IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 7, 2020, accepted March 25, 2020, date of publication March 30, 2020, date of current version April 14, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2984264

Multi-Target Defect Identification for Railway Track Line Based on Image Processing and Improved YOLOv3 Model

XIUKUN WEI[1], DEHUA WEI[2], DA SUO[2], LIMIN JIA[1], AND YUJIE LI[3]

[1]State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing 100044, China
[2]School of Traffic and Transportation, Beijing Jiaotong University, Beijing 100044, China
[3]Beijing Mass Transit Railway Operation Corporation Ltd., Beijing 100044, China

Corresponding author: Xiukun Wei (xkwei@bjtu.edu.cn)

This work was supported in part by the Chinese National Key Project of Research and Development Program under Grant 2016YFB1200402.

ABSTRACT The condition monitoring of railway track line is one of the essential tasks to ensure the safety of the railway transportation system. Railway track line is mainly composed of tracks, fasteners, sleepers, and so on. Given the requirements for rapid and accurate inspection, innovative and intelligent methods for multi-target defect identification of the railway track line using image processing and deep learning methods are proposed in this paper. Firstly, the track and fastener positioning method based on variance projection and wavelet transform is introduced. After that, a bag-of-visual-word (BOVW) model combined with spatial pyramid decomposition is proposed for railway track line multi-target defect detection with a detection accuracy of 96.26%. Secondly, an improved YOLOv3 model named TLMDDNet (Track Line Multi-target Defect Detection Network), integrating scale reduction and feature concatenation, is proposed to enhance detection accuracy and efficiency. Finally, to reduce model complexity and further improve the detection speed, with the help of dense connection structure, a lightweight design strategy for the TLMDDNet model named DC-TLMDDNet (Dense Connection Based TLMDDNet) is proposed, in which the DenseNet is applied to optimize feature extraction layers in the backbone network of TLMDDNet. The effectiveness of the proposed methods is demonstrated by the experimental results.

INDEX TERMS Railway track line defects, multi-target defect identification, image processing, deep learning, YOLOv3.

I. INTRODUCTION

In recent years, the rapid development of rail transit puts more stringent demands on transportation safety and maintenance decisions. The health state of the railway track line is critical to ensure the safe and stable operation of rail transit [1]. Railway track line is mainly composed of tracks, fasteners, sleepers, etc. Due to the influence of contact friction and vibration between the train wheels and track, coupled with the effect of the operating environment on site, defects such as rail corrugation or broken fasteners may occur on the railway track line. With the occurrence and evolution of railway track line defects, the safety of vehicle operation and passenger comfort are reduced. The maintenance cost and the difficulty of maintenance decision-making are also increased. In addition, the condition monitoring of the railway track line is mainly carried out by manual inspection or using track inspection car up till the present. Manual inspection is low detection efficient and costly, while the track inspection car has high manufacturing cost and occupies regular operating track lines during the inspection process. Therefore, there is an urgent demand to develop a railway track line detection system that uses advanced technologies such as image processing, computer vision, deep learning, and fast speed and high-resolution cameras to improve the safety and stability of the rail transit, inspect the railway track lines automatically, shorten the detection time and reduce the maintenance costs [2].

In the last decade, many researchers and institutions have worked on the development of automated railway

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

inspection methods and systems based on advanced technologies. Among them, the method for detecting fastener defect mainly include image processing-based method and deep learning-based method. In general, the image processing-based detection method has the following three key steps: 1) locating and segmenting the fastener region; 2) extracting the image features of the fastener; 3) using classification algorithms to recognize fastener defects. Existing researches effort to improve detection performance from one or more of these aspects. In [1], [3]–[5], considering the position and other characteristics of the fastener, several improved fastener positioning algorithms are proposed to improve the positioning and segmentation results of fasteners, to increase the detection accuracy of fastener defects. However, these algorithms lack robustness to complex detection scenarios. In [6]–[9], underlying features including histogram of oriented gradient (HOG), local binary pattern (LBP), and Haar-like features are used to extract the image features of fasteners. Then classification algorithms are adopted to identify fastener defects. Nevertheless, most of these algorithms can only detect the completely missing fasteners, and the recall performance of these methods is not sufficient for practical application. In [10], a novel multiple signal classification (MUSIC) algorithm is presented for fastener defect detection, which can classify the signals produced by different track components. In [11], by using line local binary pattern (LLBP), an algorithm for high-speed railway fastener detection is proposed to detect the failed fasteners in different environments. In [12], a novel vision-based fastener inspection system (VFIS) inspired by few-shot learning is presented. Nonetheless, the fasteners mentioned in these three papers are quite different from the fasteners considered in this paper. At present, deep learning-based fastener defect detection methods are gradually becoming popular in the industry. In [13] and [14], a multi-layer perception neural classifier is used to detect missing fasteners, and an algorithm for online fastener detection is implemented with the help of GPU. In [15] and [16], an MTL (Multitask Learning) framework combining multiple detectors is put forward for the detection of railway tie and fastener. Although this approach results in improved detection accuracy, the detection speed of multitasking detectors is not discussed. In [17], the detection and identification of fastener defects using image processing technologies and deep learning networks are studied, and higher recognition precision and recall are obtained. Whereas, the detection speed of fastener defects detection needs to be improved.

Besides the defects detection of the fastener, the detection of rail surface defects is also investigated by some researchers in the last decades. Here, some representative investigations are reviewed. Rail surface defects are roughly divided into two categories: corrugations and discrete defects. Rail corrugation is a periodic irregular wear phenomenon on the rail surface, while discrete defects appear on the surface of a rail head randomly. For the detection of rail corrugation, there are methods based on spatial features and frequency domain features. Spatial feature-based methods [18]–[20] mainly use spatial feature extraction followed by classification algorithms to identify rail corrugation, in which a high recognition precision is obtained. But, the method is sensitive to the setting of algorithm parameters. In [21] and [22], rail corrugation identification methods based on rail image features in the frequency domain are studied. The local frequency features used in this method can reduce the detection time effectively. Nevertheless, the automation of the detection process can be further improved. As for the rail surface discrete defect, the main detection methods are as follows. In [23]–[25], intelligent visual inspection systems are constructed for inspecting discrete defects in real-time. However, this detection method is susceptible to noise, and the recall performance of the defect detection system needs to be improved. In [26]–[28], image enhancement combined with threshold binarization is adopted to detect discrete defects. Some attempts have been made to improve detection performance, but irregularities seriously influence such methods, and threshold selection is not universal. In [29]–[31], background modeling-based detection methods offer a new way to model rail surface images to detect discrete defects. This type of method has ideal detection accuracy and can identify defects at different scales. Whereas, they face the challenge of high computational complexity. Nowadays, deep learning-based methods have gradually become the main method for detecting rail surface defects. In [32]–[36], various detection networks based on convolutional neural network (CNN) are designed to improve the detection accuracy and speed of rail surface discrete defects and to increase the intelligence of the detection process. Nonetheless, these approaches perform well in their specified tasks, and the detection process can be further simplified.

It can be concluded that both the rail surface and fastener defect inspection techniques have made significant progress. However, the researches mentioned above have only investigated the defect detection problem on the rail surface or fastener separately. In the railway track line, the track and fasteners do not appear individually and can be captured simultaneously by one camera. For practical application, it is necessary to investigate the defect detection problem of track and fasteners at the same time. It would be much more convenient, economical, and essential to use only one detection algorithm and one real monitoring system. At present, with the development of image processing and deep learning technologies, as well as the continually improving computer technology, it provides an opportunity for the detection of railway track line multi-target defect. Nevertheless, there are still some problems that need to be solved when developing a railway track line detection system that inspects track and fasteners simultaneously. The positioning accuracy and robustness are two issues need to be considered first. Furthermore, improved track and fastener defect feature extraction needs to be developed so that higher classification accuracy can be achieved. Finally, the detection time consuming for the entire process, and each image should be shortened, while the

complexity of the detection process should be reduced so that the improved method can be applied to practical application.

To solve these problems, the method based on image processing combined with traditional classification algorithm is first investigated. Specifically, The positioning of track and fasteners in the detected image is realized by using variance projection and wavelet transform. After that, Dense-SIFT [37] is used to extracting local image features of the track and fastener. The accuracy of the classification of the track and fastener status is improved by using Bag-of-Visual-Word model [38] combined with spatial pyramid decomposition technique [39]. Although this method can achieve a high detection precision, the whole detection process is complicated. In addition, the detection method based on traditional image processing technologies is not highly automated, and the feature extraction algorithm is not good enough in robustness and sensitive to image changes. Moreover, this method is only suitable for one type of scenario. Therefore, to simplify the entire detection procedure and improve the detection performance, the latest proposed deep learning method YOLOv3 [40] is considered in this paper to detect the railway track line multi-target defect. In order to improve detection accuracy and efficiency of the YOLOv3 network, and more suitable for multi-target defect identification of railway track line, an improved YOLOv3 model with scale reduction and feature concatenation is proposed, named TLMDDNet (Track Line Multi-target Defect Detection Network). Furthermore, to reduce the number of parameters of the TLMDDNet model while maintaining a desirable detection performance, DenseNet [41] is used to optimize feature extraction layers in the backbone network, named DC-TLMDDNet (Dense Connection Based TLMDDNet). Railway track line images containing multi-target status, including several different track defects and several different types of fasteners with different defects, are collected and used as input data for training the neural network models. Based on the experimental results, the method of deep learning is more efficient and robust than the traditional image recognition algorithm.

To summarize, the main contribution of this paper includes the following four aspects:

1) Based on advanced image processing technologies and deep learning networks, the problem of multi-target defect identification for the railway track line is studied and solved for the first time, and the proposed methods meet the demands for the inspection task of the railway track line.
2) A track line multi-target defect detection network (TLMDDNet) is proposed based on improved YOLOv3 with scale reduction and feature concatenation, which has better detection accuracy and efficiency than original YOLOv3 and traditional image processing-based method. To our best knowledge, this is the first research that introduces YOLOv3 to railway track line multi-target defect detection for fastener and track surface defects simultaneously.

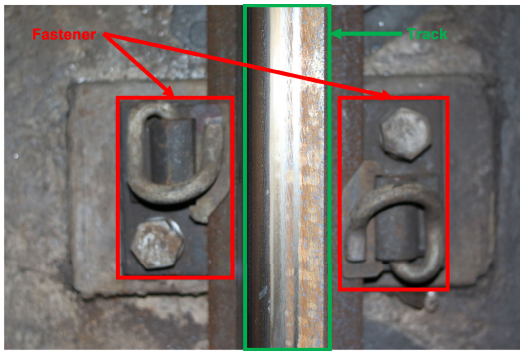

FIGURE 1. The railway track line structure of Beijing Metro Line 6.

3) With the help of dense connection structure, a lightweight TLMDDNet model is presented, named DC-TLMDDNet (Dense Connection Based TLMDDNet). DC-TLMDDNet effectively reduces the number of model parameters of TLMDDNet while maintaining a desirable detection performance.
4) The proposed detection networks can also be used to comprehensively detect the defects of track and different types of fasteners simultaneously.

The remainder of this paper is organized as follows. In Section II, the problem considered in this paper is stated. In Section III, the positioning issue of track and fastener in the railway track line image is investigated. In Section IV, the railway track line multi-target defect identification based on Dense-SIFT and SVM is presented. In Section V, the proposed railway track line multi-target defect detection model TLMDDNet is presented in detail. After that, the method of the lightweight design of TLMDDNet is investigated in Section VI. Finally, some conclusions of this paper are given in Section VII.

II. PROBLEM STATEMENT

A railway track line image contains a track and two fasteners is shown as in Fig. 1, in which the track fasteners fix the track on the ballast bed. In light of the on-site survey of the Beijing Metro Line 6 and the information provided by the maintenance engineers. The defects of the railway track lines mainly consist of three categories: broken fastener, missing fastener, and rail corrugation, as shown in Fig. 2(a) and Fig. 2(b), respectively. The broken fastener is defined as the complete or partial breakage of the elastic bar of the fastener. The missing fastener is defined as the main part missing or the complete missing of the fastener. The rail corrugation is a periodic irregular wear phenomenon on the rail surface. The dataset used in this paper is mainly made up of images that are taken from Beijing Metro Line 6. In addition, other railway track line images with different fastener types are collected and discussed to investigate the comprehensive detection capability and feasibility of the defects of the track and different types of fasteners.

The railway track line is mainly composed of a variety of key components, such as tracks, fasteners, etc. In general,

FIGURE 2. Defects of railway track line of Beijing Metro Line 6. (a) missing fastener and broken fastener; (b) rail corrugation.

in the railway track line, the track and fasteners do not appear separately and can be captured simultaneously by one camera. For practical application, it is necessary to inspect the status of multiple vital components at the same time. It would be much more convenient and economical to construct a comprehensive railway track line detection system using only one detection algorithm. However, in the existing researches reviewed above, they investigated the defect detection problem of rail surface or fastener, respectively. Therefore, it is critically important to investigate the automatic and intelligent multi-target identification of railway track line defects for practical application. In the following, the identification of railway track line multi-target defect is investigated based on advanced image processing technologies and deep learning networks, respectively.

A. RAILWAY TRACK LINE MULTI-TARGET DEFECT IDENTIFICATION BASED ON IMAGE PROCESSING

First of all, to alleviate the influence of noise and asymmetrical illumination, a filter-based image de-noising method and histogram equalization-based image enhancement algorithm are considered to improve the original image quality. After that, through the horizontal and vertical projection of the image, combined with the wavelet transform of the image, the positioning of the track and fasteners are realized to reduce unnecessary interferences. Finally, given the feature difference between the track and fastener, meanwhile, the spatial relationship between fastener parts, a BOVW model combined with spatial pyramid decomposition is proposed to classify the track and fastener defects. For the sake of convenience, this method is named as SPD_BOVW in the following.

B. RAILWAY TRACK LINE MULTI-TARGET DEFECT IDENTIFICATION BASED ON IMPROVED YOLOv3 MODELS

To simplify the process of railway track line multi-target defect detection and speed up the detection time, the advanced object detection model YOLOv3 is introduced to the railway track line multi-target defect detection issue, which can carry out both positioning and classification simultaneously. Additionally, to improve detection accuracy and efficiency, and be more suitable for railway track line multi-target defect identification, an improved YOLOv3 model with scale reduction and feature concatenation is proposed in this paper, referred to as TLMDDNet. Furthermore, to reduce the parameter number of TLMDDNet and maintain a sound detection performance, the dense block in DenseNet is considered to replace part of the residual blocks in the TLMDDNet backbone network, referred to as DC-TLMDDNet. The proposed DC-TLMDDNet based method can reduce the complexity of TLMDDNet and further improve the detection speed.

III. TRACK AND FASTENER POSITIONING BASED ON IMAGE PROCESSING

A. IMAGE PREPROCESSING

The images used in this part are captured by a handhold DSLR camera, in the meantime, the camera angle is perpendicular to the track line, and to some extent, the distance between the camera and track line is the same. The original images are RGB images with a resolution of 5472×3648 pixels, and the track is located in the middle of each image with two fasteners. Before image processing, to reduce calculations and speed up operations, the original images are resized into 540×360 pixels.

Owing to the complex environment of the metro railway system, the railway track line images collected from the field are susceptible to noise and asymmetrical illumination, which would degrade the image quality and affect the subsequent positioning process. To alleviate this problem and improve the accuracy of the positioning result, the median filtering [42] is firstly used to denoise the image, which can not only reduce the interference caused by salt and pepper noise to the image but also can protect the edge of the target object from blurring during processing. After that, the histogram equalization (HE) algorithm [43] is applied to enhance the original images.

B. POSITIONING OF THE TRACK AND FASTENERS

The railway track line images collected from the field usually contain irrelevant components such as sleepers and subgrade, which would cause unnecessary interference and increase the amount of calculation. Therefore, it is necessary to extract the track and fasteners in the original images to decrease this interference and reduce the computational cost. Taking the characteristics of the railway track line image and the positional relationship between the objects into account, a two-stage positioning strategy is proposed.

The first stage of track and fastener positioning is to position the track and backing plate. Specifically, the binarized image of the railway track line image is first obtained by using the adaptive threshold algorithm mentioned in [44] to reduce the amount of computation and simplify the positioning

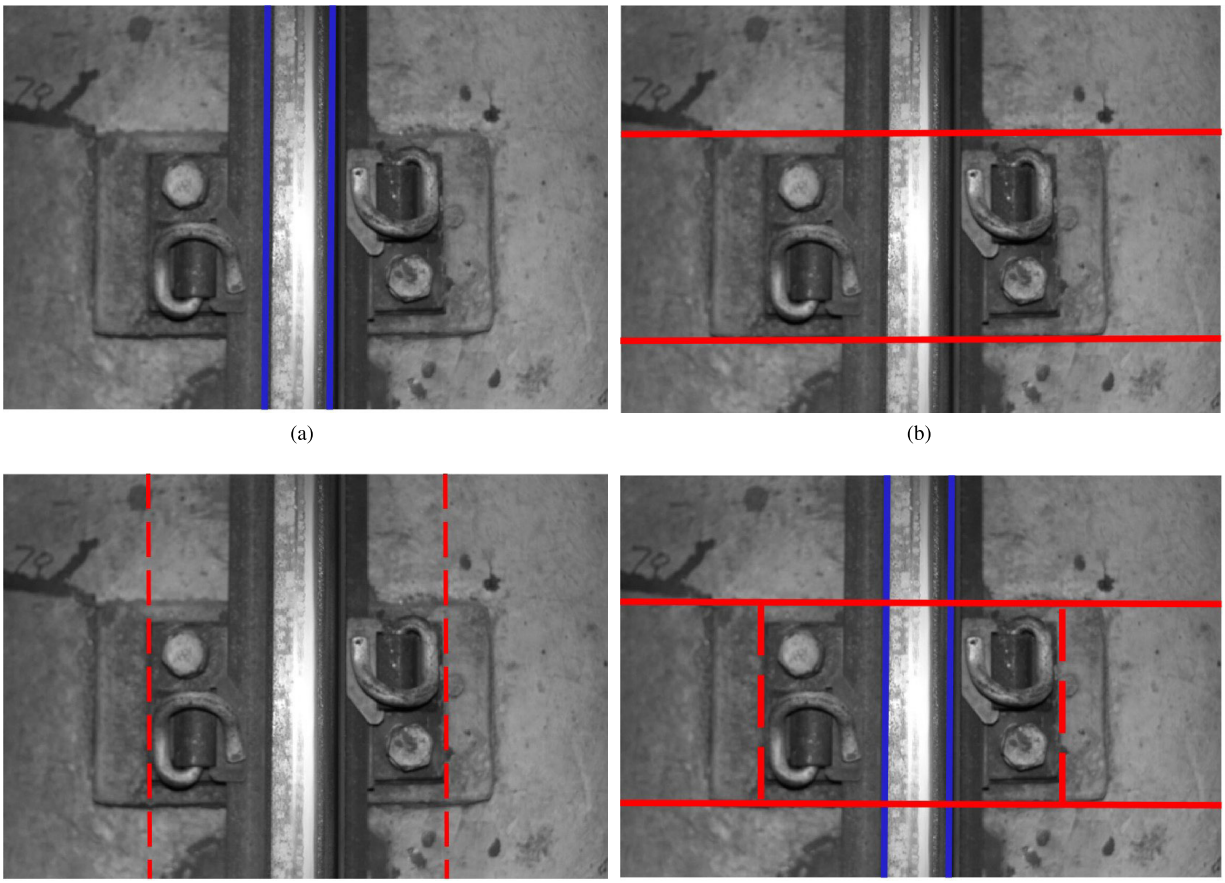

FIGURE 3. The process and results of track and fastener positioning. (a) the result of track positioning; (b) the result of backing plate positioning; (c) positioning result of fastener vertical boundaries; (d) The positioning result of track and fastener.

process of target areas. After that, the obtained binary image is processed by a vertical projection, and the difference between the projection statistics is computed to determine the starting and ending coordinates of the track, as shown in Fig. 3(a). Finally, the obtained binary image is processed by a horizontal projection, and the difference between the two adjacent projection statistics is calculated to determine the starting and ending coordinates of the backing plate, as shown in Fig. 3(b).

After locating the edge of the track and backing plate, the positioning of the fastener is performed in the second stage. First of all, given the location characteristics of the fastener, the railway track line image is processed by the vertical component of the wavelet transform [45]. Secondly, a line structural unit for morphological opening operation is designed to reduce the vertical interference and noise around the fastener profile. Finally, the vertical boundaries of the fastener are determined using the improved template matching algorithm proposed in [17]. The fastener boundary positioning results are shown in Fig. 3(c).

In summary, by combining the positioning results of the track, the backing plate, and the fastener boundary, the final positioning results of the track and fasteners are accurately obtained, as shown in Fig. 3(d). More details about the positioning of track and fastener can be found in our previous work [17]. The positioning methods used in this paper is not only applicable to subway tracks, but also high-speed rail lines, and perform better than other methods.

IV. RAILWAY TRACK LINE MULTI-TARGET DEFECT IDENTIFICATION BASED ON SPD_BOVW

In the previous section, the problem of track and fastener positioning is solved. As a result, the obtained image dataset only contains the track and fastener without other components. In the light of the results obtained before, a comprehensive detection method for track and fastener defects based on image local feature extraction and classification is proposed.

Up to now, the most widely used algorithm for local image feature extraction is the Dense Scale Invariant Feature Transform (Dense-SIFT) [37]. In this paper, the Bag-of-visual-word model [38] is first reconstructed for the Dense-SIFT feature extraction of track and fastener. Then the final feature vectors of track and fastener images are fed into SVM for multi-target classification.

A. DENSE-SIFT FEATURE

Nowadays, the widely-used method for image local feature extraction is SIFT (Scale-Invariant Feature Transform) [46].

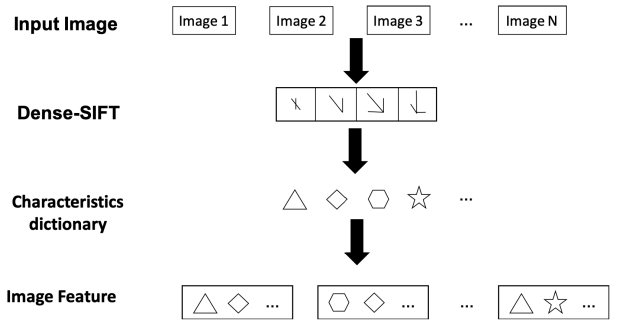

FIGURE 4. The general diagram of BOVW (Bag-of-Visual-Word) model.

In comparison with the traditional SIFT, Dense-SIFT is not necessary to construct Gaussian multi-scale space, which reduces the computation complexity. The key to the Dense-SIFT algorithm is to determine the description of the key points, which helps obtain the description that is insensitive to illumination changes. The method of extracting description from key points is defined as follows [46]

$$m(x,y) = \sqrt{(L(x+1,y)-L(x-1,y))^2+(L(x,y+1)-L(x,y-1))^2} \quad (1)$$

$$\theta(x,y) = tan^{-1}\left[\frac{L(x,y+1)-L(x,y-1)}{L(x+1,y)-L(x-1,y)}\right] \quad (2)$$

where $L(x, y)$ stands for one pixel of the input image. $m(x, y)$ and $\theta(x, y)$ are the gradient magnitude and the gradient direction of each bin, respectively. Moreover, the size of each patch is 4 bins $\times$ 4 bins, the measure of each bin is 4 pixels $\times$ 4 pixels, and the final Dense-SIFT feature of each image is a 128-dimensional vector. More details about Dense-SIFT can be found in [37].

B. BAG-OF-VISUAL-WORD MODEL

The traditional BOVW (Bag-of-Visual-Word) model [38] is developed from the BOW (Bag-of-Word) model [47] for text categorization. BOVW is a general pipeline that builds a global representation from local features. Specifically, the visual vocabulary vector is first extracted from different types of images using the Dense-SIFT algorithm. After that, the K-Means algorithm is adopted to merge visual words with similar meanings to construct a codebook containing K words. Finally, by calculating the number of times each visual word of the image appears in the codebook, the image is represented as a K-dimensional value vector. The diagram of the BOVW model is shown in Fig. 4.

C. SPATIAL PYRAMID DECOMPOSITION

Although BOVW can extract the local features of the image well, it lacks the spatial position information of the image. To overcome this inherent shortcoming of BOVW, spatial pyramid decomposition [39] of the image is considered.

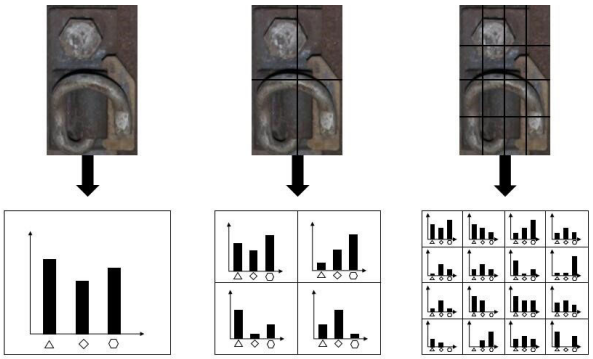

FIGURE 5. The diagram of visual characteristic word bag model based on spatial pyramid decomposition.

This method can not only extract the global information of the image but also integrates the spatial position information of track and fastener visual feature words into the BOVW model more comprehensively and effectively. In the spatial pyramid decomposition, the image is divided into spatially sub-regions that are gradually refined, and then local feature histograms are calculated from each region, as shown in Fig. 5. More details about spatial pyramid decomposition can be found in [39].

D. EXPERIMENTS AND RESULTS

1) DATA PREPARATION

After image pre-processing and positioning of track and fastener, a dataset consisting of two types of track images and three types of fastener images is finally obtained. In this dataset, there are 515 complete fasteners, 20 broken fasteners, 11 missing fasteners, 706 normal rails, and 26 rail corrugation images. Due to the small number of defective fasteners and rail images, data argumentation methods are used to expand the number of faulty samples and to ensure experimental reliability. The data enhancement methods used herein include rotation, flipping, mirroring, and noise addition. Specifically, the salt noise and the Gaussian noise with a mean of 0.1 and a variance of 0.01 are used in the noise addition. After data augmentation, the data set contains 515 complete fasteners, 210 broken fasteners, 110 missing fasteners, 706 normal rails, and 203 rail corrugation images. Since unbalanced samples could affect the classification results, subsampling is used for samples in these five categories. As a result, the data used for the experiment contains 110 samples for each type. Moreover, 70% of each type of image are used for training, and the remaining 30% are used for validation. The dataset of track and fastener is shown in Table 1.

2) EXPERIMENT PROCESS AND RESULT

Railway track line multi-target defect identification based on image feature extraction mainly consists of four steps. In the first step, the fastener images used for training and testing are adjusted to 128 $\times$ 256 pixels while the track images are adjusted to 48 $\times$ 480 pixels. Moreover, the patch size is set

TABLE 1. The number of different types of track and fastener in the dataset.

Type	rail		fastener		
	normal	corrugation	complete	broken	missing
Training	77	77	77	77	77
Testing	33	33	33	33	33

TABLE 2. The comparative classification results under different dictionary size and different kernel of SVM.

Size of dictionary	RBF kernel	Histogram intersection kernel
200	84.34%	91.92%
300	93.43%	93.43%
400	93.86%	94.44%
500	96.46%	94.95%
600	92.93%	91.36%

to 4 bins $\times$ 4 bins, the bin size is 4 pixels $\times$ 4 pixels, and the sampling step size is set to 8 pixels. Secondly, the bag of visual words with a dictionary for extracted Dense-SIFT features is constructed. After that, each image of the dataset is decomposed by a 3-layer spatial pyramid decomposition model. Finally, the visual features are fed into SVM for training and testing.

To demonstrate the effect of dictionary size and kernel to the classification result, some comparative experiments are carried out under different dictionary sizes and kernels, and the results are shown in Table 2. As can be seen from Table 2, the dictionary size is set from 200 to 600, and the larger the dictionary size is, the higher the classification accuracy will be. Nevertheless, when the size of the dictionary is greater than 500, the classification accuracy decreased. Therefore, the optimal dictionary size for the BOVW model is chosen as 500. In addition, when the size of the dictionary is set to 500, the SVM with the RBF kernel has higher classification accuracy than the SVM with the histogram intersection kernel.

To assess the proposed method quantitatively, indicators such as precision (P), recall (R), and F1 score are used, which are widely used in many scientific research fields [48]. These indicators are calculated as follows

$$P = \frac{TP}{TP + FP} \tag{3}$$

$$R = \frac{TP}{TP + FN} \tag{4}$$

$$F1 = \frac{2 * P * R}{P + R} \tag{5}$$

where TP represents the number of images that are correctly detected. FN indicates the number of images for missed inspection. FP is the number of images that are falsely reported. TN stands for the number of images that are correctly excluded.

Table 3 shows the classification results for these three indicators. It can be concluded that the identification method proposed in this paper achieves a sound performance, in which the precision and recall in the missing fastener recognition reach 100%.

TABLE 3. Detection results of track and fastener status based on pyramid decomposition.

Type	Precision(P)	Recall(R)	F1
Normal Rail (NR)	100%	93.94%	0.9688
Rail Corrugation (RC)	94.29%	100%	0.9706
Normal Fastener (NF)	96.77%	90.91%	0.9375
Missing Fastener (MF)	100%	100%	1
Broken Fastener (BF)	91.43%	96.97%	0.9412

TABLE 4. Classification accuracy of different method.

Method of feature extraction	Accuracy
HOG	92.56%
LBP	93.35%
This paper	96.36%

In addition, to prove the effectiveness of the identification method presented in this paper, a comparative experiment with other feature extraction methods (e.g., HOG feature and LBP feature) is carried out. The result is shown in Table 4.

Based on the result of the above comparison experiment, one can see that the method adopted in this study is superior to other feature extraction methods. The reason that HOG feature extraction and LBP feature extraction lead to classification errors is that these feature extraction methods ignore the spatial positional relationship of the extracted track and fastener features. Moreover, in this paper, the BOVW model and the spatial pyramid decomposition method are applied to the multi-classification problem of track and fastener for the first time, with an accuracy of 96.36%.

Although the final classification result reaches a sound detection accuracy, the entire detection process is complicated, and the result is mostly dependent on the setting and adjustment of many parameters. In addition, the detection method based on image processing is not good enough in robustness, and the feature extraction algorithm is cumbersome and sensitive to image changes. For these reasons, this method may not be competent for practical applications with complex detection scenarios. Therefore, in the following, deep learning-based methods are proposed to solve this problem.

V. RAILWAY TRACK LINE MULTI-TARGET DEFECT IDENTIFICATION BASED ON IMPROVED YOLOv3 MODEL

A. YOLOv3

The YOLOv3 network, evolved from YOLO (You Only Look Once) [49], has recently received extensive attention. It is a typical one-stage object detection network in light of the regression method. The idea of regression is used to return the target border and target category of each position directly at multiple scales of the image for a given input image. Specifically, the YOLOv3 network has made some application improvements based on YOLOv2 [50]. It not only proposes a more powerful Darknet53 network based on ResNet (Residual Network) [51] as a feature extractor but also uses multi-scale prediction for final target detection, which has excellent detection performance. The classic structure of the YOLOv3 network is shown in Fig. 6, it consists of a backbone network and detection module.

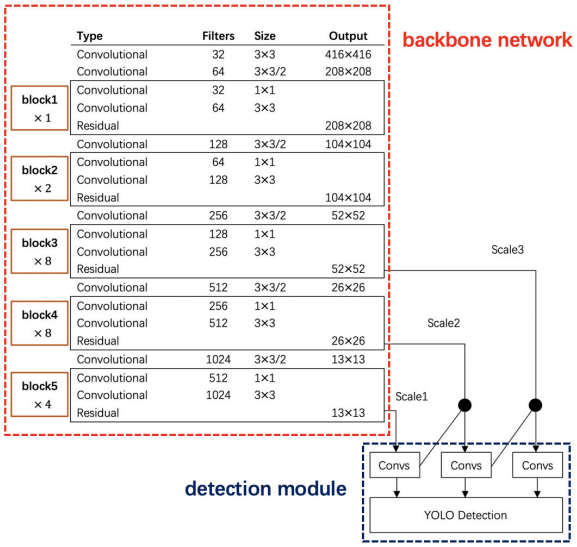

FIGURE 6. The network structure of YOLOv3.

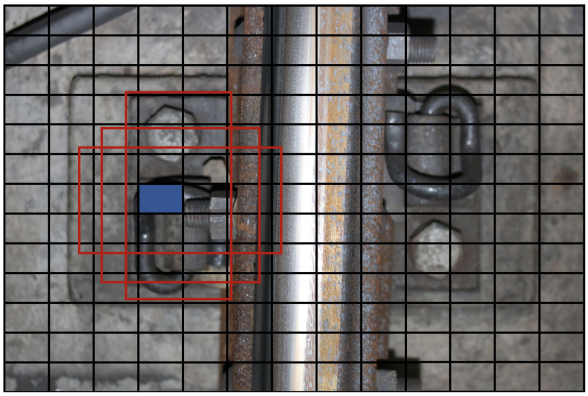

FIGURE 7. An illustration of predicted bounding boxes on 13 × 13 grids of YOLOv3.

In general, the YOLOv3 network scales the input image to a size of 416 × 416. Using a scaled pyramid structure similar to FPN (Feature Pyramid Network), the dataset image is divided into $s \times s$ grids according to the scale of the feature map. The final detection is performed on three scales of 13 × 13, 26 × 26, 52 × 52 feature map sizes, and the feature map is propagated on two adjacent scales by using the 2× upsampling. As shown in Fig. 7, each divided cell predicts 3 bounding boxes with 3 anchor boxes. More details about the YOLOv3 network can be found in [40].

B. TRACK LINE MULTI-TARGET DEFECT DETECTION NETWORK (TLMDDNet)

Although YOLOv3 network has excellent detection performance and has been successfully applied in many fields, the available YOLOv3 network can be further improved for the railway track line multi-target defect detection, directly, which is based on the following two investigations:

1) Within the railway track line image datasets, as shown in Fig. 1, the targets to be detected (e.g., rail surface and fasteners) occupy a large and consistent scale in the image. Furthermore, the relative position of fasteners and the track is fixed. Therefore, for the specific detection targets of this paper, the original YOLOv3 network detection process based on a multi-scale fusion method can be simplified.
2) Different detection targets in our datasets have significantly distinct low-level image features such as textures and edges. Moreover, as shown in Fig. 8, the shallow layers of the network have powerful low-level image features learning ability, and the effective use of the lower-layer feature map can improve the target detection accuracy, while the original YOLOv3 network fails to take full advantage of the low-level image features of the targets learned from the shallow layer.

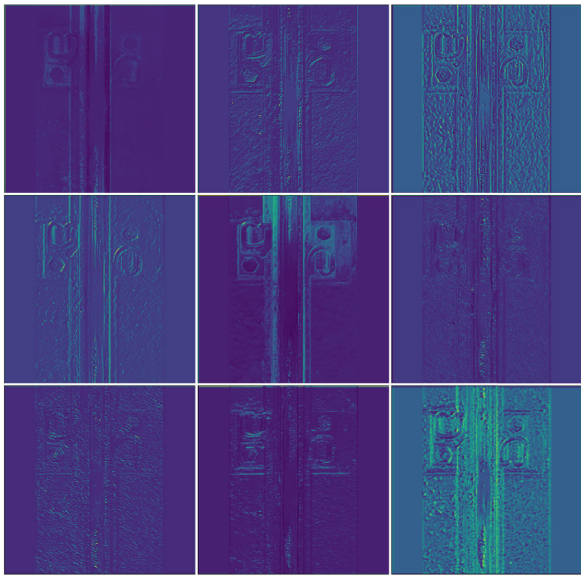

FIGURE 8. Visualization of some sample feature maps learned from shallow layer.

Therefore, to improve detection accuracy and efficiency, and be more suitable for railway track line multi-target defect identification, an improved YOLOv3 model with scale reduction and feature concatenation is proposed in this paper, called TLMDDNet (Track Line Multi-target Defect Detection Network) in brief, which mainly reconstructing the backbone of the YOLOv3 network.

More specifically, based on the former investigation, the network scale reduction that performs the final detection only on the scale of 13 × 13 feature map size is first implemented. Secondly, to take full use of the features learned from each block in the backbone network, a cross-block feature concatenation strategy is proposed, that is, the feature maps learned from each block are concatenated to all subsequent blocks as input through pooling, except for the last block. Furthermore, the feature maps of all block outputs in the backbone network are concatenated together as input to the detection module. Thus, through effective feature reuse and propagation, the input feature maps of each block can have enriched representation power and additional information for

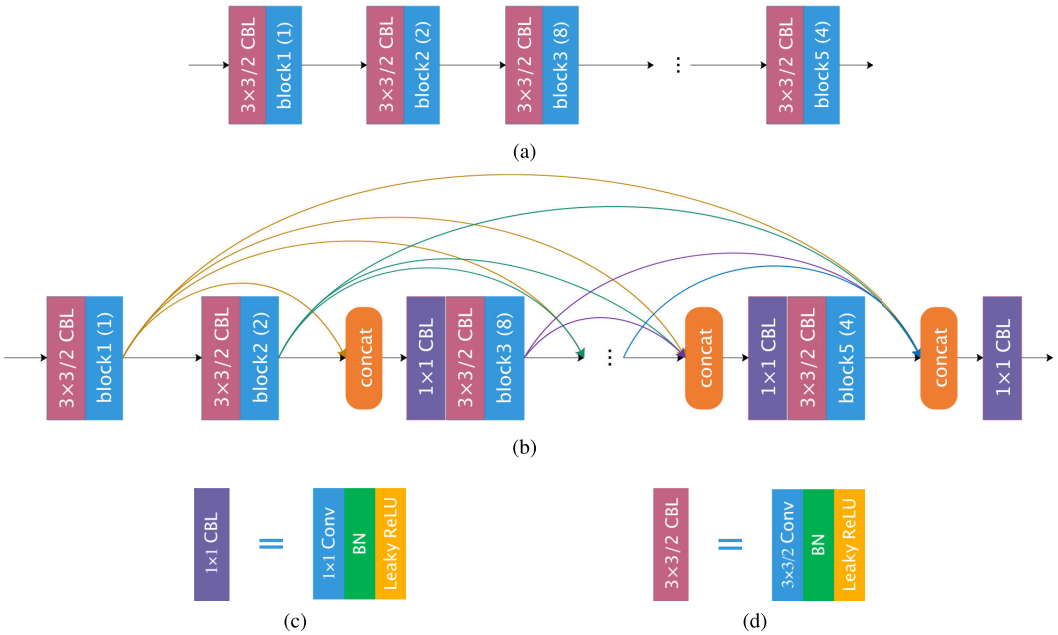

FIGURE 9. (a) the backbone structure of the original YOLOv3; (b) the backbone structure of the TLMDDNet; (c) 1 × 1 convolution module; (d) 3 × 3/2 convolution module. Note that the numbers in parentheses in the figure indicate the number of blocks.

characteristic learning of railway track line targets, which can be defined as follows

$$IBlock_i = Conc[MaxP(OBlock_1, OBlock_2, \ldots, OBlock_{i-1})], \quad i = 3, 4, 5 \tag{6}$$

$$IDBlock = Conc[MaxP(OBlock_1, OBlock_2, \ldots, OBlock_5)] \tag{7}$$

where $IBlock_i$ and $OBlock_i$ represent the input and output feature maps of the i-th block in the backbone network, respectively. $MaxP$ stands for the max-pooling operation. $Conc$ is the concatenation operation. $IDBlock$ denotes the input of the detection module. A brief diagram of the improved backbone structure is shown as in Fig. 9 (b).

Noted that after concatenating feature maps, the convolution module consisting of a 1×1 convolution layer followed by a batch normalization layer and leaky ReLU is necessary, which is used for dimension reduction and accelerate convergence. In addition, inspired by the idea on [40], the bounding box priors are redesigned by using K-means clustering, in which the Intersection over Union (IOU) of the rectangular box (represented by R_{IOU}) is used as the similarity, and the distance function of the cluster is defined as follows

$$d(B, C) = 1 - R_{IOU}(B, C) \tag{8}$$

where B and C stand for the size and center of the rectangular box, respectively. $R_{IOU}(B, C)$ represents the IOU between two rectangular boxes. Taking the network training performance and efficiency into account, 3 clusters are chosen as the bounding box priors, which are (61, 415), (78, 205), (112, 117). The specific network framework and parameters of TLMDDNet are shown in Fig. 10.

C. EXPERIMENTS AND RESULTS

1) INTRODUCTION TO EXPERIMENT DATASETS

To evaluate the performance of TLMDDNet for railway track line multi-target defect detection, two datasets are used in the experiment as follows:

Dataset 1 consists of images that are taken from Beijing Metro Line 6. We have collected 322 available images, and each image is an RGB image with a resolution of 3456 × 5472 pixels. In view of the defect features in the railway track line field, this dataset mainly consists of 5 classes: complete fastener, broken fastener, missing fastener, normal rail, and rail corrugation. These images are manually labeled by using the LabelImg [52] after image augmentation. The image data augmentation methods such as rotation, mirror, noise addition, color perturbation and etc. are applied. More specifically, the salt and pepper noise and the Gaussian noise with a mean of 0 and a variance of 0.05 are used in the noise addition. In addition, color perturbation mainly uses a random factor to comprehensively adjust the brightness, saturation, contrast, and sharpness of the image.

We end up with 1058 railway track line images containing 1679 complete fasteners, 320 broken fasteners, 100 missing fasteners, 680 normal rails, and 406 rail corrugation. 70% of the images are divided into a training set, and the rest 30% of the images are used as the test set. The data set for training and test is shown in Table 5.

Dataset 2 consists of 9 classes: complete GJ fastener, broken GJ fastener, missing GJ fastener, complete CK-1 fastener, missing CK-1 fastener, broken CK-2 fastener, normal rail, rail corrugation and rail spalling, which are shown as in Fig. 11. Noted that each image in this dataset is collected in the same

TABLE 5. Details of Dataset 1 used for experiments after data augmentation.

	Image	complete fastener	broken fastener	missing fastener	normal rail	rail corrugation
training set	740	1169	228	68	468	291
test set	318	510	92	32	212	115

TABLE 6. Details of Dataset 2 used for experiments after data augmentation.

	Image	complete GJ	broken GJ	missing GJ	complete CK-1	missing CK-1	broken CK-2	normal rail	rail corrugation	rail spalling
training set	2332	1169	228	68	1919	601	594	468	291	332
test set	1001	510	92	32	809	271	268	212	115	143

TABLE 7. Initialization parameters of TLMDDNet.

Size of input images	Batch	Momentum	Initial learning rate	Decay	Training steps
416 × 416	8	0.9	0.001	0.0005	400/600

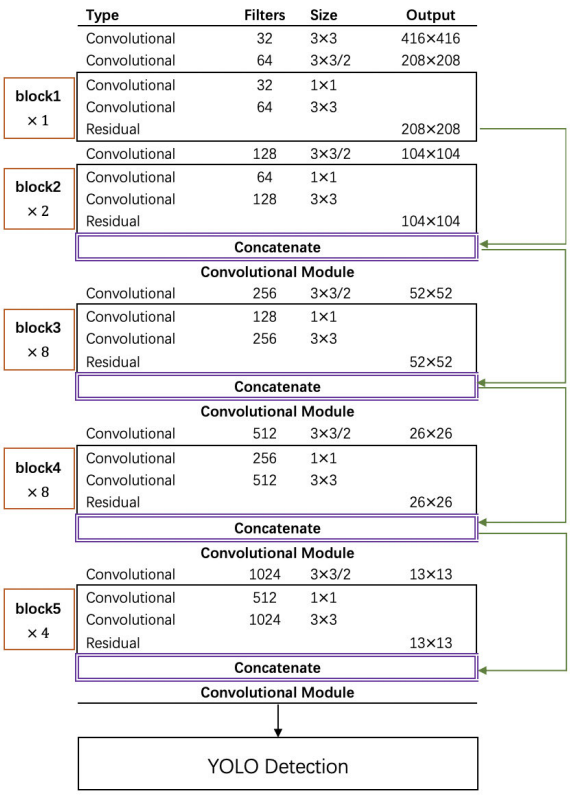

FIGURE 10. The network structure and parameters of TLMDDNet. (Different from original YOLOv3, the green solid shortcuts are used for cross-block feature map transmission.)

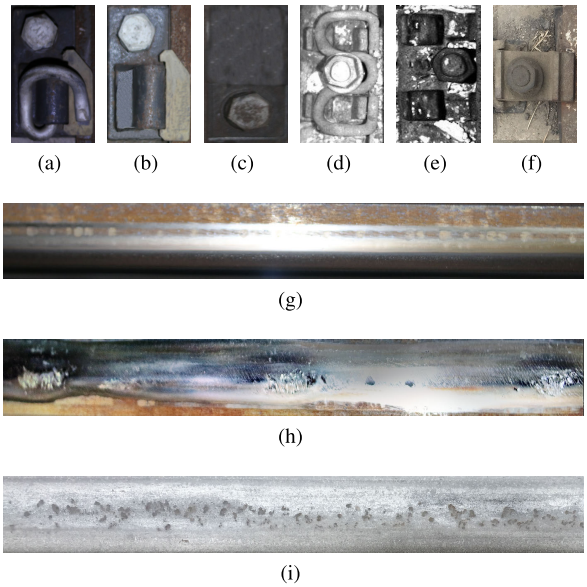

FIGURE 11. Defects sample image in Dataset 2. (a) complete GJ fastener; (b) broken GJ fastener; (c) missing GJ fastener; (d) complete CK-1 fastener; (e) missing CK-1 fastener; (f) broken CK-2 fastener; (g) normal rail; (h) rail corrugation; (i) rail spalling.

way as Dataset 1. Dataset 2 is an extension of Dataset 1. There are more fastener types in Dataset 2. It also consists of more track defects than the one in Dataset 1. Dataset 2 is applied to evaluate the performance of TLMDDNet for the comprehensive detection of railway track line multi-target defects, including more defects types. In addition, Dataset 2 is also constructed to investigate the robustness of the new proposed detection model in the face of multiple targets with diversity.

After the same image augmentation as above, there are a total of 3333 images used for experiments. 70% of the images are divided into a training set, and the rest 30% of the images are used as the test set. The data set for training and test is shown in Table 6.

2) PERFORMANCE EVALUATION FOR TRACK LINE DEFECTS IDENTIFICATION

This subsection evaluates the performance of the proposed model for railway track line multi-target defect detection. The TLMDDNet is implemented in the GOOGLE Tensorflow framework with Keras. The detection models are trained and tested on an NVIDIA Titan X [53] server. The network initialization parameters are shown in Table 7.

To adapt the input required for the model, the input images are adjusted to 416 × 416 pixels. Given the memory constraints of the server, the batch size is set to 8 in this experiment. 400 and 600 training epochs are used for training Dataset 1 and Dataset 2, respectively. The adaptive

moment estimation (Adam) [54] is adopted to update the weights of the networks. Parameters such as initial learning rate, momentum, weight decay regularization, and other parameters referred to the original parameters in the original YOLOv3. The idea of transfer learning [55] is used that the weights trained based on different datasets are used for the weight values initialization. The pre-trained weights are trained by using ImageNet [56], and only the weights of block1 and block2 in the original YOLOv3 network are loaded into our model. The model is trained after defining the training parameters and loading the pre-trained weights. During the training process, the weights of block1 and block2 in TLMDDNet are fixed.

In this experiment, the proposed model TLMDDNet is compared with SPD_BOVW, YOLOv3, and YOLOv3 with scale reduction (YOLOv3_SR). The relevant indicators for evaluating the effectiveness of these models are as follows

a. Loss Function

The loss function in YOLO is defined as follows

$$loss = \sum_{i=0}^{S^2} coord\_Error + iou\_Error + class\_Error \quad (9)$$

The coordinate prediction error *coord_Error* is defined as follows

$$\begin{aligned} &coord\_Error \\ &= \lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{I}_{ij}^{obj} [(x_i - \hat{x}_i)^2 - (y_i - \hat{y}_i)^2] \\ &\quad + \lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{I}_{ij}^{obj} [(w_i - \hat{w}_i)^2 - (h_i - \hat{h}_i)^2] \end{aligned} \quad (10)$$

where the parameter λ_{coord} is the weight of the coordinate error. S^2, B are the number of grids in the input image and the number of bounding boxes generated by each grid, respectively. According to [40], $\lambda_{coord} = 5$, $S = 13$, and $B = 3$ are selected in this paper. $\mathbb{I}_{ij}^{obj}$ indicates whether the object falls into the j-th bounding box in grid i or not. $(\hat{x}_i, \hat{y}_i)$ represents the value of the center coordinate. $(\hat{w}_i, \hat{h}_i)$ stands for the height and width of the predicted bounding box, respectively. (x_i, y_i) and (w_i, h_i) are the corresponding ground truth.

The IoU error *iou_Error* is defined as follows

$$\begin{aligned} iou\_Error = &\sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{I}_{ij}^{obj} [(C_i - \hat{C}_i)^2] \\ &+ \lambda_{noobj} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{I}_{ij}^{obj} [(C_i - \hat{C}_i)^2] \end{aligned} \quad (11)$$

where λ_{noobj} represents the weight of the IoU error, which is set to 0.5 in this study. $\hat{C}_i$ is the predicted confidence, and C_i is the corresponding true confidence.

The classification error *class_Error* is defined as follows

$$class\_Error = \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{I}_{ij}^{obj} \sum_{c \in classes} [p_i(c) - \hat{p}_i(c)]^2 \quad (12)$$

where c represents the class to which the detected target belongs. $\hat{p}_i(c)$ denotes the predicted probability value that the object belonging to class c in the grid i. $p_i(c)$ is the corresponding true probability value. In addition, note that the *class_Error* for grid i is the sum of classification errors of all the objects in the grid.

b. AP, mAP

The average-precision (AP) of each class and mean Average-Precision (mAP) are two significant indicators for quantitatively evaluating the detection performance of the model. The value of AP is the area enclosed by the precision/recall curve, and the mAP is the average of APs, that is, the mAP is defined as follows

$$mAP = \frac{1}{n} \sum_{i=1}^{n} AP_i \quad (13)$$

where n represents the number of object classes in the dataset. AP_i is the average-precision of class i.

c. Detection Rate, Parameters, FLOPs

The average detection rate for several detection models is compared and expressed in fps (frames per second). In addition to the average detection rate of the model, time complexity and spatial complexity are also two important indicators for evaluating and analyzing the model. In this experiment, the time complexity is represented by FLOPs (floating point operations), and the spatial complexity is represented by the number of parameters of the model. For a certain convolutional layer, the number of FLOPs is defined as follows

$$FLOPs_i = num\_params_i * (H_i * W_i) \quad (14)$$

where num_params_i denotes the number of parameters of the i-th convolutional layer. H_i and W_i are the height and width of the output feature maps of the i-th convolutional layer, respectively. In this paper, for convenience, BFLOPs (billion floating point operations) is used, where $1\ BFLOPs = 10^9\ FLOPs$.

For Dataset 1, the loss value of YOLOv3, YOLOv3_SR, and TLMDDNet during training is shown as in Fig. 12. The AP of each class, mAP, average detection rate, parameters, and FLOPs of the models are shown in Table 8.

Based on the above results, it can be seen that TLMDDNet has a faster convergence speed and better convergence results than YOLOv3 and YOLOv3_SR. This indicates that the training performance of the proposed detection model is slightly improved. In terms of the detection performance, the mAP of TLMDDNet is 0.9920, which is higher than the other three models. This shows that the cross-block feature concatenation strategy used in TLMDDNet can improve the accuracy of railway track line multi-target defect detection. The average detection rate of TLMDDNet is 34.25fps, which is faster than SPD_BOVW and YOLOv3 and is basically the same as the

TABLE 8. The performance of the four methods tested on Dataset 1.

Indicators	Items	SPD_BOVW	YOLOv3	YOLOv3_SR	TLMDDNet
Accuracy	complete_fastener	0.9010	0.9569	0.9961	0.9961
	broken_fastener	0.9189	0.9891	1.0000	1.0000
	missing_fastener	0.9091	1.0000	1.0000	1.0000
	normal_rail	0.8674	0.9118	0.9593	0.9815
	rail_corrugation	0.8485	0.9579	0.9391	0.9826
	mAP	0.8890	0.9631	0.9789	**0.9920**
Others	detection rate/fps	0.96	30.67	**35.97**	34.25
	parameters	\	∼58.7M	**∼53.8M**	∼56.3M
	BFLOPs	\	32.9	**27.5**	28.4

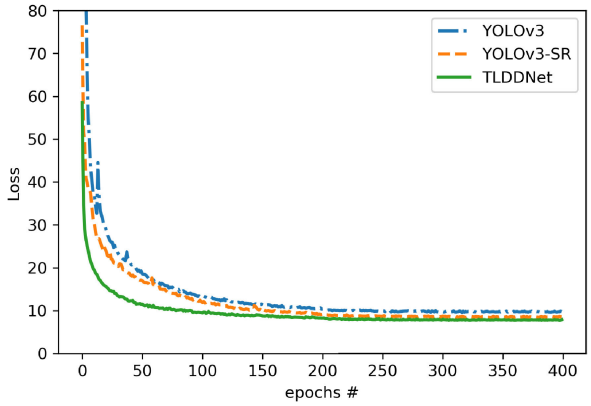

FIGURE 12. Loss curves of the three YOLO models.

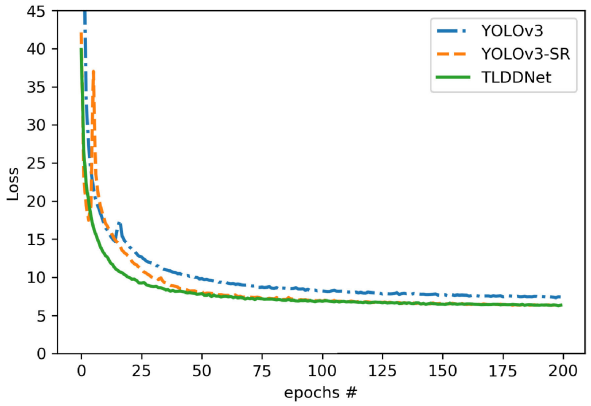

FIGURE 13. Partial loss curves for the three detection models.

YOLOv3_SR, indicating that it can provide rapid detection of railway track line multi-target defect in high-resolution images. Most importantly, as shown in Table 8, in comparison with the image processing-based detection method SPD_BOVW, the detection speed of TLMDDNet is increased by nearly 34 times, and the detection accuracy is improved by about 10.3%. These results show that deep learning-based method has much better detection performance in practice than image processing-based method. In addition, the number of parameters and time complexity of TLMDDNet are about 56.3M and 28.4 BFLOPs, respectively, which are less than YOLOv3 and slightly larger than YOLOv3_SR, indicating the validity of the TLMDDNet structure.

For Dataset 2, the loss value of YOLOv3, YOLOv3_SR, and TLMDDNet during training is shown as in Fig. 13. The AP of each class, mAP, average detection rate, parameters, and FLOPs of the models are shown in Table 9. The experiment results show that the proposed TLMDDNet has the best training performance and detection performance in these models. Specifically, for Dataset 2, the mAP and the average detection rate of TLMDDNet are 0.9947 and 33.0fps, respectively. Therefore, it is a better alternative for practical application.

VI. LIGHTWEIGHT DESIGN OF TLMDDNet

The TLMDDNet proposed in Section V achieves sound performance and is suitable for the railway track line multi-target defect detection. However, like the original YOLOv3 network, several residual blocks are used in the backbone of TLMDDNet, which achieves high detection accuracy and brings a large number of parameters to the entire network. Excessive parameters can lead to extended time training, increase the demand for data, and slow down the detection speed. Therefore, the structure of TLMDDNet can be further optimized.

A. DENSE CONNECTION IN THE TLMDDNet

DenseNet consisting of Dense Block is proposed by Huang *et al.* in 2017 [41]. It has higher computational efficiency and storage efficiency. In general, for the same prediction accuracy, DenseNet needs only half of the parameters of ResNet. This allows DenseNet to effectively alleviate gradient vanishing, strengthen feature propagation, facilitate feature reuse, and substantially reduce the number of parameters. More details about DenseNet can be found in [41].

In this paper, to decrease the complexity of the model and further speed up the detection of railway track line multi-target defect, inspired by DenseNet, a lightweight model for the TLMDDNet is proposed, named DC-TLMDDNet (Dense Connection Based TLMDDNet). In detail, as shown in Fig. 14 (a), the dense connection structure of convolutional layers, named DC block (Dense Connection block), is used to replace the residual blocks in block3, block4, and block5 located in the TLMDDNet backbone network. Each convolutional layer of the DC block in DC-TLMDDNet outputs k feature maps, that is, the growth rate is k; the l-th layer of the DC block has $k_0 + k \times (l-1)$ concatenated feature maps as input, where the number of the input feature

TABLE 9. The performance of the three methods tested on Dataset 2.

Indicators	Items	YOLOv3	YOLOv3_SR	TLMDDNet
Accuracy	complete_GJ_fastener	0.9824	0.9922	0.9961
	broken_GJ_fastener	1.0000	1.0000	1.0000
	missing_GJ_fastener	1.0000	1.0000	1.0000
	complete_CK1_fastener	0.8739	0.9975	1.0000
	missing_CK1_fastener	0.8229	0.9926	1.0000
	broken_CK2_fastener	0.9813	0.9963	0.9963
	normal_rail	0.9118	0.9907	0.9859
	rail_corrugation	0.9903	0.9417	0.9742
	rail_spalling	0.9790	0.9861	1.0000
	mAP	0.9438	0.9886	**0.9947**
Others	detection rate/fps	29.94	**34.13**	33.00
	parameters	∼58.7M	**∼53.8M**	∼56.3M
	BFLOPs	32.9	**27.5**	28.4

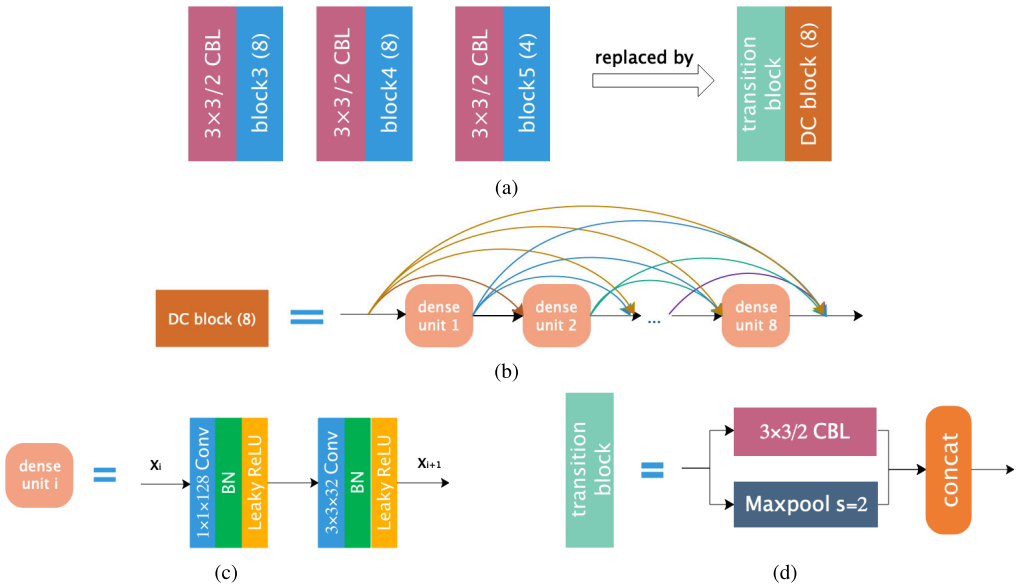

FIGURE 14. (a) The main differences between the backbone network of DC-TLMDDNet and the one of TLMDDNet; (b) the structure of DC block; (c) the structure of each dense unit; (d) the improved transition block.

maps of the first layer is k_0. As shown in Fig. 14 (b) and (c), the DC block in DC-TLMDDNet has eight densely connected units, and each unit consists of a 1×1 convolutional layer and a 3×3 convolutional layer, where each type of the convolutional layer followed by a batch normalization layer and leaky ReLU, and the growth rate k is set to 32. Therefore, the nonlinear mapping function of each dense unit can be represented as Conv(1×1)-BN-leaky ReLU-Conv(3×3)-BN-leaky ReLU. Furthermore, an improved transition block is used before each DC block, which performs the maximum pooling and convolution with a step size of 2 on the output feature maps of the previous block, respectively. It concatenates the two outputs as the input of the next block. As shown in Fig. 14 (d), the overall parameters of the new proposed network are further reduced by enhancing feature reuse.

B. EXPERIMENTS AND RESULTS

In this experiment, the proposed model is compared with YOLOv3_SR and TLMDDNet to illustrate the validity and feasibility of DC-TLMDDNet. The experimental datasets, environment, and relevant evaluation indicators are the same as discussed in the previous section. The loss value of YOLOv3_SR, TLMDDNet, and DC-TLMDDNet during training is shown in Fig. 15. The AP of each class, mAP, average detection rate, parameters, and FLOPs of the models are shown in Table 10.

The experimental results show that DC-TLMDDNet converges quickly and gets a desirable convergence result, which is almost the same as TLMDDNet. This indicates that the training performance of DC-TLMDDNet can meet the requirements. As for detection performance, compared with TLMDDNet, the detection speed of DC-TLMDDNet is increased by 7.32fps, and the detection accuracy is improved by about 0.14%. In addition, the number of DC-TLMDDNet parameters and the model complexity of DC-TLMDDNet is 25.5M and 13.0 BFLOPs, respectively, both of which are less than half of those of TLMDDNet. All these results demonstrate the effectiveness of DC-TLMDDNet, which reduces the number of parameters of the TLMDDNet while maintaining an ideal detection performance.

TABLE 10. The performance of the three methods tested on Dataset 2.

Indicators	Items	YOLOv3_SR	TLMDDNet	DC-TLMDDNet
Accuracy	complete_GJ_fastener	0.9922	0.9961	0.9961
	broken_GJ_fastener	1.0000	1.0000	1.0000
	missing_GJ_fastener	1.0000	1.0000	1.0000
	complete_CK1_fastener	0.9975	1.0000	1.0000
	missing_CK1_fastener	0.9926	1.0000	1.0000
	broken_CK2_fastener	0.9963	0.9963	0.9963
	normal_rail	0.9907	0.9859	0.9953
	rail_corrugation	0.9417	0.9742	0.9915
	rail_spalling	0.9861	1.0000	0.9861
	mAP	0.9886	0.9947	**0.9961**
Others	detection rate/fps	34.13	33.00	**40.32**
	parameters	∼53.8M	∼56.3M	**∼25.5M**
	BFLOPs	27.5	28.4	**13.0**

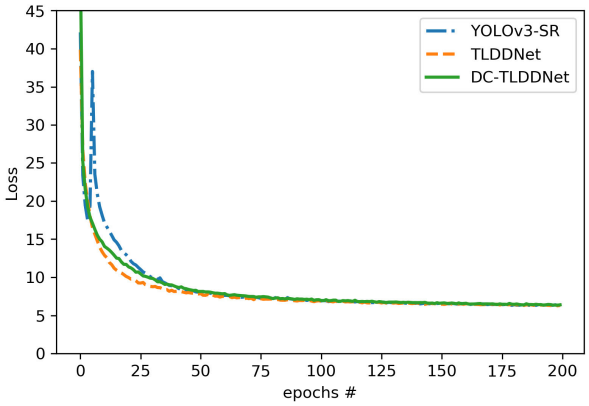

FIGURE 15. Partial loss curves for the three detection models.

VII. CONCLUSION

In this paper, the railway track line multi-target (mainly focused on the track and different types of fasteners) defect identification issues are concerned the first time. First of all, methods based on image pre-processing, feature extraction, and classifier are investigated. A sound positioning method for track and fastener is proposed. Based on the positioning results, a BOVW model combined with spatial pyramid decomposition is applied to the classification. Secondly, to simplify the detection procedure and improve the detection accuracy, the TLMDDNet based on YOLOv3 is proposed for the considered issues. This model simultaneously enables the positioning and classification of track and fasteners. In TLMDDNet, the final detection is performed only on the scale of 13 × 13 feature map size, and a cross-block feature concatenation strategy is utilized for learning object features comprehensively. The quantitative experiments on Dataset 1 and Dataset 2 show that TLMDDNet achieves markedly better performance than the other methods. Finally, the TLMDDNet is improved by incorporating the DenseNet method for reducing the parameter number of TLMDDNet. The DC-TLMDDNet proposed in this paper uses dense blocks and improved transition blocks to optimize the backbone network by enhancing feature propagation and improving network performance. The experimental results demonstrate that the DC-TLMDDNet based method can reduce the complexity of TLMDDNet and further improve the detection speed of the railway track line multi-target defect, which has a better performance compared to the methods available in the literature. In addition, based on the experimental results, it can be concluded that the TLMDDNet and DC-TLMDDNet can also be used to detect the defects of track and different types of fasteners simultaneously and comprehensively.

In the future, the defect detection of other components of the railway track line (e.g., sleeper and ballast) will be carefully investigated and integrated to develop a comprehensive railway track line detection system. The methods proposed in this paper will be further validated in a field test.

REFERENCES

[1] Y. Li, H. Trinh, N. Haas, C. Otto, and S. Pankanti, "Rail component detection, optimization, and assessment for automatic rail track inspection," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 2, pp. 760–770, Apr. 2014.
[2] L. Zuwen, "Overall comments on track technology of high-speed railway," *J. Railway Eng. Soc.*, vol. 1, pp. 41–54, Jan. 2007.
[3] Y. Xia, F. Xie, and Z. Jiang, "Broken railway fastener detection based on AdaBoost algorithm," in *Proc. Int. Conf. Optoelectron. Image Process.*, vol. 1, Nov. 2010, pp. 313–316.
[4] J. LIU, Y. XIONG, B. LI, and L. LI, "Research on automatic inspection algorithm for railway fastener defects based on computer vision," *J. China Railway Soc.*, vol. 38, no. 8, p. 11, 2016.
[5] Y. Li, C. Otto, N. Haas, Y. Fujiki, and S. Pankanti, "Component-based track inspection using machine-vision technology," in *Proc. 1st ACM Int. Conf. Multimedia Retr. (ICMR)*, 2011, p. 60.
[6] J. Yang, W. Tao, M. Liu, Y. Zhang, H. Zhang, and H. Zhao, "An efficient direction field-based method for the detection of fasteners on high-speed railways," *Sensors*, vol. 11, no. 8, pp. 7364–7381, 2011.
[7] H. Feng, Z. Jiang, F. Xie, P. Yang, J. Shi, and L. Chen, "Automatic fastener classification and defect detection in vision-based railway inspection systems," *IEEE Trans. Instrum. Meas.*, vol. 63, no. 4, pp. 877–888, Apr. 2014.
[8] L. Yong-Bo, L. Bai-Lin, and Y. Xiong, "Railway fastener state detection based on hog feature," *Transducer Microsyst. Technol.*, vol. 1, pp. 110–113, Jun. 2013.
[9] J. Liu, B. Li, Y. Xiong, B. He, and L. Li, "Integrating the symmetry image and improved sparse representation for railway fastener classification and defect recognition," *Math. Problems Eng.*, vol. 2015, Nov. 2015, Art. no. 462528.
[10] E. Resendiz, J. Hart, and N. Ahuja, "Automated visual inspection of railroad tracks," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 2, pp. 751–760, May 2013.
[11] H. Fan, P. C. Cosman, Y. Hou, and B. Li, "High-speed railway fastener detection based on a line local binary pattern," *IEEE Signal Process. Lett.*, vol. 25, no. 6, pp. 788–792, Jun. 2018.
[12] J. Liu, Y. Huang, Q. Zou, M. Tian, S. Wang, X. Zhao, P. Dai, and S. Ren, "Learning visual similarity for inspecting defective railway fasteners," *IEEE Sensors J.*, vol. 19, no. 16, pp. 6844–6857, Aug. 2019.
[13] F. Marino, A. Distante, P. L. Mazzeo, and E. Stella, "A real-time visual inspection system for railway maintenance: Automatic hexagonal-headed bolts detection," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 37, no. 3, pp. 418–428, May 2007.

[14] P. De Ruvo, A. Distante, E. Stella, and F. Marino, "A GPU-based vision system for real time detection of fastening elements in railway inspection," in *Proc. 16th IEEE Int. Conf. Image Process. (ICIP)*, Nov. 2009, pp. 2333–2336.

[15] X. Gibert, V. M. Patel, and R. Chellappa, "Robust fastener detection for autonomous visual railway track inspection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, Jan. 2015, pp. 694–701.

[16] X. Gibert, V. M. Patel, and R. Chellappa, "Deep multitask learning for railway track inspection," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 1, pp. 153–164, Jan. 2017.

[17] X. Wei, Z. Yang, Y. Liu, D. Wei, L. Jia, and Y. Li, "Railway track fastener defect detection based on image processing and deep learning techniques: A comparative study," *Eng. Appl. Artif. Intell.*, vol. 80, pp. 66–81, Apr. 2019.

[18] C. Mandriota, M. Nitti, N. Ancona, E. Stella, and A. Distante, "Filter-based feature selection for rail defect detection," *Mach. Vis. Appl.*, vol. 15, no. 4, pp. 179–185, Oct. 2004.

[19] F. Marino and E. Stella, "Visyr: A vision system for real-time infrastructure inspection," in *Vision Systems: Applications*. London, U.K.: IntechOpen, 2007.

[20] D. Wei, X. Wei, Y. Liu, L. Jia, and W. Zhang, "The identification and assessment of rail corrugation based on computer vision," *Appl. Sci.*, vol. 9, no. 18, p. 3913, 2019.

[21] Q. Li, H. Zhang, and S. Ren, "Detection method for rail corrugation based on rail image feature in frequency domain," *China Railway Sci.*, vol. 37, no. 1, pp. 24–30, 2016.

[22] Q. Li, Z. Shi, H. Zhang, Y. Tan, S. Ren, P. Dai, and W. Li, "A cyber-enabled visual inspection system for rail corrugation," *Future Gener. Comput. Syst.*, vol. 79, pp. 374–382, Feb. 2018.

[23] Q. Li and S. Ren, "A visual detection system for rail surface defects," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 42, no. 6, pp. 1531–1542, Nov. 2012.

[24] Q. Li and S. Ren, "A real-time visual inspection system for discrete surface defects of rail heads," *IEEE Trans. Instrum. Meas.*, vol. 61, no. 8, pp. 2189–2199, Feb. 2012.

[25] Y. Min, B. Xiao, J. Dang, B. Yue, and T. Cheng, "Real time detection system for rail surface defects based on machine vision," *EURASIP J. Image Video Process.*, vol. 2018, no. 1, p. 3, Dec. 2018.

[26] Q. Li, Y. Huang, Z. Liang, and S. Luo, "Thresholding based on maximum weighted object correlation for rail defect detection," *IEICE Trans. Inf. Syst.*, vol. E95.D, no. 7, pp. 1819–1822, 2012.

[27] Y. Xiao-cui, W. Lu-shen, and C. Hua-wei, "Rail image segmentation based on otsu threshold method," *Opt. Precis. Eng.*, vol. 24, no. 7, pp. 1772–1781, 2016.

[28] Z. He, Y. Wang, F. Yin, and J. Liu, "Surface defect detection for high-speed rails using an inverse P-M diffusion model," *Sensor Rev.*, vol. 36, no. 1, pp. 86–97, Jan. 2016.

[29] J. Gan, Q. Li, J. Wang, and H. Yu, "A hierarchical extractor-based visual rail surface inspection system," *IEEE Sensors J.*, vol. 17, no. 23, pp. 7935–7944, Dec. 2017.

[30] H. Yu, Q. Li, Y. Tan, J. Gan, J. Wang, Y.-A. Geng, and L. Jia, "A coarse-to-fine model for rail surface defect detection," *IEEE Trans. Instrum. Meas.*, vol. 68, no. 3, pp. 656–666, Mar. 2016.

[31] J. Gan, J. Wang, H. Yu, Q. Li, and Z. Shi, "Online rail surface inspection utilizing spatial consistency and continuity," *IEEE Trans. Syst., Man, Cybern. Syst.*, early access, May 4, 2018, doi: 10.1109/TSMC.2018.2827937.

[32] S. Faghih-Roohi, S. Hajizadeh, A. Nunez, R. Babuska, and B. De Schutter, "Deep convolutional neural networks for detection of rail surface defects," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2016, pp. 2584–2589.

[33] A. James, W. Jie, Y. Xulei, Y. Chenghao, N. B. Ngan, L. Yuxin, S. Yi, V. Chandrasekhar, and Z. Zeng, "TrackNet—A deep learning based fault detection for railway track inspection," in *Proc. Int. Conf. Intell. Rail Transp. (ICIRT)*, Dec. 2018, pp. 1–5.

[34] L. Shang, Q. Yang, J. Wang, S. Li, and W. Lei, "Detection of rail surface defects based on CNN image recognition and classification," in *Proc. 20th Int. Conf. Adv. Commun. Technol. (ICACT)*, Feb. 2018, pp. 45–51.

[35] Z. Liang, H. Zhang, L. Liu, Z. He, and K. Zheng, "Defect detection of rail surface with deep convolutional neural networks," in *Proc. 13th World Congr. Intell. Control Autom. (WCICA)*, Jul. 2018, pp. 1317–1322.

[36] S. Yanan, Z. Hui, L. Li, and Z. Hang, "Rail surface defect detection method based on YOLOv3 deep learning networks," in *Proc. Chin. Autom. Congr. (CAC)*, Nov. 2018, pp. 1563–1568.

[37] C. Liu, J. Yuen, and A. Torralba, "Sift flow: Dense correspondence across scenes and its applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 3, no. 5, pp. 978–994, May 2010.

[38] H. Kato and T. Harada, "Image reconstruction from Bag-of-Visual-Words," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 955–962.

[39] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.(CVPR)*, vol. 2, Jun. 2006, pp. 2169–2178.

[40] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*. [Online]. Available: http://arxiv.org/abs/1804.02767

[41] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.

[42] R. Hodgson, D. Bailey, M. Naylor, A. Ng, and S. McNeill, "Properties, implementations and applications of rank filters," *Image Vis. Comput.*, vol. 3, no. 1, pp. 3–14, Feb. 1985.

[43] H. D. Cheng and X. J. Shi, "A simple and effective histogram equalization approach to image enhancement," *Digit. Signal Process.*, vol. 14, no. 2, pp. 158–170, Mar. 2004.

[44] N. Ostu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-9, no. 1, pp. 62–66, Jan. 1979.

[45] J. M. Niya and A. Aghagolzadeh, "Edge detection using directional wavelet transform," in *Proc. 12th IEEE Medit. Electrotech. Conf.*, vol. 1, Jun. 2004, pp. 281–284.

[46] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. 7th IEEE Int. Conf. Comput. Vis.*, vol. 99, Sep. 1999, pp. 1150–1157.

[47] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," *Commun. ACM*, vol. 18, no. 11, pp. 613–620, Nov. 1975.

[48] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, Jun. 2006.

[49] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[50] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7263–7271.

[51] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[52] L. Tzutalin. *LabelImg Overviews, GitHub*. Accessed: Apr. 2, 2020. [Online]. Available: https://github.com/tzutalin/labelImg

[53] D. Kirk, "NVIDIA cuda software and gpu parallel computing architecture," in *Proc. 6th Int. Symp. Memory Manage. (ISMM)*, vol. 7, 2007, pp. 103–104.

[54] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

[55] H. Chen, Y. Wang, Y. Shi, K. Yan, M. Geng, Y. Tian, and T. Xiang, "Deep transfer learning for person re-identification," in *Proc. IEEE 4th Int. Conf. Multimedia Big Data (BigMM)*, Sep. 2018, pp. 1–5.

[56] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

XIUKUN WEI received the Ph.D. degree from Johannes Kepler University, Linz, Austria, in 2002. From 2006 to 2009, he was a Postdoctoral Researcher with the Delft Center for System and Control, Delft University of Technology, Delft, The Netherlands. From 2002 to 2006, he was a Research Assistant with the Institute of Design and Control of Mechatronical Systems, Johannes Kepler University. He is currently a Professor with the State Key Lab of Rail Traffic Control and Safety, Beijing Jiaotong University, China. His research interests include fault diagnosis and its applications, intelligent transportation systems, and condition monitoring and its applications in a variety of fields, such as rail traffic control, safety, and transportation.

DEHUA WEI received the master's degree from Fuzhou University, Fuzhou, China, in 2017. He is currently pursuing the Ph.D. degree with the School of Traffic and Transportation, Beijing Jiaotong University, Beijing, China. His research interests include computer vision, image processing, defect detection, and deep learning method and application.

DA SUO received the bachelor's degree from Beijing Jiaotong University, China, in 2018. His research interests include image processing, fault diagnosis, and deep learning.

LIMIN JIA is currently a Professor with the State Key Lab of Rail Traffic Control and Safety, Beijing Jiaotong University, China. His research interests include intelligent control, system safety, and fault diagnosis and their applications in a variety of fields, such as rail traffic control and safety and transportation.

YUJIE LI received the Ph.D. degree in civil engineering from Beijing Jiaotong University, China, in 2003. He is currently a Senior Engineer with Beijing Mass Transit railway Operation Corporation Ltd. His work mainly focus on condition monitoring for railway and other industrial applications.