of identifying small objects. Further, in order to share the fully connected layers of the pretraining network and reduce the complexity of the model, Tucker tensor decomposition is used to extract features from the fused-feature map. The operation greatly reduces training time. Through the detection of images collected on the Lanxin railway line, experiments result show that the proposed multiview Faster R-CNN based on tensor decomposition had lower miss probability and higher detection accuracy for the three types faults. Compared with object-detection methods YOLOv3, SSD, and the original Faster R-CNN, the average miss probability of the improved Faster R-CNN model in this paper is decreased by 37.83%, 51.27%, and 43.79%, respectively, and average detection accuracy is increased by 3.6%, 9.75%, and 5.9%, respectively.

Keywords High-speed railway catenary · Fault detection · Faster R-CNN · Multiview feature fusion · Tensor decomposition

1 Introduction

The railway is the most critical part of basic transportation facilities. As a key of people's livelihood, it has an important influence on Chinese economy [1]. In recent years, China's high-speed railway construction has rapidly developed. This puts forward higher requirements on the safety and reliability of the power-supply equipment of high-speed railways.

Overhead contact systems (OCSs) are important devices for electrified railways that are mainly used to provide electrical support to electric locomotives. They are laid over a high-speed railway line and are mainly composed of contact-suspension devices, support devices, positioning devices, pillars, and other devices [2]. Because the OCSs receive wind and sun in the external environment, and there is no backup system, this becomes a weak link in the

railway-traction power-supply system. Malfunctions of each component in each device of the OCSs are likely to occur. The three types of malfunctions, namely, unstressed droppers, foreign-body invasions, and pole number-plate malfunctions, are the most common.

For these three types of malfunctions, there are already some image monitoring methods. Karakose et al. [3] proposed a new approach using image processing-based tracking to diagnose faults in the pantograph-catenary system. Liu et al. [4] proposed a unified deep learning architecture for the detection of all catenary support components. Qu et al. [5] used a genetic optimization method based on a deep neural network to predict pantograph and catenary comprehensive monitor status. Zhong et al. [6] introduced a CNN-based defect inspection method to detect catenary split pins in high-speed railways. For foreign bodies with irregular-edge shapes, such as bird's nests, plastic bags and tree branches, Scholars have put forward some detection methods of railway foreign-body intrusion. Yang Pei [3] used dual discriminators to generate a generative adversarial network (GAN) to locate and identify bird's nest in railway OCSs. First, the position of the bird's nest in the image was located by fast region CNN, and then the GAN was used to classify

✉ Wenfeng Jing
wfjing@xjtu.edu.cn

¹ China Railway First Survey and Design Institute Group Co., Ltd, Xi'an 710043, China

² National Engineering Laboratory for Big Data Analytics, Xi'an Jiaotong University, Xi'an 710049, China

the objects in the location area, so as to judge whether the bird's nest exists or not. Petitjean et al. [4] proposed a top-down dropper detection method. First, they used a priori knowledge to extract the reliable position of the dropper, and then used MLP to classify dropper-malfunction. The above research was applied in an ideal scenario or experiment environment with good results. It cannot deal with the task of dropper-malfunction detection in actual situations, and it is hard to efficiently and accurately locate and identify the dropper. The use of traditional image methods for OCSs defect identification has certain limitations. In recent years, some studies showed that deep learning methods can achieve better results. This is mainly because a convolutional neural network can mine deeper features of the image, and has better effects of malfunction detection. It has become a trend to replace traditional methods with deep learning techniques. For example, Jiang et al. [5] proposed a method for detecting bird's nests and foreign bodies on the high-speed railway catenary. The candidate regions were extracted by a line-detection method, and bird's nest detection was performed by deep-network object-detection algorithm YOLOv3 [6] with good results. The literature [8] used deep learning and traditional image methods for the detection of dropper defects. First, a Faster R-CNN was used to detect the positioning clamp of the dropper; then, the position of the dropper was divided by the clamp; lastly, the Canny edge was used to identify defects of the extracted droppers. Wu et al., on the basis of semi-supervised learning, first used Lenet-5 to learn and extract deep features of the image [7]; then, they trained the extracted image features from the CNN through the support-vector-data-description (SVDD) algorithm; lastly, they identify whether rods in the image were abnormal. Guo et al. [9] proposed an improved Faster R-CNN algorithm that can accurately identify and locate droppers. The method consisted of two parts. First, a balanced attention feature pyramid network (BA-FPN) was used to predict the detection anchor. Based on the attention mechanism, BA-FPN performs feature fusion on feature maps of different levels of the feature pyramid network to balance the original features of each layer. After that, a center-point rectangle loss (CR Loss) is designed as the bounding box regression loss function of Faster R-CNN. Through a center-point rectangle penalty term, the anchor box quickly moves closer to the ground-truth box during the training process.

In the above-mentioned methods, the parts that may have been faulty in the image were located through the deep learning method, divided into separate pictures; then, the malfunction detection was performed on the divided images. Separating the positioning and detection process increases training and testing time. In addition, as the number of convolutional-network layers increases, the pixels of the feature map gradually decrease. When the size of the malfunction area in the map is small, the deep-feature map of the output

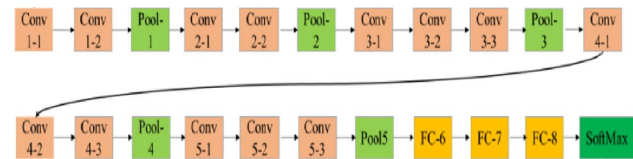


Fig.1 VGG16 network architecture

cannot retain malfunction details. Deepening the network layer increases the receptive field, so the malfunction area extracted by the pretraining network in the deep layer is mapped to a wider area in the input image. This also contains information about the natural environment and buildings around the object. It is easier to cause interference to the fault location and detection, and missed detection occurs. Inspired by Sparse PARAFAC2 decomposition of literature [16], this paper takes the fusion of shallow- and deep-feature maps in the Faster R-CNN network to retain more detailed information of the malfunction area, and then, reduces the dimension of the fusion feature map to improve training efficiency. Experiments showed that this method greatly reduced the miss probability of the model.

The remainder of this paper is organized as follows. Section 2 introduces the idea of multiview feature fusion and Tucker tensor decomposition, and describes our proposed multiview Faster R-CNN based on tensor decomposition. Section 3 applies the method into the dropper detection, and shows the effect of our method compared with YOLO v3, SSD and Faster R-CNN in training time, miss-probability, and detection accuracy. The relevant conclusions are given in Sect. 4.

2 Multiview faster R-CNN based on tensor decomposition

2.1 Multiview feature fusion

The Faster R-CNN method usually achieves higher recognition accuracy for object-detection on natural image datasets. Objects in such natural images generally occupy a larger area of the image. However, video images acquired by the 2C detection system of the high-speed railway have relatively low resolution, and the target objects (droppers, pole number plates, foreign objects) in the image are usually small, so it is difficult to identify them directly with Faster R-CNN. This is because the deep convolutional layer of the pretraining network in the standard Faster R-CNN has a larger receptive field. The most common pretraining network, VGG16, the structure of which is shown in Fig. 1, contains 5 convolution modules: conv1, conv2, conv3, conv4, and conv5. In order to reduce the number of parameters in the model, the maximal

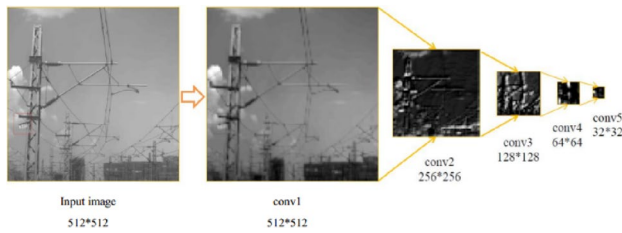


Fig.2 Feature maps of different layers in VGG16

pooling layer is used to reduce the feature map of each module after the convolution operation. Multiple convolution pooling operations result in one pixel on the feature map corresponding to multiple pixels in the input image, and the receptive field refers to the size of this corresponding area. The deep-feature map in the network had a larger receptive field and lower resolution, and the contained features were also more abstract, while the shallow feature map had a higher resolution and a smaller receptive field, which could retain more details of the image and local features.

Figure 2 shows the five feature maps in VGG16. The size of the input image was 512×512 pixels. The size of the feature map became increasingly smaller as the number of network layers increased. Lastly, the feature map output by the conv5 module was only 32×32 pixels. In the shallow features of the network, more details of the input image could be retained, and the contour of the fault area could also be seen from the conv2 layer. In the process of gradually deepening the number of network layers, the receptive field also increased, the resolution of the feature map became increasingly smaller, the extracted feature semantic-discrimination ability became stronger and more abstract, and the fault area, as a small object, had almost no response in the deep-feature map. If the size of a dropper in an original image is 64×64 pixels, its output in the last convolutional layer conv5 becomes 4×4 pixels, making it difficult to extract effective information features from 4×4 pixels. In addition, as the deep receptive field increases, each corresponding pixel in the feature map contains more convolution information outside the region-of-interest (RoI). Therefore, if the RoI is small, then the proportion of information outside it contained in the corresponding feature map is higher, which interferes with detection. On the basis of the two problems above, the feature map output by the last convolutional layer was less representative of the fault area than it was of the small area.

Although the shallow feature map in the pretraining network contained more noise, it also retained more detailed information such as position, shape, and texture in the fault area, and had higher resolution. Although deep-feature maps have stronger semantic discrimination than that of shallow feature maps, features obtained by the image after

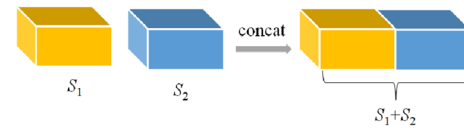


Fig.3 Concat fusion

layer-by-layer convolution are more abstract, resolution is lower, and detailed information of the fault area is ignored. If deep and shallow features can be effectively combined in the model's pretraining network, combined with global and local features, that is, multiview feature fusion can enhance global and local more details and local information of the fault region contained in the shallow layer of the network, which can help the model to better detect a fault area.

In order to improve network performance, one can consider fusing the first few layers of feature maps of the shallow pretraining network and the last layer of conv5 feature maps, and then putting them into the RoI pooling layer. In the experiment stage, this paper used the concat method to fuse features of different fields of view. As shown in Fig. 3, concat means to directly fuse the feature map along the dimension of the number of channels. This fusion method can add feature maps that describe the details of the fault area. During the experiment, the best combination was found by comparison: conv3 + conv4 + conv5.

2.2 Tucker tensor decomposition

Compared with the feature map output by the conv5 layer, the number of channels in the feature map after fusion increased. In order to share the fully connected layer of the pretraining network to reduce model complexity, the number of channels in the feature map should be reduced to be the same as that before fusion. The method in this paper added a convolutional layer after fusion to reduce the number of channels in the feature map. However, because this layer is new, there are no pretraining parameters to initialize. If the parameters of this convolutional layer are initialized in a random way, the instability of the network may increase, and subsequent parameter updates in training require more time. Therefore, in this paper, we directly decomposed the feature tensor after fusion to reduce the dimension.

At present, there are two mainstream tensor decomposition models, namely, the CANDECOMP/ PARAFAC model (CP decomposition) [9, 10] and the Tucker model [11, 12]. The result of CP decomposition is to decompose an N-order tensor into the sum of R tensors of rank 1, and determine the size of R according to the needs of the decomposition. Tucker tensor decomposition decomposes an N-order tensor into a product of a kernel tensor and N-factor matrices, where the kernel tensor represents

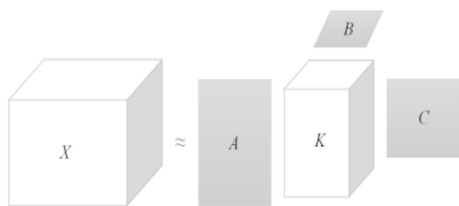


Fig.4 Tucker tensor decomposition

the main features of the original tensor, and the factor matrices represent the importance of each feature in the kernel tensor. From this perspective, CP decomposition and Tucker tensor decomposition can be regarded as the higher-order expansion of matrix singular-value decomposition (SVD) and principal-component analysis (PCA) [13]. In this paper, we used the Tucker model for dimensionality reduction of image-fusion features, performed Tucker decomposition on the fusion feature map tensor, and took the obtained kernel tensor as the input in the next stage. Moreover, the decomposition process showed that Tucker-1 decomposition used in this study was equivalent to 1 × 1 convolution, so the number of network parameters could be reduced on the basis of precise feature extraction.

The process of Tucker decomposition is shown in Fig. 4.

$X \in \mathbb{R}^{H \times W \times S}$ represents a third-order tensor of size $H \times W \times S$. The result of the Tucker decomposition of X is:

$$X = K \times_1 A \times_2 B \times_3 C \tag{1}$$

$$X_{h,w,s} = \sum_{r_1=1}^{R_1} \sum_{r_2=1}^{R_2} \sum_{r_3=1}^{R_3} K_{r_1,r_2,r_3} A_{h,r_1} B_{w,r_2} C_{s,r_3} \tag{2}$$

where K represents the third-order kernel tensor of $R_1 \times R_2 \times R_3$. Similar to the principal-component factor in principal-component analysis (PCA), it actually represents the main feature of original tensors A , B , and C , respectively, representing factor matrices of size $H \times R_1$, $W \times R_2$, and $S \times R_3$. Factor matrices are usually orthogonal, indicating the importance degree of each feature in the kernel tensor. When R_1 , R_2 and R_3 values are less than H , W , and S values, the kernel tensor can be used as the result of the compression of original tensor X . The \times_n symbol in Formula (2) means matrixing and multiplying on the n -th dimension of the tensor:

$$X \times_n U \iff UX_{(n)} \tag{3}$$

where $X_{(n)}$ means matrixing the tensor in the n -th dimension. Matrixing refers to the process of rearranging the elements of a tensor to obtain a matrix. For third-order tensor X , it is matrixed at the third-order to obtain matrix $X_{(3)}$, and the expansion process is shown in the following Fig. 5.

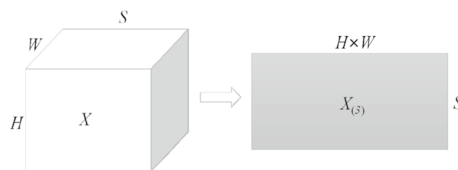


Fig.5 Tensor matrixed in third-order

The most common method for solving the third-order Tucker decomposition is the higher-order SVD (HOSVD) method [14]. This method first needs to matrix the tensor in each dimension; then, for each matrix $X_{(n)}$, it performs SVD decomposition, and the left singular value matrix is the corresponding factor matrix in this dimension. After the factor matrix is obtained, the kernel matrix can be determined by calculating $K = X \times_1 A^T \times_2 B^T \times_3 C^T$. However, the HOSVD method generally cannot obtain a good approximate result, so the result is usually used as the initial value of the high-order orthogonal iteration (HOOI) [15]. Then, we used the HOOI method to solve it.

In Tucker decomposition, not every dimension must be decomposed, and it is possible to determine which dimensions are decomposed according to needs. For example, for the feature map after fusion, both H and W are related to the spatial dimension, indicating the length and width, and usually have a small value, so there is no need to decompose these two dimensions. S represents the number of channels. Taking pretraining network VGG16 as an example, if output feature maps of the last three convolution modules are selected for fusion, the number of channels of resulting fusion feature map X is 1280, and it needs to be reduced to 512 channels. Then, only the third dimension needs to be decomposed:

$$X = K \times_3 C \tag{4}$$

$$X_{h,w,s} = \sum_{r_3=1}^{R_3} K_{h,w,r_3} C_{s,r_3} \tag{5}$$

This variant of Tucker decomposition is called Tucker-1 decomposition. Kernel tensor K here is a tensor of size $H \times W \times R_3$, which is feature map of the next stage of the network that we want to get. When solving, some constraints are usually added to ensure a unique solution. The most common one is to add unit orthogonal constraint to the factor matrix. Therefore, solved factor matrix C is an orthogonal matrix, and the above formula can be transformed into:

$$K_{h,w,r_3} = \sum_{r_3=1}^{R_3} C_{s,r_3}^T X_{h,w,s} \tag{6}$$

The above formula shows that the process of obtaining output feature map K from input fusion feature map X is completely consistent with the principle of 1×1 convolution, which is essentially pixel-level linear reorganization of the input map. However, in contrast, 1×1 convolution can only reduce dimensionality, does not have the function of feature selection, and Tucker decomposition has more feature-extraction functions than 1×1 convolution does.

Generally, 3×3 convolution is used in a pretraining network, and the RPN network in the whole network. If 3×3 convolution is also used here, the number of parameters is greatly increased, and the efficiency of network training is reduced. Using Tucker decomposition also reduces the number of parameters and of channels in the input layer. Taking pretraining network VGG16 as an example, the size of the feature map obtained after using three-layer feature fusion was $1280 \times H \times W$, namely, 1280 channels. The feature map had to be reduced to 512 channels to be consistent with the original conv5 feature map. Using Tucker tensor decomposition, the number of parameters was only 1280×512 . If a 3×3 convolution operation is used, the number of involved parameters is $(1280 \times 3 \times 3 + 1) \times 512$, which is 9 times the number of tensor decomposition parameters. Tucker tensor decomposition can not only extract tensor features and reduce dimensionality, but also reduce the number of calculated parameters, which can improve the efficiency of network training. Subsequent experiments also showed the advantages of this method. In the actual calculation process, factor matrix C on the dimension is solved by SVD for the third dimension of feature map X ; then, the projection of the feature map tensor on the dimension was calculated, which was kernel tensor K . In order to ensure the uniformity of the network and consistency of the parameter update, convolution layer 1×1 was still set in the actual operation; then, the value of the obtained kernel tensor was used as the parameter to initialize the newly added convolutional layer, which greatly reduced training time.

2.3 Multiview faster R-CNN based on tensor decomposition

In summary, in view of the fact that some of the three types of faults occupy a relatively small area, and it is not easy to capture the detailed information of the deep field of view of the pretraining network, this paper enhanced global and local information by fusing deep- and shallow feature maps in the pretraining network, and then, using Tucker decomposition to extract the fusion feature map and the core features into the subsequent ROI pooling layer. The method is called the multiview Faster R-CNN based on Tensor decomposition (MV-FRCNN-TD).

The structure of the pretraining network in this method is shown in Fig. 6. For the five shared modules included in

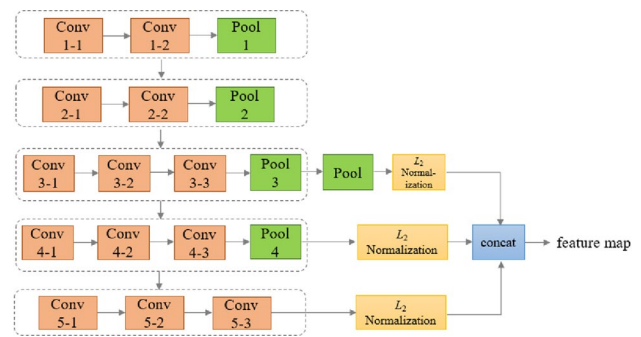


Fig. 6 VGG16 structure diagram of feature fusion

pretraining network VGG16, outputs of three convolution modules conv3, conv4, and conv5 were combined. Since the conv3 module's output feature map was larger, a pooling layer needed to be added to ensure that the size of feature map was consistent with conv5-3. After that, L_2 normalization was carried out for the output feature map of each layer; then, normalized results were connected together to obtain the fused-feature map.

In order to share the fully connected layer of VGG16, the feature map after fusion needed to be restored to the number of channels consistent with conv5. In the model, a convolutional layer is added after the pretraining network. The Tucker-1 decomposition of Formula (4) was solved by the HOOI method to realize feature extraction of the fusion feature map. The obtained factor matrix was used to initialize the weight of the convolutional layer, and the kernel tensor obtained was the feature map after dimension reduction. The flowchart of this model is shown in Fig. 7.

The RPN network that is part of this method was consistent with the RPN part of Faster R-CNN, mainly to improve the fast R-CNN part to detect smaller objects. After feature fusion, Tucker decomposition, and dimensionality reduction, the feature map was used as input to the RoI pooling layer and the RPN network. The RPN network determines whether the candidate region was the background or foreground, and performed preliminary regression on the border of the candidate region. Then, the feature map connected the last two fully connected layers, and lastly connected the layer used for object classification to output the probability that the candidate region belongs to each category, and the regression layer used for the fine positioning of the object's bounding box.

3 Numerical experiments

3.1 Data introduction

The images used in this article were collected by the 2C system on the section of Lanxin railway from January 2017 to

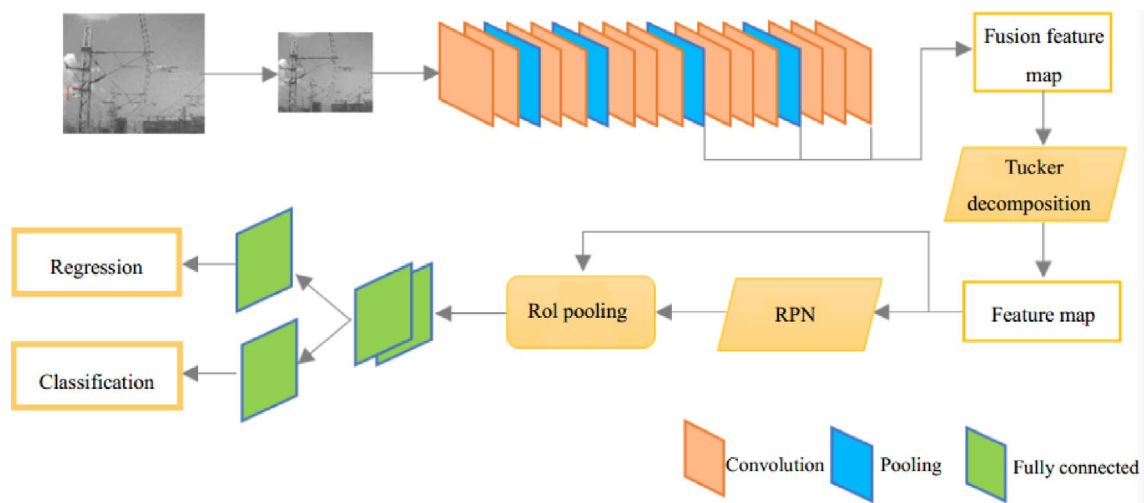


Fig. 7 Multiview Faster R-CNN model based on tensor decomposition

Table 1 Type criteria of small, medium, and large objects

Type	Minimal area	Maximal area
Small object	0	96×96
Medium object	97×97	192×192
Large object	193×193	–

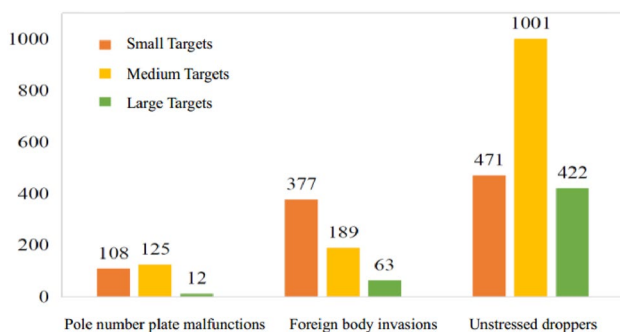


Fig. 8 Comparison of the number of three malfunction images

December 2018. According to MS COCO,¹ type criteria for small objects in OCSs malfunction-detection image dataset are as follows Table 1.

In accordance with the above standards, the statistical results are shown in Fig. 8. A total of 245 malfunction images were in the pole number plates in the dataset, including mall objects and only 12 large objects. For foreign-body invasions, there were 629 malfunction objects, including 63 large objects and 377 small objects, accounting for

59.94% of the total. For unstressed droppers, the total number of objects is 1894, of which small and medium objects accounted for 77.72%. Overall, the number of large objects in the three types of failures was relatively less, while small and medium objects accounted for a relatively high proportion.

Due to the small amount of data, images were randomly flipped, translated, and randomly cropped to expand the data set. During training, 70% of the images were randomly selected from the labeled dataset as training data. The training images of unstressed droppers, foreign-body invasions, and malfunction of the pole number plate were 2056, 848, and 338, respectively. The remaining images were used as test data, and the number of images included in the test set were 441, 183, and 73, respectively.

3.2 Evaluation of model detection effect

In the experiment, the configuration of the computer we used is: CPU main frequency, 3.0 GHz; The memory, 128 GB; GPU, NVIDIA Tesla P100 with 16 GB display memory. When training, we take batch size = 20, LR = 0.003. And the used epochs were 3370, 20,560, 4160 for the three malfunctions of pole number-plate malfunctions, unstressed droppers, and foreign-body invasions, respectively.

3.2.1 Training time comparison

Figure 9 compares the training times of using Tucker decomposition and 3×3 convolution to reduce the fusion feature dimension. On the right side of the figure, we can see that training time was 817.6, 4791.1, and 1907.3 s by using 3×3 convolution for the three malfunctions of pole number-plate

¹ <http://cocodataset.org/>

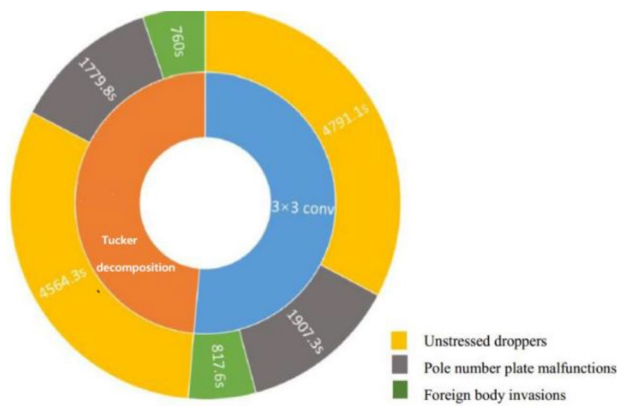


Fig. 9 Comparison of training time using Tucker decomposition and 3×3 convolution

malfunctions, unstressed droppers, and foreign-body invasions, respectively. On the left side, we can see that.

Training time was 760, 4564.3, and 1779.8 s by using Tucker decomposition for the three malfunctions, respectively. Obviously, the reduced training times of the latter method were 7.04%, 4.73%, and 6.68% less than the former method, respectively. This was mainly because that the numbers of parameters of Tucker vector decomposition and using a 3×3 convolution operation were 1280×512 and (1280×3×3+1)×512. The latter was 9 times as much as the former. Tucker decomposition performed dimensionality reduction on the fusion feature map that was equivalent to a 1×1 convolution operation. The resulting factor matrix was used to initialize the parameters of the convolution layer. This could make network converge faster than with random initialization. Therefore, training efficiency of using Tucker decomposition was improved.

3.2.2 Miss probability comparison

In order to verify the effectiveness of the proposed method in the experiment of the same training set and test set, YOLO v3, SSD, Faster R-CNN (FRCNN) and multiview Faster R-CNN method based on tensor decomposition (MV-FRCNN-TD) in the study were compared.

Table 2 Miss-probability comparison

Method	Pole number-plate malfunctions			Unstressed droppers			Foreign-body invasions		
	TP	FN	MDR	TP	FN	MDR	TP	FN	MDR
YOLOv3	67	6	8.22%	504	51	9.19%	173	14	7.49%
SSD	65	8	10.96%	490	65	11.71%	170	17	9.09%
FRCNN	67	6	8.22%	513	42	7.57%	165	22	11.76%
MV-FRCNN-TD	71	2	2.74%	517	38	6.85%	176	11	5.88%

TP True positive; FN False negative

For the method proposed, the feature-fusion method was conv3 + conv4 + conv5, and other comparison methods were set with reference to the optimal parameters given by the author. The comparison results of the missed detection rate (MDR) are shown in Table 2. The average miss detection rate of the three types of malfunctions using YOLOv3, SSD, and FRCNN methods was 8.3%, 10.59% and 9.18%, respectively. The average missed detection rate of this model was 5.16%. In comparison, there was a significant decrease of 37.83%, 51.27%, and 43.79%, respectively. For pole number-plate malfunctions, the MV-FRCNN-TD model had a missed detection rate of only 2.74%. Of the 73 pole number-plate failures in the test set, only 2 were not detected, compared with 6 for YOLOv3, 8 for SSD, and 6 for FRCNN. The missed detection rate of the model dropped by 66.67%, 75%, and 66.67%, respectively. For the malfunction of unstressed droppers, the number of missed detections of this model in 555 unstressed droppers was 38, and miss probability was 6.85%, which was higher than that of the three other types of models on dropper-malfunction. Rates dropped by 25.46%, 41.5%, and 9.51%, respectively. The miss probability of foreign-body-invasion malfunctions obviously dropped. The miss probability of the model was 5.88%, compared with the miss probability of 7.49%, 9.09%, and 11.76% of the YOLOv3, SSD, and FRCNN models, a decrease of 21.50%, 35.31%, and 50%, respectively. In the malfunctions of 187 foreign-body invasions in the test set, only 11 of them were not identified. This was slightly different from the dropper and pole number plates. Malfunctions such as foreign-object invasions are part of the normal size and small objects on the image, and the feature map after MV-FRCNN-TD multiview feature fusion had more detailed features. This part of the foreign-body with a smaller area was identified, so miss probability was significantly reduced.

3.2.3 Detection accuracy comparison

The accuracy P-R(precision-recall) curves in Fig. 10, 11, 12 show that the MV-FRCNN-TD method in this paper is much better than YOLOv3, SSD, and FRCNN. For three types of malfunctions, the method proposed has a higher recall rate.

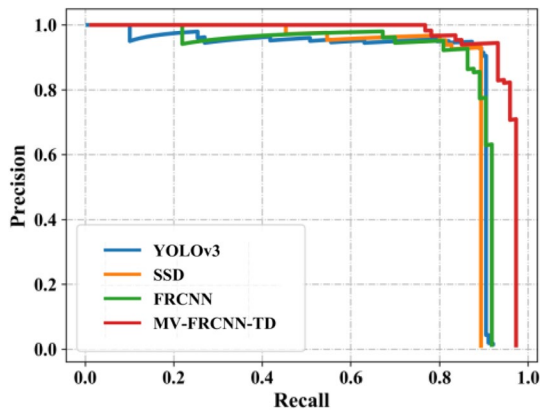


Fig. 10 Accuracy rate–recall rate (P–R) curve of pole number-plate malfunctions

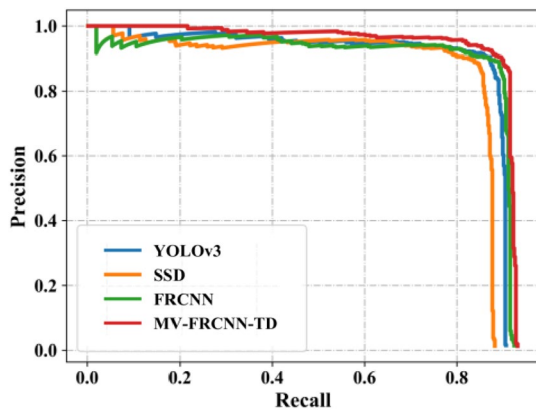


Fig. 11 P–R curve of unstressed droppers

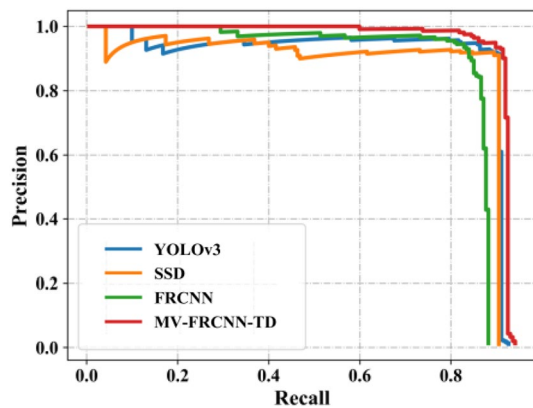


Fig. 12 P–R curve of foreign-body invasion

In general, the P–R curve of the unstressed dropper is smoother than the ones of the other two types of malfunctions. As shown in Fig. 10, the most obvious contrast is at the beginning. This model did not have jagged jitter like

the other three types of methods. In addition, in the P–R curve of pole number-plate malfunction and foreign-body invasion, the three remaining types of models contain more jagged turns than this model does. This is because there are fewer training and test samples for these two types of malfunctions. When drawing curves to set different threshold points, the number of malfunctions correctly identified by the model and the number of malfunctions identified as background malfunctions increase at the same time, resulting in a decrease in accuracy and an increase in recall rate; fewer test samples make the trend change of the P–R curve more obviously. On the other hand, these show that the performance of the model is more stable. The model in this paper can better cope with challenging conditions such as the low resolution and complex background of the object, and can find more malfunctions thanks to the Tucker tensor decomposition into volumes. The build-up layer provided initial parameters that enable the model to achieve a faster performance. Therefore, the MV-FRCNN-TD model used in this study has a higher recall value.

The comparisons of mean average precision (mAP) and F1 Score also show that the MV-FRCNN-TD method has significant performance improvement. As shown in Table 3, The mAP and F1 score of our model on the test set are 89.61% and 93.76%, respectively. Compared with YOLOv3, SSD, and FRCNN, the former is improved by 3.6%, 9.75%, and 5.9%, respectively, and the later is increased by 4.33%, 7.29%, 4.41%, respectively.

Obviously, MV-FRCNN-TD is more suitable for datasets with many small objects in the malfunction area.

The original FRCNN is better at capturing the global features of large and medium objects in this situation, and features in the deep-network are more abstract, and thus ignoring detailed features of small objects. MV-FRCNN-TD enhance pixels through the multiview feature-fusion strategy, so that the local features of small objects extracted by the shallow network can be retained, and the core part of the fusion feature is retained by Tucker decomposition. This is more in line with actual application scenarios. In reality, due to different shooting angles and different distances, the size of the malfunctions in the images is different, and the multiview feature-fusion model can flexibly detect malfunctions of various sizes.

In general, the multiview feature-fusion model proposed in this paper shows a good detection effect. The improvement in network resolution brought by feature fusion has greatly improved detector performance, especially in the detection of small objects. Figure 13 shows the detection effect of the MV-FRCNN-TD method. The green box in the figure is the model detection result, and the red box is the manual annotation result. Compared with unstressed droppers and pole number-plate malfunctions, the model has the most obvious improvement in the average detection accuracy

Table 3 Detection accuracy comparison

Method	Pole number-plate malfunctions	Unstressed droppers	Foreign-body invasions	mAP/mean F1 score
YOLOv3	0.8757/0.9015	0.8440/0.8823	0.8752/0.9121	0.8650/0.8986
SSD	0.8072/0.8681	0.7825/0.8743	0.8599/0.8791	0.8165/0.8738
FRCNN	0.8727/0.9065	0.8622/0.9115	0.8038/0.8756	0.8462/0.8979
MV-FRCNN-TD	0.9012/0.9565	0.8862/0.9245	0.9010/0.9317	0.8961/0.9375

mAP Average detection accuracy

Fig. 13 Detection effect of MV-FRCNN-TD model



of malfunctions from foreign-body invasions. On the one hand, this is due to the large number of small and medium objects in the malfunction area invaded by foreign objects,

accounting for 59.94%, which is the largest proportion of the three types of malfunctions. On the other hand, the characteristics of droppers and pole number plates are relatively

simple, while foreign objects have different shapes. Therefore, the feature fusion of multiple fields of view retain more detailed features of the malfunction area, which play a key role in reducing miss probability and improving average detection accuracy.

4 Conclusion

This paper focuses on the characteristics of the three faults of high-speed railway OCSs, namely, unstressed droppers, foreign-body invasions, and pole number-plate malfunctions. We find that there are many abnormal images that occupy a relatively small area in the original image, it is not easy to capture their features, and using only features of the deep field of view is prone to missed detection. From the perspective of multiview feature fusion, a more accurate detection method of Faster R-CNN is proposed by fusing features extracted from the shallow and deep layers of the feature-extraction-network part of Faster R-CNN to enhance detailed features of the object. Moreover, in order to reduce model complexity, Tucker tensor decomposition is adopted to extract features of the fused-feature map, and thus reducing the training time of the model.

The most important think in system-fault detection is detection accuracy, and it is better to have a number of wrong detections rather than missing detections, especially for a system that threatens the national economy and people's livelihood once a fault occurs in the overhead contact system. This research is based on improving classical object-detection algorithm Faster R-CNN. Feature fusion effectively retain the detailed information of the fault area in the original image, and Tucker decomposition accurately extracts its core features. Experiments show that the improved method has a significant effect on reducing miss probability and promoting detection accuracy. Compared with object-detection methods YOLOv3, SSD, and the original Faster R-CNN, the average miss probability of the improved Faster R-CNN model is decreased by 37.83%, 51.27%, and 43.79%, respectively, and average detection accuracy is increased by 3.6%, 9.75%, and 5.9%, respectively. In particular, in the case of dense droppers, the algorithm could still effectively detect droppers that are not stressed, and accurately detect the positions of pole number-plate malfunctions and foreign-body invasions, which has important reference significance for the detection of other types of faults in overhead contact systems.

Acknowledgements This paper is supported by Major Special Project (18-A02) of China Railway Construction Corporation in 2018 and Science and Technology Program (201809164CX5J6C6, 2019421315KYPT004 JC006) of Xi'an.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Yuan, G.-K., Li, K.-Q.: Supporting economic and social upgrading development with the construction of railway transportation artery. *China Emerg. Manag.* **06**, 24 (2016)
2. Li, C.-H.: Talking about common fault points of catenary and technical measures to improve performance. *Guide Sci-tech Mag.* **02**, 182 (2012)
3. Karakose, E., Gencoglu, M., Karaköse, M., Aydin, I., Akin, E.: A new experimental approach using image processing-based tracking for an efficient fault diagnosis in pantograph–catenary systems. *IEEE Trans. Ind. Informat.* **13**(2), 635–643 (2017)
4. Liu, Z., Liu, W., Nunez, A., Han, Z.: Unified deep learning architecture for the detection of all catenary support components. *IEEE Access* **8**, 17049–17059 (2020)
5. Qu, Z., Yuan, S., Chi, R., Chang, L., Zhao, L.: Genetic optimization method of pantograph and catenary comprehensive monitor status prediction model based on adadelta deep neural network. *IEEE Access* **7**, 23210–23221 (2019)
6. Zhong, J., HanLiu, Z., Han, Y., Zhang, W.: A CNN-based defect inspection method for catenary split pins in high-speed railway. *IEEE Trans. Instrum. Meas.* **68**(8), 2849–2860 (2019)
7. Yang, P.: Generative adversarial networks based on dual discriminator and its application in OCS nest detection [D]. Southwest Jiaotong University, Chengdu (2018)
8. Petitjean, C., Heutte, L., Kouadio, R., et al.: A top-down approach for automatic dropper extraction in catenary scenes. *Proc IbPRIA Póvoa de Varzim Portugal* **5524**, 225–232 (2009)
9. Jiang, X.-L., Jia, W.-B.: Machine vision detection method for foreign object intrusion in high-speed rail contact net. *Comput. Eng. Appl.* **55**(22), 250–257 (2019)
10. Redmon J, Farhadi A 2018 YOLOv3: An Incremental Improvement
11. Wu, J.-F., Jin, W.-D., Tang, P.: Catenary pillar image anomaly detection combined with SVDD and CNN. *Comput. Eng. Appl.* **55**(10), 193–198 (2019)
12. Xu, Y.-B.: Application of image processing in detecting defects of catenary hanger. Southwest Jiaotong University, Chengdu (2018)
13. Guo, Q.-F., Liu, L., Xu, W.-J., et al.: An improved faster R-CNN for high-speed railway dropper detection. *IEEE Access* **8**, 105622–105633 (2020). <https://doi.org/10.1109/ACCESS.2020.3000506>
14. Carroll, D., Chang, H.J.: Analysis of individual differences in multidimensional scaling via an n-way generalization of Eckart-Young decomposition. *Psychometrika* **35**(3), 283–319 (1970)
15. Harshman, R.A., Lundy, M.E.: Parafac: Parallel factor analysis. *Comput. Stat. Data Anal.* **18**(1), 39–72 (1994)
16. Tucker, L.R.: Some mathematical notes on three-mode factor analysis. *Psychometrika* **31**(3), 279–311 (1966)

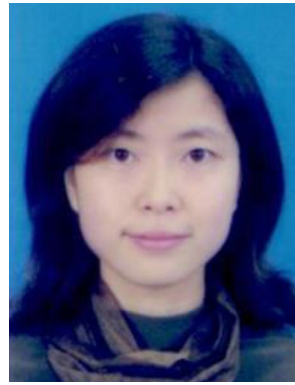
Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations



Xuewu Zhang Xuewu Zhang was born in Gansu Province, China, in 1984, and graduated from Beijing Jiaotong University in 2009 with a master's degree in electrical engineering and automation. His areas of expertise include the contact line technology of the electrified railways in complex and challenging environments, and the intelligent monitoring technology of the contact line.



Yansheng Gong Yansheng Gong was born in Shandong Province, China in 1965 and graduated from Southwest Jiaotong University in Sichuan Province in 1987 with a Bachelor of Engineering degree in design of electric locomotive. His area of research focuses on electrified railway traction power supply technology in complex and difficult environment, and rail transportation intelligent power supply technology.



Chen Qiao Chen Qiao was born in Xi'an, China, in 1976. She received her PhD in Applied mathematics from Xi'an Jiaotong University in 2009. Her current research interests are focused on Machine Learning, Deep Learning, Neuroimaging and Biomedical Image Analysis.



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