Attention Network for Rail Surface Defect Detection via Consistency of Intersection-over-Union(IoU)-Guided Center-Point Estimation

Xuefeng Ni, Ziji Ma, Jianwei Liu, Bo Shi, and Hongli Liu

***Abstract*—Rail surface defect inspection based on machine vision faces challenges against the complex background with interference and severe data imbalance. To meet these challenges, in this article, we regard defect detection as a key-point estimation problem and present the proposed attention neural network for rail surface defect detection via consistency of Intersection-over-Union(IoU)-guided center-point estimation (CCEANN). The CCEANN contains two crucial components. The two components are the stacked attention Hourglass backbone via cross-stage fusion of multiscale features (CSFA-Hourglass) and the CASIoU-guided center-point estimation head module (CASIoU-CEHM). Furthermore, the CASIoU-guided center-point estimation head module integrating the delicate coordinate compensation mechanism regresses detection boxes flexibly to adapt to defects' large-scale variation, in which the proposed CASIoU loss, a loss regressing the consistency of intersection-over-union (IoU), central-point distance, area ratio, and scale ratio between the targeted defect and the predicted defect, achieves higher regression accuracy than state-of-the-art IoU-based losses. The experiments demonstrate that the CCEANN outperforms competitive deep learning-based methods in four surface defect datasets.**

***Index Terms*—Attention, convolutional neural network, key-point estimation, multilevel feature fusion, rail surface defects.**

Manuscript received February 28, 2021; revised May 8, 2021; accepted May 24, 2021. Date of publication June 1, 2021; date of current version December 6, 2021. This work was supported in part by the National Nature Science Foundation of China under Grant 61771191 and Grant 61971182, in part by the Hunan Provincial Natural Science Foundation of China under Grant 2020JJ4213, and in part by Changsha City Science and Technology Department Funds under Grant KQ2004007. Code will be [Online]. Available: https://github.com/Xuefeng-Ni/CCEANN. Paper no. TII-21-1007. *(Corresponding author: Hongli Liu.)*

The authors are with the College of Electrical and Information Engineering, Hunan University, Changsha 410082, China (e-mail: nixuefeng@hnu.edu.cn; zijima@hnu.edu.cn; ljw1990@hnu.edu.cn; boshi@hnu.edu.cn; hongliliu@hnu.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TII.2021.3085848.

Digital Object Identifier 10.1109/TII.2021.3085848

I. Introduction

RAILWAY maintenance is of great importance in guaranteeing passengers' safety and the efficiency of rail transportation. Since rolling contact fatigue (RCF) damage [1] has caused plenty of railway accidents and has become increasingly common, it is critical for railway workers to efficiently obtain the information of the RCF damage so as to repair damaged rails in time.

To date, the rail surface defect inspection still mainly relies on manual inspection. For the purpose of alleviating the disadvantages of time-consuming and labor-intensive manual inspection, nondestructive testing techniques are applied in this field, such as ultrasonic inspection [2] and eddy current pulsed thermography inspection [3]. Compared with the aforementioned techniques, the visual inspection technique [4] has the advantages of low cost and high efficiency.

A. Related Detection Algorithms for Rail Surface Defects Based on Visual Inspection

In the past few years, rail surface defect detection based on machine vision has become a hot field. In terms of conventional detection approaches, some deployed image enhancement and statistical methods to segment defects [4], [5]. Later, researchers explored the morphological method [6] and detection strategies combing background modeling [7] with extraction of several features in frequency and spatial domains [8]–[10]. These methods have made achievements in treating uneven illumination. However, it is hard for them to extract abundant high-level semantic features from rail images.

Recently, the application of data-driven frameworks in defect detection has attracted scholars' attention. The existing machine learning-based defect inspection strategies can be summarized into five categories on the whole: defect image classification, defective pixel segmentation, defective patch detection with sliding windows, defective line recognization, and anchor-based defect region detection.

Regarding defect image classification approaches, [11] and [12] separately identified defective images by the deep convolutional neural network (DCNN) and feature extractors of texture and frequency. These studies have done forward-looking

work for recognizing rail damage, while they are difficult to detect multiple defects on one rail image and locate accurate positions of defects. The defective pixel segmentation strategy exploits a classification framework to identify each pixel [13], [14] or superpixel [15], [16] belonging to the defective pixel. This strategy can highly segment defect contours, while pixel classification is sensitive to the background's local gray-level change. In addition, fixed superpixels' number is not quite conducive to the scale adaptability of defect segmentation. Defective patch detection with sliding windows represents the strategy of dividing rail images into small patches, and then, identifying defective patches [17]. However, the fixed sliding window's size will lead to the location error of multiscale defects to some extent. Defective line recognition [18] refers to the strategy recognizing each line of rail images is defective. This strategy deploys the vertical consistency of rail images and can reduce the labeling work. However, it is sensitive to the local variation of rail surfaces with rust interference. As for anchor-based defect region detection algorithms, faster-Region-CNN [19] and YOLOv3 [20] are utilized in this field. Nevertheless, given that the anchors' ratios are limited to discrete types, defects that do not match anchors will be missed.

In summary, despite some institutions have studied rail surface defect detection based on data-driven frameworks, the main problem in this field, multiscale rail surface defect detection in a variety of the complex railway environment, has not been discussed in depth.

B. Challenges of Detection for Rail Surface Defects Based on the DCNN Model

The challenges of rail surface defect detection [1] based on machine vision and machine learning are as follows.

1) Foreground-Background Imbalance, Scale Imbalance, and Class Imbalance of Rail Surface Defect Detection: Defects are commonly sparse on the rail surface; big and severe defects are even rare. Furthermore, the different formation possibility of distinct classes of defects brings about severe class imbalance. Moreover, defects' scale is various, and even some defects possess extreme aspect ratios.

2) Variation of Reflection Property on the Rail Surface: There is a high contrast between defects and the wheel-rail contact area, but defects in the coarse metal area have low contrast with the background, which leads to uneven illumination for rail surface defect detection.

3) Plentiful Interference and the Changeable Weather in the Complex Railway Environment: Raindrops, rust, grinding marks, and stains on the rail surface have similar characteristics with defects. Moreover, the external environment, such as changes in natural light, shadow, and rain, results in the reduction of imaging quality.

C. Outline of Our Work

In order to address the aforementioned problems, based on the monorail image acquisition system (MIAS), this article puts forward a novel detection solution for two kinds of the most common rail surface defects, namely spalling and cracks. The core contributions are summarized as follows.

1) Aiming at subduing the data imbalance in the rail surface defect detection and resisting various interference on the complex rail surface, we propose a novel solution combining the targeted data augmentation (TDA) and an anchor-free defect detection framework, which outperforms related state-of-the-art methods in four different surface defect datasets. It verifies CCEANN's effectiveness and good generality.

2) We propose a new backbone, CSFA-Hourglass, in which double single-stage attention Hourglass structures balancing the network depth and feature fusion are cascaded by the cross-stage fusion of multiscale features (CSF), which not only efficiently transmits multiscale features between Hourglass modules (HGMs) of different stages, but also enhances the feature representation ability for fine center-point estimation of defects.

3) The convolutional block attention module with variable receptive fields (VRF-CBAM) is developed in the CSFA-Hourglass to effectively suppress the spatial noise of multiscale feature maps, in which the variable convolution filters strengthen the spatial attention's scale adaptability so as to enhance the detection accuracy for multiscale defects, especially large defects.

4) The presented anchor-free consistency of Intersection-over-Union(IoU)-guided center-point estimation head module (CASIoU-CEHM) flexibly detects defects with various scales and dynamically compensates coordinate offsets of predicted defect regions so as to alleviate the scale imbalance and precisely locate defects with extreme aspect ratios. Besides, in the training process of the CASIoU-CEHM, the proposed CASIoU loss achieves higher convergence accuracy than state-of-the-art IoU-based losses.

The rest of this article is organized as follows. Section II introduces the related works. Section III describes the proposed solution in detail. Section IV analyzes the experimental results. Section V discusses the wide applicability of the CCEANN. Finally, Section VI concludes this article.

II. Related Works

In this section, related works of CCEANN's three main components, namely attention mechanisms, multiscale feature extraction, and DCNN-based detectors, are reviewed.

The attention mechanism is an effective way to dig DCNN's important features. Self-attention [21] and single-dimension attention mechanisms [22], [23] merely considered strengthening one-aspect feature representation. Recently, multidimension attention mechanisms, BAM [24] and CBAM [25], combined channel and spatial attention to reweight DCNN's features. However, the convolution kernel size is fixed in their spatial attention modules, which does not work well in low-resolution feature maps. In this article, VRF-CBAM is presented to solve the aforementioned problems for multiscale feature extraction, in which aiming at feature maps with varying resolution, the spatial attention submodule (SAS) utilizes convolutional filters with variable receptive fields to generate spatial attention maps.

To precisely detect small defects and defective center points, it is crucial to effectively obtain the final high-resolution feature

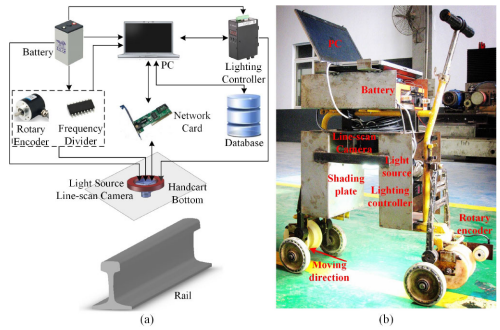

Fig. 1. Experimental platform MIAS. (a) Schematic diagram of the MIAS. (b) Physical diagram of the MIAS.

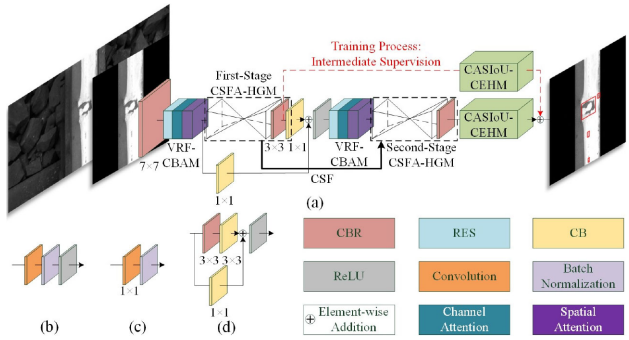

Fig. 2. Overall architecture of the CCEANN and related modules. (a) Structure of the CCEANN. (b) CBR. (c) CB. (d) RES. The color represented by each module or layer is shown at the bottom.

maps fusing abundant multiscale features. Feature pyramid networks (FPN) [26], cascaded pyramid network (CPN) [27], and deep layer aggregation (DLA) [28] with one-stage encoder–decoder structures adopted different feature fusion strategies to enhance the effectiveness of multiscale feature extraction. Stacked-Hourglass network [29] strengthens the feature representation by stacking multiple encoder–decoder architectures, whereas the fine-grained feature extraction and the feature transmission between different stages of this backbone are weak. Moreover, the large depth gap of multiscale feature fusion in the aforementioned backbones blur the feature representation to some extent. Based on the two-stage encoder–decoder structure, the proposed CSFA-Hourglass not only integrates a single-stage attention Hourglass structure balancing the network depth and lateral feature fusion, but also introduces CSF, which strengthens the multiscale feature representation for defects.

DCNN-based detectors play vital roles in regressing defects' location and size. The most representative two-stage detectors, R-CNN series [30] generate region proposals and detect objects by various pre-defined anchors. One-stage detectors [31], [32] can directly regress objects' location on final feature maps. Nevertheless, it is challenging for the aforementioned anchor-based detectors to regress defects with large-scale variation. Recently, flexible anchor-free detectors [28], [29] have become increasingly popular. However, the anchor-free strategy's model fitting is more inclined to defects with common scales in the dataset. In this article, CASIoU-CEHM is designed to precisely locate rail surface defects with large-scale variation. It not only owns the anchor-free advantage of flexibly regressing defects' size and location but also dynamically compensates defective coordinate offsets to alleviate the scale imbalance, in which CASIoU loss enhances the regression accuracy.

III. Proposed Detection Solution

A. Overview of the MIAS

This article deploys MIAS, a handcart equipped with integrated hardware devices, to capture rail images online in the outdoor railway. The schematic and physical diagrams of MIAS are shown in Fig. 1. One of the most critical components of the MIAS is the DALSA Linea high-speed line-scan camera with a line rate up to 80 kHz and the resolution of 1×2048 pixels. Besides, the circular light source installed around the camera provides stable illumination. Its power is controlled by the lighting controller. When the MIAS goes ahead, the rotary encoder coaxially connected with the MIAS's rear-wheel sends signals to the PC to record the mileage and triggers the camera to capture rail images. Every frame is stitched by 560 lines of pixels, and the resolution of each pixel is 0.39×0.39 mm. After a panoramic grayscale image with the resolution of 2048×560 is captured, it will be transferred to the PC via Ethernet. It is noteworthy that the image quality is unavoidably affected by environmental factors, namely sunlight and shadow.

B. Overview of CCEANN

Aiming to detect multiscale defects in the complex railway, CCEANN is proposed to offline process rail images captured by the MIAS. As shown in Fig. 2, after the histogram-based track extraction [4] extracts the rail region of interest to filter unrelated regions such as sleepers and ballets, the CCEANN detects defects in the cropped rail image, which contains two primary parts: 1) CSFA-Hourglass and 2) CASIoU-CEHM. The first effectively extracts multiscale features of defects, and then, fuses these features as the input of the second part. Especially, VRF-CBAM and CSF are presented in this part. For the second part, defects are located by flexible bounding boxes based on center-point estimation combined with CASIoU-guided fine coordinate compensation, in which the gradient propagation mechanism and CASIoU loss will be emphasized.

C. Basal Backbone

CSFA-Hourglass's architecture is inherited from Hourglass-104 (HG-104) [29], a high-performance backbone used in the human pose estimation. However, the situation in rail surface defect detection is different. In detail, effective feature extraction for small objects and precise location for irregular multiscale objects become the main tasks. In order to develop an effective backbone for the further center-point estimation of multiscale defects, the thoughts of attention mechanism, modular improvement, and multiscale feature fusion are developed to design the novel backbone, CSFA-Hourglass.

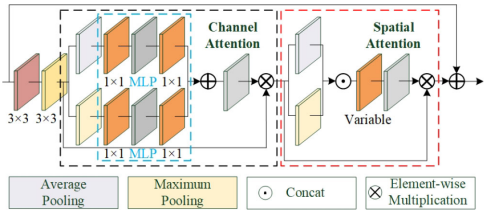

Fig. 3. Diagram of VRF-CBAM's structure.

In CSFA-Hourglass, when a rail image with the fixed resolution of 560 × 188 and its labeled defects is fed into the network, it will be initially padded and resized to achieve the size of 512 × 512. Then, the resolution of the resized image is reduced four times to generate initial feature maps. Regarding the further feature extraction, CSFA-Hourglass mainly consists of five kinds of modules, namely Conv-(BN) batch normalization module (CB), Conv-BN-ReLU module (CBR), the stacked attention HGM via cross-stage fusion of multiscale features (CSFA-HGM), VRF-CBAM, and CASIoU-CEHM. The CB is made up of a convolution layer and a batch-normalization layer. The CBR consists of a convolution layer, a batch-normalization layer, and a ReLU layer. CSFA-HGM fuses low-level and high-level features by a two-stage encoder–decoder structure with multiple VRF-CBAMs and skip layers so as to extract multiscale features. The VRF-CBAM consists of a residual module (RES) and attention-based submodules. Moreover, for optimizing the parameter update in the shadow layers, the intermediate supervision in the training process allows the loss of intermediate heatmaps to be updated. Subsequently, a number of crucial components, including VRF-CBAM, CSFA-HGM, and CASIoU-CEHM, will be described concretely in the next subsections.

D. VRF-CBAM

The resultant heatmaps of the center-point estimation posses high resolution and multiple channels. To improve CCEANN's representation ability and suppress noise in multiscale feature maps, VRF-CBAM is introduced, making the framework highlight important features by weight assignment. The VRF-CBAM consists of three parts, which utilizes RES with the channel attention submodule (CAS) and improves the SAS in the CBAM. A diagram illustrating this module's structure is shown in Fig. 3.

CAS's structure includes two streams with pooling operations and the multilayer perceptron (MLP). In MLP, a 1 × 1 convolutional layer with a ReLU layer first reduces feature maps' channel dimension. Then, the further 1 × 1 convolutional layer recovers the output channel dimension to the original dimension. The two streams squeeze the spatial information, and then, find the dependence degree on each channel so as to give more attention to more informative channels. The output feature maps F_c of the CAS can be expressed as

$$F_c = F * \delta_S(\mathrm{MLP}(\mathrm{Pool}_{\mathrm{Avg}}(F)) \oplus \mathrm{MLP}(\mathrm{Pool}_{\mathrm{Max}}(F))) \quad (1)$$

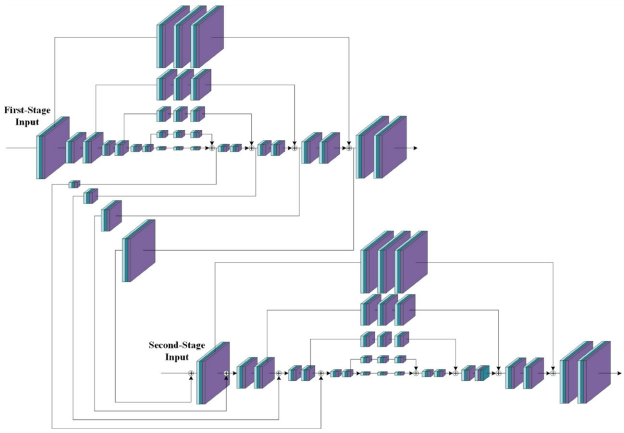

Fig. 4. Structure of the CSFA-HGM.

where F denotes input feature maps of the CAS. $\mathrm{Pool}_{\mathrm{Avg}}$ and $\mathrm{Pool}_{\mathrm{Max}}$ refer to the average and max pooling operation, respectively. $*$ and $\oplus$ are the element-wise multiplication and addition operators, receptively; and δ_S represents the Sigmoid activation function.

SAS's function is to highlight informative parts and suppress the spatial noise. After the pooling operations of two streams squeeze the channel information, convolution layers with large receptive fields help the framework find noteworthy regions in the spatial dimension. However, convolution filters with too large receive fields beyond the effective area will lead to low feature extraction effectiveness in low-resolution feature maps for large defect detection. To address that, we constantly reduce kernel sizes of SAS's convolution filters along with the decline in resolution. In detail, the convolution filter's size for processing 128 × 128 and 64 × 64 feature maps is 7 × 7, the size at the resolution of 32 × 32 and 16 × 16 is reduced to 5 × 5, and that at the lowest resolution (8 × 8) is 3 × 3. In this way, the SAS gets the ability to adapt to the spatial scale change. SAS's reweighted output feature maps F_s can be expressed as

$$F_s = F_c * \delta_S(\mathrm{conv}_{\mathrm{var}}(\mathrm{Pool}_{\mathrm{Avg}}(F_c) \odot \mathrm{Pool}_{\mathrm{Max}}(F_c))) \quad (2)$$

where $\odot$ is the concatenation operator; $\mathrm{conv}_{\mathrm{var}}$ denotes the convolution operation with variable receptive fields. Last, the skip connection operation is utilized between the input of VRF-CBAM and the final output of the SAS to strengthen the effectiveness of the deep feature extraction.

E. CSFA-HGM

CSFA-HGM is the most critical component of the whole backbone. We consider introducing a novel two-stage attention Hourglass structure and the multiscale feature transmitting between multistage encoder–decoder modules to enhance this component's feature extraction capacity for defects. The structure of the proposed module is depicted in Fig. 4.

First of all, VRF-CBAMs are stacked at every resolution to strengthen the feature representation ability. In each stage, the resolution is reduced four times in the encoder part to reach the lowest resolution of 8 × 8, which is a suitable feature map

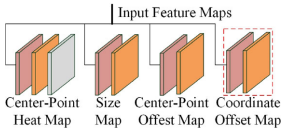

Fig. 5. Structure of the CASIoU-CEHM.

size for detecting large defects. Meanwhile, the channel number along this way is set to [256, 384, 384, 384, 512].

Instead of just deepening the network as the original HGM, a tradeoff strategy for network depth and feature fusion is developed. We moderately deepen the network in the encoder and decoder parts, reduce the number of convolutional layers at the lowest resolution and increase that in the skip layers at other resolution to pay more attention to extract fine-grained features for small and medium defect detection as well as enhance the center-point location accuracy. In this way, the gap between the depth of output feature maps in the skip layers and that of fused feature maps in the decoder part is also narrowed, which can enhance the effectiveness of feature fusion at the same resolution.

Eventually, in order to strengthen the feature dependence between different stages and transmit richer multiscale information to the next stage, in CSF, multiscale feature maps after skip connections of the first-stage HGM primarily pass through VRF-CBAMs, and then, connect with input feature maps of the second-stage HGM at corresponding resolution by the element-wise summation.

F. CASIoU-CEHM

Detection for multiscale rail surface defects, especially slender defects with random lengths, needs flexible detection boxes with variable scales and ratios. Inspired by [28], we first introduce the center-point estimation branch, the size prediction branch, and the center-point offset prediction branch to output initial defective detection boxes in an anchor-free way. Subsequently, to further improve the location accuracy, a new coordinate prediction offset branch in parallel with the aforementioned branches is designed to associate the key evaluation index IoU [33] with the center-point estimation results so as to finely compensate the size and location of detection boxes. CASIoU-CEHM's structure is shown in Fig. 5.

1) Center-Point Estimation: In the head module, considered class imbalance problem, two-channel heatmaps of two classes of defects' center points are separately generated by a CBR with a 1×1 convolution layer and a Sigmoid layer. For avoiding the foreground-background imbalance and efficiently digging hard samples' key features, Focal loss L_c is used to train the center-point estimator

$$L_c = -\frac{1}{N}\sum_{c}\begin{cases}(1-H_c)^{\alpha}\log(H_c) & \text{if } H_c^{gt}=1\\ (1-H_c^{gt})^{\beta}(H_c)^{\alpha}\log(1-H_c) & \text{otherwise}\end{cases} \tag{3}$$

where α and β represent the hyperparameters that control the training weights of easy samples and negative samples, respectively. Here, α and β are separately set to 2 and 4. N denotes the ground-truth defective center points' number in the whole image. H_c^{gt} and H_c refer to the ground truth and the prediction point of the center-point heatmap, respectively.

Moreover, both the center-point offset prediction branch and the size prediction branch are made of CBR and a 1×1 convolution layer. Note that only outputs at the ground-truth center-point position of these branches are trained by L1 loss

$$L_s = \frac{1}{N}\sum_{k=1}^{N}\left|S_c - S_c^{gt}\right| \tag{4}$$

$$L_o = \frac{1}{N}\sum_{k=1}^{N}\left|O_c - O_c^{gt}\right| \tag{5}$$

where L_s and L_o represent the size loss and center-point loss, respectively. S_c^{gt} and S_c separately denote the actual and predicted size of defects. O_c^{gt} and O_c, respectively, refer to the ground-truth and predicted offset of the defect's center point.

2) CASIoU-Guided Coordinate Compensation: Model fitting of the anchor-free strategy is more inclined to defects with common scales in the dataset, which leads to the scale imbalance. Thus, the accuracy of detecting partial defects with big scale differences is declined. This article presents a fine adjustment way to improve the location accuracy by training offsets of two corner-point coordinates by CASIoU loss, which can further minimize the error of the two bounding boxes' matching degree in the training process. Regarding the prediction process, predicted boxes' location and size will be finely compensated by the output coordinate offsets.

The structure of the coordinate offset prediction branch is similar to that of the size prediction branch, while the difference is this branch's output channel size, which is set to 4. IoU-based losses [34], [35] can be used in this process. Nevertheless, these losses did not combine the four key conditions, including the consistency of IoU, center-point distance, area, and scale together to regress loss. Thereby, an effective CASIoU loss related to the four conditions is presented to regress the error of the center-point estimation. Especially when the predicted defect having area difference with the ground truth is located accurately, losses in [34] and [35] will degenerate into IoU loss, but CASIoU loss still can be distinguishable.

Regarding CASIoU loss, we initially employ CIoU loss's error calculation of the center-point distance and scale. Afterward, the area error is added in the loss function as follows:

$$L_{\text{CASIoU}} = 1 - \text{IoU} + \frac{\rho^2(b,b^{gt})}{cl^2} + k_s \cdot s + k_a \cdot a. \tag{6}$$

Its value domain is [0,4], and it consists of four penalty terms, including IoU penalty term, center-point distance penalty term $\frac{\rho^2(b,b^{gt})}{cl^2}$, scale penalty term $k_s \cdot s$, and area penalty term $k_a \cdot a$. Specifically, $\rho^2(b,b^{gt})$ denotes the Euclidean distance between the center points of the ground truth and the detection result; cl refers to the length of the catercorner line of the minimum external rectangle containing the two boxes. Especially, the

location of the detection box B is shown as follows:

$$\begin{aligned}&(x_i+\delta_{x_i}x_i-w_i/2+c_{tl}^{x_i},y_i+\delta_{y_i}y_i-h_i/2+c_{tl}^{y_i}\\&x_i+\delta_{x_i}\mathrm{x}_i+w_i/2+c_{br}^{x_i},y_i+\delta_{y_i}y_i+h_i/2+c_{br}^{y_i})\end{aligned}\tag{7}$$

where (x_i, y_i) and $(\delta_{x_i}, \delta_{y_i})$ are the center point and the center-point offset of the detection box, respectively; (w_i, h_i) denotes the detection box's size; $(c_{tl}^{x_i}, c_{tl}^{y_i}, c_{br}^{x_i}, c_{br}^{y_i})$ refer to the coordinate offset in the top-left and bottom-right corners of the detection box. Then, the parameter k_i associating the scale factor and the area factor with IoU is defined as

$$k_i=\frac{i}{(1-IoU)+i}i=s,a.\tag{8}$$

And the scale factor s can be represented as

$$s=\frac{4}{\pi^2}(\arctan\frac{w}{h}-\arctan\frac{w^{gt}}{h^{gt}})^2.\tag{9}$$

Then, a piecewise function satisfying the continuous derivable condition is designed to fit the area penalty, which gives greater punishment to the result that has a larger area difference with the ground truth. The area factor a is depicted as

$$a=\begin{cases}\left(\frac{A}{A^{gt}}-1\right)^2 & 0<\frac{A}{A^{gt}}\le 1\\ \frac{1}{(i-1)(i-1+\pi)}\left(\frac{A}{A^{gt}}-1\right)^2 & 1<\frac{A}{A^{gt}}\le i\\ \frac{4}{i-1+\pi}\arctan\left(\frac{A}{A^{gt}}-i+1\right)+\frac{i-1-\pi}{i-1+\pi} & \frac{A}{A^{gt}}>i\end{cases}\tag{10}$$

where the tunable parameter i controls the steepness of the area penalty term's curve distribution in CASIoU loss, and $i\in(1,+\infty)$. As i increases, the steepness of this distribution reduces. $\frac{A}{A^{gt}}$ is the area ratio between the detection box and the ground truth. Note that CASIoU loss is only implemented to regress the error of coordinate offsets so that the center-point estimation part merely has forward propagation in this branch. In this way, this loss can play a fine-tuning role instead of being the main task of the detection box regression. Ultimately, Adam is used to optimize the total loss in the training process

$$L=L_c+L_o+\lambda_sL_s+\lambda_cL_{\mathrm{CASIoU}}\tag{11}$$

where λ_s and λ_c are both set to 0.1.

In the prediction process, the heatmap value for each pixel is regarded as the confidence score of a suspected defect, and the top 100 suspected defects are screened by the soft nonmaximum suppression (Soft NMS) [28]. Next, an area filter is designed as follows to filter small noise of detection results:

$$T=\begin{cases}1 & \text{if}(\text{cls}=\text{spalling})\&(A>T_a)\\ 0 & \text{otherwise}\end{cases}.\tag{12}$$

The length of one pixel captured by the MIAS is around 0.39 mm. According to the demand of the railway maintenance, we speculate that spalling detection boxes whose area less than 12mm^2 are unlikely to be defects that influence railway safety. Therefore, T_a is set to 81 via the conversion.

IV. Experimental Results and Analyses

In this section, the excellent performance of CSFA-Hourglass and CASIoU-CEHM are verified in ablation studies. In addition, the CCEANN is compared with related deep learning-based detection methods to validate its effectiveness in different rail surface defect datasets. Finally, the running time and the computation complexity of the CCEANN are offered.

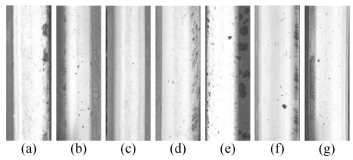

Fig. 6. Samples of rail images in our rail surface defect dataset. (a) Severe spalling. (b) Small defects covered by rust. (c) Spalling with extreme aspect ratios. (d) Dense cracks. (e)–(g) Rail images with interference, including raindrops, grinding marks, and stains, respectively.

A. Datasets

Three rail surface defect datasets, including our rail surface defect dataset based on the MIAS, rail surface discrete defects dataset (RSDDS), and northeastern university (NEU) rail surface defect dataset (NEU RSDDS-113), are utilized to evaluate the effectiveness and the generality of the CCEANN. Note that based on the pixel-level labels, the ground truths of bounding boxes for RSDDS and NEU RSDDS-113 are generated.

1) Our Rail Surface Defect Dataset: Based on rail images captured by the MIAS, we built a real-world rail surface dataset that contains multiscale spalling and cracks on rainy and sunny days. Various kinds of negative samples, including rail slots, grinding marks, rust, stains, raindrops, and shadow, are also collected. Some typical samples in our dataset are illustrated in Fig. 6. In this dataset, 1 773 images are augmented to 5 027 images to train the framework, and the other 1 581 rail images are selected based on stratified sampling for testing. Note that all defect regions are labeled by experts. For the TDA utilized in the training set, brightness change, horizontal flip, vertical flip, mirroring operation, Gaussian, and salt noise addition are adopted. Considered the scale and class imbalance, the sample number of severe defects and cracks is increased by 200% and 600%, respectively. For native samples and hard samples with low contrast, their number is expanded by 200%. In this way, the proportion in the training set among spalling, cracks, and natives is approximately 3:3:1. Note that in the test set, crack images, spalling images, and negative samples separately occupy 23%, 63%, and 14%. Especially, in spalling samples, severe and small spalling samples separately account for about 7% and 24%, hard samples and defects with extreme aspect ratios take up 15% and 17%, respectively.

2) RSDDS: RSDDS [9] consists of Type-I subdataset with 67 express rail images and Type-II subdataset with 128 common rail images. The following two ways are carried out to build the training set and the test set. First, rail images in Type-I and Type-II are vertically clipped to two and eight parts, respectively. Then, all clipped images are resized to 560 × 188. Subsequently, 92 defective images and 42 defect-free images in Type-I, as well

as 290 defective images and 734 defect-free images in Type-II, are collected, in which 75 defective images and 24 defect-free images in Type-I, as well as 231 defective images and 364 defective-free images in Type-II, are put into the training set, and the other rail images consist of the test set. All defective images in the training set are augmented by 500%. In detail, one of the scaling operation, contrast change, and brightness change, is mixed with one of the horizontal flip, vertical flip, and mirroring operation for DA of every defective image. Moreover, defective images in the test set are augmented one time by the vertical and horizontal mirror operation. To sum up, the sample number in the training set and the test set are 2224 and 540, respectively.

3) NEU RSDDS-113: This dataset [16] consists of 113 defective images on the rail surface with multiresolution, in which 91 and 22 images are randomly selected to build the training set and the test set, respectively. After that, we augment the training set by 900%. Specifically, a mixture of one of the scaling operation, contrast change, brightness change, rotation operation, and clipping operation with one of the horizontal flip, vertical flip, and mirroring operation is adopted for each image. Similarly, the sample number of the test set is increased by 600% based on the horizontal flip, vertical flip, mirroring operation, contrast change, brightness change, and clipping operation. Thus, the sum of samples in the training set and the test set are 910 and 154, respectively.

TABLE I
DETECTION PERFORMANCE WITH DIFFERENT DATA AUGMENTATION STRATEGIES

DA	Sample Size	Spalling			Cracks		
		PR(%)	RC(%)	F_1(%)	PR(%)	RC(%)	F_1(%)
	1773	76.58	77.97	77.27	38.15	56.28	45.47
GDA	5319	86.87	87.43	87.15	67.98	68.97	68.47
TDA	**5073**	**90.97**	**91.04**	**91.01**	**89.27**	**87.81**	**88.53**

Bold entities in Table I represent the proposed solution's experimental data.

TABLE II
DETECTION PERFORMANCE WITH DIFFERENT ATTENTION MECHANISMS

ATTN	PR(%)	RC(%)	F_1(%)	AP^{50}(%)	AR^{50}(%)	AP_L(%)	AR_L(%)
	87.13	89.85	88.47	90.77	96.06	69.10	70.83
SE	88.93	89.31	88.85	90.85	95.44	74.38	75.67
GC	87.73	89.24	88.48	90.07	94.97	58.68	66.50
ECA	87.35	89.57	88.44	90.87	95.71	68.86	72.67
CBAM	88.63	91.41	90.00	91.79	96.16	62.48	64.83
VRF-CBAM	**90.48**	**90.10**	**90.29**	**92.45**	**96.52**	**76.74**	**78.83**

Bold entities in Table II represent the proposed solution's experimental data.

B. Evaluation Criteria

In the following experiments, we use eight evaluation indexes, namely precision (PR), recall (RC), F1-measure (F_1), average precision (AP^{50}), average recall (AR^{50}), average precision, and average recall on large defects (AP_L, AR_L) in the common objects in context (COCO) style [36], and parameters (Para.), to quantitatively estimate the performance of CCEANN and related methods.

C. Implementation Details

The CCEANN is implemented in Pytorch, and all experiments are carried out on a server on the Ubuntu 16.04 system with an Intel(R) Core (TM) i7 CPU (3.50 Hz), one NVIDIA RTX 2080Ti GPU, and 32-GB memories. Note that all parameters in the CCEANN are randomly initialized without pretraining. This network is trained for 70 epochs with a batch size of 4. The initial learning rate is set to 1.25×10^{-4} and is decayed by 0.1 at 45 and 65 epochs. The soft NMS threshold is set to 0.5. i and the confidence score thresholds in our rail surface defect dataset, RSDDS, and NEU RSDDS-113, are separately set to [1.5, 2.25, 1.5] and [0.35, 0.25, 0.25]. Particularly, in NEU RSDDS-113, T_a is set to 0, and a small regularization term (1.71×10^{-6}) is added to avoid the overfitting.

D. Ablation Studies

In order to verify the effectiveness of the TDA, VRF-CBAM, CSFA-Hourglass, and CASIoU-CEHM, the results of distinct settings for the CCEANN are evaluated by quantitative and visualized analyses. Note that ablation studies are carried out in our rail surface defect dataset.

1) Impact of the TDA: Here, we investigate the detection results on different data augmentation (DA) strategies. Besides the training results with the original training dataset, we compare the results of the TDA with that of the general data augmentation strategy (GDA). In the GDA, vertical flip, mirroring operation, and Gaussian noise addition are used for every sample to increase the sample number of the original training dataset by three times to 5 319. The experimental results in Table I illustrate that although the GDA possesses the largest sample size, the TDA still achieves a better training effect, especially in crack detection, because the TDA alleviates the severe data imbalance so as to obtain the model with a better generalization capability.

2) Impact of CSFA-Hourglass: In this article, the effectiveness of three crucial components in the CSFA-Hourglass is estimated one by one. Then, the performance of several related backbones is compared with that of the CSFA-Hourglass.

a) Effects of VRF-CBAM: To start with, we perform comparative experiments to estimate the influences of different attention mechanisms on the CCEANN. Four attention modules are compared with VRF-CBAM, including SENet [22], ECA-Net [23], GCNet [21], and CBAM [25]. From Table II, it can be seen that VRF-CBAM outperforms all competitive attention mechanisms. Especially, the VRF-CBAM achieves the highest F_1 score, which represents that it gets a better fitting effect. Meanwhile, compared with the performance without attention mechanism, the enhancement of AP^{50} (1.68%) and AR^{50} (0.46%) demonstrates that the VRF-CBAM is helpful to produce more potentially correct detection results. Finally, the

TABLE III
DETECTION PERFORMANCE WITH DIFFERENT HGMS

Module Type	Depth	PR(%)	RC(%)	F_1(%)	AP^{50}(%)	AR^{50}(%)
A-HGM	54	89.48	90.20	89.84	91.94	95.91
EDFFA-HGM	22	87.46	90.08	88.75	90.87	95.53
CSFA-HGM	**36**	**90.48**	**90.10**	**90.29**	**92.45**	**96.52**

Bold entities in Table III represent the proposed solution's experimental data.

TABLE IV
DETECTION PERFORMANCE OF CSFA-HGM WITH DIFFERENT STRUCTURES

Stage	CSF	VRF-CBAM	PR(%)	RC(%)	F_1(%)	AP^{50}(%)	AR^{50}(%)
1			85.86	85.90	85.88	88.09	94.35
2			87.88	90.38	89.11	91.68	95.98
2	✓		88.57	90.16	89.36	91.70	96.00
2	✓	✓	**90.48**	**90.10**	**90.29**	**92.45**	**96.52**

Bold entities in Table IV represent the proposed solution's experimental data.

results in this table's last two rows show that VRF-CBAM has outstanding performance in large defect detection.

b) Effects of CSFA-HGM: In this experiment, under the precondition that the downsampling times and the computational complexity of each HGM are the same, we explore the performance of HGMs with different structures in the CCEANN. The depth of attention HGM (A-HGM) with 4× downsampling and VRF-CBAMs is 54, and that of the CSFA-HGM at the single stage is 36. The A-HGM pays more attention to deepening the network, but the CSFA-HGM attaches more importance to effective feature fusion. How to find the best tradeoff way between the depth and the feature fusion? Besides these two modules, the HGM with a shallower depth (22) and more convolution in skip layers that we called A-HGM with equal depth feature fusion (EDFFA-HGM) is also compared. In this module, VRF-CBAMs' number in the skip layer along the increase of resolution are 3, 4, 5, and 6 in turn, which promotes the feature maps that have undergone downsampling to fuse with the features of the same depth in the decoder part so that it possesses better feature fusion strategy. In addition, the VRF-CBAMs' number at each resolution in the encoder and decoder are all set to 1 to reduce its computation complexity.

As shown in Table III, the CSFA-HGM achieves the best results, which improves precision while maintaining recall and decreasing the network depth. This can be interpreted by the following three factors.

1) The feature fusion of A-HGM with large depth differences blurs the final feature representation.
2) More VRF-CBAMs at high resolution enhances the center-point location accuracy and the effectiveness of feature extraction for small defects.
3) Properly deepening the network is an essential factor in enhancing this module's performance.

c) Effects of CSF: The third experiment compares the results of different stages of the CSFA-HGM with and without CSF. Especially, the situations that skip layers in the CSF with and without VRF-CBAM are also discussed to evaluate the effectiveness of this strategy without increasing the computational complexity. Table IV shows that the performance of the two-stage encoder–decoder structure is better than that of the single-stage encoder–decoder structure, which indicates the superiority of the cascaded Hourglass structure. Moreover, the CSF promotes the framework's performance with the improvement of F_1 score (1.18%) and maintains the recall simultaneously.

d) Backbone Comparison: In Table V, the CSFA-Hourglass is compared with three multiscale feature extraction backbones, namely CPN-ResNet101 [27], HG-104 [29], and DLA-102 [28], as well as the simple baseline ResNet-101 (SB-ResNet101) [37]. Overall, the CSFA-Hourglass achieves the best performance among the five backbones in Table V.

TABLE V
DETECTION PERFORMANCE BASED ON DIFFERENT BACKBONES

Backbone	PR(%)	RC(%)	F_1(%)	AP^{50}(%)	AR^{50}(%)
SB-ResNet101	83.18	84.00	83.59	80.03	87.58
DLA-102	74.76	62.47	68.07	61.96	81.86
CPN-ResNet101	88.00	89.63	88.81	90.06	95.16
HG-104	85.35	88.49	86.89	89.14	95.20
CSFA-Hourglass	**90.48**	**90.10**	**90.29**	**92.45**	**96.52**

Bold entities in Table V represent the proposed solution's experimental data.

TABLE VI
DETECTION PERFORMANCE OF CASIOU-CEHM WITH DIFFERENT STRUCTURES AND IOU-BASED LOSSES

Box Offset Branch	IoU-based Loss	PR(%)	RC(%)	F_1(%)	AP^{50}(%)	AR^{50}(%)
		88.98	90.78	89.87	91.76	96.15
Scale	CASIoU	89.22	90.08	89.65	91.56	96.02
Coordinate	IoU	88.84	90.24	89.54	91.47	95.87
Coordinate	GIoU	88.52	89.69	89.10	91.57	95.85
Coordinate	DIoU	88.54	91.07	89.79	91.62	96.18
Coordinate	CIoU	89.09	90.84	89.96	91.90	96.07
Coordinate	**CASIoU**	**90.48**	**90.10**	**90.29**	**92.45**	**96.52**

Bold entities in Table VI represent the proposed solution's experimental data.

3) Impact of CASIoU-CEHM: Through the following two experiments, the importance of the coordinate offset prediction branch and CASIoU loss is verified.

a) Effects of the coordinate offset prediction branch: In order to analyze the coordinate offset branch's contribution, three variants are compared together, including center-point estimation without the detection box offset branch, with scale offset branch, and with coordinate offset branch. Note that the scale offset branch only outputs two-channel feature maps that contain height offsets and width offsets of detection boxes. Table VI (the first, second, and eighth rows) demonstrates that the coordinate offset branch evenly improves CCEANN's performance in multiple evaluation indexes, which verifies its effectiveness.

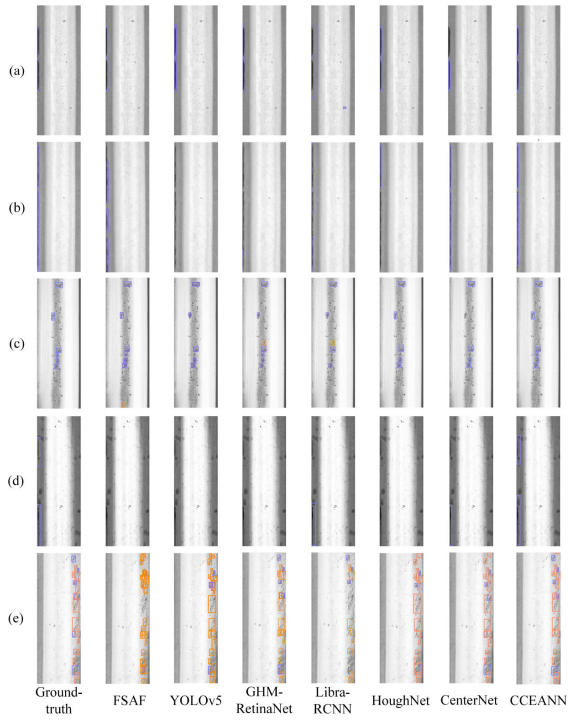

Fig. 7. Visualization results in our dataset. (a) and (b) Results of spalling with large scale differences. (c) and (d) Results of samples that are difficult to detect. (e) Results of dense cracks with area differences. Blue boxes and orange boxes denote detected spalling regions and crack regions, respectively.

b) Effects of CASIoU loss: In order to evaluate the convergence accuracy of IOU-based losses, the experiments of the coordinate offset prediction branch with different IoU-based losses, namely IoU loss [33], GIoU loss [34], DIoU loss, CIoU loss [35], and CASIoU loss, are carried out. With the results presented in Table VI (the fourth–eighth rows), although the recall of the CASIoU loss is slightly lower than that of the CIoU loss, the CASIoU loss gets better performance in the other four evaluation indexes. It decreases false positives to reach the highest F_1 score. Besides, it generates more potentially correct detection boxes, which means that the CASIoU loss possesses the best regression accuracy among the five IoU-based losses.

E. Comparisons With Related Frameworks

We compare the CCEANN against six state-of-the-art deep learning-based frameworks, including Libra-RCNN [30], YOLOv5 [31], GHM-RetinaNet [32], CenterNet [28], FSAF [38], and HoughNet [39]. The backbones of CenterNet and HoughNet are HG-104, and that of Libra-RCNN, GHM-RetinaNet, and FSAF are ResNet101-FPN. Besides, YOLOv5x in the YOLOv5 series is selected. Because we have augmented the training sets, no extra DA is used for the aforementioned frameworks.

1) Detection Results in Our Rail Surface Defect Dataset: As the intuitional detection results of the CCEANN and the other six related detection approaches portrayed in Fig. 7, even though all the comparative frameworks lead to false detections or missing inspections, the CCEANN can successfully detect spalling with large-scale differences and dense cracks in complex rail surfaces with disturbances. The quantitative results presented in Table VIII (the second–seventh columns) further illustrate that the performance of the CCEANN is superior to that of other frameworks. Ultimately, the P/R curves of the seven related approaches are compared in Fig. 8(a) and (b). It can be observed that CCEANN's P/R curves of two classes of defects almost completely wrap the other curves, which shows that the proposed framework outperforms other state-of-the-art frameworks with the highest AP^{50} (93.66% and 91.25%) in detection for spalling and cracks, respectively.

2) Detection Results in RSDDS: From Fig. 9's visual detection results, it is observed that although the other methods are sensitive to noise or rust as well as result in missdetections for defects with scale variation, the CCEANN obtains ideal detection results. The numeric results presented in Table VIII (the eighth–tenth columns) indicate that the CCEANN achieves the best performance. Fig. 8(c) illustrates that the P/R curve of the CCEANN is closer to the upper right corner of the coordinate system than that of other methods, which indicates that the CCEANN achieves better performance in the defect detection of RSDDS.

3) Detection Results in NEU RSDDS-113: From Fig. 10, we can observe that in multiscale rusty rail images with different illumination, the CCEANN achieves the best visualization results compared with the related six frameworks. In detail, although interference on complex backgrounds brings challenges to the other frameworks, the CCEANN can detect all the defect regions without false detection. The quantitative results in Table VIII (the 11th–13th columns) show that the CCEANN gets the highest F_1 among this comparison. Eventually, it can be observed from Fig. 8(d) that the P/R curves of the CCEANN and YOLOv5 are closer to the upper right corner of the coordinate system than that of other methods. Furthermore, the P/R curve's upper right corner of the CCEANN is closer to the coordinate system's top than that of YOLOv5, which shows CCEANN's superiority in maintaining a high precision in this task.

F. Parameters and Runtime Analyses

Based on our rail surface defect dataset, the computation complexity and the computing time of the proposed framework and CCEANN with single-stage CSFA-HGM (single-stage CCEANN) in Section V-D2(c) are compared, and then, test images' resolution and parameters of the CCEANN and related six deep learning-based frameworks are given. Note that the input images' resolution of all comparison methods is set to the default value in their public codes.

At first, as shown in Table VIII, we can see that CCEANN can run at seven frame/s. Besides, the parameters and the computing time of the single-stage CCEANN are merely half of those of the CCEANN. Nevertheless, the detection performance of the CCEANN is markedly superior to that of the single-stage CCEANN in Section V-D2(c). In addition, as the improvement of detection performance, the detection efficiency of the CCEANN is relatively lower than that of other related methods. However, it is worthy to note that compared with HoughNet

TABLE VII
DETECTION PERFORMANCE FOR DIFFERENT FRAMEWORKS IN DIFFERENT DATASETS

Method	Our Dataset						RSDDS			NEU RSDDS-113			KolektorSDD		
	Spalling			Cracks			Defects			Defects			Defects		
	PR(%)	RC(%)	F_1(%)	PR(%)	RC(%)	F_1(%)	PR(%)	RC(%)	F_1(%)	PR(%)	RC(%)	F_1(%)	PR(%)	RC(%)	F_1(%)
FSAF	72.62	79.97	76.12	52.79	67.64	59.30	84.16	86.29	85.21	91.15	91.96	91.55	93.22	91.67	92.44
YOLOv5	87.74	72.49	79.39	81.62	87.40	84.41	88.94	87.62	88.28	96.88	96.88	96.88	83.93	78.33	81.03
GHM-RetinaNet	79.20	82.92	81.02	76.11	83.11	79.46	86.34	87.62	86.98	89.96	91.96	90.95	86.67	75.36	80.62
Libra-RCNN	74.73	82.36	78.36	75.07	74.73	74.90	89.29	86.63	87.94	91.74	89.29	90.50	93.33	93.33	93.33
HoughNet	85.45	89.68	87.51	79.26	83.04	81.10	85.93	84.65	85.29	97.69	94.20	95.91	94.74	90.00	92.31
CenterNet	87.73	89.10	88.41	80.19	86.85	83.39	87.74	88.61	88.18	96.43	96.43	96.43	86.67	86.67	86.67
CCEANN	**90.97**	**91.04**	**91.01**	**89.27**	**87.81**	**88.53**	**92.93**	**91.09**	**92.00**	**99.09**	**96.88**	**97.97**	**95.08**	**96.67**	**95.87**

Bold entities in Table VII represent the proposed solution's experimental data.

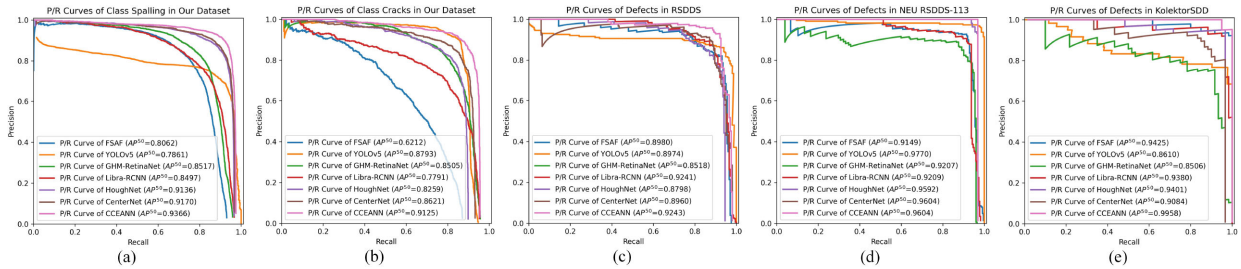

Fig. 8. P/R curves in different rail surface defect datasets. (a) Curves of spalling in our dataset. (b) Curves of cracks in our dataset. (c) Curves in RSDDS. (d) Curves in NEU RSDDS-113. (e) Curves in KolektorSDD.

TABLE VIII
COMPUTATION COMPLEXITY AND RUNNING TIME OF DIFFERENT FRAMEWORKS

Method	Resolution	Para. (M)	Time (s)
FSAF	640×400	55.00	0.03
YOLOv5	600×600	84.30	0.02
GHM-RetinaNet	1333×800	55.12	0.06
Libra-RCNN	1333×800	60.39	0.06
HoughNet	512×512	191.25	0.08
CenterNet	512×512	191.24	0.10
Single-stage CCEANN	512×512	83.45	0.07
CCEANN	512×512	**167.28**	**0.14**

Bold entities in Table VIII represent the proposed solution's experimental data.

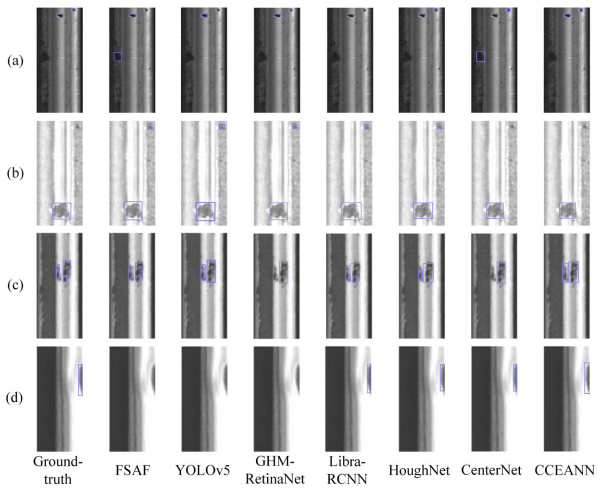

Fig. 9. Visualization results in RSDDS.

and CenterNet, the frameworks with two-stage encoder–decoder architectures, the parameters of the CCEANN are approximately 24 M smaller than that of them. Besides, as an offline rail surface defect detection method, the CCEANN with outstanding detection accuracy is promising and competent.

V. DISCUSSION

In this section, CCEANN's wide applicability is discussed in the Kolektor surface-defect dataset (KolektorSDD) [40]. Except that the regularization term is not added, the other parameter settings of the CCEANN in this dataset are similar to that in NEU RSDDS-113. KolektorSDD consists of 52 crack images and 347 defect-free images of the plastic surface, in which 37 defective images and 174 defect-free images are randomly selected to build the training set, and the others are chosen to build the test set. Based on the pixel-level ground truths, region-level ground truths are generated. Defective images' sample number in the training set is increased by 1100% based on the horizontal

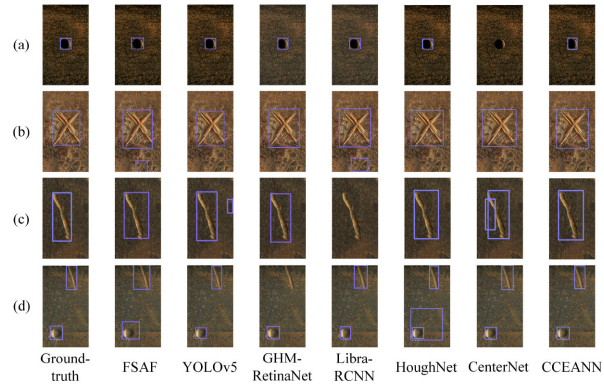

Fig. 10. Visualization results in NEU RSDDS-113.

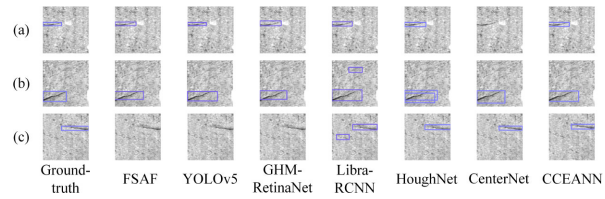

Fig. 11. Visualization results in KolektorSDD.

flip, vertical flip, mirroring operation, scaling operation, contrast change, brightness change, rotation operation, clipping operation, Gauss noise addition, and salt noise addition. Moreover, defective images in the test set are augmented three times by the vertical flip, horizontal flip, and mirror operation. In all, the sample size of the training set and the test set are 618 and 233, respectively.

The visual results in Fig. 11 indicate that compared with the other related frameworks, the CCEANN successfully locates all thin cracks in noisy backgrounds. As shown in Table VIII (the 14th–16th columns), the performance of the CCEANN is superior to that of other frameworks. Besides, it can be observed in Fig. 8(e) that CCEANN's P/R curve completely wraps the other curves, which shows that the CCEANN outperforms other competitive frameworks with the highest AP^{50} (99.58%) in the plastic surface crack detection.

VI. Conclusion

For addressing the data imbalance and complex situations in rail surface defect detection, this article put forward a defect detection framework, CCEANN, in which not only the TDA strategy was introduced, but also the CSFA-Hourglass and CASIoU-CEHM were integrated. In the proposed backbone, on one hand, a novel HGM's structure balanced the network depth and the feature fusion enhanced the effectiveness of feature extraction at each stage, and on the other hand, the CSF transmitted abundant multiscale information between different stages of HGMs. Furthermore, the VRF-CBAM was proposed to highlight representative feature parts, which enhanced CCEANN's performance in detection for multiscale defects, especially large defects. Regarding the flexible anchor-free CASIoU-CEHM, the dynamic coordinate compensation mechanism effectively fine tuned the location and size of detection results in the center-point estimation, in which the CASIoU loss improved the convergence accuracy than other state-of-the-art IoU-based losses according to the consistency of IoU, center-point distance, scale, and area. The experimental results showed that the CCEANN achieved a better detection performance than competitive deep learning-based frameworks in four different defect datasets, which verified its effectiveness, superior generality, and potential of practical application for the intelligent railway inspection.

Although the CCEANN has an obvious performance enhancement in rail surface defect detection, its efficiency can be improved. Hence, model pruning and accelerated calculation methods will be explored to reduce CCEANN's computational complexity and realize online detection in the future.

References

[1] F. Sadeghi, B. Jalalahmadi, T. S. Slack, N. Raje, and N. K. Arakere, "A review of rolling contact fatigue," *J. Tribol.*, vol. 131, pp. 041403-1–041403-15, Jul. 2009.

[2] D. A. Ramatlo, C. S. Long, P.W. Loveday, and D.N. Wilke, "A modelling framework for simulation of ultrasonic guided wave-based inspection of welded rail tracks," *Ultrasonics*, vol. 108, 2020, Art. no. 106215.

[3] M. Karakose and O. Yaman, "Complex fuzzy system based predictive maintenance approach in railways," *IEEE Trans. Ind. Informat.*, vol. 16, no. 9, pp. 6023–6032, Sep. 2020.

[4] 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.

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

[6] M. Nieniewski, "Morphological detection and extraction of rail surface defects," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 9, pp. 6870–6879, Sep. 2020.

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

[8] Z. He, Y. Wang, F. Yin, and J. Lui, "Surface defect detection for high-speed rails using an inverse P-M diffusion model," *Sensor. Rev.*, vol. 36, pp. 86–97, 2016.

[9] H. Yu *et al.*, "A coarse-to-fine model for rail surface defect detection," *IEEE Trans. Instrum. Meas.*, vol. 68, no. 3, pp. 656–666, Mar. 2019.

[10] J. Gan, J. Wang, H. Yu, Q. Li, and Z. Shi, "Online rail surface inspection utilizing spatial consistency and continuity," *IEEE Trans. Syst., Man., Cybern., Syst.*, vol. 50, no. 7, pp. 2741–2751, Jul. 2020.

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

[12] S. Hajizadeh, A. Núñez, and D. M. J. Tax, "Semi-supervised rail defect detection from imbalanced image data," *IFAC*, vol. 49, no. 3, pp. 78–83, 2016.

[13] 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. 1593–1608, Jul. 2018.

[14] D. Zhang *et al.*, "MCnet: Multiple context information segmentation network of no-service rail surface defects," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–9, 2021, doi: 10.1109/TIM.2020.3040890.

[15] J. Wang, Q. Li, J. Gan, H. Yu, and X. Yang, "Surface defect detection via entity sparsity pursuit with intrinsic priors," *IEEE Trans. Ind. Informat.*, vol. 16, no. 1, pp. 141–150, Jan. 2020.

[16] M. Niu, K. Song, L. Huang, Q. Wang, Y. Yan, and Q. Meng, "Unsupervised saliency detection of rail surface defects using stereoscopic images," *IEEE Trans. Ind. Informat.*, vol. 17, no. 3, pp. 2271–2281, Mar. 2021.

[17] L. Zhuang, L. Wang, Z. Zhang, and K.L. Tsui, "Automated vision inspection of rail surface cracks: A double-layer data-driven framework," *Transp. Res. C, Emerg. Technol.*, vol. 92, pp. 258–277, 2018.

[18] D. Zhang, K. Song, Q. Wang, Y. He, X. Wen, and Y. Yan, "Two deep learning networks for rail surface defect inspection of limited samples with line-level label," *IEEE Trans. Ind. Informat.*, to be published, doi: 10.1109/TII.2020.3045196.
[19] X. Jin *et al.*, "DM-RIS: Deep multimodel rail inspection system with improved MRF-GMM and CNN," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 4, pp. 1051–1065, Apr. 2020.
[20] H. Yuan, H. Chen, S. Liu, J. Lin, and X. Luo, "A deep convolutional neural network for detection of rail surface defect," in *Proc. IEEE Veh. Power Propulsion Conf.*, 2019, pp. 1–4.
[21] Y. Cao *et al.*, "GCNET: Non-local networks meet squeeze-excitation networks and beyond," in *2019 IEEE/CVF ICCVW*, 2019, pp. 1971–1980, doi: 10.1109/ICCVW.2019.00246.
[22] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.
[23] Q. Wang *et al.*, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *2020 IEEE/CVF CVPR*, 2020, pp. 11531–11539, doi: 10.1109/CVPR42600.2020.01155.
[24] J. Park *et al.*, "BAM: Bottleneck attention module," in *Proc. BMVC*, 2018.
[25] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. IEEE Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.
[26] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2117–2125.
[27] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, "Cascaded pyramid network for multi-person pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7103–7112.
[28] X. Zhou *et al.*, "Objects as points," 2019, *arXiv:1904.07850*.
[29] H. Law and J. Deng, "Cornernet: Detecting objects as paired keypoints," in *Proc. IEEE Eur. Conf. Comput. Vis.*, 2018, pp. 734–750.
[30] J. Pang, K. Chen, J. Shi, H. Feng, W. Ouyang, and D. Lin, "Libra R-CNN: Towards balanced learning for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 821–830.
[31] G. Jocher *et al.*, "Yolov5," *Code Repository*, 2020. [Online]. Available: https://github. com/ultralytics/yolov5
[32] B. Li, Y. Liu, and X. Wang, "Gradient harmonized single-stage detector," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 8577–8584.
[33] J. Yu, Y. Jiang, Z. Wang, Z. Cao, and T. Huang, "Unitbox: An advanced object detection network," in *Proc. 24th ACM Int. Conf. Multimedia*, 2016, pp. 516–520.
[34] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 658–666.
[35] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-IoU loss: Faster and better learning for bounding box regression," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 12993–13000.
[36] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. IEEE Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.
[37] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," in *Proc. IEEE Eur. Conf. Comput. Vis.*, 2018, pp. 466–481.
[38] C. Zhu, Y. He, and M. Savvides, "Feature selective anchor-free module for single-shot object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 840–849.
[39] N. Samet, S. Hicsonmez, and E. Akbas, "Houghnet: Integrating near and long-range evidence for bottom-up object detection," in *Proc. IEEE Eur. Conf. Comput. Vis.*, 2020, pp. 406–423.
[40] D. Tabernik, S. Šela, J. Skvarč, and D. Skočaj, "Segmentation-based deep-learning approach for surface-defect detection," *J. Intell. Manuf.*, vol. 31, no. 3, pp. 759–776, 2020.

Xuefeng Ni received the B.Sc. degree in electronic information engineering in 2017 from Hunan University, Changsha, China, where he is currently working toward the Ph. D. degree in electronic science and technology with the College of Electrical and Information Engineering.

His research interests include computer vision, image processing, machine learning, and railway inspection.

Ziji Ma received the B.Sc. degree in electronic information engineering from Hunan University, Changsha, China, in 2001, and the Ph.D. degree in information science from the Nara Institute of Science and Technology, Nara, Japan, in 2012.

He is currently an Associate Professor with the College of Electrical and Information Engineering, Hunan University. His research interests include machine vision, signal processing, and V2V communication.

Jianwei Liu received the B.S. degree in electronic information engineering and the M.S. degree in electronic science and technology, in 2014 and 2017, respectively, from Hunan University, Changsha, China, where he is currently working toward the Ph. D. degree in electronic circuit and system with the College of Electrical and Information Engineering.

His research interests include computer vision, image processing, machine learning, and railway fastener inspection.

Bo Shi received the M.S. degree in 2020 from Hunan University, Changsha, China, where he is currently working toward the Ph.D. degree with the College of Electrical and Information Engineering, both in electronic science and technology.

His research interests include machine vision, optical 3-D measurement, and railway inspection.

Hongli Liu received the B.Sc. degree in electrical engineering and the Ph.D. degree in control theory and engineering from Hunan University, Changsha, China, in 1985, and 2000, respectively.

He is currently a Professor with the College of Electrical and Information Engineering, Hunan University. His current research interests include intelligent information processing and transmission technology.