MRSDI-CNN: Multi-Model Rail Surface Defect Inspection System Based on Convolutional Neural Networks

Hui Zhang, *Member, IEEE*, Yanan Song, Yurong Chen, *Member, IEEE*, Hang Zhong, Li Liu, *Member, IEEE*, Yaonan Wang, Thangarajah Akilan, *Member, IEEE*, and Q. M. Jonathan Wu, *Senior Member, IEEE*

***Abstract*— Defects on rail surfaces, which have become critical problems, need to be detected and removed as quickly as possible to ensure the fast, safe, and stable operation of trains. At present, although many solutions have been proposed to address these problems, the comprehensiveness, rapidity, and accuracy of defect detection remain unsatisfactory. This study aims to resolve these existing problems and accordingly proposes a multi-model rail surface defect detection system based on convolutional neural networks (MRSDI-CNN) from the standpoint of studying the squat on the rail surface. The convolutional neural networks utilized include the improved Single Shot MultiBox Detector (SSD) and You Only Look Once version 3(YOLOv3)—two types of one-stage networks. We expounded and analyzed the performance of the convolutional neural networks as well as their applicability to rail surface defect detection. We used a diverse range of rail defect sizes to improve the detection performance of the two deep learning networks, following which they could identify three types of squats in parallel with improved accuracy and without reduction of the detection speed. The experimental results confirm the effectiveness and superiority of the proposed method over those of previous studies.**

***Index Terms*— Rail surface defect, convolutional neural network, multi-model, improved SSD network, improved YOLOv3 network, one-stage.**

I. Introduction

With the prosperity of the Chinese railway industry, the mileage, speed, and density of operations have continued to increase and requirements for rail detection have further increased. When a train travels on a rail, it produces friction, rolling contact, and elastic deformation on the inner surface of the rail. Over time, rail surface defects will occur in the form of rail surface wear, corrugation, depressions, shelling, local batter, and other phenomena, which seriously affect or even endanger the safety of passengers [1]. Therefore, to improve railway safety, train run stability, and passenger safety and comfort, it is extremely important to detect rail surface defects.

A. Physical Methods for Detecting Rail Surface Defects

The traditional method of detecting defects on a rail surface is manual inspection [2]. However, this method has many problems including its low efficiency, high time and labor costs, and subjectivity and it does not produce auditable visual records. At present, there are many methods for detecting rail surface defects, such as radiation detection, ultrasonic inspection, ultrasonic guided waves, electromagnetic ultrasonic, laser ultrasonic, and eddy current testing. We classify these methods as physical detection technology. Physical detection is the use of sound, light, electromagnetic, and other physical field action on the rail produced by the physical phenomenon for testing the quality of rail. Physical detection technology serves physics as the basis and hardware as the core, because the quality of the hardware directly affects the detection effect. Another parallel term is machine vision detection. In the machine vision system, the image processing unit is the important component, which focuses more on model and algorithm capabilities. It is significant to improving the accuracy and real-time performance of vision algorithms for the overall competitiveness of machine vision systems.

Alahakoon *et al.* [3] gave an in-depth review of the latest technology related to rail flaw detection and summarized various flaw detection technologies based on the acoustic, photoelectric and magnetic field in the reported scientific literature. This work provided a good platform for future researchers to start their work in the field of orbital detection. Sensors play an unshakable role in these detection technologies, which is the first line of defense to ensure the performance of defect detection technology. Hodge *et al.* [4] reviewed sensor underpower solutions for monitoring systems deployed in remote

Manuscript received November 29, 2019; revised February 19, 2021; accepted June 30, 2021. This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFB1308200, in part by the National Natural Science Foundation of China under Grant 61971071 and Grant 6202780012, in part by the Changsha Science and Technology Project under Grant kq1907087, in part by the Special Funds for the Construction of Innovative Provinces in Hunan Province under Grant 2020SK3007, and in part by the Postdoctoral Innovative Talent Support Program under Grant BX20200122. The Associate Editor for this article was D. Chen. *(Corresponding authors: Hui Zhang; Yurong Chen.)*

Hui Zhang, Yurong Chen, Hang Zhong, Li Liu, and Yaonan Wang are with the National Engineering Laboratory of Robot Visual Perception and Control Technology, School of Robotics, Hunan University, Changsha 410082, China (e-mail: zhanghuihby@126.com; chenyurong1998@outlook.com).

Yanan Song is with the College of Electrical and Information Engineering, Changsha University of Science and Technology, Changsha 410012, China.

Thangarajah Akilan is with the Department of Software Engineering, Lakehead University, Thunder Bay, ON P7B 5E1, Canada.

Q. M. Jonathan Wu is with the Department of Electrical and Computer Engineering, University of Windsor, Windsor, ON N9B 3P4, Canada.

Digital Object Identifier 10.1109/TITS.2021.3101053

and hazardous environments and detailed the requirements of wireless sensor networks to avoid transmission errors, data loss, or data corruption. Trends in fractures and structural defects can also be analyzed and identified in advance based on state monitoring data generated by various wireless sensors.

Zhang *et al.* [5] proposed a method for detecting track surface defects in combination with machine vision and three-dimensional linear laser detection technology. Although this method reduces the missed detection rate to a certain extent, its detection accuracy, and robustness are poor. Xiong *et al.* [6] proposed a new type of three-dimensional (3D) laser profiling system to image rails using a laser scanner with k-means and decision trees to detect and classify track surface defects; the error recognition rate is improved but the detection accuracy and classification accuracy is low. Somalraju *et al.* [7] designed a cost effective yet robust solution to the railway crack detection system, which includes a GPS module, GSM Modem, and LED-LDR based crack detector assembly. Qian *et al.* [8] proposed a rail surface defect detection system based on a linear array charge coupled device (CCD). The system uses a linear CCD as an image sensor through an A/D converter, timing generator, data storage, and other control logic for image acquisition and digital signal processing to achieve rail surface defect detection. This method can only detect rail surface cracks. Although it improves the detection accuracy, it is prone to false detection and missed detection.

Chen *et al.* [9] designed an ultrasonic defect detection system consisting of a conventional probe, phased array probe, and a field programmable gate array module that improved detection speed and accuracy. Based on ultrasonic sensors, the detection system improves the detection speed and accuracy for small cracks but it cannot detect medium and large cracks. Sun *et al.* [10] established a real-time photoacoustic imaging system for track nondestructive testing. By reconstructing the image via photoacoustic signals, it is possible to effectively identify the appearance of the track defects, extension trend, and depth of the damage information; however, the recognition accuracy is low. Scalea *et al.* [11], [12] used traditional electro-ultrasound to identify and detect high-speed defects of tracks using an unsupervised learning algorithm. This method can only detect small transverse cracks in the rail head. Rose *et al.* [13] proposed using guided waves for rail defect detection. However, the applicability of guided waves was only discussed and experiments on and analysis of rail defect detection were not involved. Hesse [14], [15] used a contact wheel probe to excite low-frequency ultrasonic surface waves to detect straight and oblique cracks on rail treads. This method can reliably detect defects, but its robustness is poor. Sebko *et al.* [16] detected rail defects using the specular shading method and echo method of volume shear and surface waves; the rail defect recognition rate was improved. Dixon *et al.* [17] developed a rail flaw detection technique based on ultrasonic detection technology that detects surface cracks on the rail surface by exciting surface waves via electromagnetic sound transducers; however, it is easy to miss detection. Nielasen *et al.* [18] proposed an automatic laser detection system using a non-contact ultrasonic method to detect defects on the orbital surface and horizontal and vertical defects on the railhead; although the detection speed was improved, the detection accuracy is low.

Wei *et al.* [19] proposed a method for detecting track defects based on vibration acceleration signals that can identify the type and damage degree of track defects. The method can recognize defects, but the recognition accuracy is low. Chen [20] used the finite element method to study the velocity effect of railway magnetic flux leakage detection and the identification of orbital oblique cracks. The test results show that the method still has the phenomenon of missed detection. Wilson and Tian [21] used a magnetic flux leakage test to detect rail tread cracks and the rail defect recognition rate was improved. Based on the magnetic flux leakage (MFL) guideway defect detection technology, Gao *et al.* [22] proposed a track leakage magnetic detection system based on AT89C51. The system is stable and has the advantages of high detection accuracy and high speed, but it is susceptible to the experimental environment. Antipov and Markov [23] proposed a new magnetization system. The system detects the depth of the rail defect using a magnetic mechanical method. The system can only detect the depth of rail defects and cannot show the defect location information. Although these methods can achieve the purpose of detection, because of various external interferences, their signals are difficult to process and may lead to detection of blind areas [24].

B. Machine Vision Method for Detecting Rail Surface Defects

Currently, machine vision is typically used to detect rail surface defects. Santur *et al.* [25] analyzed four phenomena including rail cracks, scratches, wave wear, and railhead cracking and proposed to use a 3D laser camera and deep learning method to check the railway surface defects, improving the detection speed and accuracy. Deutschl *et al.* [26] proposed a new vision-based orbital surface defect detection technique. Defects on the track surface were automatically detected by a color line scan camera and spectral image difference method; the detection speed was improved, but the detection accuracy was easily affected by the type of rail defects. Nitti *et al.* [27] used the DALSA series scanner model SP-12 to acquire a guide rail image and used a neural network to perform gradient approximation to extract the rolling plane surface and detect track defects, the accuracy of the track defect identification is very low. Mandriota [28] performed an experimental comparison between three filtering methods based on the texture characteristics of the track surface to detect surface ripple defects. Li *et al.* [24] proposed a real-time visual inspection system based on discrete surface defects. The sub-image of the track is cut by the trajectory extraction algorithm, the contrast in the track image is enhanced using the local normalization method, and the defect is detected using the defect contour based on the projection contour, which is robust in terms of noise and run speed but the detection accuracy is not good. Min *et al.* [29] proposed a real-time detection method for track surface defects based on machine vision and designed an image acquisition device equipped with an LED auxiliary light source and blackout box. Based on the morphological process, the surface defects of the track were optimized

and tracked. The direction chain code obtained the defect characteristics. Although this method improves the detection speed and meets real-time requirements, the detection accuracy is low. Rezaeitabar *et al.* [30] analyzed a method based on pixel similarity and histogram similarity for images acquired using a high-speed 3-D laser rangefinder and proposed a fusion method. This method employs lesser computation and can be used for real-time monitoring. Tastimur *et al.* [31] used techniques based on contact image processing to diagnose railway components and defects, including pre-processing, morphological feature extract1ion, fault, and defect detection steps.

Tim *et al.* [32] used long-term and short-term memory and cyclic neural networks to diagnose faults in railway track circuits. The detection performance is better than that only using a single network. The Washington Research Center Hoang Trinh *et al.* [33] proposed a real-time automatic vision orbit detection system for important rail components such as tie bars, tie plates, and anchors. The method has high detection accuracy and high efficiency, but can only detect rail components, such as anchors, and has poor expandability. Arun *et al.* [34] explores the computer vision methods based on drone imagery to detect various abnormal conditions in railway tracks. Aiming at the low brightness and contrast caused by the shaking of drones and the high reflection characteristics of railways, Wu *et al.* [35] proposed a local Weber-like contrast image enhancement and gray-scale stretching maximum entropy segmentation method to improve defect detection performance. The recall rate of the LWLC-GSME model for rail defects reached 93%. References [36] explores the usage of drones in railways for early warning, situation assessment and decision support applications. In addition, [37] introduces various challenges of dealing with drone data contains.

The Dutch Delft University of Technology [38] proposed a method for analyzing images using deep convolutional neural networks to detect surface defects, improving the robustness in terms of background noise. The Beijing Jiaotong University research team [39] used the coarse extractor after image preprocessing to find the approximate location of defects in the track surface image. They used the fine extractor to determine whether the point of the obtained outlier is a real defect or another noise point and discarded the noise point to obtain real results. This method has a high recognition rate, low missed detection rate, but poor detection accuracy. Zhang *et al.* [40] of the Harbin Institute of Technology proposed an improved orbital health monitoring method based on convolutional neural network and acoustic emission event probability analysis. The convolution neural network (CNN) deep learning method was used to classify the defects and improved the classification accuracy. Shang *et al.* [41] proposed a method for detecting surface defects based on CNN image recognition and classification. This method has good robustness and high detection accuracy, but its detection speed is slow and it cannot meet the real-time requirements.

In this paper, we focus on the most crucial defect, the squat defect, which is the most representative and detrimental defect on the surface of railway and has a sufficient data to support us to do analysis. Other defects, such as scratches, are trivial and there is not an ample dataset. Squat defects can be classified based on size; the damage degree of the defects varies, which in turn results in different effects on smooth train operation. The current deep learning method is divided into one-stage and two-stage approaches. A one-stage algorithm such as YOLOv1 [42], YOLOv2 [43], YOLOv3 [44], and SSD [45] is fast and can meet real-time detection requirements but the detection accuracy is not satisfactory; a two-stage algorithm such as R-CNN [46], Fast-RCNN [47], and Faster-RCNN [48] has a low false positive rate and missed detection rate but cannot satisfy the real-time detection scene. At the same time, for small rail defects, most algorithms cannot achieve better detection results. In this study, considering the speed and accuracy index, a fast, high-precision, and real-time method is proposed, based on improved SSD and YOLOv3 deep learning networks. SSD uses Visual Geometry Group (VGG16) as the backbone to extract multi-scale feature map. The model produces predictions of different scales from feature maps of different scales to achieve high detection accuracy. Small feature maps with deep information are responsible for detecting large targets, while large feature maps with shallow information are responsible for detecting small targets. The architecture can be visualized in the Fig.3 (top) line. Using darknet19 [43] and darknet53 [44], which are the backbone network which proposed in object detection network of YOLOv2 and YOLOv3, and it is mainly used to extract feature representations, which are feed to detector part and then can be predicted accurate result of bounding box regression and classification, as backbone for feature extraction, YOLOv3 use multi-scale which include 3 different scales for box prediction. Due to this strategy, it improves meaningful semantic information learning and benefits to accuracy of small target detection. The architecture can be visualized in the Fig.3 (bottom) line. In this study, considering the speed and accuracy index, we propose a fast, high-precision, and real-time method based on improved Single Shot MultiBox Detector (SSD) [43] and You Only Look Once version 3(YOLOv3) [44] deep learning networks.

The main work completed is as follows:

1. We reasonably combine the advantages of the SSD algorithm and the YOLOv3 algorithm. The first one is excellent for medium and large squat detection and the YOLOv3 shows good performance for small squats detection.

2. Furthermore, the SSD network is improved for detecting rail small squats without losing the detection capability of medium and large squats, by modifying the network output layer and the corresponding configuration parameters.

3. We propose a new target detection network based on the YOLOv3 network. In particular, based on the network model of darknet19 and darknet53, a new feature extraction network (basic network) is proposed; secondly, following the basic network, a multi-scale network is added according to the YOLOv3 network structure to improve small target detection accuracy; finally, in multi-scale detection, because the anchors of different data sets are different, we used the k-means clustering algorithm to regenerate the anchor box that conforms to the rail data set and improved the intersection over-union (IOU).

The structure of this paper is as follows. Section II provides an overview of the rail surface defect detection system. Section III details two improved convolutional neural network structures. Section IV focuses on the test process of the rail defect dataset used and analyzes the superiority of the proposed method through several experiments and comparisons. Section V draws conclusions and presents future research priorities.

II. System Overview

1) Overview of Rail Surface Defects: A rail surface defect refers to breaking, cracking, and other conditions that affect and restrict the performance of rails in use. There are many types of rail surface defects: indentations, deformations, fissuring, head checking, breakout, etc. In the context of rail surface defect detection, research on defect detection methods is mainly based on physical and machine vision detection methods.

Among them, the detection techniques using physical methods mainly include ray, ultrasonic, ultrasonic guided wave, electromagnetic ultrasonic, laser ultrasonic, eddy current and pulse eddy current, magnetic flux leakage, and magnetic particle detection. The machine vision method mainly describes the traditional image processing method and the deep learning method. At present, machine vision detection results are generally better than those of detection techniques using physical method. Therefore, rail defect detection methods are gradually changing from physical hardware to software, that is, from a physical method to a machine vision method. On the one hand, this weakens our rigid requirements for a physical experimental environment, but on the other hand, it can increase an algorithm's robustness and anti-interference ability. Although the detection performance of machine vision is better, particularly deep learning algorithm detection, the performance of an algorithm is different for different environments.

Different algorithms have different detection performances because of different defect sizes and rail defects of different sizes often affect the timely replacement and repair of rails. For minor defects, the railway department can detect and implement certain deployment plans in time to prevent further expansion; for medium defects, the rail should be marked with a key, and real-time detection of the impact of the defect on railway operation is needed; for large squats, the railway department discovers and makes corresponding replacements in time.

To timely monitor and test the influence of the size of the rail defects on smooth train operation, this study classified squats into three categories according to the degree of defect damage, namely small, medium, and large squats. Small squats can be observed with the naked eye and the defect radius is approximately 1.5 mm; a medium squat defect radius is between 1.5 and 3mm; a rail defect radius greater than 3 mm is classified as a large squat. Figure 1 shows samples of the three classes of rail squats.

2) Rail Surface Defect Detection System: With the development of railway transportation lines, the current rail inspection systems have been unable to meet the technical requirements of railway rail inspection. For the comprehensive requirements

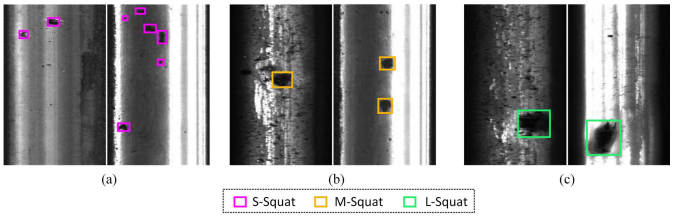

Fig. 1. Examples of rail defect images. (a) Small squats(the radius is approximately 1.5 mm). (b) Medium squats (radius is between 1.5 and 3mm). (c) Large squats (the radius is greater than 3 mm).

of rails, i.e. full mileage, high precision, and cost saving, rail inspection technology must solve the following problems:

(1) Improve the detection speed. A low detection speed will lead to difficulties in railway dispatching and lengthen detection times. If the detection speed is higher, it is easier to grasp the rail state in real time, improving railway transportation safety.

(2) Using multi-modal technology to improve detection accuracy. Non-destructive testing technology has its advantages and limitations during the process of identifying squats. It is difficult to achieve high-speed, accurate, and comprehensive detection using a single detection method. In practical application, two or more detection technologies are combined to complement each other's advantages and improve the squat detection rate, recognition rate, and other indicators.

(3) Apply a wireless sensor network system and vision module (audible and visual alarm) to a rail defect detection system to achieve all-around inspection of rail status. A wireless sensor network system integrates various sensors, such as those for temperature and gravity, that are distributed throughout the rail lines; real-time monitoring of the status of rail locations; visualizing hazardous rail defect locations; and quantitative evaluation of rail stress distribution and life span to ensure the safe operation of rails.

To better solve the aforementioned problems, we propose a defect detection system as shown in Fig. 2. This system includes two main steps: image acquisition and image processing. The specific process of image acquisition includes installing a camera above the rail for acquiring images and adding light source devices on both sides of the camera for 1) enhancing light and eliminating interference from other light sources and 2) avoiding blurring of the acquired image because of illumination problems. This satisfies fast image acquisition when the train is running. The specific process of image processing includes the server setting the control system on the train. When the picture is acquired, it is sent to the server through the control system, and the deep learning algorithm in the server can quickly identify it. The two aspects coordinate and work together to form a real-time detection system for rail surface defects. To achieve better detection requirements, we combined two one-stage deep learning algorithms: the improved SSD and improved YOLOv3 algorithms, which significantly improved the detection accuracy.

3) Rail Surface Defect Detection Overall Framework: Figure 3 shows the rail surface defect inspection process. Aiming at addressing the problem of squat detection in rails, we propose a multi-model rail surface defect inspection system

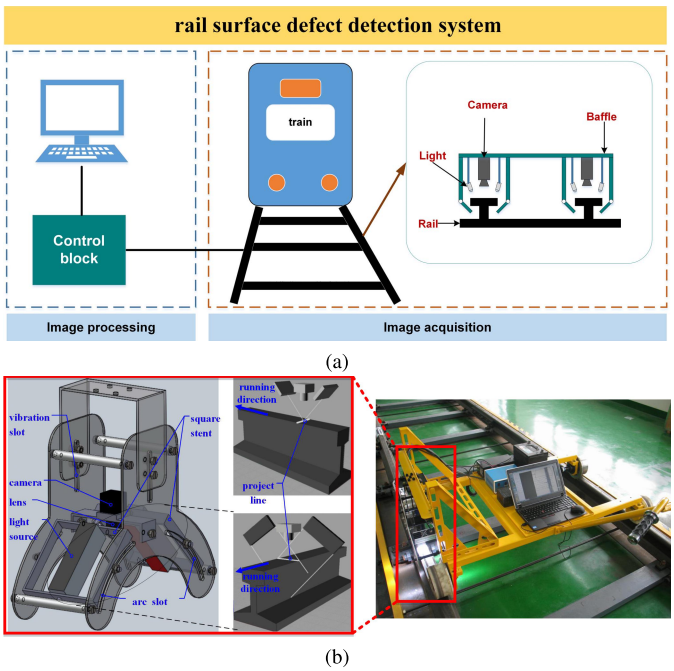

Fig. 2. Rail surface defect detection system. (a) System Components. (b) 3-D illustration and physical map of imaging equipment.

based on convolutional neural networks (MRSDI-CNN). The system primarily aims to solve the problems of slow speed and low precision rail defect detection. It consists of three aspects as follows: When inputting the rail defect image,

(I) In upward route of network, we use VGG16 [48] as backbone network for the improved SSD to extract task-related feature to object detection or classification framework. In order to obtain more useful feature map for detection task before calculating the confidence of rail squats and location information of bounding box, we add more convolution layers on the basis of VGG16. In this paper, we use the subnetwork1 (the improved SSD network) as mainly part to detect medium and large squats, and we observed that the improved model has achieved good results when detecting the small squats.

(II) in the downward path, the improved YOLOv3 first extracts features through a series of convolution and pooling operations in the feature extraction network and then detects rail squats on three scales via convolution, up-sampling, and feature fusion operations in a multi-scale network. Finally, the defect images with maximum IOU are selected. Then, the confidence is calculated, and the location of the bounding box is predicted. We chose to use subnetwork 2 (improved YOLOv3 network), which is mainly used for the detection of small squats, and this network also obtained good results in medium and large squat detection.

(III) Finally, by comparing two output results, the maximum of the two detection areas are the real defects. As shown in Figure 3, the original rail has two defects (small and medium squats). After inputting the original rail image, one defect was detected by subnetwork 1 in (I) and two defects were detected by subnetwork 2 in (II). By comparing the original images, it shows that the number of defects detected in (II) is the largest, and they are real defects; thus, the two networks finally output the images in (II).

During the process of squat detection, utilizing the superiority of the detection performance for the rail defect size, the performance of subnetwork 1 (improved SSD network) is better for detecting medium- and large-scale rail defects. Subnetwork 2 (improved YOLOv3 network) is better in detecting small squats. The two improved algorithms which are both one-stage identify the rail defects in parallel, which effectively ensures the accurate detection of small, medium, and large squats of the rail without slowing the detection speed, ensuring the comprehensiveness of defect detection.

III. Network Model

A. Using SSD to Detect Rail Surface Defects

The SSD algorithm does not generate a proposal process and it directly predicts the coordinates and categories of the bounding box. The SSD is based on the VGG16 network. The 6-th fully connected layer and 7-th fully connected layer in VGG16 basic network are replaced by convolutional layers following which four convolutional layers are added to form a new network. It uses two 3×3 convolution cores to convolute the output of five different convolution layers, one confidence for the output classification and four confidences for each default box (for the rail defect data set with three object categories), and a localization for output regression and each default box generates four coordinate values. The five convolutional layers generate a default box through the prior box layer. Finally, the front three calculation results are combined and passed to the loss layer [45].

The default box is generated by feature maps in different convolutional layers with different aspect_ratios [49]. In the SSD algorithm, the default box is determined by min_sizes, max_sizes, aspect_ratios, and receptive field steps. One core of the SSD algorithm is to use both the lower and upper feature maps for detection as shown in Figure 5. Each center point of the feature map of each layer in the SSD output 6-layer convolution layer will generate 4, 6, 6, 6, 4, and 4 default boxes, respectively. For each default box, three categories (the L-Squat, M-Squat, and S-Squat) and four offsets (x, y, w, and h) need to be predicted. Because the improved output layer adds conv3_3 instead of conv9_2, we modified the original aspect_ratio, min_sizes, max_sizes, receptive field steps, and regular initialization. The default box has different scales in different feature layers and different aspect_ratios in the same feature layer; thus, it can basically cover the detection targets of various shapes and sizes in the input image.

When the default box and ground truth match, then the default box is defined as a positive example. If they do not match, it is defined as a negative example; far more negative samples are produced than positive samples. Then, the negative samples are sorted according to the confidence loss and some negative samples sorted in the front are selected for training such that the ratio of negative samples to positive samples is approximately 3:1, the ratio which follow the setting of YOLOv3 to overcomes imbalance of ratio for positive boxes and negative boxes.

In the original SSD network, the region of influence (ROI) between the prior box and the ground truth reached 0.5 and was placed into the network for training. Large targets may

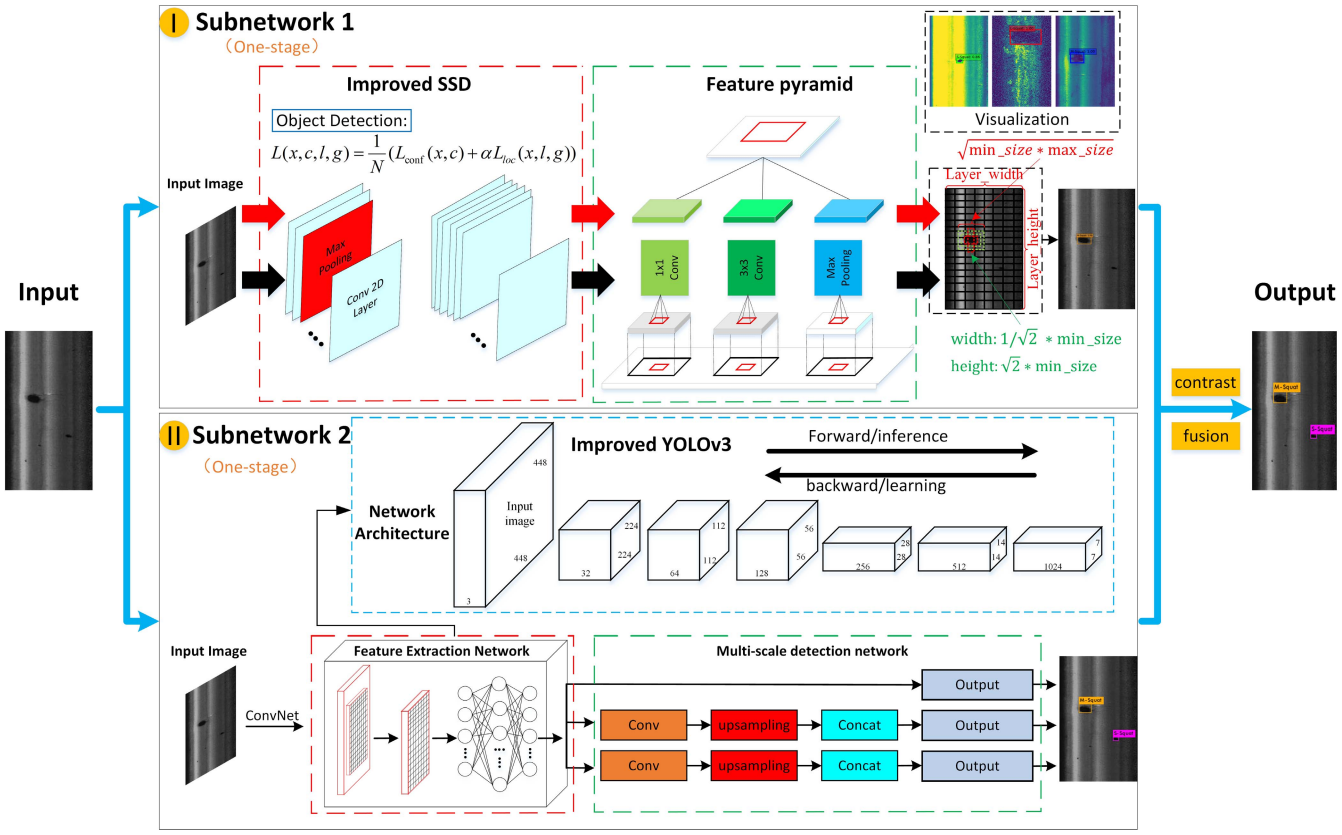

Fig. 3. Overall framework for rail surface defect detection. in (I), the improved SSD uses VGG16 as the basic model to obtain image features; then adds convolution layers on the basis of VGG16 to obtain more feature maps for detection. In (II), the improved YOLOv3 first extracts features through a series of convolution and pooling operations in feature extraction network; then detects rail squats on three scales by convolution, up-sampling and feature fusion operations in multi-scale network. Finally, by comparing two output results, the maximum of the two Detection areas are the real defects.

have a much larger ROI value so there are more boxes and it can be adequately trained. In contrast, the small goal used for training will be much less and it cannot be sufficiently trained. Because these small targets contain quite limited information at the top of the network, the detection of small targets is worse than that of large targets, which shows that increasing the size of the input image is helpful for small targets. In addition, using random cropping and rotation operation, such as contrast adjustment, not only increases the number of images but also enlarges the image to a certain extent and increases the small target network feature information [45].Therefore, to better detect small squats, we made the following modifications to the original SSD network:

(1) Input image size: the SSD original input pixel is 300×300. Because the image size of the rail is 160×250, the defect size is approximately 15×15. To better detect the rail defect, first resize the image input size by 320×500.

(2) Output layer: The original SSD output has six layers: Conv4_3(38×38), Conv7(19×19), Conv6_2(10×10), Conv7_2(5×5), Conv8_2(3×3), and Conv9_2(1×1), where the full connection layers fc6 and fc7 in the VGG-16 network are replaced by convolution layers Conv6 and Conv7. Because the shallow layer acquires the small target feature information, the deep layer acquires the large target feature information. Therefore, we make adjustments for the network output layer. The modified network output layers are Conv3_3(125×80), Conv4_3(63×40), Conv7(16×10), Conv6_2(8×5), Conv7_2(4×3), and Conv8_2(2×2). A comparison of the network framework before and after the modification is shown in Figure 4.

(3) Small target default box: Conv9_2 detects large targets and Conv3_3 is used to detect small targets. When the output layer is modified, the corresponding default frame aspect_ratios, receptive field, min_sizes, max_sizes, and regularization scale are further modified, as previously described.

B. Use of YOLOv3 to Detect Rail Surface Defects

The YOLOv3 model can be divided into a feature extraction layer (Darknet-53) and a processing output layer. The feature extraction layer is a model that combines Darknet-19 and a Resnet-like network. The multi-scale target detection layer is similar to the Feature Pyramid Network (FPN). Each bounding box contains five position coordinates, including box coordinates (x, y, w, and h) and confidence. The YOLOv3 model can be divided into 106 layers (counting from the 0 layer), of which 75 layers are convolution layers, 23 layers are shortcut layers, 4 layers are route layers, 3 layers are yolo layers, and 2 layers are up-sample layers. Among them, the 1×1 and 3×3 filters are mainly used in the convolutional layers, the 3×3 convolutional layer is used to reduce the

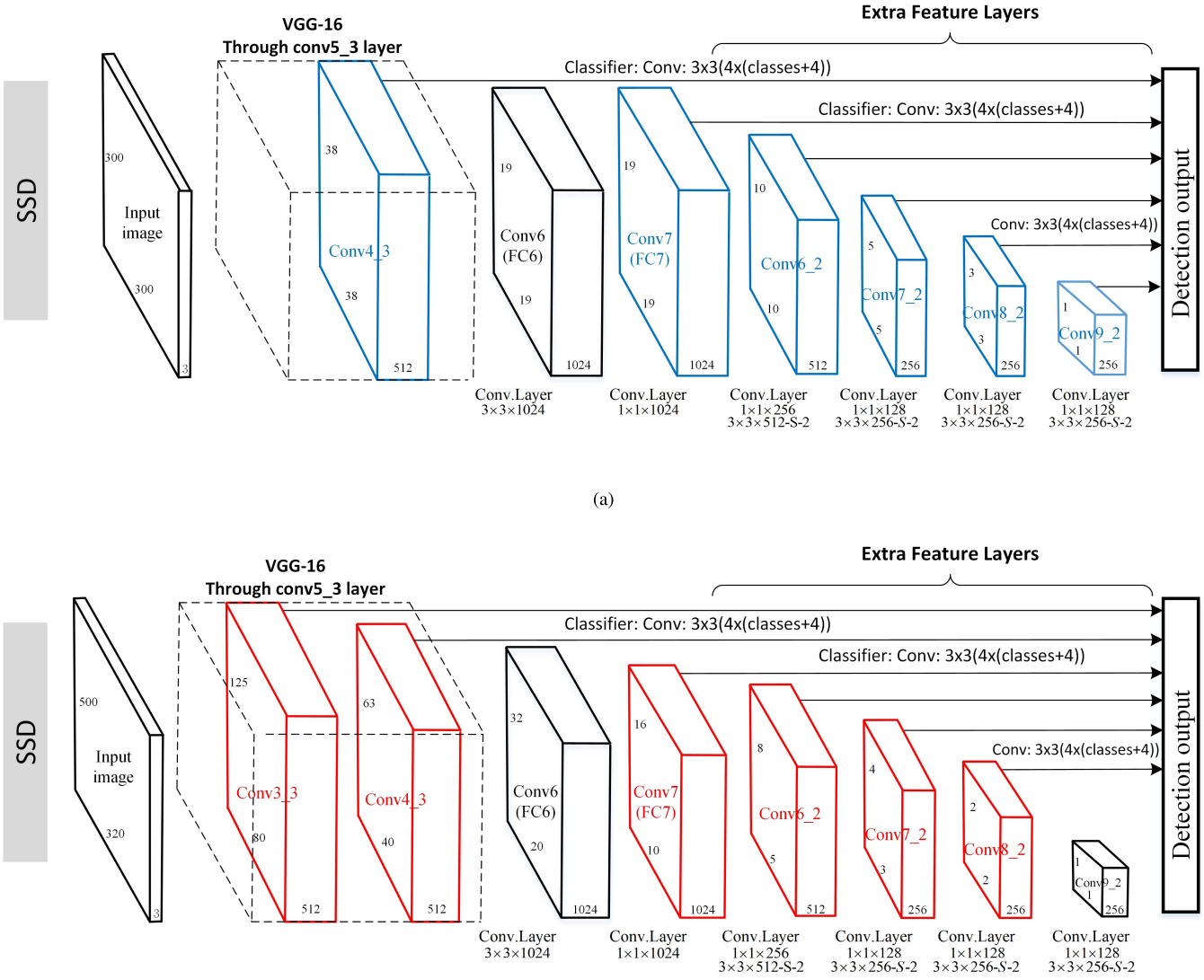

Fig. 4. CNN architecture of SSD Framework. (a) SSD network architecture in [45]. (b) Modified SSD network architecture. The main difference between the two SSD network structures is the network structure and configuration parameters of the detection layer.

width and height, the number of channels is increased, and the 1×1 convolution layer is used to compress the 3×3 volume of the characteristic representation. The greater the number of network layers, the more difficult it is to train. The shortcut layer uses the Shortcut of the Resnet network, which greatly reduces the training difficulty and improves the training accuracy. The route layer realizes a cross-layer connection which is conducive to the integration of many different features and common learning. The yolo layer is used to finally output the coordinates and categories of the predicted targets. The up-sample layer uses the up-sample layer of the FPN network and fusion decision. Using two rounds of upsampling, the combination of high-resolution feature maps with low-resolution feature maps enhances the recognition of small targets.

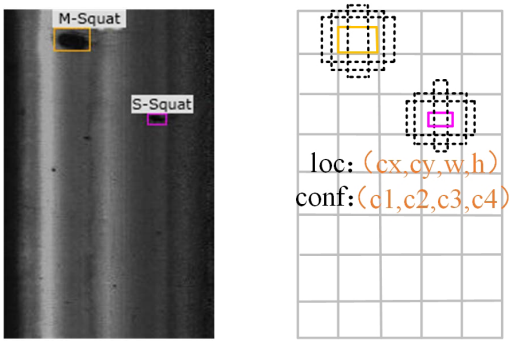

Fig. 5. Default box of the SSD framework. (a) Input with ground truth boxes. (b) 8×5 feature map.

In the multi-scale target detection network, the last layer extraction network with a 13×13 feature map is used to predict the object bounding box coordinates, object confidence value, and class probability of the three anchor boxes. The two times upsampling layers of the last layer are also used to predict the test results. The detection results of the three sizes

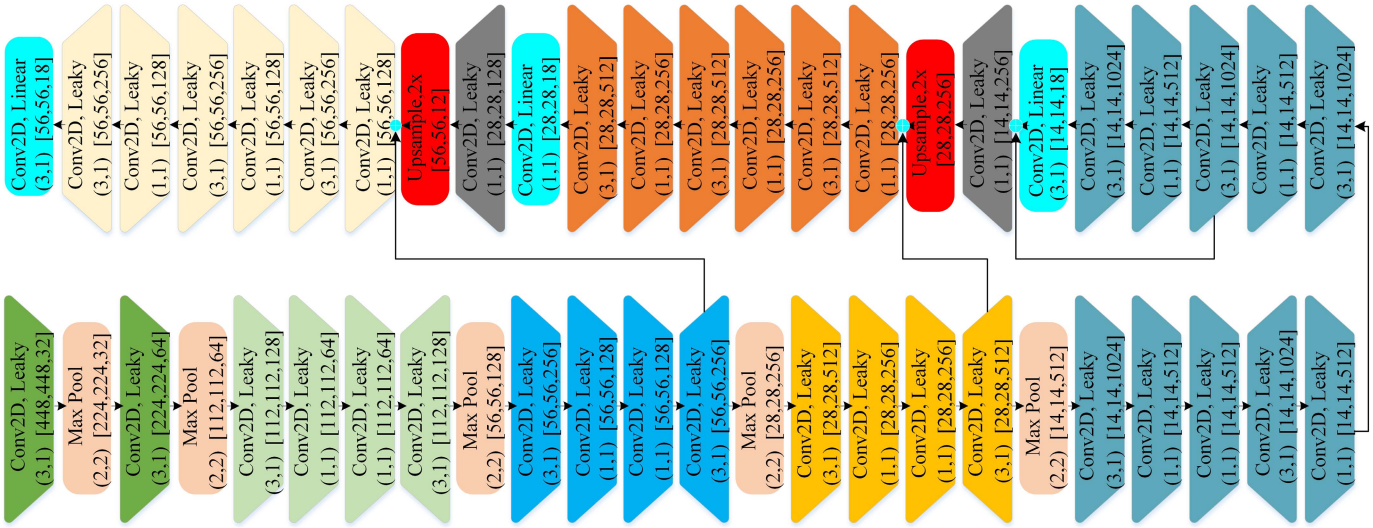

Fig. 6. The main features of Subnetwork 2: (1) After the input image, two pooling layers are used to preserve the feature information of small targets. (2) different filters adopt continuous 3×3, 1×1, 1×1, 3×3 convolutional layers to achieve descending dimension, increasing dimension, and reducing dimension to more extraction of target features; (3) Adding Multi-scale Networks. The shallow feature and deep feature are fused by up-sampling, and detected independently on the fused feature map of three scales. (4) Using k-means to select the appropriate anchor, extracting shallow target feature information, and improving IOU.

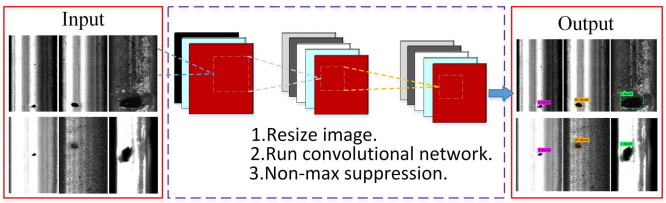

Fig. 7. YOLOv3 detection system. (1) resizing the input image size. (2) running the convolution network according on the image. (3) thresholding the resulting detections by the models confidence.

are compared and some thresholds are used to determine the final result. The use of a small box to detect objects on large feature maps effectively improves the detection accuracy of small objects.

1) End-to-End Detection: The core idea of YOLOv3 is to use the whole picture as the input of the network and directly return the position of the bounding box and its category in the output layer to achieve end-to-end detection. For any convolutional network, we input an image, train through the network, and finally output the predicted image to achieve end-to-end processing, greatly improving the detection speed. The YOLOv3 end-to-end detection system is shown in Figure 7.

2) Dimension Clustering: Traditional algorithms generally use artificial marquees but these can result in lower precision. To better choose the prior network, YOLOv3 inherits the method of YOLOv2 to calculate the anchor box and chooses to use the clustering method for the box. The traditional K-means clustering method uses the Euclidean distance function, which means that larger boxes will produce more errors than those of smaller boxes. The results of using traditional clustering methods may lead to bias. To obtain a network better and more reasonable in prediction, the K-means clustering method was chosen to train bounding boxes. Anchors boxes were used to predict the bounding box. YOLOv3 is divided into three scales: 13×13, 26×26, and 52×52. By clustering the VOC datasets, nine anchors are obtained to predict the bounding box, which improves the detection rate. However, there were 20 YOLOv3 detection categories and the target size is very different. The anchors obtained by clustering are not suitable for the detection of small targets, particularly for pixel-level detection targets. Therefore, the anchors obtained via K-means clustering calculation of the rail defect data set were used for the detection of rail defects, which can improve the detection accuracy to some extent. The height and width of the obtained anchor boxes were (272.2909, 23.6288), (336.5818, 35.9424), (264.7274, 13.312), (370.6182, 27.6224), (378.1818, 46.9248), (325.2364, 61.4016), (200.4364, 16.1408), (340.3636, 21.632), and (264.7273, 18.304).

For rail defect detection, deeper networks can result in excessive computational costs and a slower detection speed. Therefore, after analyzing the network configuration, we constructed a feature extraction network based on rail surface defect detection. Analysis of the characteristics of the rail defect data indicates that the defect size is much smaller than the entire image. Meanwhile, if the network is deep, then the characteristic information of tiny objects will be lost. A good approach is to learn the effective features from the shallow layers and then fuse the features between the shallow and deep networks.

YOLOv3 uses a deeper convolutional network and three output layers to predict detection targets. For rail defect detection, deeper neural networks are not suitable for small objects. Therefore, affected by the Darknet19 and Darknet53 network frameworks, we modified the YOLOv3 network as follows:

(1) Input image size: First, the original Darknet53 network input is 416×416. To improve the detection accuracy of the small target, the network input image size is resized to 448×448.

(2) New feature extraction network: The greatest difference from the YOLOv3 network is that we proposed a new feature extraction network. The network has only a convolutional layer and pooling layer and we removed the original shortcut layer. The specific network description is as follows: After inputting the image, two pooling layers are continuously used to save the original image information as much as possible while reducing the image size. After using two max poolings in succession, we used 3×3, 1×1, 1×1, and 3×3 convolution to achieve a dimensionality reduction, dimension increment, and descending dimension. After the 3×3 convolutional layer, the 1×1 convolutional layer can be used as retention feature information. The receptive field size is related to the network depth. As the convolution kernel is larger, it will create a larger receptive field of a single node on the feature map. As the depth of the network increases, the receptive field of the nodes on the deeper feature map becomes larger, thus the feature becomes more abstract. Therefore, we need to use the pooling to compress the input feature map following convolution. On the one hand, the feature map is reduced and the network computation complexity is simplified. On the other hand, feature compression is performed to extract the main features.

(3) Multi-scale network: After proposing a new feature extraction network, we added a multi-scale target detection network. Different from the YOLOv3 multi-scale network, according to the difference in the data set, we modified the corresponding anchor box to adapt to the size of the rail defect using the k-means algorithm. The anchor value was previously shown. The network structure is shown in Figure 6.

IV. Experiment and Analysis

A. Training Process

To improve the robustness of the algorithm and construct better data sets, we selected the rail images in six different scenes and divide them into five groups according to the scene (G1: original image, G2: rotated image, G3: Brightness enhancement, G4: Reduce contrast, and G5: Noise image, including Gaussian noise and salt and pepper noise. There were 1,000 images of rails in each group). The rail sample images were equally divided into three categories: S-Squat, M-Squat, and L-Squat. In each group, 700 images were used for training and 300 images were used as test sets. Similarly, the train set and test sets were distributed equally according to category. The pixels of the rail defect images in the data set are 250×160.

The experimental training environment was as follows: (Subnetwork1-improved SSD network: deep learning open source framework Caffe), (Subnetwork2 - improved YOLOv3 network: open source framework Darknet), Ubuntu 16.04, 32-GB RAM, CPU clocked at 3.7Hz, GTX 1080ti graphics processing unit (GPU), and 11G graphics memory.

The experimental steps were as follows:

Step 1: For each group of experiments manually mark 1000 rail defect images. At the same time, the data sets were reasonably divided (training with 700 sets in an improved deep learning network and testing results with 300 sets).

TABLE I

Based on Improved SSD Feature Extraction Strategy

Input: Rail defect data set X
Output:Feature Model M_weights of Rail Defect Dataset X
1. Data preprocessing. Use labelimg to generate true coordinate information of rail defects.
2. Load the pre-training model. Initializing the rail defect data set feature model M_weights
3. Positioning defect candidate area.
3. Calculate default box width and height:
i. $s_k = s_{min} + \frac{s_{max}-s_{min}}{m-1}(k-1)$;
ii. $w_k^a = S_k\sqrt{a_r}$; $h_k^a = S_k/\sqrt{a_r}$;
5. Calculate the bounding box size.$B = (X_b, Y_b, W_b, H_b)$.
6. Get the defect detection loss value.$L(x, c, l, g) = \frac{1}{N}((L_{conf})(x, c) + \alpha L_{loc}(x, l, g))$.
7. Updating feature weight models M_weights using stochastic gradient descent.

TABLE II

Based on Improved YOLOv3 Feature Extraction Strategy

Input: Rail defect data set X
Output:Feature Model M_weights of Rail Defect Dataset X
1. Data preprocessing. Use labelimg to generate true coordinate information of rail defects.
2. Load the pre-training model. Initializing the rail defect data set feature model M_weights
3. Positioning the center of the defect candidate area.
4. Calculate the bounding box size.$B = (b_x, b_y, b_w, b_h)$ where
$b_x = \sigma(t_x) + c_x$;$b_y = \sigma(t_y) + c_y$;
$b_w = p_w e^{t_w}$; $b_h = p_h e^{t_h}$;
5. Calculate confidence score.$confidence_{score} = P_r(object) * IOU_{pred}^{truth}$.
6. Forecast defect category.$c = -\frac{1}{n}\sum_{x=1}^{n}(ylog(\hat{y}) + (1-y)log(1-\hat{y}))$,
7. Updating feature weight models M_weights using stochastic gradient descent.

Step 2: Train the proposed network model using the pre-training model.

Step 3: Use the trained network to detect the unlabeled 300 rail defect sample images. To meet the accuracy requirement the value of confidence is greater than 0.5.

Step 4: When the number of iterations increases, the network loss is low and stable, the detection accuracy meets the performance requirements, the training ends. The network specific training process is shown in Tables I and Tables II.

During the training process, according to the performance of the equipment, the improved YOLOv3 model takes the batch-size as 64, subdivisions as 16, learning rate as 0.001, and the weight decay and momentum as default values, respectively, of 0.0005 and 0.9. The improved SSD model takes the batch-size as 16, and initial learning rate as 0.00001 to stabilize the network. After 1000 iterations, the learning rate enlarged 10 times to 0.0001, changed to 0.001 over 2000 iterations, reduced to 0.0001 over 50000 iterations, and decreased to 0.00001 over 70,000 iterations.

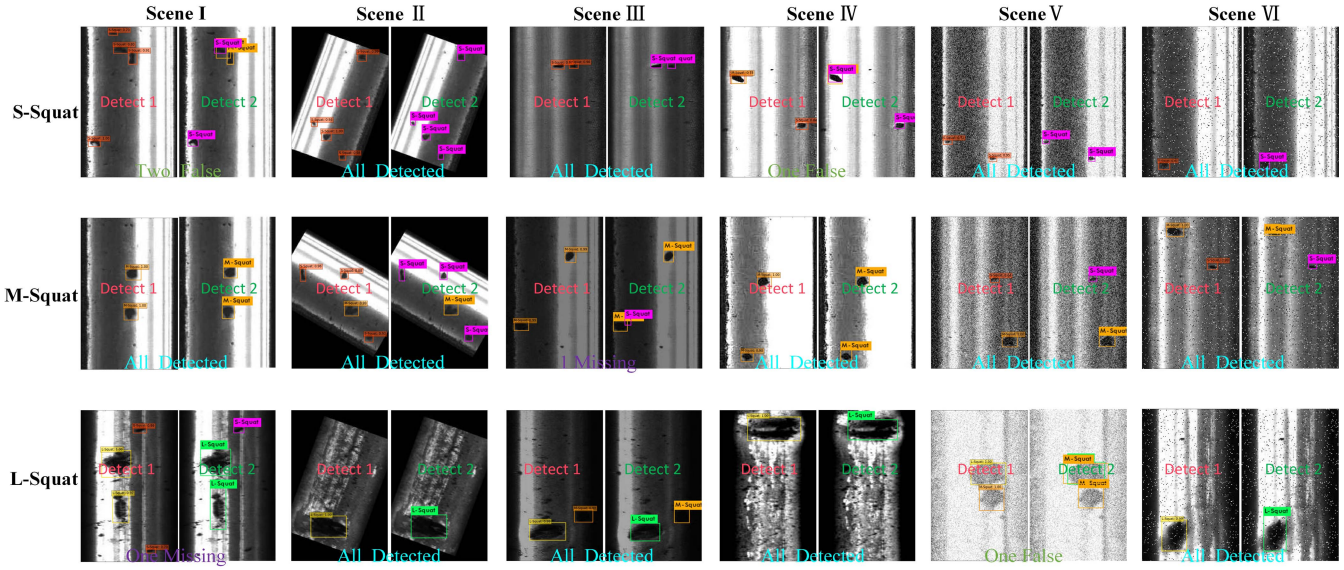

Fig. 8. The detection results of three kinds of squats in subnetwork 1 and subnetwork 2 under different scenes. Scenes I-VI are the original image, rotating image, brightness enhancement, contrast reduction, Gauss noise and salt and pepper noise respectively. Three types of squats include small squats, medium squats, and large squats. The detect results of subnetwork 1 are named as "Detect 1" in orange, and those of subnetwork 2 are named as "Detect 2" in green. The comparison results are marked below the figure, including all detected, false, and missing. Since a single image contains large, medium and small squats inevitably, multiple types of rail defects appear on one image when selecting images for comparison and display.

B. Experimental Results and Discussion

1) Evaluation of Experimental Results: The trained data were not used for testing. The test images were randomly selected from different scenes in the test set under the same environment. To detect the effect of modifying the network, a random selection in each scene was made for comparison, as shown in Figure 8. At the same time, to compare the performance of different networks, 15 images were randomly selected in the six scenes of the test set with equal proportions, and then we tested them five runs, as shown in Figure 9.

In the six scene types, the results of each type of rail defect detection marked were compared to subnetwork1 and subnetwork2 in figure 8. Small squats are difficult to identify because of the small volume in the detection of the rail defect. It was found that when detecting small and medium squats, false and missed detections are common; when detecting large squats, because of their large volume and the easy identification of features, false detection is not common but occasionally missed detection occurs.

A comparison of the detection results of various algorithms (YOLOv2, YOLOv3, SSD, Faster-RCNN, Subnetwork1, Subnetwork2) is shown in figure 9. We used ground truth as a reference, and at the same time provide the detection confidence below the image, marking the rail defect image that is an error detected or undetected. It can be seen from the test results that YOLOv2 demonstrates missed detection during the process of identification medium squats in Line(c) of Figure 4. During the process of identifying medium squats and large squats, there is a problem of errored detection and missed detection in Line(c) of Figure 7, 8, 10. At the same time, the confidence of the detection results shows that the YOLOv2 confidence levels are low when detecting rail defects. Although YOLOv3 is sensitive to small targets because of multi-scale training, error detections may also occur. Compared to YOLOv2, the detection confidence of YOLOv3 is greatly improved. SSD has a high rate of missed detection of small squats in the detection of randomly selected rail defect images of Line (f), accompanied by less error detection, but a low detection confidence. The Faster-RCNN detection has few error detections and missed detections and the confidence level is higher. A comparison of the detection results of various algorithms (YOLOv2, YOLOv3, SSD, Faster-RCNN, Subnetwork1, Subnetwork2) is shown in Figure 9.

The proposed method shows good performance in detecting all three types of squats. Notably, although a phenomenon of missed detection is observed during detection of small squats, the rate of missed detection is decreased and lower than that of the other methods, which is consistent with the multi-model method proposed. At the same time, the confidence level of the proposed method is higher than that of the others, showing certain advantages in detecting rail defect images.

In addition, to verify the validity of the experiment, some indicators during the experimental process were evaluated, including average accuracy, accuracy of each class, detection time (FPS), and comparison a with other algorithms was drawn. Where, T_p is a true case, F_p is a false positive, F_N is a false negative. The real case (T_p), the false positive case (F_p), and the false negative case (F_N) are used in the following list to evaluate the accuracy and recall rate. Following the indexes of [50], we use average accuracy (mAP) as performance of model:

$$P = \frac{T_p}{T_p + F_p}, \tag{1}$$

$$R = \frac{T_p}{T_p + F_N}, \tag{2}$$

$$mAP = \int_0^1 P(R)dR, \tag{3}$$

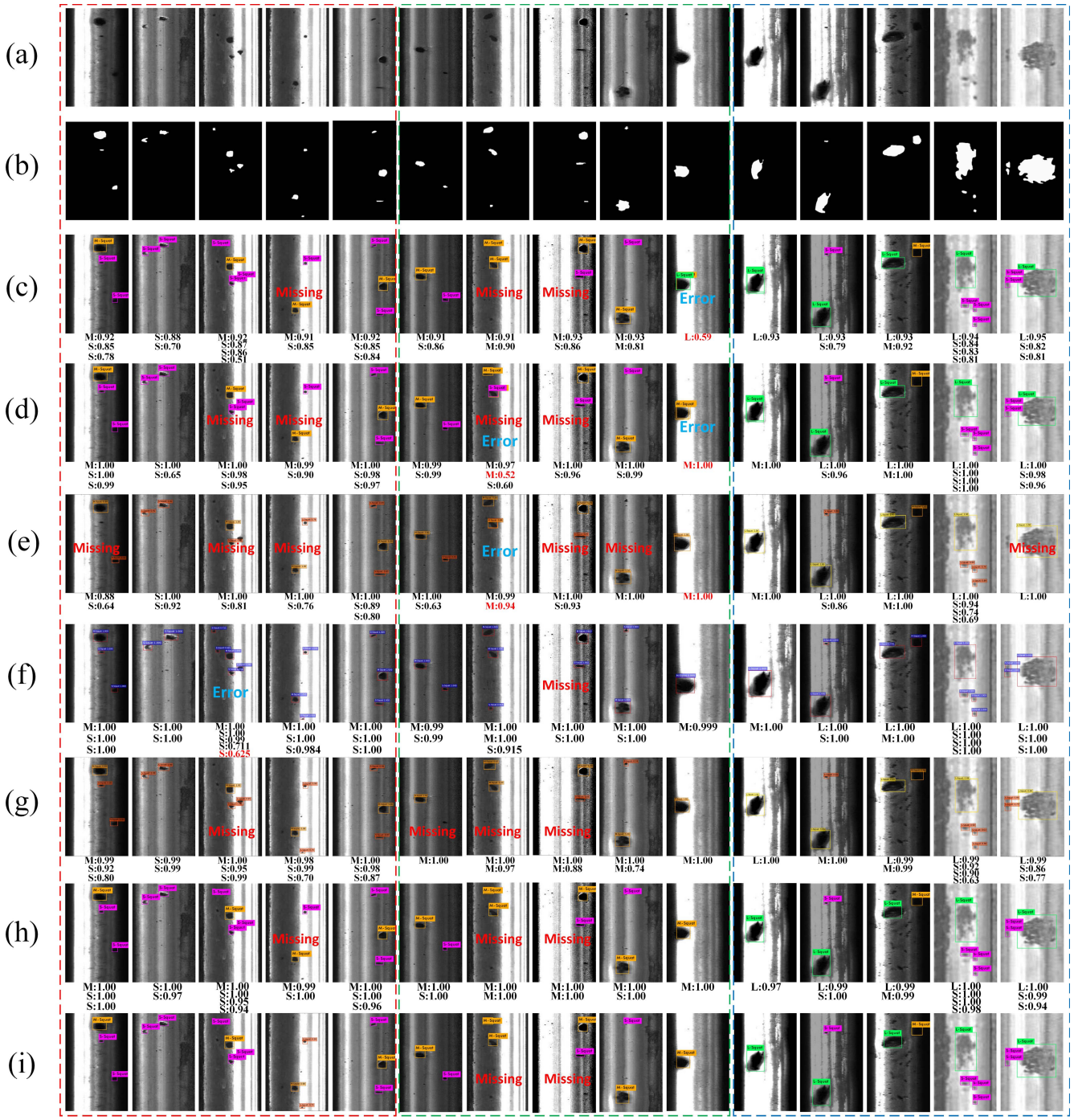

Fig. 9. Comparison of experimental results of different detection algorithms. (a) The original image. (b) Ground Truth. Detected results of rail squats detected by (c) YOLOv2. (d) YOLOv3. (e) Faster-RCNN. (f) SSD. (g) Subnetwork1. (h) Subnetwork2. (i) our method.

where T_p denotes true positive, F_p means false positive, F_N represents false negative, and P, R means precision and recall respectively. Following [53], the precision can be denoted as the number of retrieved Relevant items (i.e., right classified samples) as a proportion of the number of retrieved items, and the recall is the number of retrieved Relevant items as a proportion of all Relevant items.

2) Modify the Impact of the Output Layer in the SSD Network: In the improved SSD experiment, we compared the effects of different output layers on the detection of rail defects, as shown in Table III, where alternatives are comparison experiments that we design four kinds of network which use different convolution layer as output layers respectively. Although the original SSD is effective for different aspect_ratio detection targets, the rail defects data set is different from the VOC data set, and the L-Squat (large scratch) only occupies approximately 70×40 pixels, which is approximately 1/14 of the image. To preserve the characteristics of the front layer network (the shallow layer mainly contains small target features and the deep layer mainly contains large target features),

TABLE III
COMPARISON OF DETECTION ACCURACY OF DIFFERENT OUTPUT LAYERS

Method	Output layers								mAP
	conv2_2	conv3_3	conv4_3	conv7	conv6_2	conv7_2	conv8_2	conv9_2	
Proposed SSD		√	√	√	√	√	√		89.92
Original SSD			√	√	√	√	√	√	86.46
Alternative 1	√	√	√	√	√	√	√	√	88.73
Alternative 2	√	√	√	√	√	√			85.98
Alternative 3		√	√	√	√	√	√	√	87.26
Alternative 4	√	√	√						72.84

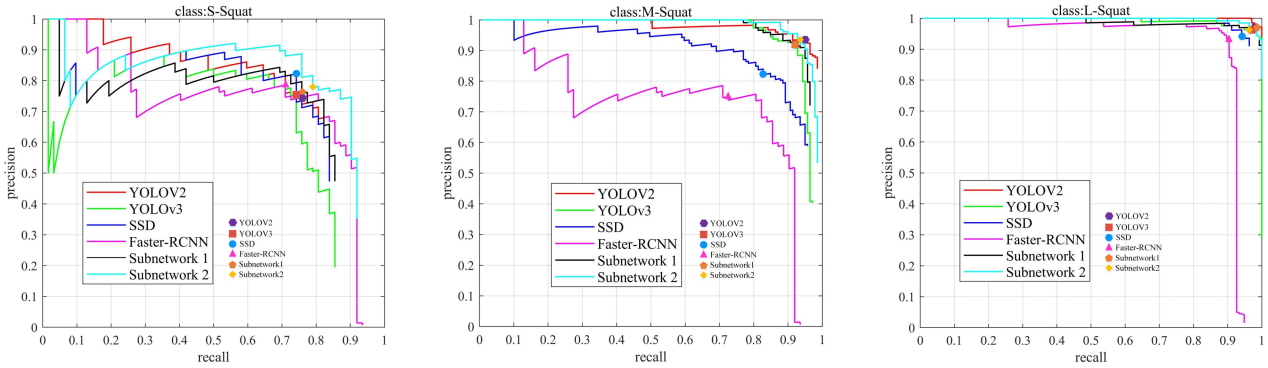

Fig. 10. P-R curves of three types of squats. Detection results of different algorithms for (a) small squats. (b) medium squats. (c) large squats.

TABLE IV
BASIC NETWORK COMPARISON

network	Resnext50	Resnet101	Densenet201	Darknet19	Darknet53
mAP	84.73	89.26	88.52	87.95	90.49
Time(ms/Img)	29.34	26.87	27.17	11.20	19.98
S-Squat	69.56	76.0	72.98	70.82	77.57
M-Squat	90.72	93.63	95.30	95.23	95.51
L-Squat	93.92	98.15	97.26	97.80	98.40

the Conv2_2 and Conv3_3 convolution layers in VGG16 were introduced. A comparison between the proposed model and the original model shows that the accuracy is improved when the low-level convolution layer is used; that is, the modified output layer shows good robustness in the rail defect data set.

3) Comparison of the Effects Between Backbones: In the improved YOLOv3 experiment, the Resnext50, Resnet101, Densenet201, Darknet19, and Darknet53 frameworks; the time used and the accuracy of each class were compared in the detection of rail defect data sets, as shown in Table IV (the original weights were extracted, and the initialization weight model was regenerated. In the network, the corresponding backbone is combined with the multi-scale network and the anchor is modified to the numerical value obtained by the K-means algorithm). We chose and compared the best results among them. It can be seen from the table that the original Darknet53 network, after modifying the anchor value, is far superior to other methods in detecting small, medium, and large squats. Because Darknet19 has a lighter network and fewer network parameters, its detection speed is faster than that of the other basic models.

4) Comparison Between Different Algorithm Performances: We plotted the accuracy-recall rate (P-R) curve using MATLAB2017b to visualize the performance of different detection algorithms. As shown in Figure 10, it can be seen from the figure that when the curve reaches the equilibrium point, which is a comprehensive evaluation index for evaluating quality of Precision-Recall curve, of subnetwork1 and subnetwork2 are superior in accuracy compared to the other methods in detecting the small squat. YOLOv2, YOLOv3, and Faster-RCNN have poor detection of medium squats and the detection accuracy of subnetwork1 is higher than that of the other algorithms. In detecting large squats, the detection performance of each method is better because of the obvious features. In general, because of the difficulty in identifying small squats, most algorithms have poor robustness. There are obvious differences in defect characteristics between medium and large squats; thus, it is easy to detect the defects and have high accuracy. This phenomenon is most obvious in the P-R curve of large squats.

Similarly, to evaluate the effectiveness of the improved method, we integrated the training results of five different scenes, as shown in Table V. Comparing subnetwork1 and subnetwork2 to the method in [51], [52], YOLOv2, YOLOv3, and Faster-RCNN were assessed for accuracy, detection time, and training time. Overall, our model achieved good results with respect to all small, medium, and large squats. In a comparison

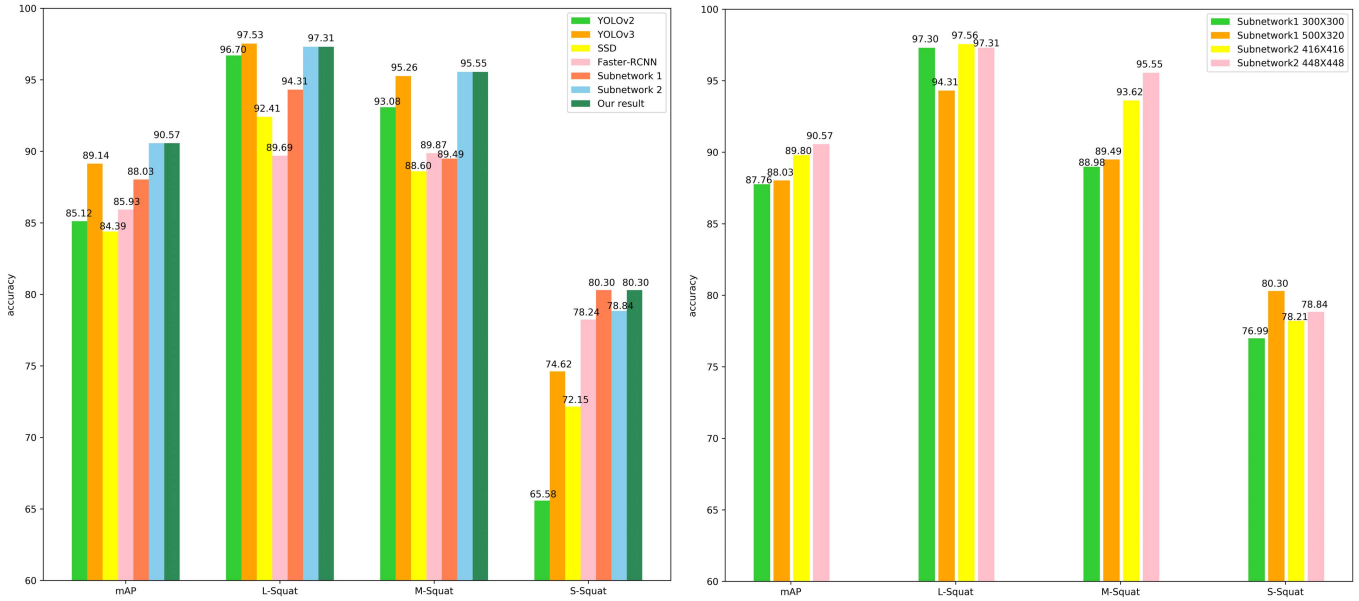

Fig. 11. Effect analysis of modified network. (a) Performance comparison of different algorithms. (b) Performance comparison of different input sizes of modified network.

TABLE V

COMPARISON OF THE COMPREHENSIVE PERFORMANCE OF THE PROPOSED METHOD AND SOTA METHODS ON ORIGINAL DATASETS

Group	Method	mAP	L-Squat	M-Squat	S-Squat	FPS	Train time
G1	Method in [51]	57.01	64.36	56.67	50.00	\	\
	Method in [52]	67.07	77.52	70.46	53.24	4	\
	YOLOv2(448)	87.52	**98.40**	95.19	68.97	**102**	12h
	YOLOv3(416)	**89.13**	96.8	**96.5**	74.09	49	18h
	SSD(300)	82.95	89.37	88.71	70.77	70	16h
	Faster-RCNN	85.73	89.91	88.07	**79.22**	7	**10h**
	Subnetwork 1	85.37	90.55	89.37	76.20	70	16h
	Subnetwork 2	88.39	97.15	95.58	72.43	88	14h

of 5 groups, we can see the average accuracy (mAP) of subnetwork2 (G1: 88.39, G2: 89.88, G3: 90.98, G4: 91.43, and G5: 92.15) is higher than that of the other algorithms in the other groups except it is slightly lower than YOLOv3 in G1. The average accuracy (mAP) of Subnetwork1 is second only to Subnetwork2 and YOLOv3. Among large squats, the highest detection accuracy of YOLOv2 is 98.4 in G1, and YOLOv3 has the highest detection accuracy of 97.7, 97.6, 97.7, 97.87 in G2-G5. For medium squats, YOLOv3 has the highest detection accuracy of 96.5 and 96.23 in G1 and G4, respectively; Subnetwork 2 has the highest detection accuracy in G2, G3, and G5 of 93.99, 96.43, and 96.48, respectively. Among small squats, Faster-RCNN has the highest detection accuracy of 79.22 and 83.48 in G1 and G3, respectively; subnetwork 1 has the highest detection accuracy of 82.11 in G2; and Subnetwork 2 has the highest detection accuracy of 81.36 and 82.23 in G4 and G5, respectively. The aforementioned data were shown in comparison: (1) the traditional method of detecting rail squats has poor performance. (2) YOLOv2 and SSD have poor detection performance when detecting small squats; (3) As a two-stage algorithm, Faster-RCNN has lower precision when detecting large-scale defects, but it has better detection performance when detecting small squats. (4) Subnetwork 1 and the Subnetwork 2 improved algorithm is better than the original algorithm when detecting different types of defects in the rail, particularly in the detection of small squats. It is undeniable that YOLOv3 also achieved satisfactory results for the detection of medium and large rail defects, but for small rail defects, its detection capability is not as good as that of subnetwork 2. In addition, YOLOv2 has the highest detection speed in terms of detection time, followed by that of Subnetwork2, Subnetwork1, and SSD, which also shows the performance of our proposed algorithm with real-time monitoring in detection time.

At the same time, to fully reflect the recognition ability of the algorithm in different scenes, we calculated the mAP of each algorithm and the average precision in each category as shown in Figure 11(a). It shows that the mAP and average precision of small and medium rail squat detection in the improved network is higher than that of the other algorithms. The drawback is that the average precision of the proposed networks (Subnetwork1, Subnetwork2) are respectively slightly lower than SSD and YOLOv3 in large rail squat detection, but they do meet the accuracy requirements. We also compared the accuracy of the improved algorithm under different input image sizes. As shown in Figure 11(b), the mAP of Subnetwork1 (improved SSD algorithm) under 500 $\times$ 320 pixels and the detection accuracy of small and medium squats are higher than that of the original SSD (300 $\times$ 300). For detecting large squats, the detection accuracy is insufficient. Subnetwork2

TABLE VI

COMPARISON OF THE COMPREHENSIVE PERFORMANCE OF THE PROPOSED METHOD AND OTHER METHODS

Group	Method	mAP	L-Squat	M-Squat	S-Squat	FPS	Train time
G2	Method in [52]	62.48	70.53	68.16	48.75	2	\
	YOLOv2(448)	83.25	**95.34**	92.23	62.17	**104**	12h
	YOLOv3(416)	86.14	97.7	93.56	67.17	54	18h
	SSD(300)	82.77	89.22	88.51	70.57	70	16h
	Faster-RCNN	85.79	89.50	87.10	80.77	8	**10h**
	Subnetwork 1	89.35	96.6	89.34	**82.11**	70	16h
	Subnetwork 2	**89.88**	96.65	**93.99**	79	88	14h
G3	Method in [52]	68.91	78.52	68.49	59.73	2	\
	YOLOv2(448)	86.39	96.3	90.38	72.5	**103**	12h
	YOLOv3(416)	89.55	**97.60**	94.19	76.86	50	18h
	SSD(300)	87.34	97.40	88.20	76.42	70	16h
	Faster-RCNN	87.68	90.04	89.51	**83.48**	8	**10h**
	Subnetwork 1	88.99	97.60	89.43	79.93	66	16h
	Subnetwork 2	**90.98**	97.34	**96.43**	79.17	88	14h
G4	Method in [52]	66.42	76.34	72.75	50.17	5	\
	YOLOv2(448)	82.56	96.40	92.95	58.34	**104**	12h
	YOLOv3(416)	89.88	**97.70**	**96.23**	75.72	50	18h
	SSD(300)	85.63	96.49	88.75	71.65	70	16h
	Faster-RCNN	84.63	89.9	89.3	74.7	7	**10h**
	Subnetwork 1	87.20	90.76	89.54	81.31	70	16h
	Subnetwork 2	**91.43**	97.68	95.25	**81.36**	87	14h
G5	Method in [52]	74.59	83.57	76.73	63.47	3	\
	YOLOv2(448)	85.88	97.07	94.64	65.92	**103**	12h
	YOLOv3(416)	90.98	**97.87**	95.80	79.28	54	18h
	SSD(300)	83.24	89.59	88.82	71.32	70	16h
	Faster-RCNN	85.84	89.10	95.39	73.01	8	**10h**
	Subnetwork 1	89.26	96.04	89.79	81.95	70	16h
	Subnetwork 2	**92.15**	97.73	**96.48**	**82.23**	88	14h

(improved YOLOv3 algorithm) has a higher mAP at 448×448 pixels and detection accuracy for small and medium squats compared to that of the original YOLOv3 (416×416); for large squats, it is slightly less than the original YOLOv3 (416×416).

V. CONCLUSION

This study proposes a multi-network combined rail surface defect detection system. Through parallel detection of two subnetworks, it is able to accurately identify and detect three types of squats on a rail surface so as to timely determine the state of the rail damage and the degree of influence on the smooth operation of the railway. Because the algorithm has different advantages in detecting the surface defects of a rail, to ensure a fast detection speed and in addition (1) for the sake of accurately detecting large squats in the rail, an SSD network is used; (2) for detection of small squats, the YOLOv3 network is applied. In addition, to enable the two networks to detect small, medium, and large squats, respectively, the two networks are further optimized. The detection process is completed by two improved CNNs. Through the interaction of two one-stage algorithms, the error detection rate is reduced, and the detection accuracy is improved. The experimental results show that the detection method has good applicability for squats of rail surfaces.

REFERENCES

[1] M. Molodova, Z. Li, A. Núñez, and R. Dollevoet, "Automatic detection of squats in railway infrastructure," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 5, pp. 1980–1990, Oct. 2014.

[2] 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.

[3] S. Alahakoon, Y. Q. Sun, M. Spiryagin, and C. Cole, "Rail flaw detection technologies for safer, reliable transportation: A review," *J. Dyn. Syst., Meas., Control*, vol. 140, no. 2, pp. 1–17, Feb. 2018.

[4] V. J. Hodge, S. O'Keefe, M. Weeks, and A. Moulds, "Wireless sensor networks for condition monitoring in the railway industry: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 3, pp. 1088–1106, Jun. 2015.

[5] C. Zhang, Z. Su, Q. Li, and M. Chen, "Inspection system for detection of defects on rail surface based on LED and linear laser," *Sci. Technol. Eng.*, vol. 12, no. 36, pp. 9877–9880, 2012.

[6] Z. Xiong, Q. Li, Q. Mao, and Q. Zou, "A 3D laser profiling system for rail surface defect detection," *Sensors*, vol. 17, no. 8, p. 1791, Aug. 2017.

[7] S. Somalraju, V. Murali, G. Saha, and V. Vaidehi, "Robust railway crack detection scheme (RRCDS) using LED-LDR assembly," in *Proc. Int. Conf. Recent Trends Inf. Technol.*, Chennai, India, Apr. 2012, pp. 477–482.

[8] Q. Jian-Hua, L. Lin-Sheng, and Z. Jing-Gang, "Design of rail surface crack-detecting system based on linear CCD sensor," in *Proc. IEEE Int. Conf. Netw., Sens. Control*, Sanya, China, Apr. 2008, pp. 1626–1631.

[9] C. Ling, G. Jianqiang, G. Xiaorong, W. Zeyong, and L. Jinlong, "Research on rail defect detection system based on FPGA," in *Proc. IEEE Far East NDT New Technol. Appl. Forum (FENDT)*, NanChang, China, Jun. 2016, pp. 195–200, doi: 10.1109/FENDT.2016.7992023.

[10] M. Sun, X. Lin, Z. Wu, Y. Liu, Y. Shen, and N. Feng, "Non-destructive photoacoustic detecting method for high-speed rail surface defects," in *Proc. IEEE Int. Instrum. Meas. Technol. Conf. (I MTC)*, Montevideo, Uruguay, May 2014, pp. 896–900.

[11] D. Scalea *et al.*, "High-speed defect detection in rails by noncontact guided ultrasonic testing," *Transp. Res. Rec.*, vol. 16, no. 1, pp. 66–77, 2005.

[12] P. Rizzo *et al.*, "Unsupervised learning algorithm for high-speed defect detection in rails by laser/air-coupled non-contact ultrasonic testing," *Smart Struct. Materials. Int. Soc. Opt. Photon.*, vol. 6174, Apr. 2006, Art. no. 61742G.

[13] J. L. Rose, M. J. Avioli, P. Mudge, and R. Sanderson, "Guided wave inspection potential of defects in rail," *NDT&E Int.*, vol. 37, no. 2, pp. 153–161, Mar. 2004.

[14] D. Hesse and P. Cawley, "Surface wave modes in rails," *J. Acoust. Soc. Amer.*, vol. 120, no. 2, pp. 733–740, Aug. 2006.

[15] D. Hesse and P. Cawley, "Excitation of surface wave modes in rails and their application for defect detection," in *Proc. AIP Conf.* College Park, MD, USA: American Institute of Physics, 2006, pp. 1593–1600.

[16] V. P. Sebko, G. M. Suchkov, and V. M. Kamardin, "Sensitivity of the electromagnetic acoustical technique for testing railway rails by the mirror-shadow method," *Russian J. Nondestruct. Test.*, vol. 40, no. 3, pp. 170–177, Mar. 2004.

[17] S. Dixon, R. S. Edwards, and X. Jian, "Inspection of rail track head surfaces using electromagnetic acoustic transducers (EMATs)," *Insight-Non-Destructive Test. Condition Monitor.*, vol. 46, no. 6, pp. 326–330, Jun. 2004.

[18] S. Nielasen *et al.*, "Automatic laser ultrasonics for rail inspection," in *Proc. 16th World Conf. NDT*, Montreal, QC, Canada, 2004, pp. 1–8.

[19] Q. Wei, X. Zhang, Y. Wang, N. Feng, and Y. Shen, "Rail defect detection based on vibration acceleration signals," in *Proc. IEEE Int. Instrum. Meas. Technol. Conf. (IMTC)*, May 2013, pp. 1194–1199.

[20] Z. Chen, J. Xuan, P. Wang, H. Wang, and G. Tian, "Simulation on high speed rail magnetic flux leakage inspection," in *Proc. IEEE Int. Instrum. Meas. Technol. Conf.*, May 2011, pp. 1–5.

[21] J. W. Wilson and G. Y. Tian, "3D magnetic field sensing for magnetic flux leakage defect characterisation," *Insight-Non-Destructive Test. Condition Monitor.*, vol. 48, no. 6, pp. 357–359, Jun. 2006.

[22] J. Gao, G. Du, and H. Wei, "The research of defect detection test system based on magnetic flux leakage," in *Proc. Int. Forum Strategic Technol.*, 2011, pp. 1225–1229.

[23] A. G. Antipov and A. A. Markov, "Evaluation of transverse cracks detection depth in MFL rail NDT," *Russian J. Nondestruct. Test.*, vol. 50, no. 8, pp. 481–490, Aug. 2014.

[24] 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.
[25] Y. Santur, M. Karakose, and E. Akin, "A new rail inspection method based on deep learning using laser cameras," in *Proc. Int. Artif. Intell. Data Process. Symp. (IDAP)*, Sep. 2017, pp. 1–6.
[26] E. Deutschl, C. Gasser, A. Niel, and J. Werschonig, "Defect detection on rail surfaces by a vision based system," in *Proc. IEEE Intell. Vehicles Symp.*, Jun. 2004, pp. 507–511.
[27] M. Nitti *et al.*, "Real time classification of rail defects," *WIT Trans. Built Environ.*, vol. 61, pp. 335–344, May 2002.
[28] 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.
[29] 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, pp. 1–11, Dec. 2018.
[30] Ç. Aytekin, Y. Rezaeitabar, S. Dogru, and I. Ulusoy, "Railway fastener inspection by real-time machine vision," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 45, no. 7, pp. 1101–1107, Jul. 2015.
[31] C. Tastimur, O. Yaman, M. Karakose, and E. Akin, "A real time interface for vision inspection of rail components and surface in railways," in *Proc. Int. Artif. Intell. Data Process. Symp. (IDAP)*, Sep. 2017, pp. 1–6.
[32] T. de Bruin, K. Verbert, and R. Babuska, "Railway track circuit fault diagnosis using recurrent neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 3, pp. 523–533, Mar. 2017.
[33] H. Trinh, N. Haas, Y. Li, C. Otto, and S. Pankanti, "Enhanced rail component detection and consolidation for rail track inspection," in *Proc. IEEE Workshop Appl. Comput. Vis. (WACV)*, Jan. 2012, pp. 289–295.
[34] A. K. Singh, A. Swarup, A. Agarwal, and D. Singh, "Vision based rail track extraction and monitoring through drone imagery," *ICT Exp.*, vol. 5, no. 4, pp. 250–255, Dec. 2019.
[35] Y. Wu, Y. Qin, Z. Wang, and L. Jia, "A UAV-based visual inspection method for rail surface defects," *Appl. Sci.*, vol. 8, no. 7, p. 1028, Jun. 2018.
[36] F. Flammini, C. Pragliola, and G. Smarra, "Railway infrastructure monitoring by drones," in *Proc. Int. Conf. Electr. Syst. Aircr., Railway, Ship Propuls. Road Vehicles Int. Transp. Electrific. Conf. (ESARS-ITEC)*, Nov. 2016, pp. 1–6.
[37] O. I. Chumachenko and A. V. Gilevoy, "Image processing in UAV," in *Proc. IEEE 2nd Int. Conf. Actual Problems Unmanned Air Vehicles Develop. (APUAVD)*, Oct. 2013, pp. 75–76.
[38] X. Gibert, V. M. Patel, and R. Chellappa, "Robust fastener detection for autonomous visual railway track inspection," in *Proc. Appl. Comput. Vis.*, 2015, pp. 694–701.
[39] 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.
[40] X. Zhang, K. Wang, Y. Wang, Y. Shen, and H. Hu, "An improved method of rail health monitoring based on CNN and multiple acoustic emission events," in *Proc. IEEE Int. Instrum. Meas. Technol. Conf. (I MTC)*, May 2017, pp. 1–6.
[41] 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.
[42] 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.
[43] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6517–6525.
[44] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*. [Online]. Available: http://arxiv.org/abs/1804.02767
[45] W. Liu *et al.*, "SSD: Single shot MultiBox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.
[46] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.
[47] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.
[48] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
[49] J. Chen, Z. Liu, H. Wang, A. Nunez, and Z. Han, "Automatic defect detection of fasteners on the catenary support device using deep convolutional neural network," *IEEE Trans. Instrum. Meas.*, vol. 67, no. 2, pp. 257–269, Feb. 2017.
[50] J. Zhong, Z. Liu, Z. Han, Y. Han, and W. Zhang, "A CNN-based defect inspection method for catenary split pins in high-speed railway," *IEEE Trans. Instrum. Meas.*, vol. 68, no. 8, pp. 2849–2860, Aug. 2019.
[51] 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.
[52] H. Zhang, X. Jin, Q. M. J. Wu, Y. Wang, Z. He, and Y. Yang, "Automatic visual detection system of railway surface defects with curvature filter and improved Gaussian mixture model," *IEEE Trans. Instrum. Meas.*, vol. 67, no. 7, pp. 1–16, Feb. 2018.
[53] M. Buckland and F. Gey, "The relationship between recall and precision," *J. Amer. Soc. Inf. Sci.*, vol. 45, no. 1, pp. 12–19 2014.

Hui Zhang (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in pattern recognition and intelligent systems from Hunan University, Changsha, China, in 2004, 2007, and 2012, respectively. He was a Visiting Scholar with the CVSS Laboratory, Department of Electrical and Computer Engineering, University of Windsor. He is currently a professor with the School of Robotics, Hunan University. His research interests include machine vision, image processing, and visual tracking.

Yanan Song received the bachelor's degree from Henan University of Technology in 2017. He is currently a Graduate Student with the School of Electrical and Information Engineering, Changsha University of Science and Technology. His main research interests are rail visual inspection and image processing.

Yurong Chen (Member, IEEE) was born in 1998. He received the B.S. degree in electrical and computer engineering from Changsha University of Science and Technology in 2019 and the M.S. degree in electrical and computer engineering from the University of Pittsburgh, PA, USA, in 2020. He is currently pursuing the Ph.D. degree with the National Engineering Laboratory of Robot Visual Perception and Control Technology, Hunan University. His current research interests include image processing, machine learning, and domain adaption.

Hang Zhong received the B.S., M.S., and Ph.D. degrees in automation from the College of Electrical and Information Engineering, Hunan University, Changsha, China, in 2013, 2016, and 2020, respectively. He currently holds the post-doctoral position with the National Engineering Laboratory of Robot Visual Perception and Control Technology, Hunan University. His research interests include robotics modeling and control, visual servo control, and path planning of the aerial robots.

Li Liu (Member, IEEE) was born in 1984. He received the B.S. degree in measurement and control technology and instrument from Southeast University, Nanjing, China, in 2006, and the Ph.D. degree in automation from the College of Electrical and Information Engineering, Hunan University, in 2020. His current research interests include robot vision measurement, robot path planning, and intelligent control.

Yaonan Wang received the Ph.D. degree in electrical engineering from Hunan University, Changsha, China, in 1994. He was a Post-Doctoral Research Fellow with the Normal University of Defence Technology, Changsha, from 1994 to 1995. From 1998 to 2000, he was a senior humboldt fellow in Germany. From 2001 to 2004, he was a Visiting Professor with the University of Bremen, Bremen, Germany. From 2001 to 2020, he was the Dean of the College of Electrical and Information Engineering, Hunan University, China. Since 1995, he has been a Professor with Hunan University. His current research interests include intelligent control, robotics, and image processing. He was the Principle Leader with the National Engineering Laboratory of Robot Visual Perception and Control Technology, Hunan, China. He is currently the President of China Society of Image and Graphics, Beijing, China. He is a fellow of Chinese Academy of Engineering.

Thangarajah Akilan (Member, IEEE) received the Ph.D. degree in electrical and computer engineering from the University of Windsor, Windsor, ON, Canada. He is currently an Assistant Professor with the Department of Software Engineering, Lakehead University, Thunder Bay, ON, Canada. His research interests include object/action recognition, image/video processing and segmentation, and data fusion using statistical techniques, machine/deep learning, and natural language processing. He was a recipient of the 2015–2016 Golden Key's premier Graduate Scholar Award and the 2013–2014 His Majesty the King's Scholarship of Royal Thai Government. He was a Secretary of Young Professionals, IEEE Winnipeg Section, Canada, and a Reviewer for several journals, including IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, and IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS.

Q. M. Jonathan Wu (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from the University of Wales, Swansea, U.K., in 1990. He was affiliated with the National Research Council of Canada for ten years beginning, in 1995, where he became a senior research officer and a group leader. He is currently a Professor with the Department of Electrical and Computer Engineering, University of Windsor, Windsor, ON, Canada. He has published more than 300 peer-reviewed papers in computer vision, image processing, intelligent systems, robotics, and integrated microsystems. His current research interests include 3-D computer vision, active video object tracking and extraction, interactive multimedia, sensor analysis and fusion, and visual sensor networks. He is a fellow of Canadian Academy of Engineering. He held the Tier 1 Canada Research Chair in Automotive Sensors and Information Systems from 2005 to 2019. He has served on technical program committees and international advisory committees for many prestigious conferences. He is an Associate Editor for the IEEE TRANSACTION ON CYBERNETICS, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the *Journal of Cognitive Computation*, and the *Neurocomputing*.