International Journal of Machine Learning and Cybernetics (2023) 14:1903–1920
https://doi.org/10.1007/s13042-022-01736-y

ORIGINAL ARTICLE

An inspection method of rail head surface defect via bimodal structured light sensors

Jiajun Zheng[1] · Le Wang[2] · Junbo Liu[2] · Hao Wang[2] · Shengchun Wang[2] · Liang Wang[1] · Jiaxu Zhang[1]

Received: 7 October 2021 / Accepted: 28 November 2022 / Published online: 21 December 2022

Abstract
Rail defects have long threatened the safety of railway vehicles. Existing inspection methods still have some problems and flaws that can not meet practical application. In this paper, we propose a rail surface defect inspection method based on bimodal structured light sensors, termed Rail Surface Defect Inspection Network (RSDINet), which can detect and measure defect in bimodal rail images. To verify effect of the method, we establish a bimodal image dataset of intensity and depth images collected by the constructed bimodal structured light sensors data acquisition system. To solve the irregularity of rail surface defect shape, we propose a feature extraction backbone network by introducing deformable convolution. Moreover, RSDINet adopts a parallel feature extraction strategy to process bimodal images respectively. Specifically, we apply different backbone networks to bimodal images respectively for different image characteristics to enhance the feature representation ability of network. Then, our RSDINet fuses multi-scale feature of bimodal images respectively and carries out multi-scale rail surface defect detection and measurement. It is worth noting that the proposed RSDINet can accomplish these two tasks end-to-end simultaneously. Experiments demonstrate that based on the established dataset, the method achieves 87.17 mAP and 39.07 mSAP for detection and measurement respectively at 6.2 FPS on a single GPU, which has a better performance than previous SOTA methods and shows a promising potential for application in high-speed railway.

Keywords Rail surface defect inspection · Defect severity measurement · Bimodal structured light sensors · Bimodal image fusion · Deep convolutional neural network · High-speed railway

1 Introduction

High speed railways have entered a period of rapid development, and the mileage of high-speed railways in various countries has increased over the past decade. With the increase of train speed and train load, rail deterioration speed is unavoidably accelerated in the long-term daily operation of high-speed railway, which results in rail damage and defects. If rail defects are not detected at early stage and dealt with in time, the maintenance cost in the later period is very high, posing a huge threat to the safety of railway traffic. However, due to existing many problems of traditional manual inspection, such as low efficiency, high subjective component, high rate of missing inspection and high risk coefficient, with the continuous development of high-speed railway, automatic inspection of random distribution,

Jiajun Zheng and Le Wang contributed equally to this work.

✉ Shengchun Wang
wangshengchun@rails.cn

✉ Liang Wang
wangliang7@mail.sysu.edu.cn

Jiajun Zheng
zhengjj55@mail2.sysu.edu.cn

Le Wang
wl91@rails.cn

Junbo Liu
liujunbo@rails.cn

Hao Wang
wanghao@rails.cn

Jiaxu Zhang
zhangjiaxu402@163.com

[1] School of Electronics and Communication Engineering, Sun Yat-Sen University, No. 66, Gongchang Road, Guangming District, Shenzhen 518107, Guangdong, China

[2] Railway Infrastructure Research Institute, China Academy of Railway Science Corporation Limited, No. 2 Daliushu Road, Haidian District, Beijing 100081, China

various categories and sizes defects on rail with huge railway mileage is an important issue related to national economic development and people's travel safety, which is of great significance.

In recent years, with the rapid development of machine vision and image processing technology, many scholars and researchers have carried out extensive researches on rail defect inspection [1]. In general, in the field of rail surface defect detection, these inspection techniques can be roughly divided into two categories: one is based on two-dimensional rail surface image data; the other is from three-dimensional perspective, such as structured light, point cloud and so on. The basic idea of two dimension is based on the image data collected by the high-speed linear CCD (charge coupled device) camera installed under the test vehicle, employing traditional hand-crafted and pre-defined feature or neural network for feature extraction, so as to achieve the purpose of inspecting rail surface defect, such as defect localization based on projection profile (DLBP) [2], Gabor filter [3], coarse-to-fine model [4] and so on [5–10]. The three-dimensional methods usually employ the line structured light emitted by the laser scanner combined with the CCD cameras to study rail surface defect, such as [11, 12].

However, by analyzing the two-dimensional rail intensity images shown in Fig. 1, it is not difficult to see that only from the two-dimensional intensity point of view, different categories of defects are definitely similar and some defects are very difficult to identify. Moreover, rail often happens stains and oxide scale in the process of railway operation, which are so similar to rail surface defects in the gray intensity images that existing methods can't distinguish commendably. In practice, due to the influence of uneven illumination and the difference of surface properties in different areas of rail, some existing methods [13, 14] are difficult to effectively detect the real rail surface scratch only by intensity rail images. All of the above are the reasons for high false alarm rate and low accuracy of two-dimensional methods. All the studies reviewed so far suffer from these problems. Thus, researchers consider to obtain depth information of rail surface defects and solve these problems from a three-dimensional perspective. However, existing technologies [11, 12, 15] only stay at the level of obtaining rail profile information for analysis, with slow speed and low accuracy, which is far from being able to meet a specific criterion of practical application in railway daily inspection.

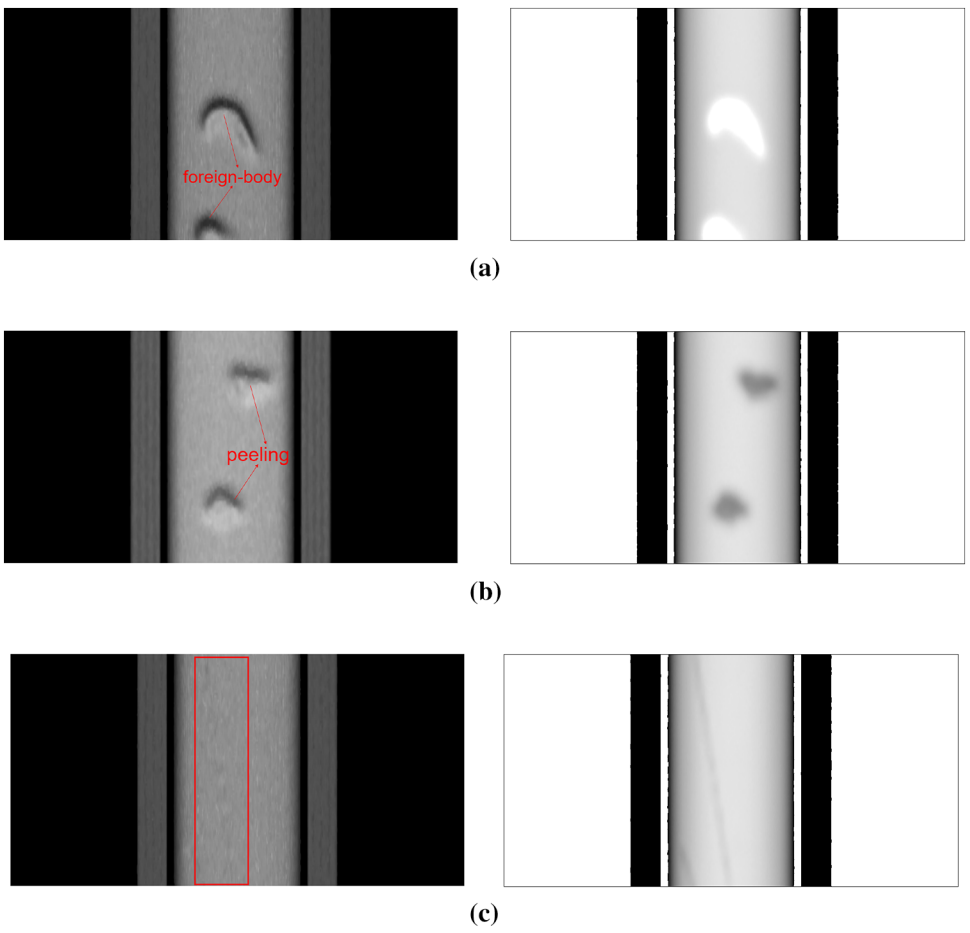


Fig. 1 In **a**, **b**, it shows the two "foreign-body" and "peeling" rail surface defects intensity images and corresponding depth images. We can see that different types of defects may resemble each other only from the perspective of two dimensions. However, from the perspective of three-dimensional depth, the difference is obvious. In **c**, for the defect that is not easy to see in the intensity image, it will be easier to identify in the depth image

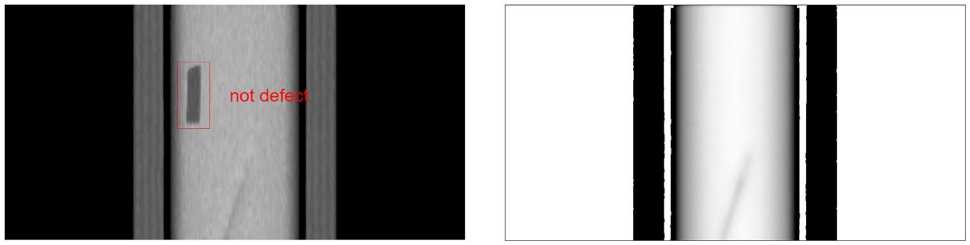


Fig. 2 Example of stains or oxide scale in intensity images but not exist in depth images

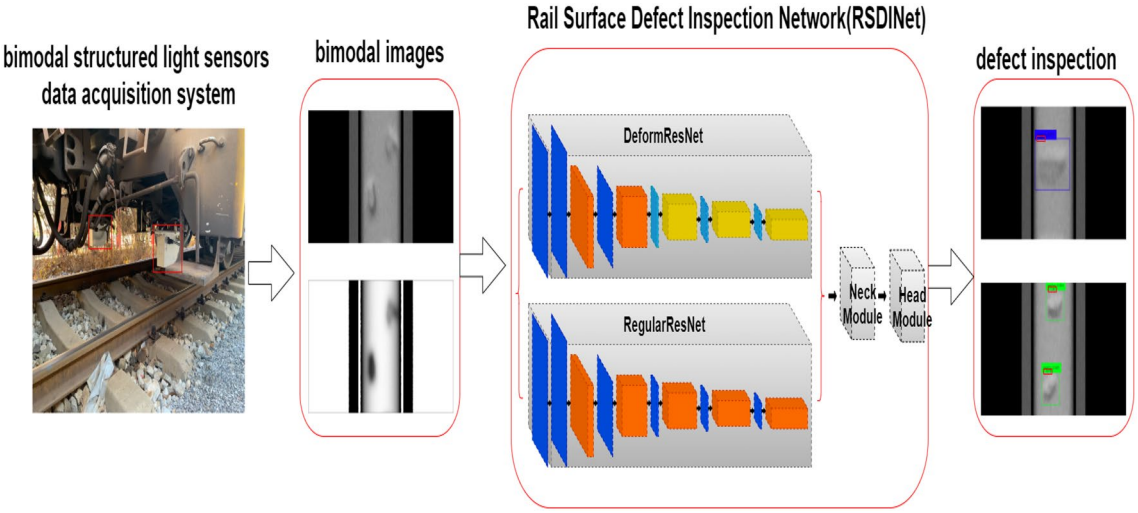


Fig. 3 The flow chart of rail surface data acquisition and defect inspection. In railway application, the rail surface data is collected by bimodal structured light sensors, and the bimodal images are obtained after processing. Then, the bimodal images are input into Rail Surface Defect Inspection Network (RSDINet) to get defect inspection results

How to effectively employ the three-dimensional information of rail surface defects is an urgent problem to be solved.

Compared with other defect detection systems like steel, copper and so on, detection process of other defect detection systems usually occur in the production process with relatively simple environmental conditions. Yet the inspection process of rail surface defect inspection system occurs in outdoor open air conditions, which may be sunny, rainy or even night. Besides, rails are in operation every day. As a result, it is very difficult to collect rail surface data, and the quality of the data may be greatly reduced due to the influence of lighting and other conditions. During the daily operation of the railway, some problems may cause the rail to be observed with defects in the intensity images, but not in depth images, as shown in Fig. 2. Therefore, the rail surface defect inspection system needs to have strong robustness and can distinguish false defects caused by external environment and daily operation.

Accordingly, in this paper, in order to solve the problems existing in existing researches and applications, we propose a rail surface defect inspection method based on bimodal image fusion, termed Rail Surface Defect Inspection Network (RSDINet), which can effectively employ 3D information to detect rail surface defects on the inspection vehicle using the rail surface data collected by constructed bimodal structured light sensors data acquisition system, and creatively measure the severity of rail surface defects end-to-end, so as to quantify the extent of rail damage (Fig. 3). The main contributions of this paper are summarized as follows:

(1) Aiming at the limitation of existing inspection methods based on only the two-dimensional rail information, based on bimodal structured light sensors, we apply a 3D camera to construct a bimodal image data acquisition system based on the principle of three-dimensional measurement of linear structured light to collect rail surface three-dimensional data. Our constructed bimodal image data acquisition system not only has the function of obtaining three-dimensional rail surface data, but also can reduce the impact of the external environment, so as to obtain high-quality rail surface data.
(2) Even if the bimodal images we get have good quality, traditional defect inspection methods still have the disadvantages of poor generalization ability and low robustness. However, DCNN-based methods generally

have better generalization ability and robustness, and are more suitable for railway application scenarios. Therefore, we propose a rail surface defect inspection method RSDINet based on deep convolutional neural network (DCNN) that achieves a satisfactory result based on the established limited number dataset. The inspection method proposes a new backbone network for performing feature extraction which applies deformable convolution to adapt to the irregularity of rail surface defects, and adopts a parallel feature extraction strategy to employ bimodal images meanwhile. Experiments show that the proposed method can effectively employ 3D information and has advantages of lower false alarm rate and higher accuracy, which can meet the specific criterion of inspection vehicle application.

(3) To verify the effectiveness of the proposed inspection method, we establish a new bimodal rail surface defect image dataset that contains gray-scale intensity images and corresponding depth images collected by the constructed system. The established dataset contains four categories of typical rail surface defects, through which we can research on rail surface defects from a three-dimensional perspective.

(4) A measurement approach of rail surface defect is proposed innovatively, which realize defect severity measurement through the end-to-end prediction on the basis of detection. Then, we prove the effectiveness of the method through a series of experiments, which shows a promising application prospect and considerable research significance.

This paper is organized as follows. In Sect. 2, we review some related work on rail defect inspection and DCNN. In Sect. 3, we describe the overview of bimodal structured light sensors data acquisition system for rail surface defect and analyze organization of the established dataset. In Sect. 4, we describe the architecture and operation of the proposed Rail Surface Defect Inspection Network (RSDINet) that can realize rail surface defect inspection, including detection and measurement. Section 5 presents the experimental results together with a comparison of different training and organization strategies and compares RSDINet with previous methods. Section 6 concludes the paper with a brief discussion.

2 Related work

In this section, we briefly review the related work about rail defect inspection and deep convolutional neural network (DCNN).

2.1 Rail defect inspection

In the past decades, knowledge-based methods are usually applied to defect inspection, which is a pattern of a feature extractor to obtain hand-crafted features and followed by a feature classifier. Recently, whether it is based on two-dimensional or three-dimensional method is also like that. For instance, Huber-Mörk et al. [3] applied Gabor filter groups to 2D surface texture description, and classes are modeled by Gaussian mixtures. Then, Bayesian classifier is used to differentiate between surface texture classes. In [2], a real-time visual inspection system (VIS) for discrete surface defects is proposed by Li and Ren. VIS acquires a rail image by the image acquisition system, cuts the sub-image of rail track by the track extraction algorithm and enhances the contrast of rail image using local normalization (LN) method. Then, detects defects using defect localization based on projection profile which is called as DLBP by the authors. In addition, Li and Ren [5] presents an intelligent vision detection system (VDS) for discrete surface defects that contains the local Michelson-like contrast (MLC) measure to enhance rail images and the new automatic thresholding method-proportion emphasized maximum entropy (PEME) thresholding algorithm. Yu et al. [4] and Gan et al. [10] both applied coarse-to-fine strategy to inspection of rail surface defect, which uses coarse extractors to roughly locate defects in the track surface image, and then uses refined extractors to distinguish the true category of points with outliers. Xu [11] applied the technique of 3D detection base on structured light to surface defect of steel defect during production. Li and Wang et al. [16] put forward a method to obtain rail corrugation data by using three-dimensional structured light, and then use wavelet analysis to detect rail corrugation.

Since the creative work of [17], DCNN has been demonstrated to be robust to feature extraction and learning through the continuous development of researchers. DCNN has fine generalization and transferability so that there are some defect inspection methods based on DCNN which are applied to two-dimensional rail images. For example, Faghih-Roohi et al. [18] used a deep convolutional neural network solution to the analysis of image data for rail surface defect detection, which is obtained from many hours of automated video recordings. Shang et al. [19] proposed a novel two-stage pipeline method for rail defect detection by localizing and classifying rail images. Song et al. [20] achieved a real-time algorithm to realize rail surface defect detection based on YOLOv3. Moreover, in [21, 22], DCNNs have been used for steel surface defects, and achieved great results. With continuous in-depth researches on DCNN, the performance has been further improved.

2.2 Deep convolutional neural network

Over the past few years, Krizhevsky et al. [17] proposed AlexNet to classify 1.2 million ILSVRC images in 1000 classes, which uses sequential pipeline architecture. Subsequently, more and deeper neural networks have been developed, such as VGGNet, GoogleNet. Those achieved state-of-the-art (SOTA) results at that time. Recently, the most popular DCNN architecture may be residual architecture, instead, which is used as backbone network. The representation of residual architecture is ResNet which uses residual blocks to make networks deeper without overfitting and achieve better accuracy. In addition, residual architecture has been applied in various backbone networks.

Meanwhile, some researchers concentrate on how to use more effective methods to improve accuracy. Due to the lack of regular convolution, atrous convolution [23] is proposed to increase receptive field and capture multi-scale context information. Dai et al. [24] developed deformable convolution and deformable RoI pooling to solve the problem that the traditional method of fixed size is difficult to adapt to new and unknown deformation.

Recently, various object detection algorithms based on DCNN have been proposed and achieved good results in the vision benchmark [25]. They are mainly divided into two methods: one is based on region proposal, and another is based on regression. The most representative region-based object algorithm is the R-CNN [26] series, including fast R-CNN [27], faster R-CNN [28]. Faster R-CNN is superior in precision with an end-to-end framework which unifies region proposal generation and object classification network. As for regression algorithm, the most representative method are YOLO [29–32]series, SSD [33]. YOLO [29] and SSD [33] are both fast single-shot detection methods that directly divide the input image into grids and predicts bounding boxes and probabilities for each grid simultaneously, which then regressed to the groundtruth. Later, the author of YOLO and other researchers has proposed v2 [32], v3 [30] and v4 [31] editions, which offers great detection accuracy improvements compared to previous editions while keeping a very high detection speed.

Moreover, 3D object detection algorithm also achieved good results, according to the data is divided into three categories, namely monocular image, multi-view image, point cloud. The method based on monocular image, for example, [34], uses monocular camera to complete 3D object detection. A multi-view based method, for example, [35], can obtain a depth map using parallax obtained from images of different views. The method based on point cloud, such as [36, 37], is a more intuitive and accurate three-dimensional object detection method.

3 System and dataset overview

The section below describes details of the constructed data acquisition system based on bimodal structured light sensors that is installed under the comprehensive inspection vehicle to collect rail surface data in real time and organization of the established bimodal rail surface defect image dataset.

3.1 Data acquisition system overview

The dataset which we establish to verify the effectiveness of the proposed inspection method is collected based on the principle of linear structured light three-dimensional measurement. Perspective projection geometric model for line structured light profile measurement technology is shown in Fig. 4a. A set of line structured light contour measuring components is composed of a line laser, a lens and a camera. The line structured light is incident on the surface of the measured object, and is modulated into a light strip reflecting the contour information of the measured object. The laser section image of the measured object is obtained by the camera. Pixel coordinates of the center of the light bar are extracted from the image. Combining the pixel coordinates of the center of the light bar with the system calibration parameters, the actual contour of the measured object can be calculated. The contour data of the measured object can be obtained with equal spacing in conjunction with the scanning motion, thus realizing the three-dimensional measurement of the whole measured object.

Based on the principle of three-dimensional measurement of linear structured light, a 3D measurement system of linear structured light is composed of a 3D camera, a machine vision lens and a linear laser. As shown in Fig. 4b, the plane of the optical knife of the linear structured light is incident vertically on the surface of the rail, and the laser cross-section images of the rail and other components are acquired by the 3D camera. Then, we scan rail at equal spacing to obtain a series of rail contour data, and the rail contour data are arranged at equal spacing according to the actual sampling interval to obtain the three-dimensional point cloud data of the rail.

Finally, we will process the three-dimensional point cloud data to obtain bimodal images meanwhile, which contain the gray-scale intensity image and corresponding depth image of rail surface. The intensity images are rich in rail surface texture information, and the corresponding depth images contain three-dimensional information of rail surface.

3.2 Dataset organization

In total, the established bimodal rail surface defect image dataset contains 400 gray-scale intensity images and 400

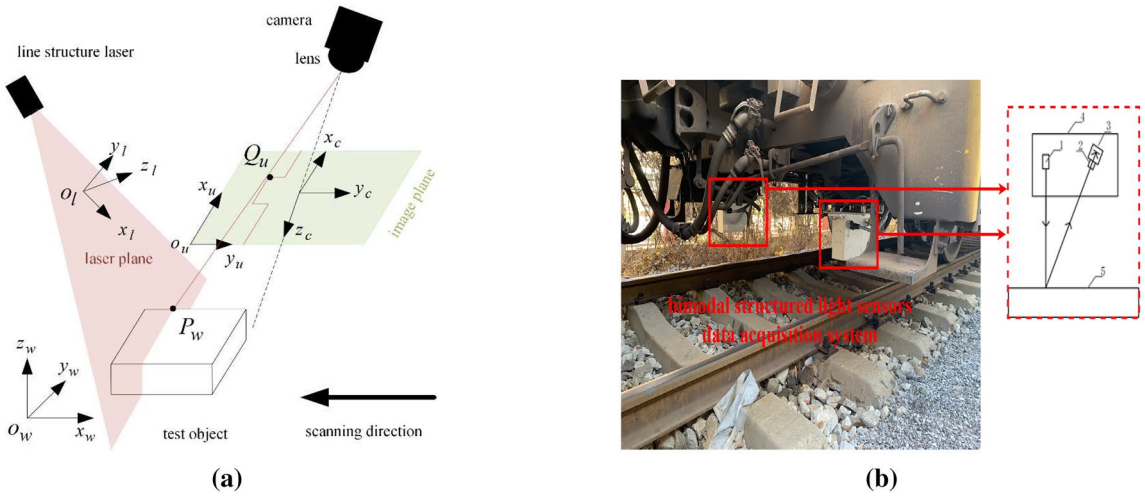


Fig. 4 **a** Schematic diagram of line structured light 3D measurement. **b** 3D measurement data acquisition system installed under the comprehensive inspection vehicle (1: line structure laser, 2: lens, 3: camera, 4: 3D measurement module, 5: rail)

corresponding depth images(range images) over four categories of typical defects (i.e., abrasion, foreign-body, scratch and peeling). The each depth image contains the depth information of rail surface, which is essentially the distance image of rail surface from the structured light source. The size of each pixel represents its distance value. And the size of images in our dataset is 1280×654 pixels. Some sample images of four categories of rail surface defects are shown in Fig. 5.

Moreover, to build the training set for defect detection and measurement tasks, we manually draw the ground truth bounding boxes and assign the depth labels and class labels of approximately these images. The bounding boxes and class labels indicate location information and categories of rail surface defect respectively.

Meanwhile, in order to measure severity of rail defect, we design depth labels based on the ground truth bounding boxes and depth information of depth images. We measure severity of rail surface defects by calculating the difference between the statistical depth of defects in the ground truth bounding boxes and the depth of normal rail surface. Code-level details are presented in Algorithm 1. Then, to perform defect detection and measurement tasks, we save these annotations in corresponding XML files.

Algorithm 1 *How to generate depth labels for measurement task*: According to depth image and corresponding bounding box of each rail surface defect in the image. $(x_n, y_n)_{n=1...4}$ represents 4 corners of each bounding box.

Input: input depth image: I ;
bounding boxes corner position: $(x_n, y_n)_{n=1...4}$;

Select max corner pixel: $Max\_corner = \max(I[x_n, y_n])$
Crop depth image based on bounding box: $Crop$
for each $pixel$ in $Crop$ **do**
    Select minimum or maximum 1% pixels in $Crop$: $Value\_list$
    (Only class **foreign-body** using maximum)
end for
Calculate $Value\_list$ mean value: $Mean\_value$
Calculate: $Diff = Max\_corner - Mean\_value$
Output: Defect severity: $Diff$

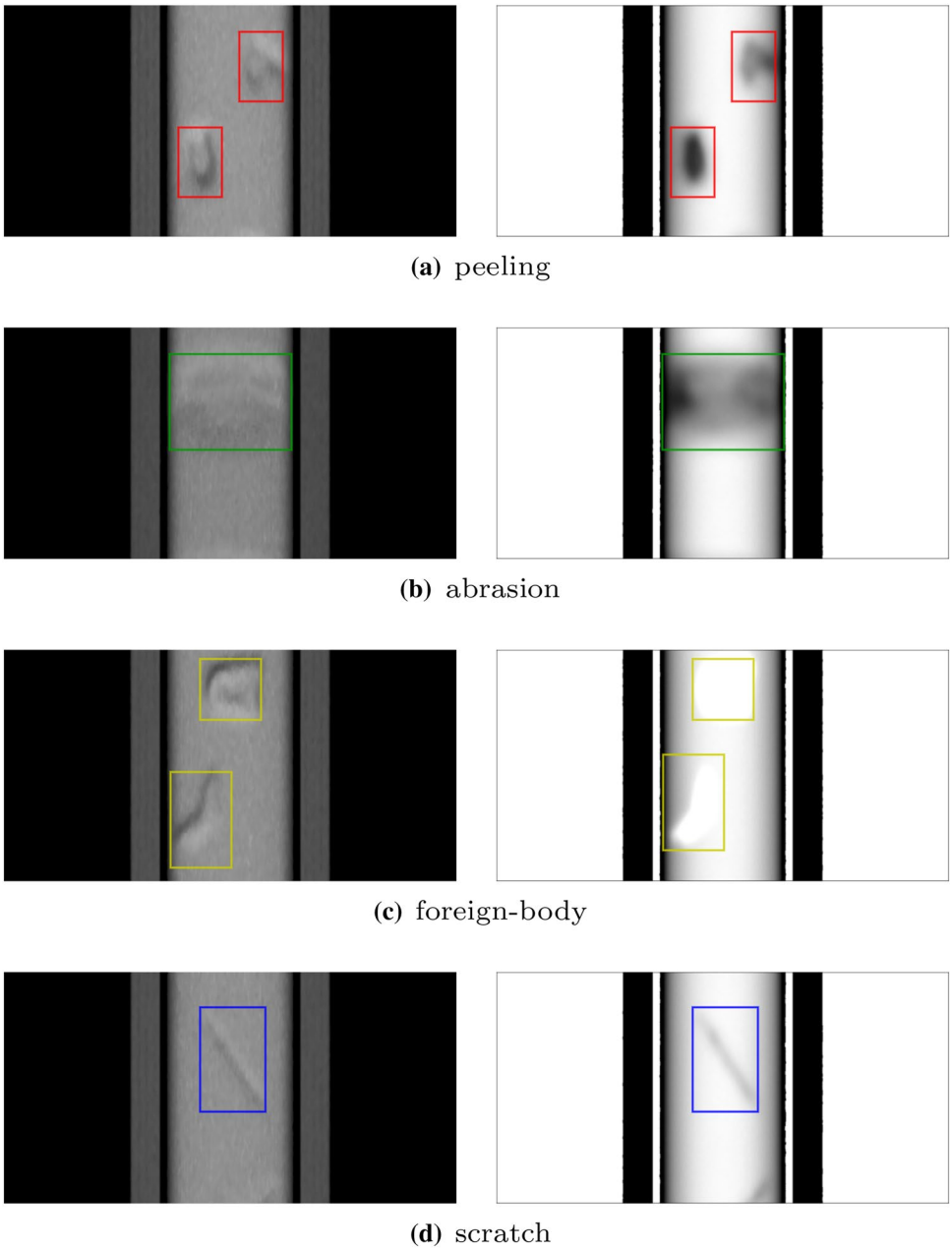


Fig. 5 Samples of various categories and sizes of defects in the rail surface defect data set (left: intensity image, right: depth image). **a** Peeling, **b** abrasion, **c** foreign-body, **d** scratch

4 Rail surface defect inspection network architecture

In this section, we explain the architecture of Rail Surface Defect Inspection Network (RSDINet) that we propose for rail surface defect inspection. The network is composed of three modules which can be called as backbone, neck and head.

4.1 Backbone: deformable convolution and residual block

As we all know, backbone network as the feature extractor is of great significance to achieve optimal performance of algorithms. It is found that deformable convolution can improve the generalization ability of model for irregular rail surface defects, and the residual block has been proved to be effective in improving deep network performance.

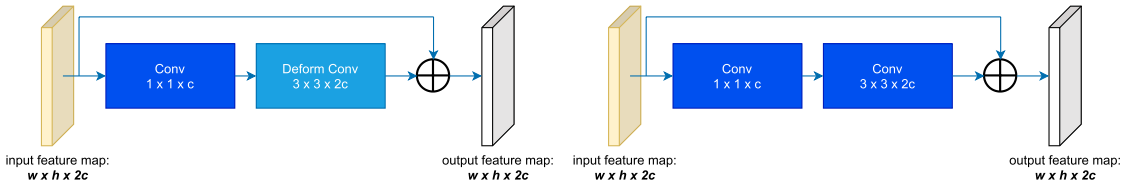
By using jump connection, residual block can increase the network depth and improve the accuracy, and at the same time, alleviate the gradient disappearance problem caused by the increase of network depth. So while designing our backbone for defect feature extraction, we also choose residual block.

Deformable convolution is designed to deal with the complex deformation of the objects in some special situation. Since the regular convolution kernel is that it has poor

adaptability to unknown changes and poor generalization ability. For objects with more complex deformations, this convolution may not work very well. However, deformable convolution introduces an offset in the convolution kernel, and the offset is learnable. The kernel of deformable convolution can adjust its shape according to the actual situation and better extract the characteristics of the input.Rail surface defects, as everyone knows, are random and irregular. This coincides with the characteristics of deformable convolution.

Inspired by these above factors and [30], we propose a new backbone network for performing feature extraction of rail surface defect.

The network input size is assumed to be 512×512, and the overall backbone network structure and detailed parameters shown in Fig. 7 and Table 1, called DeformResNet. DeformResNet is composed of 6 Conv layers and 5 Res blocks, including RegularConv, DeformConv, RegularRes and DeformRes, which are mainly composed of regular



(a) DeformRes **(b)** RegularRes

Fig. 6 Deformable residual block (DeformRes) and regular residual block (RegularRes) used in Rail Surface Defect Inspection Network. RegularRes contains 1 regular convolution with 1×1, 1 regular convolution with 3×3 and 1 shortcut layer, and each convolution layer is followed by batch normalization layer and LeakyReLU activation. DeformRes replaces the 3×3 regular convolution with deformable convolution

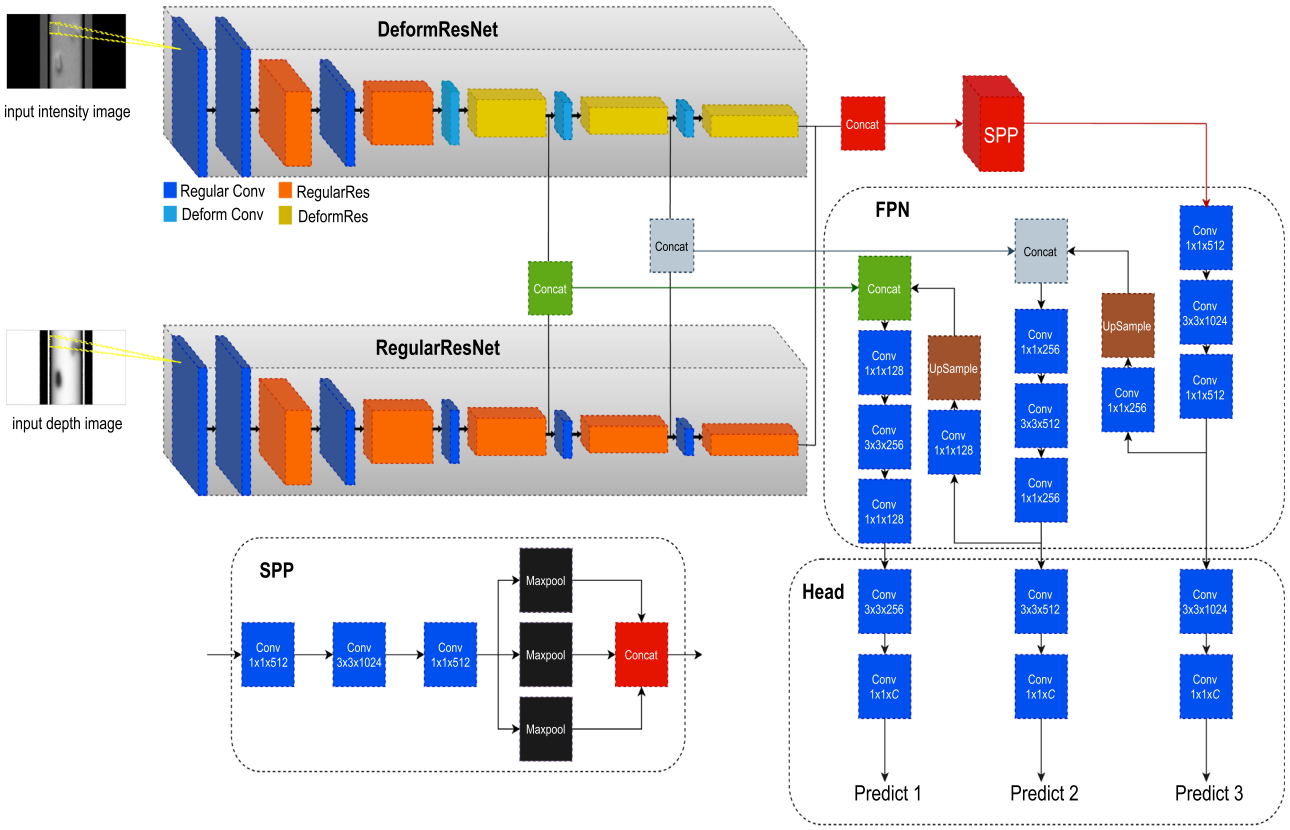


Fig. 7 The architecture of Rail Surface Defect Inspection Network. RSDINet adopts a parallel feature extraction strategy, which creates two different backbone networks to extract features of intensity image and depth image respectively. Then, output three-scale feature maps and concatenate respectively. And then through the neck module processing, head module predicts the bounding boxes, classification probability and severity of rail surface defect

Table 1 Detailed parameters of DeformResNet

	Type	Filters	Size	Output
Conv0	Conv	32	3×3	512×512
Conv1	Conv	64	$3 \times 3/2$	256×256
Res1(RegularRes $\times$1)	Conv	32	1×1	
	Conv	64	3×3	
	Residual			256×256
Conv2	Conv	128	$3 \times 3/2$	128×128
Res2(RegularRes $\times$1)	Conv	64	1×1	
	Conv	128	3×3	
	Residual			128×128
Conv3	Deform-Conv	256	$3 \times 3/2$	64×64
Res3(DeformRes $\times$ N)	Conv	128	1×1	
	Deform-Conv	256	3×3	
	Residual			64×64
Conv4	Deform-Conv	512	$3 \times 3/2$	32×32
Res4(DeformRes $\times$ N)	Conv	256	1×1	
	Deform-Conv	512	3×3	
	Residual			32×32
Conv5	Deform-Conv	1024	$3 \times 3/2$	16×16
Res5(DeformRes $\times$ N)	Conv	512	1×1	
	Deform-Conv	1024	3×3	
	Residual			16×16

convolution layers, shortcut layers, and deformable convolution layers. As shown in Fig. 6, RegularRes contains a 1×1 and 3×3 regular convolution layer and a shortcut layer, while DeformRes replaces the 3×3 regular convolution layer with a 3×3 deformable convolution layer. We add batch normalization layer and LeakyReLU activation on all of the regular and deformable convolution layers in DeformResNet to avoid overfitting and accelerate the convergence speed of network. The shortcut layer is similar to the shortcut layer of ResNet network, which can greatly reduce training difficulty and improve training accuracy. At the same time, the backbone network performs five (32-times) downsampling operations on the image by setting convolution stride to 2 as shown in Table 1, and finally outputs three-scale feature maps, that is, 8-times, 16-times, and 32-times downsampling feature maps.

Moreover, to better utilize bimodal-sensors data, we employ both intensity image and depth image as network input in RSDINet at the same time. So we adopt a parallel feature extraction strategy which creates two backbone networks to extract feature of the two inputs respectively. However, the depth image contains more rail depth information and less deformation feature information of rail defects. Therefore, we replace the deformable convolution in DeformResNet with regular convolution in depth image backbone network, called RegularResNet, as shown in Fig. 7.

4.2 Neck: enhance receptive field and multi-scale features fusion

In order to enhance receptive field and get multi-scale information, inspired by YOLOv4 [31], we adopt spatial pyramid pooling (SPP) to convolute parallel sampling of maximum downsampling feature map at different sampling rates. Meanwhile, referring to feature pyramid networks (FPN), we simplify YOLOv3 [30]'s FPN framework to reduce the model parameters and computation without affecting the accuracy. Generally in principle, FPN adopts top-down processing method to fuse upsampled high-level features and low-level features, which contain the better structural information of low-level features and stronger semantic information of high-level features at the same time. Therefore, we adopt FPN to encourage the reusing of low-level features which commonly contain more information about small defects which are easy to ignore, so that model can adapt to various sizes of rail surface defects and enhance the generalization of the model. Then, the output of FPN is three-scale feature maps with different receptive fields to define different head modules which can predict bounding boxes, classification probability and severity of rail surface defects with various sizes, as shown in Fig. 7.

4.3 Head: multi-scale prediction

After the feature extraction of backbone network and neck module, the multi-scale feature maps are directly fed into head module to detect defects of different sizes for adapting to the unpredictability of rail surface defect size. In order to further utilize depth information to achieve end-to-end measurement of defect severity, we mend the detection algorithm of YOLO to carry out defect detection and measurement at the same time.

When the input feature maps are given, the feature maps are divided into $S \times S$ grid cells according to the resolution after two convolution layers processing. Thus, each grid cell predicts classification probabilities for four-class rail surface defects and three candidate bounding boxes with the confidence score and defect severity. Each bounding box contains four position indicators, including the box coordinates (x, y, w, h), the corresponding defect severity s and the object confidence. Overall, the network output is a series of tensor of $S \times S \times 3 \times (4 + 1 + 1 + 4)$, as shown in Fig. 8.

The bounding box coordinates (x, y, w, h) and defect severity measurement s prediction formula is as follows:

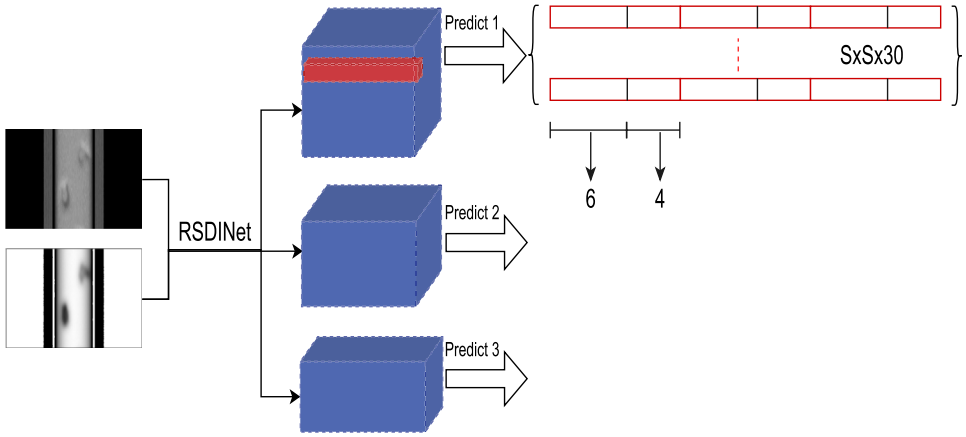


Fig. 8 The makeup of prediction output of Rail Surface Defect Inspection Network. The number "6" contains (x, y, w, h, s, c), which means bounding box coordinates, defect severity and object confidence. The number "4" represents the shape of a tensor determined by the amount of rail surface defect class, which contains probability value of each class, i.e. $p(c)$. $S \times S$ represents the size of the divided grid cell, which depends on the resolution of the input feature map

$$\begin{cases} b_x = \sigma(t_x) + c_x \\ b_y = \sigma(t_y) + c_y \\ b_w = p_w e^{t_w} \\ b_h = p_h e^{t_h} \\ \sigma(b_s) = t_s. \end{cases} \tag{1}$$

Among the formula, the parameter of network prediction is $(t_x, t_y, t_w, t_h, t_s)$, which represents the coordinates, width and height of the center point of the bounding box and severity of defect in the bounding box; (c_x, c_y) indicates the coordinate offset of the divided grid cells; (p_w, p_h) represents the width and height of bounding box prior; $(b_x, b_y, b_w, b_h, b_s)$ represents the position and size of the bounding box relative to the feature map and the true severity of rail surface defect.

RSDINet predicts an objectness score for each bounding box by using logistic regression referring to YOLOv3, as shown in formula 2. First, calculate the size of Intersection over Union (IoU), which represents how much each bounding box prior overlaps a ground truth object. Then, calculate the delta of true and predicted severity (DoS). The two factors contribute to the calculation of objectness score $P_r(object)$. This should be 1 if the bounding box prior's IoU greater than any other bounding box prior and the pre-set threshold (we set it to 0.5), and meanwhile the minimum DoS less than 0.5; otherwise, set it to 0. Therefore, if a bounding box prior is not assigned to a ground truth object it incurs no loss for coordinate, severity or class predictions, only calculate objectness loss.

$$P_r(object) = \begin{cases} 1, & \max(IoU) \geq threshold \\ & \bigcap \min(DoS) \leq 0.5 \\ 0, & otherwise. \end{cases} \tag{2}$$

During testing, RSDINet uses bounding box priors and the coordinate offsets of predicted bounding boxes to calculate candidate bounding boxes. Meanwhile, obtain the corresponding classification probability value and severity measurement value.Finally, after NMS algorithm screening, RSDINet outputs the optimal object bounding box.

During training, the parameters of RSDINet model are updated by minimizing the loss function defined in formula 3, which comprised of four parts that contains coordinate loss L_{coord}, severity loss $L_{severity}$, object loss L_{obj} and class loss L_{cls}. We apply sum of squared error loss to L_{coord} and $L_{severity}$ as shown in formulas 4 and 5. Meanwhile, we use sigmoid cross-entropy loss for object loss L_{obj} as shown in formula 6. Considering that this is a class mutually exclusive classification task, we use softmax cross-entropy loss for class loss L_{cls} as shown in formula 7.

$$Loss = L_{coord} + L_{severity} + L_{obj} + L_{cls} \tag{3}$$

$$\begin{aligned} L_{coord} = & \lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} 1_{i,j}^{obj} (2 - w_i \times h_i)[(x_i - \hat{x}_i)^2 \\ & + (y_i - \hat{y}_i)^2 + (w_i - \hat{w}_i)^2 + (h_i - \hat{h}_i)^2] \end{aligned} \tag{4}$$

$$L_{severity} = \lambda_{severity} \sum_{i=0}^{S^2} \sum_{j=0}^{B} 1_{i,j}^{obj} (s_i - \hat{s}_i)^2 \tag{5}$$

$$L_{obj} = -\left(\lambda_{obj}\sum_{i=0}^{S^2}\sum_{j=0}^{B}1_{i,j}^{obj}(\hat{c}_i log(c_i) + (1-\hat{c}_i)log(1-c_i)\right) + \lambda_{noobj}\sum_{i=0}^{S^2}\sum_{j=0}^{B}1_{i,j}^{noobj}(\hat{c}_i log(c_i) + (1-\hat{c}_i)log(1-c_i))) \tag{6}$$

$$L_{cls} = -\lambda_{cls}\sum_{i=0}^{S^2}\sum_{j=0}^{B}1_{i,j}^{obj}\sum_{c\in classes}(\hat{p}_i(c)log(p_i(c))). \tag{7}$$

Here, $x_i, y_i, w_i, h_i, s_i, c_i, p_i(c)$ represent the prediction of model that denotes the box center point coordinates, its width and height, the defect severity, the objectness confidence score and the class score of object. Variables $x, \hat{x}$ denote the prediction and the ground truth respectively (like y, w, h, s, c, $p(c)$). λ is an adjustable parameter to control the importance of each loss. 1^{obj}, 1^{noobj} is computed by the formula 2. S and B represent the number of grid cell and bounding box prior respectively.

5 Experiment and analysis

The above description explains the principle and feasibility of the proposed Rail Surface Defect Inspection Network (RSDINet). We conduct a series of detection and measurement experiments to evaluate the effect of RSDINet and compare RSDINet with previous methods.

5.1 Experimental setup

The compilation environment is Ubuntu 16.04 in the experiment, and we use NVIDIA RTX 2080Ti graphical processing unit (GPU) with 12-GB memory to accelerate training processing. The batch size is set to 8. We adopted Adam optimizer with a initial learning rate of 0.001, and trained a total of 200 epochs. The learning rate is used to control the rate of gradient descent of the training loss, and the training loss guides the training process. When the training loss of 5 consecutive epochs does not decrease, the learning rate decrease 10 times.

Table 2 Numbers of various defects in each set

	Train	Valid	Test
Abrasion	93	22	23
Foreign-body	137	30	24
Scratch	134	19	32
Peeling	141	31	30

The bimodal rail surface defect image dataset mentioned in Sect. 3.2 is employed to validate the proposed network. A total of 400 intensity and 400 depth corresponding images with the size of 1280×654 in the dataset were divided into training set, validation set and test set in the ratio of 0.7:0.15:0.15. Therefore, among these rail images, the training set contains 280 intensity and 280 depth corresponding images, 60 intensity and 60 depth corresponding images in the validation set and others in the test set. Meanwhile, each subset also contains the corresponding annotation XML files. Finally, the numbers of various defects contained in each subset are shown in Table 2.

Moreover, to enhance robustness, we train the model with multi-scale instead of fixing the input image size to 512×512, every 10 batches of training, randomly select a multiple of 32 from [416, 640] as the network input size of next 10 batches. Meanwhile, to enlarge the number of training samples and avoid overfitting, we also apply several data augmentation methods while training such as rotation, random crop and flip, which solve the problems of data imbalance and small amount of data.

5.2 Evaluation indicator

Mean average precision (mAP) and mean severity average precision (mSAP) are used as the evaluation indicators of the experiments to evaluate the accuracy of detection and measurement respectively.

The true positive (TP), false negative (FN), and false positive (FP) of each class are counted to compute the following statistical indicators precision and recall. Then, we calculate average precision (AP) by precision and recall. The average precision is computed according to the relationship P(R) of precision (P) and recall (R), which is a good tradeoff between the two significant detection indexes. These indexes are defined as follows:

$$Precision = \frac{TP}{TP+FP} \times 100\% \tag{8}$$

$$Recall = \frac{TP}{TP+FN} \times 100\% \tag{9}$$

$$AP = \int_0^1 P(R)dR. \tag{10}$$

For AP calculation, different object detection datasets have different calculation methods. We apply the method proposed in VOC2010 [25] to select the maximum precision value for each different recall value (including 0 and 1), and then calculate the area under the PR curve as the AP value.

Moreover, among all true positive(TP) results, we compute severity precision (Ps) of each class by means of the delta of predicted and true severity less than 0.5, and then calculate severity average precision(SAP) by severity precision and recall, as shown in formulas 11 and 12. Proposed SAP index comprehensively considers the severity precision and recall, which can effectively avoid the false excellent model effect caused by low recall rate and high severity precision, and can better characterize the effectiveness of the model.

$$P_s = \frac{TP_s}{TP_s + FP_s} \times 100\% \tag{11}$$

$$SAP = P_s \times R \times 100\%. \tag{12}$$

Finally, mAP and mSAP are average values of AP and SAP of each class, which are defined as follows:

$$mAP = \frac{1}{N}\sum_{n=1}^{N} AP_n \tag{13}$$

$$mSAP = \frac{1}{N}\sum_{n=1}^{N} SAP_n. \tag{14}$$

5.3 Experiments results and analysis

To verify the effectiveness of the proposed rail surface defect inspection method, a series of experiments are carried out to evaluate the proposed method in terms of mAP, mSAP and the processing time costs (frames/s, FPS). Experiments employ the same hyper-parameter as the default setting in Sect. 5.1 and test in a single GPU environment. Figure 9 shows several examples of visualized detection and severity measurement results.

5.3.1 Effects of the proposed DCNN architecture

The proposed DCNN architecture's backbone can be scaled to adapt to different application requirements, according to Table 1, which means different numbers of residual blocks at different scales. We design three scales backbone networks, including light, medium and large, which differed from each other by the number of residual blocks in Res3, Res4, Res5. In the comparative experiment, the three scales DCNN architectures are trained and tested on the same dataset.

Table 3 shows the results of this experiment. We can observe that the "large" achieves a best mAP of 87.17 while the mSAP is also the highest at 39.07. As a whole, the mAP and mSAP indicators of "large", "medium" and "light" both show a downward trend. This result demonstrates that with

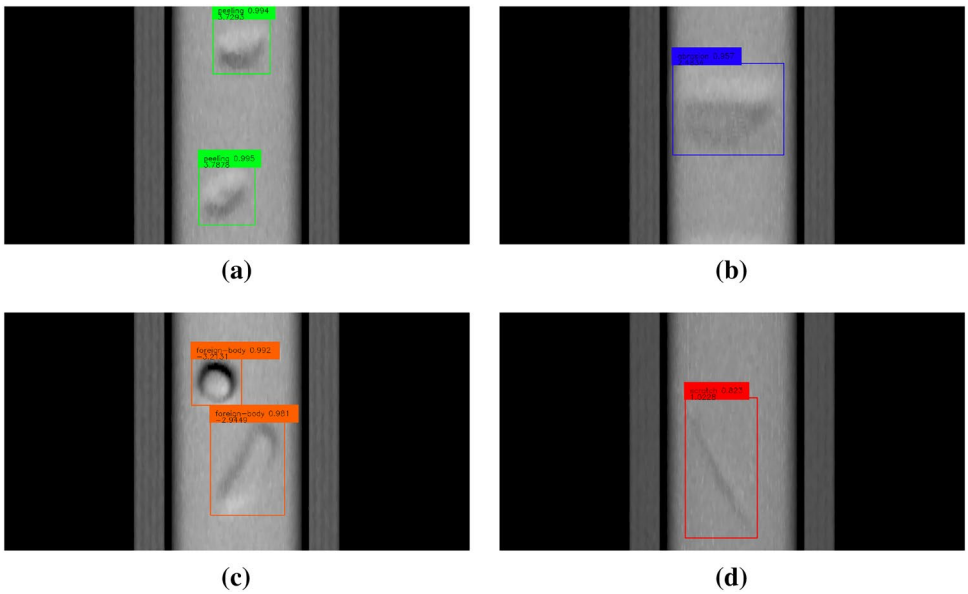


Fig. 9 Examples of detection and predicted severity results shown on intensity images in bimodal rail surface defect image dataset. For each defect, the box with color is the bounding box indicating its location and the information in the upper-left corner contains defect label, class score and severity measurement result. The subset to which the image belongs **a** peeling, **b** abrasion, **c** foreign-body, **d** scratch

Table 3 Comparison of the three scales DCNN architectures

Configuration	Numbers of residual block contained					mAP	mSAP	FPS	Parameters
	Res1	Res2	Res3	Res4	Res5				(million)
Light	1	1	1	1	1	81.79	32.69	9.1	60.0
Medium	1	1	2	2	1	85.49	35.09	8.0	64.8
Large	1	1	4	4	2	**87.17**	**39.07**	6.2	89.5

Bold represents the best results of different experiments under the same indicator

the increase of network depth, it is beneficial to both the prediction accuracy of defect severity measurement and the localization accuracy of model. However, compared with the other two network architectures, the "large" has the lowest FPS according to Table 3. This is mainly due to the use of deformable convolution in Res3, Res4, and Res5 of DeformResNet, which is applied to process intensity image input in the proposed method. Deformable convolution has a larger number of parameters than regular convolution because it adds additional direction parameters to each element of convolution kernel to learn how to deform convolution kernel. While the "large" employs more deformable convolution layers and more residual blocks in Res3, Res4 and Res5, so it has more model parameters than the other two network architectures, resulting in an increase in image processing time and a decrease in FPS.

In summary, the "large" network decreases the speed of the task compared to the other two to a certain extent, but achieves a better model accuracy.

5.3.2 Effectiveness of single deformable convolution

As mentioned above, the proposed RSDINet adopt a parallel feature extraction strategy to extract feature respectively, which applies deformable convolution to intensity image, but not depth image(Y-N). To verify the effectiveness of the method, therefore, we compare the method with both intensity image and depth image with no deformable convolution(N-N) and both intensity image and depth image with deformable convolution(Y-Y) by the proposed "large" DCNN architecture. The statistical results are summarized in Table 4.

According to Table 4, the N-N mode only achieves a mAP of 78.24 and a mSAP of 25.18, 8.93 in mAP and 13.89 in mSAP lower than the Y-N mode respectively. Meanwhile, mAP and mSAP of the Y-Y mode is 85.85 and 33.64, which is 1.32 and 5.43 off the Y-N mode. Besides, the N-N mode apparently outperforms the Y-N mode in speed according to Table 4. This is not surprising as the use of deformable convolution will increase the amount of parameters, which will lead to the decrease of speed. As a result of the increasing use of deformable convolution, FPS of the N-N, Y-N and Y-Y modes decreases gradually.

Table 4 Network with deformable convolution comparison

Network	N-N	Y-N	Y-Y
mAP	78.24	**87.17**	85.85
mSAP	25.18	**39.07**	33.64
FPS	30.8	6.2	4.1

Bold represents the best results of different experiments under the same indicator

However, even if the processing speed of Y-N mode is not optimal, the Y-N mode performs better both on mAP and mSAP than the backward two. The statistical results show that it get 11.4% and 1.5% improvement in mAP and 55.2% and 16.1% improvement in mSAP than the N-N mode and the Y-Y mode respectively.

Figure 10a lists the AP result of four-class defects which represents the localization accuracy of each class. We can observe that the Y-N mode is superior to the N-N mode in terms of all classes and shows a little poorer performance than the Y-Y mode in terms of "foreign-body" and "scratch". Figure 10b lists the SAP result of four-class defects. Plainly, the Y-N mode shows the best performance among the three modes. Moreover, it can be seen that no matter which mode we use, the SAP of "abrasion" is far lower than the others. We consider that this may be due to the large defect area of defect "abrasion" and will be our future research work.

Comparison of three-mode results is shown in Fig. 11. We can observe that the Y-N mode can correctly detect most defects while some of the defects are missed and considered as false class through the other two modes, even if the Y-N mode still fails to find the whole defects existing.

Overall, the proposed DCNN architecture with Y-N mode is effective to improve the accuracy of defect severity measurement and the localization accuracy at the expense of some speed. Through the previous defect localization and severity measurement experiments, it is proven that deformable convolution contributes to improve feature extraction on intensity images, so as to improve the localization accuracy, and moreover, has a negative impact on defect severity measurement while being applied on depth images. That means that the parallel feature extraction strategy with the Y-N mode is reasonable, effective and optimal, and verifies our view of deformation convolution in Sect. 4.1.

5.3.3 Comparison with previous methods

To better understand the performance of the proposed method and verify the superiority of our design, we compare with other methods, DCNN method in [18], YOLOv3 [30] and YOLOv4 [31], based on the established rail surface defect data set. Based on the above algorithms, we experiment on intensity images and depth images respectively to compare the localization effect of the proposed RSDINet with the "large" DCNN architecture and the parallel feature extraction strategy (RSDINet-parallel). Moreover, we also test the effect of the RSDINet with single processing strategy using DeformResNet backbone (RSDINet-single) on intensity images and depth images respectively.

Detection method To better measure the performance of different methods, we also consider *F-score* evaluation

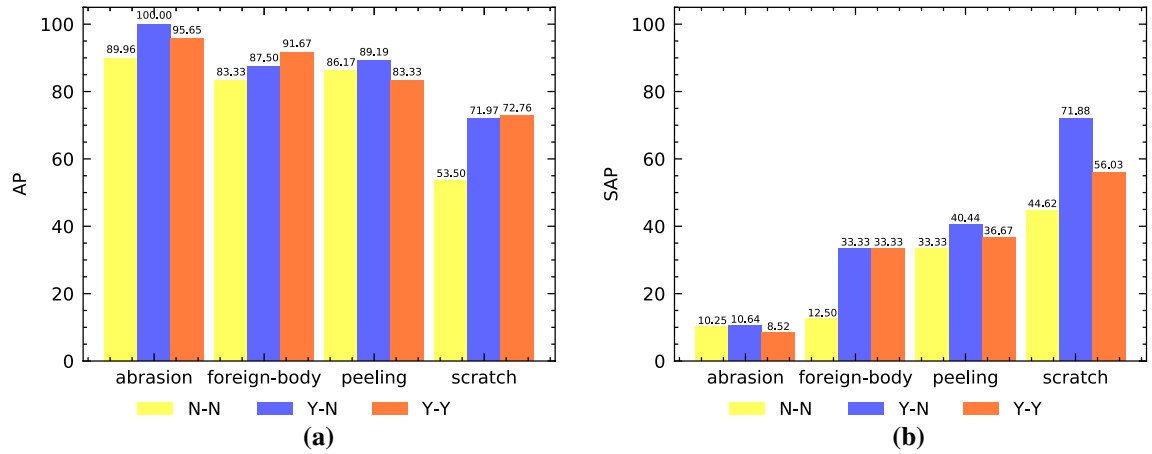


Fig. 10 Effect analysis of single intensity image with deformable convolution on the proposed "large" DCNN architecture. "N-N": both intensity image and depth image with no deformable convolution; "Y-N": single intensity image with deformable convolution; "Y-Y": both intensity image and depth image with deformable convolution

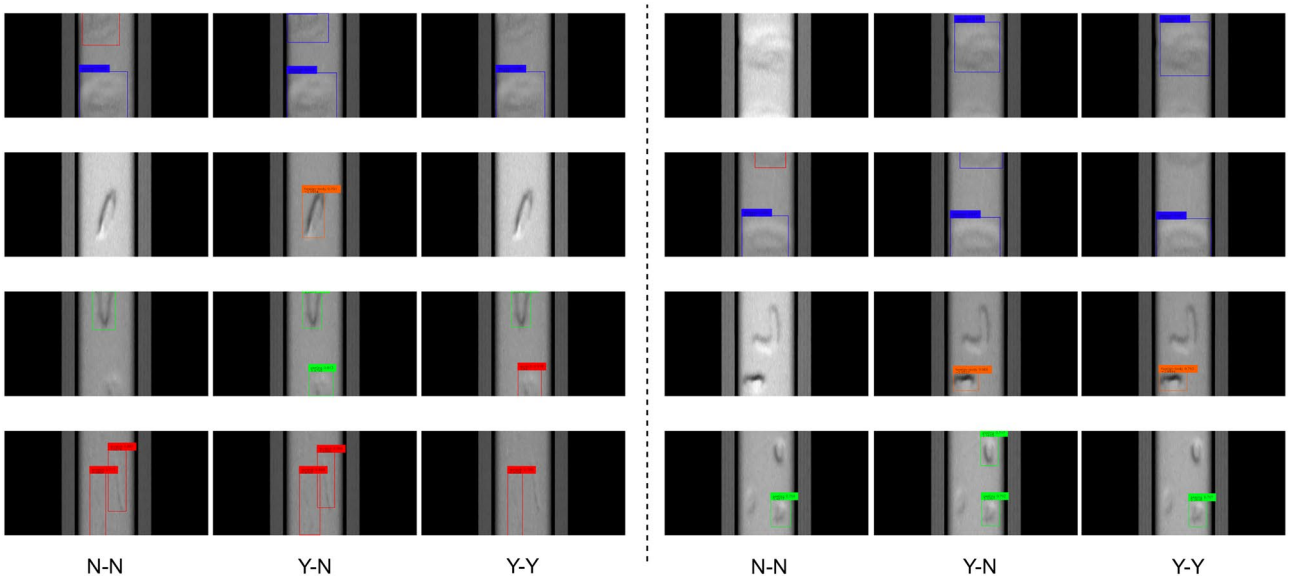


Fig. 11 Comparison examples of three-mode visualized results. Three pictures per line show the results of the same input in three modes

Table 5 Comparison results with previous detection methods

Serial Number	Methods	Precision	Recall	F_1-score	F_2-score	mAP	FPS
1	RSDINet-parallel-Bimodal	**75.49**	**88.96**	**81.67**	**85.89**	**87.17**	6.2
2	YOLOv3-Intensity [20, 30]	45.06	60.62	51.45	56.71	52.31	37.3
3	YOLOv4-Intensity [31]	68.21	82.76	74.78	79.37	79.67	25.9
4	RSDINet-single-Intensity	69.69	87.19	77.46	83.02	84.73	6.7
5	YOLOv3-Depth [20, 30]	64.72	58.11	61.24	59.32	53.09	35.5
6	YOLOv4-Depth [31]	55.25	53.65	54.43	53.95	50.38	23.2
7	RSDINet-single-Depth	64.66	49.86	56.30	52.25	44.57	6.5

• YOLOv4 [31] is the first application in the field of rail surface defect inspection which we firstly propose
Bold represents the best results of different experiments under the same indicator

indicator in this experiment, as shown in formula 15. There is an inverse relationship between precision and recall. The higher *F-score* is, the method is better. When $\beta = 1$, *F-score* is harmonic mean of precision and recall. If β is greater than 1, recall is more important than precision, such as $\beta = 2$, which meet the requirement of higher recall rate in defect detection field.

$$F_\beta = (1 + \beta^2) * \frac{Precision * Recall}{\beta^2 Precision + Recall}. \quad (15)$$

• *Comparison with Original YOLO Methods:* As shown in Table 5, the results are summarized. Combining all of the experimental results, our method with parallel feature extraction strategy with the Y-N mode shows the best performance than previous SOTA methods and single processing strategy methods in terms of *Precision, Recall,* F_1*-score,* F_2 *-score and mAP*.

Meanwhile, through the contrast experiments of intensity images, we can see that our RSDINet-single method get 58.1% improvement than YOLOv3 [30] and 6.3% improvement than YOLOv4 [31]. On the other hand, the performance of our RSDINet-single method lags behind the previous methods on depth images, which also proves our previous view in Sect. 4.1 and experimental results of the previous section that the effect of deformation convolution on depth images is not obvious or even has a negative effect.

Moreover, according to Table 5, we can make out that our RSDINet-parallel method doesn't decrease much in speed by comparing with our RSDINet-single methods, but has a great improvement in mAP. So it proves that the parallel feature extraction strategy with the Y-N mode is superior to the usual single processing strategy.

• *Comparison with other DCNN Methods:* Other existing DCNN methods ,such as VGG16,ResNet50 and DenseNet, are used for common object detection in common scenes. However, our task is to inspect rail surface defects commendably in our special our bimodal images. The following Table 6 summarizes the experimental comparison results of DCNNs replacing our proposed DeformResnNet in RSDINet-single.

Table 6 Comparison with other DCNN methods

Backbone	Precision	Recall	mAP
VGG16	48.24	55.49	50.82
ResNet50	61.01	81.91	78.09
DenseNet121	64.62	83.12	79.01
GoogleNet	68.20	85.44	83.60
DeformResNet(ours)	**69.69**	**87.19**	**84.73**

Bold represents the best results of different experiments under the same indicator

Table 7 Comparison with previous defect detection method

Methods	F_1-score	F_2-score	mAP
DDN in [21]	74.75	72.68	75.21
RSDINet	**81.67**	**85.89**	**87.17**

Bold represents the best results of different experiments under the same indicator

As shown in Table 6, we test the detection results of the above-mentioned DCNNs as the backbone network only on the intensity images. As shown in Table 5, intensity image is very important for detecting defects. Based on its detection effect, we can prove the superiority of different DCNNs.

Finally, from Table 6, it can be seen that the performance of our DeformResNet is superior to other DCNN methods in all indicators.

• *Comparison with Previous Defect Detection Method:* Defect Detection Network(DDN) in [21] proposes a defect detection method in steel surface defect detection field. We train DDN [21] model and test based on our dataset.

Table 7 shows the comparison result with DDN in [21]. We can obviously see that RSDINet outperforms DDN [21] in all F_1*-score,* F_2*-score and mAP*.

Combining results of the above experiments, it is proven that our RSDINet's special design for rail surface defect detection has better performance than previous outstanding detection methods.

Classification method The DCNN method in [18] is essentially a classification method. So we train the model and test based on our dataset in method [18]. And we compare the accuracy results of different categories in method [18] with our RSDINet method.

Table 8 shows the accuracy comparison results. We can obviously see that RSDINet has much better performance that the method in [18], especially "foreign-body" and "scratches". As shown in Fig. 12, experiments find that the method in [18] is difficult to distinguish the "scratch" defects from normal rail surface. Moreover, it is easy to misjudge the "foreign-body" and "peeling" because of the similarity between the two categories like Fig. 1a, b.

Table 8 Accuracy comparison with previous classification method (%)

Methods	Abrasion	Foreign-body	Peeling	Scratch
Method in [18]	80.00	25.00	81.25	5.88
RSDINet	**93.33**	**91.67**	**93.75**	**64.70**

Bold represents the best results of different experiments under the same indicator

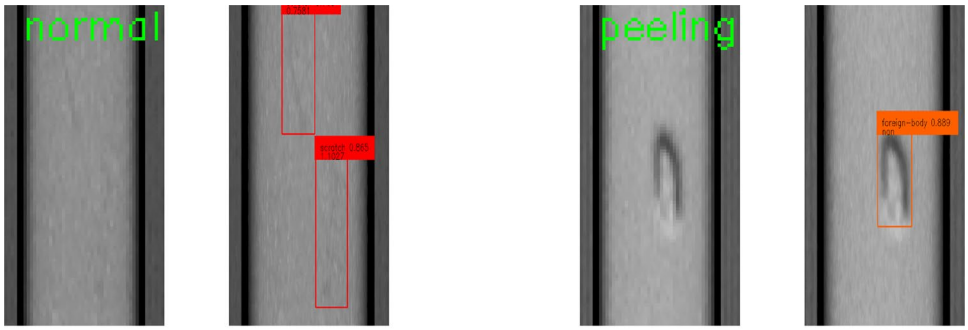


Fig. 12 Comparison results examples of method in [18] (left) and RSDINet(right)

Table 9 Improvement analysis of tricks

Methods	Parallel	DeformConv-Intensity	DeformConv-Depth	mAP	improvement
Best-Intensity	×	×	×	79.67	+0
Best-Depth	×	×	×	53.09	−26.08
RSDINet-single	×	✓	×	84.73	+5.06
RSDINet-single	×	×	✓	44.57	−35.1
RSDINet(N-N)	✓	×	×	78.24	−1.43
RSDINet(Y-N)	✓	✓	×	**87.17**	**+7.5**
RSDINet(Y-Y)	✓	✓	✓	85.85	+6.18

Bold represents the best results of different experiments under thesame indicator

5.3.4 Improvement analysis of different tricks

Our proposed method RSDINet is based on the characteristics of bimodal data, and mainly adopts the tricks of the Y-N parallel feature processing strategy and DeformResNet backbone network designed by deformable convolution. To verify the performance improvement brought by different techniques, we have made statistics in Table 9.

The effectiveness comparison results of different tricks are shown in the Table 9. Compared with the effect of the original best YOLO algorithm only on intensity images, our RSDINet-single with DeformResNet backbone brings 5.06 mAP improvement. In addition, the Y-N parallel feature processing strategy brings extra 2.44 mAP improvement.

Through the above experimental comparison, in summary, our RSDINet method using Y-N parallel processing strategy can effectively extract the features of different images, make full use of the differences of intensity and depth images, and achieve much better results than previous outstanding methods.

5.3.5 Error analysis of measurement result

As shown in Fig. 13, it shows the measurement results of severity of the detected rail surface defects in the test set. We can observe that the minimum defect severity measured is 0.73114 mm from Fig. 13a. On the other hand, Fig. 13b shows the error of severity measurement compared with the ground truth. It shows that the minimum error can achieve 0.0545 mm, which is much lower than 0.3 mm of the theoretical depth resolution parameter of sensors.

According to the empirical parameters set by analyzing the mean values of various defects of samples, we can see from Fig. 13c that the average error is about 0.80253 mm. Besides, the figure also shows the average error under different effect ratios, which varies within [95%, 70%], and most of errors are about 0.5 mm. Although the experimental results are not perfect due to the irregular rail surface defects, insufficient data set size and artificial errors in data annotation, experiments show that the method can detect most defects with controllable error.

Generally speaking, the end-to-end defect severity measurement method which we proposed is of great significance for quantifying the damage degree of rail surface defects, and provides a new approach for solving the measurement of rail surface defect severity.

6 Conclusion

We present a end-to-end rail surface defect inspection method for defect detection and measurement tasks based on bimodal image fusion. In terms of data acquisition, we construct a data acquisition system to collect rail surface

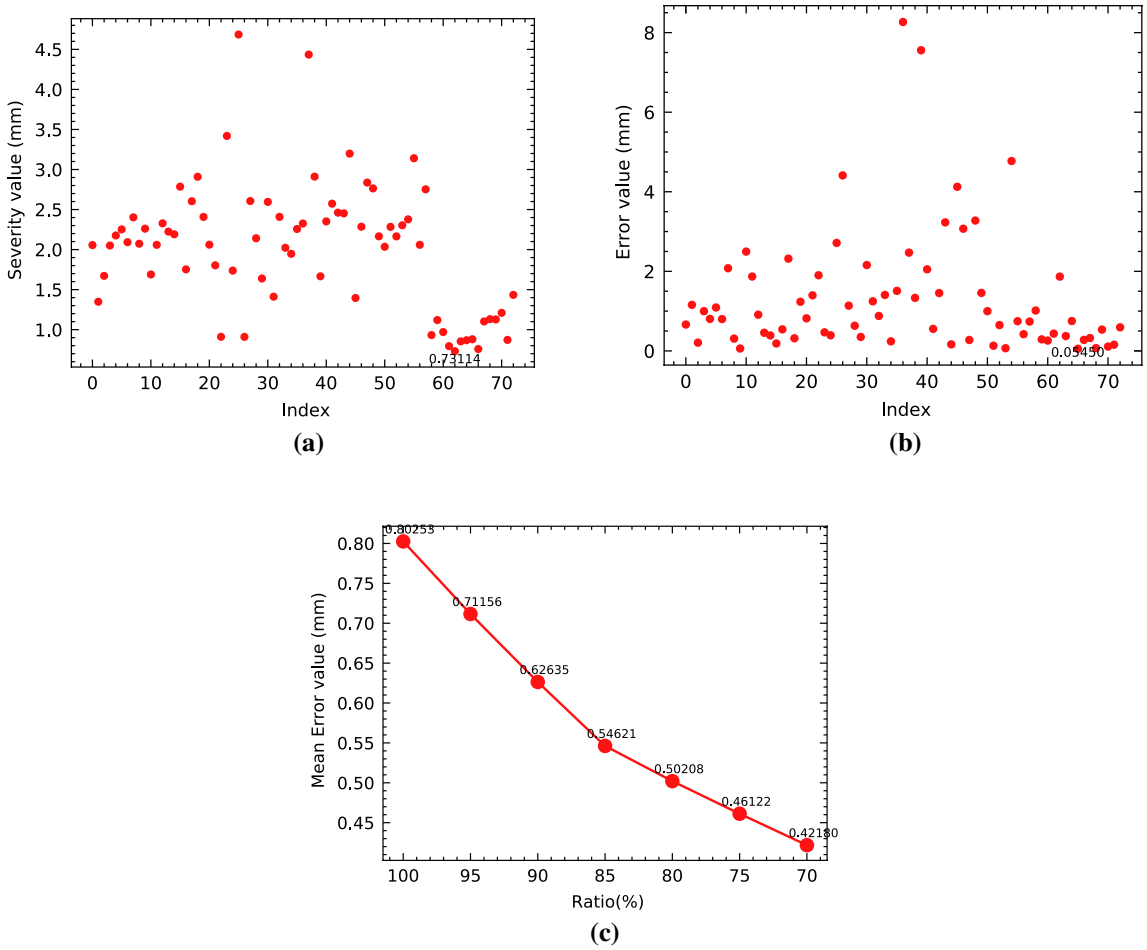


Fig. 13 Error analysis of measurement result. **a** Severity of prediction results; **b** error of severity compared with the ground truth; **c** the mean error under different effect ratios

three-dimensional data and establish a bimodal image dataset of defect intensity image and depth image based on the constructed system to verify the effectiveness of the proposed inspection method. We adopt a parallel feature extraction strategy to extract feature respectively and construct RSDINet for defect detection and measurement tasks. The proposed method RSDINet based on DCNN can obtain the category, detailed location and severity of a defect by collected rail surface data through the data acquisition system. A series of experiments demonstrate that RSDINet can achieve 87.17 mAP for detection task and 39.07 mSAP for measurement task, and in addition, RSDINet has a better performance than previous methods both in detection and classification. In terms of rail surface defect detection and measurement, our method has the advantages of low false alarm rate, high accuracy, and provide a new technical solution for reference. Overall, the proposed inspection method RSDINet shows a promising application prospect and can be in real time to run on a inspection vehicle.

In the future, we will test our method on the high-speed inspection vehicle, and improve the sensors and algorithm model according to the test results of actual lines.

Acknowledgements Work described in this paper was supported by National Natural Science Foundation of China—China National Railway Group Co., LTD. High-speed Railway Basic Research Fund under Grant no. U1934215, Research and Development Plan of China Academy of Railway Sciences Co. LTD under Grant no. 2021IMXM04.

Data availability Some or all data, models, or code generated or used during the study are available from the corresponding author by request (Shengchun Wang).

References

1. Dou Y, Huang Y, Li Q, Luo S (2014) A fast template matching-based algorithm for railway bolts detection. Int J Mach Learn Cybern 5(6):835–844
2. Li Q, Ren S (2012) A real-time visual inspection system for discrete surface defects of rail heads. IEEE Trans Instrum Meas 61(8):2189–2199

3. Huber-Mörk R, Nölle M, Oberhauser A, Fischmeister E (2010) Statistical rail surface classification based on 2d and 21/2d image analysis. In: International conference on advanced concepts for intelligent vision systems. Springer, Berlin, pp 50–61
4. Yu H, Li Q, Tan Y, Gan J, Wang J, Geng YA, Jia L (2019) A coarse-to-fine model for rail surface defect detection. IEEE Trans Instrum Meas 68(3):656–666
5. Li Q, Ren S (2012) A visual detection system for rail surface defects. IEEE Trans Syst Man Cybern Part C (Appl Rev) 42(6):1531–1542
6. Molodova M, Li Z, Núñez A, Dollevoet R (2014) Automatic detection of squats in railway infrastructure. IEEE Trans Intell Transp Syst 15(5):1980–1990
7. Wang J, Li Q, Gan J, Yu H, Yang X (2020) Surface defect detection via entity sparsity pursuit with intrinsic priors. IEEE Trans Ind Inform 16(1):141–150
8. Chen L, Liang Y, Wang K (2010) Inspection of rail surface defect based on machine vision system. In: The 2nd international conference on information science and engineering. IEEE, Hangzhou, China, pp 3793–3796
9. Molodova M, Li Z, Nunez A, Dollevoet R (2013) Monitoring the railway infrastructure: detection of surface defects using wavelets. In: International IEEE conference on intelligent transportation systems. IEEE, The Hague, pp 1316–1321
10. Gan J, Li Q, Wang J, Yu H (2017) A hierarchical extractor-based visual rail surface inspection system. IEEE Sens J 17(23):7935–7944
11. Xu K (2010) 3D detection technique of surface defects for steel rails based on linear lasers. Chin J Mech Eng. https://doi.org/10.3901/JME.2010.08.001
12. Ke X, Zhou P, Hu C (2012) 3D detection technique of surface defects for heavy rail based on binocular stereo vision. Proc SPIE Int Soc Opt Eng 8417:07
13. Ren S, Li Q, Xu G, Han Q, Feng Q (2011) Research on robust fast algorithm of rail surface defect detection. Zhongguo Tiedao Kexue/China Railw Fence 32(1):25–29
14. Zhao HW, Huang YP, Wang SC, Qing-Yong LI (2014) Rail surface defect detection algorithm based on spatial filtering. Comput Sci 41(1):130–137
15. Gao JQ, Liu GH (2017) 3D defect detection technology for rail surface with multi-camera line structure light. Mach Des Manuf 3:170–172
16. Li P, Wang P, Chen P, Xu H (2018) Rail corrugation detection based on 3D structured light and wavelet analysis. Railw Stand Des 62(8):33–38
17. Krizhevsky A, Sutskever I, Hinton G (2017) Imagenet classification with deep convolutional neural networks. Commun ACM 60(6):84–90
18. Faghih-Roohi S, Siamak H, Núez A, Babuska R, Schutter BD (2016) Deep convolutional neural networks for detection of rail surface defects. In: International joint conference on neural networks (IJCNN 2016). IEEE, Vancouver, pp 2584–2589
19. Shang L, Yang Q, Wang J, Li S, Lei W (2018) Detection of rail surface defects based on CNN image recognition and classification. In: 2018 20th International conference on advanced communication technology (ICACT). IEEE, Chuncheon, pp 45–51
20. Song Y, Zhang H, Liu L, Zhang H (2019) Rail surface defect detection method based on YOLOv3 deep learning networks. In: 2018 Chinese automation congress (CAC). IEEE, Xi'an, pp 1563–1568
21. He Y, Song K, Meng Q, Yan Y (2020) An end-to-end steel surface defect detection approach via fusing multiple hierarchical features. IEEE Trans Instrum Meas 69(4):1493–1504
22. Yi L, Li G, Jiang M (2017) An end-to-end steel strip surface defects recognition system based on convolutional neural networks. Steel Res Int 88(2):176–187
23. Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL (2018) DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans Pattern Anal Mach Intell 40(4):834–848
24. Dai J, Qi H, Xiong Y, Li Y, Zhang G, Hu H, Wei Y (2017) Deformable convolutional networks. In: 2017 IEEE international conference on computer vision (ICCV). IEEE, Venice, pp 764–773
25. Everingham M, Gool LV, Williams CKI, Zisserman WA (2010) The Pascal visual object classes (VOC) challenge. Int J Comput Vis 88(2):303–338
26. Girshick R, Donahue J, Darrell T, Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: 2014 IEEE conference on computer vision and pattern recognition. IEEE, Columbus, pp 580–587
27. Girshick R (2015) Fast R-CNN. In: 2015 IEEE international conference on computer vision (ICCV). IEEE, Santiago, pp 1440–1448
28. Ren S, He K, Girshick R, Sun J (2017) Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell 39(6):1137–1149
29. Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: unified, real-time object detection. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR). IEEE, Las Vegas, pp 779–788
30. Redmon J, Farhadi A (2018) Yolov3: an incremental improvement. arXiv e-prints. arXiv:1804.02767 [cs.CV]
31. Bochkovskiy A, Wang CY, Liao HYM (2020) Yolov4: optimal speed and accuracy of object detection. arXiv e-prints. arXiv:2004.10934 [cs.CV]
32. Redmon J, Farhadi A (2017) Yolo9000: better, faster, stronger. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR). IEEE, Honolulu, pp 6517–6525
33. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, Berg AC (2016) SSD: single shot multibox detector. In: Computer vision—ECCV 2016. Springer, Cham, pp 21–23
34. Li B, Ouyang W, Sheng L, Zeng X, Wang X (2019) Gs3d: an efficient 3d object detection framework for autonomous driving. In: 2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR). IEEE, Long Beach
35. Chen X, Kundu K, Zhu Y, Ma H, Fidler S, Urtasun R (2018) 3D object proposals using stereo imagery for accurate object class detection. IEEE Trans Pattern Anal Mach Intell 40(5):1259–1272
36. Qi CR, Liu W, Wu C, Su H, Guibas LJ (2018) Frustum pointnets for 3D object detection from RGB-D data. In: 2018 IEEE/CVF conference on computer vision and pattern recognition (CVPR). IEEE, Salt Lake City
37. Zhou Y, Tuzel O (2018) Voxelnet: End-to-end learning for point cloud based 3D object detection. In: 2018 IEEE/CVF conference on computer vision and pattern recognition (CVPR). IEEE, Salt Lake City