IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 19 September 2023, accepted 18 October 2023, date of publication 26 October 2023, date of current version 9 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3327910

RESEARCH ARTICLE

Crack Detection of Track Slab Based on RSG-YOLO

TANGBO BAI[1,2], BAILE LV[1,2], YING WANG[3], JIALIN GAO[1,2], AND JIAN WANG[1,2]
[1]School of Mechanical-Electronic and Vehicle Engineering, Beijing University of Civil Engineering and Architecture, Beijing 100044, China
[2]Beijing Key Laboratory of Performance Guarantee on Urban Rail Transit Vehicles, Beijing 100044, China
[3]Beijing Institute of Aerospace Control Device, Beijing 100094, China

Corresponding author: Tangbo Bai (baitangbo@bucea.edu.cn)

This work was supported in part by the Beijing Natural Science Foundation under Grant L211007, and in part by the General Project of Scientific Research Program of Beijing Municipal Education Commission under Grant KM202010016003.

ABSTRACT The surface cracks on high-speed railway ballastless track slabs directly influence their lifespan, while the efficiency of damage detection and maintenance is crucial for ensuring operational safety. Leveraging deep learning image processing technology can significantly enhance detection efficiency. Therefore, in response to the specific attributes of ballastless track slab crack detection, this paper introduces the RSG-YOLO model. By implementing a reparameterized dual-fused feature pyramid structure, we bolster the network's feature extraction capacity and curtail the loss of crack features during extraction. SIoU is used to replace CIoU to optimize the bounding box regression loss function, reduce the degrees of freedom of the loss function, and improve the convergence speed The GAM attention mechanism is integrated to heighten the model's responsiveness to diverse channel information. The proposed RSG-YOLO model was evaluated against mainstream models in the field of crack detection. The results demonstrated improved detection accuracy and recall rates. Specifically, when compared to baseline models, our approach exhibited significant advancements in reducing both missed detections and false alarms. These improvements were quantified by a 4.34% increase in crack detection accuracy and a 3.08% rise in mAP_0.5. Consequently, the RSG-YOLO model effectively enables precise detection of track slab cracks.

INDEX TERMS High speed railway, track slab cracks, YOLO, crack detection, image processing.

I. INTRODUCTION

With the rapid development of high-speed railways, the track structure has emerged as an indispensable component of the high-speed rail system, constantly exposed to a complex external environment. Throughout its service life, the ballastless track bed and other supporting structures often encounter challenges related to cracking. The development of cracks in these components can potentially lead to a reduction in the load-bearing capacity of the railway structure, consequently impacting the smooth operation of high-speed trains. thereby, conducting research on crack formation in high-speed railway track panels holds significant scientific value and practical importance.

In recent years, numerous studies [1], [2], [3], [4], and [5] on crack detection. Li et al. [6] utilized infrared imaging for crack detection and achieved effective localization of cracks with a width as small as 0.14mm when the ambient temperature was above 20° Wu et al. [7] employed morphological wavelet operators to decompose road surface images containing cracks, and then extracted the cracks through traditional binarization methods, thereby achieving effective crack extraction in asphalt pavement. Salman et al. [8] proposed an automatic method for identifying cracks in road surface images using Gabor function-based image analysis. This method achieved an initial accuracy of 95% in crack recognition and demonstrated its applicability in road surface crack identification. CHAMBON [9] used a two-dimensional matched filter to define an adaptive mother wavelet and incorporated the detection results into a Markov random field (MRF) to segment and detecting crack structures. Traditional image processing algorithms have demonstrated good performance in crack detection. However, high-speed railway track panel images exhibit characteristics such as low contrast and

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

complex backgrounds, making crack features less prominent and place greater demands on algorithm performance.

Methods based on deep learning can effectively extract deep features of the target and generally achieve better detection results. Chen and Jahanshahi [10] employed a deep learning network that combines Convolutional Neural Networks (CNN) [11] with a Naive Bayesian data fusion approach to detect crack information in video frames. They improved the CNN for crack detection and utilized the Naive Bayesian decision method to eliminate false positives, resulting in enhanced crack detection performance. Fan et al. [12] utilized a deep learning network to identify cracks in images. They applied bilateral filtering to smooth crack images and performed crack segmentation using adaptive thresholding, achieving good extraction of crack information. Xiang et al. [13] combined YOLOv5 and Transformer modules for crack detection to improve the model's crack detection capability. The aforementioned methods show better detection performance when dealing with clear features and simple backgrounds. However, in practical crack detection scenarios, cracks often coexist with complex and diverse backgrounds, leading to suboptimal detection results with a higher frequency of false positives and false negatives.

In order to address these challenges, this study takes into account the actual characteristics of track panel cracks and proposes a crack detection method that combines YOLOv6 3.0 [14] and YOLOv7 [15], utilizing the RSG-YOLO model. The main contributions are as follows:

Firstly, we enhanced the structure of the feature fusion component of our model, thereby augmenting its capability for feature extraction. This enhancement enables us to acquire more detailed information regarding rail plate cracks. Secondly, the loss function has been refined to improve the model's accuracy in crack localization, thereby reducing instances of false alarms and missed detections. Lastly, a novel attention mechanism has been introduced to heighten the model's awareness of channel and spatial information, thus further refining its ability to detect crack-related data. These contributions significantly enhance overall performance and the facilitation of practical applications. The proposed method is evaluated using a custom track panel crack dataset.

II. RELATED WORKS

A. YOLOv7 MODEL

The YOLOv7 network model consists of four main components: Input, Backbone, Neck, and Head. Firstly, the input undergoes preprocessing operations, including data augmentation, to prepare the images for further processing. These preprocessed images are then passed through the Backbone, which extracts relevant features from them. Subsequently, the extracted features are fused using the Neck module to generate feature maps of different sizes, namely large, medium, and small. Finally, the fused features are fed into the Head, where object detection is performed, resulting in the output of detection results.

The Backbone module of the YOLOv7 network consists of several components, namely the CBS convolutional module, Efficient Layer Aggregation Networks (ELAN), MP module, and SPPCSPC module. The CBS convolutional module comprises convolutional layers, batch normalization (BN) layers, and the SiLU activation function. The ELAN module consists of multiple CBS convolutional modules and enhances the model's learning efficiency and convergence speed by controlling the longest and shortest gradient paths. The SPPCSPC module introduces parallel MaxPool layers at four different scales within a sequence of convolutional modules. This allows the module to adapt to feature maps of different resolutions and addresses the issue of repeated feature extraction in the model. In the MPConv module, the maximum pooling layer expands the receptive field of the current feature layer and combines it with the processed feature information from the convolutional module, thereby improving the network's generalization capability. In the Neck module, YOLOv7 adopts the traditional PAFPN structure, which is the same as the YOLOv5 [16] network. In the detection head, the baseline YOLOv7 model employs detection heads that represent three different target sizes: large, medium, and small.

B. FEATURE PYRAMID STRUCTURE

The Feature Pyramid (FP) structure is a state-of-the-art detector commonly used for detecting objects at different scales. It extracts spatial features from the last feature layer, allowing strong semantic features to propagate along a top-down pathway, significantly improving the accuracy of object detection. However, due to the pooling effect, the top-down pathway cannot accurately preserve the positional information of cracks. To accurately transmit accurate positional information of cracks, a bottom-up pathway is needed to compensate for the lost information from the bottom-level feature maps. YOLOv7, similar to YOLOv5, adopts the traditional PAFPN (Path Aggregation Feature Pyramid Network) structure, which combines the top-down Feature Pyramid Network (FPN) and the bottom-up Pyramid Attention Network (PAN) to perform multi-scale feature fusion. In this structure, FPN propagates strong semantic features from higher levels to enhance the entire pyramid, but it only enhances semantic information without transmitting localization information. To address this issue, a bottom-up pyramid is added after FPN to complement it by propagating low-level localization features upwards. This combined pyramid incorporates both semantic and localization information, greatly improving the detection performance of the model.

C. INTERSECTION OVER UNION (IoU) LOSS

In object detection tasks, the loss function for bounding box regression is crucial, and IoU loss [17] is used to measures the overlap between predicted and ground truth bounding boxes to accurately localize the detected objects. In 2019, to address the issue of measuring the performance of the bounding box regression loss when the predicted and ground truth boxes do

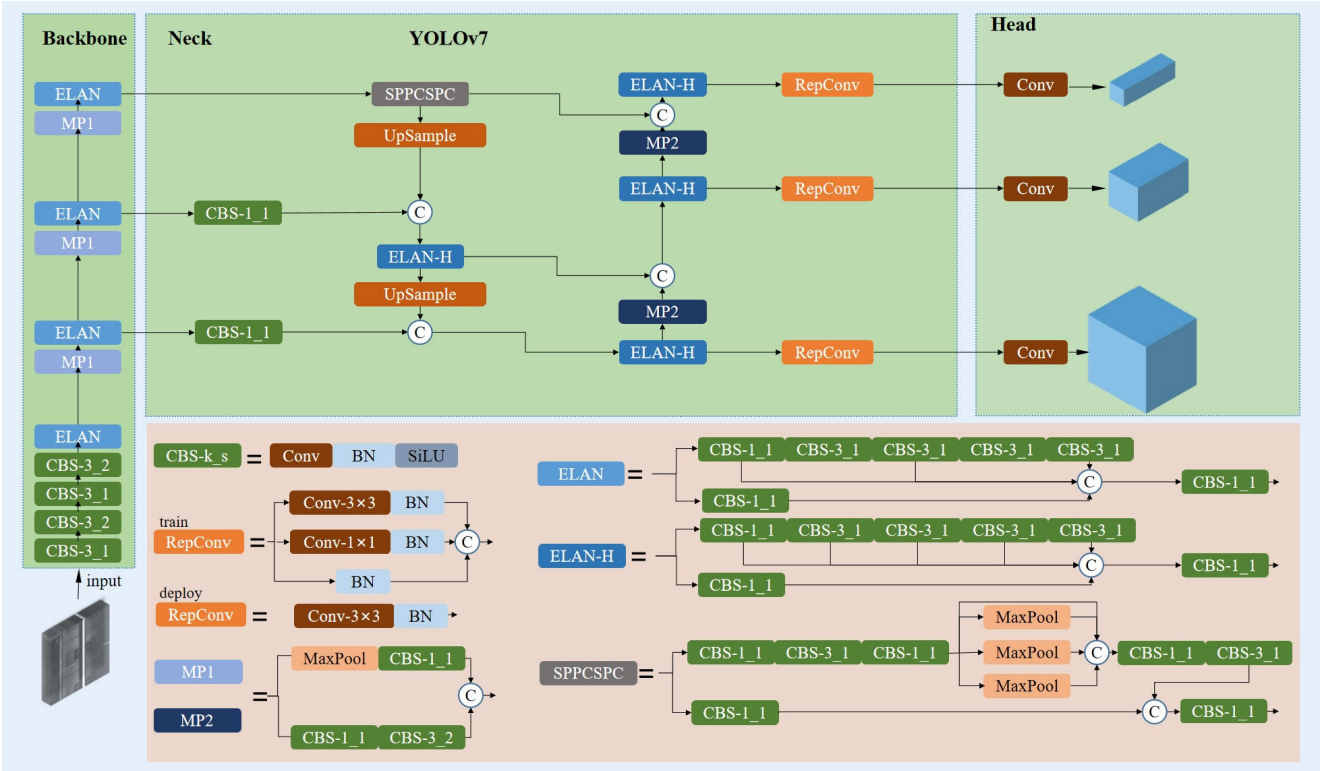

FIGURE 1. YOLOv7 structure diagram.

not intersect in IoU loss, GIoU loss [18] introduced the concept of the minimum enclosing box (the smallest rectangle that can enclose both the predicted and ground truth boxes) to obtain the proportion of the predicted and ground truth boxes within the enclosing area. In 2020, to handle the case when the predicted and ground truth boxes are aligned horizontally or vertically, where GIoU loss degrades to IoU loss, DIoU loss [19] introduced the distance between the centers of the predicted and ground truth boxes, minimizing the distance between the two boxes and improving convergence speed. Additionally, CIoU loss builds upon DIoU loss by incorporating the aspect ratio of the bounding boxes into the loss function, thereby further enhancing the regression accuracy.

However, when the aspect ratio of the predicted and ground truth boxes is the same, the aspect ratio penalty term in the CIoU loss remains constant at 0. Additionally, poor samples can have a significant impact on the regression loss, leading to larger fluctuations during the convergence process. Therefore, in this paper, a more refined representation of IoU loss is proposed to reduce the fluctuations during the convergence process and improve the localization accuracy of bounding boxes.

D. THE ATTENTION MECHANISM

Attention mechanisms, as a data processing approach, have been widely employed in deep learning networks. Introducing attention mechanisms allows the model to focus on the target regions and gather more detailed information. This helps to reduce interference from irrelevant information and ultimately improves the overall detection performance of the model.

In this study, multiple experiments were conducted, and it was observed that some crack information was not fully utilized, which had a significant impact on the crack detection results. Therefore, this paper aims to enhance the detection capability of the model by enhancing the attention to crack information in the network, thus reducing the occurrence of crack omission.

III. PROPOSED METHOD

A. TECHNICAL ROUTE

Figure 2 illustrates the technical approach employed for track slab crack detection. Initially, specialized equipment is utilized to capture images of the track slabs. The images containing crack information undergo cutting and data enhancement processes. These procedures aim to retain crack information, eliminate invalid data, and enhance the network training speed. Subsequently, the preprocessed track slab surface crack dataset is fed into the model proposed in this paper for training purposes. The trained model is then employed to predict cracks on the track slab's surface. Finally, the detection results for surface cracks on the track slab have been obtained.

B. REPBI-PAN STRUCTURE FUSION IMPROVEMENT

In object detection networks, effective multi-scale fusion networks play a crucial role in improving detection performance. YOLOv7's feature fusion module adopts the traditional

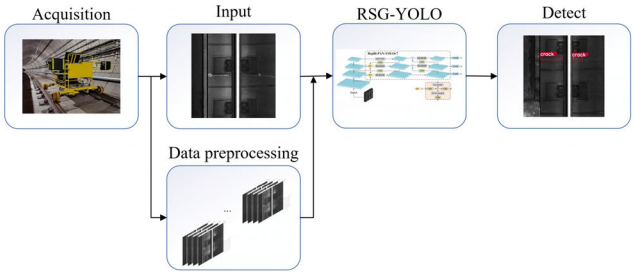

FIGURE 2. Technical route of the proposed method.

PAFPN (Path Aggregation Feature Pyramid Network) structure, which adds a top-down pathway into the FPN structure. BiFPN (Bi-directional Feature Pyramid Network with Learnable Weights) introduces learnable weights for different input features. PRB-FPN (Parallel Residual Bi-directional Feature Pyramid Network) utilizes a parallel residual feature pyramid structure with dual fusion to extract high-quality features, thereby achieving precise localization.

In this study, the YOLOv7 model was used for multiple crack detection experiments. It was observed that during the feature fusion process, some crucial crack contour information and shallow texture information were not fully utilized. This has a certain impact on crack detection and may lead to missed or false detections. Therefore, in this paper, improvements were made to the feature fusion part (Neck) of YOLOv7. The goal was to fully leverage the crack information extracted by the backbone network while considering the input-output relationship. These modifications were made without affecting any other network structures. The aim was to enhance the network's feature fusion capability and reduce the occurrence of false detections and missed cracks.

The YOLOv6 3.0 version, released in January 2023, introduced the RepBi-PAN structure, which possesses stronger feature extraction capabilities. In the backbone network, shallow features have higher resolution and contain abundant spatial location information, which is beneficial for accurate object localization. To aggregate these shallow features into the network, a common approach is to add fusion layers and corresponding detection heads into the Feature Pyramid Network (FPN). However, this can result in increased computational costs.

The improved Neck component utilizes the RepBi-PAN bidirectional linking structure, which introduces the shallower features from the backbone network in the top-down pathway. This allows for more efficient multi-scale fusion of shallow features, thereby enhancing the expressive power of the feature fusion module. This particular structure preserves more accurate crack position information, which is crucial for precise crack localization.

C. SIoU LOSS

The original network loss function CIoU loss calculation formula is as follows:

$$L_{CIoU} = 1 - I_{IoU} + \frac{\rho^2\left(b, b^{gt}\right)}{c^2} + \alpha v \tag{1}$$

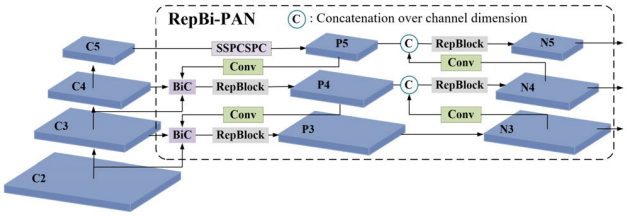

FIGURE 3. Structure of RepBi-PAN.

$$v = \frac{4}{\pi^2}\left(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h}\right)^2 \tag{2}$$

$$\alpha = \frac{v}{(1 - I_{IoU}) + v} \tag{3}$$

where:

ρ represents the Euclidean distance between the center points of the two boxes,

b is the center point of the predicted box,

b^{gt} is the central point of the ground truth box,

c is the diagonal length of the minimum enclosing box of the predicted and ground truth boxes,

a is a balancing parameter,

v is a parameter that measures the consistency of aspect ratios,

w^{gt} is the width of the ground truth box,

h^{gt} is the height of the ground truth box,

w is the width of the predicted box,

h is the height of the predicted box.

It can be observed that when the aspect ratio of the ground truth box and the predicted box is the same, the penalty term for aspect ratio, v becomes 0. In such cases, the stability of this loss function is compromised.

In this paper, the original CIoU loss function is replaced with the SIoU (Segmentation Intersection over Union) loss [20], which introduces an angle penalty term to effectively reduce the overall degrees of freedom in the loss function. The SIoU loss function consists of four components: angle cost, distance cost, shape cost, and IoU cost.

1) ANGULAR COST

During the convergence process, the model initially attempts to minimize the angle between the predicted and ground truth boxes. If the angle exceeds 45 degrees, the angle cost is calculated as follows:

The formula for calculating the angle cost is as:

$$\Lambda = 1 - 2 * \sin^2\left(\arcsin(x) - \frac{\pi}{4}\right) \tag{4}$$

where:

$$x = \frac{c_h}{\sigma} = \sin(\alpha) \tag{5}$$

$$\sigma = \sqrt{\left(b_{c_x}^{gt} - b_{c_x}\right)^2 + \left(b_{c_y}^{gt} - b_{c_y}\right)^2} \tag{6}$$

$$c_h = \max\left(b_{c_y}^{gt}, b_{c_y}\right) - \min\left(b_{c_y}^{gt}, b_{c_y}\right) \tag{7}$$

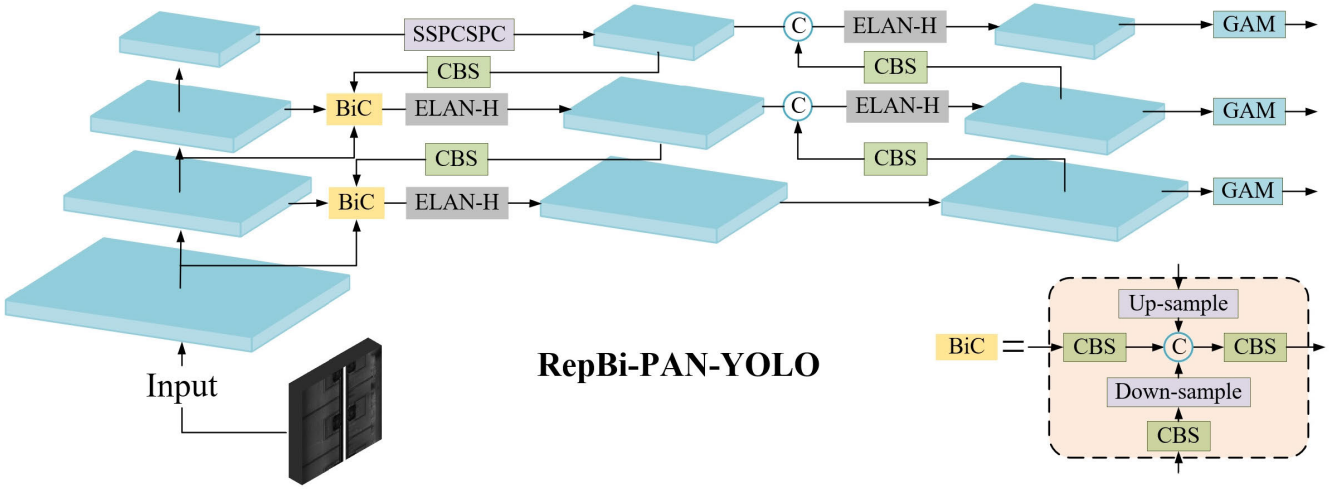

FIGURE 4. RSG-YOLO structure diagram.

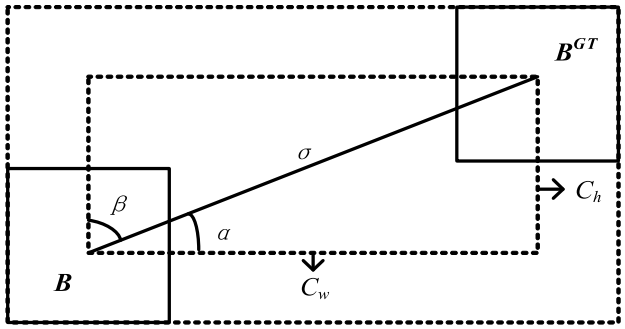

FIGURE 5. Angular costing process.

2) DISTANCE COST

We redefine the distance cost Δ, taking into account the previously defined angle cost Λ:

$$\Delta = \sum_{t=x,y} \left(1 - e^{-\gamma \rho_t}\right) \tag{8}$$

where:

$$\rho_x = \left(\frac{b_{c_x}^{gt} - b_{c_x}}{c_w}\right)^2 \tag{9}$$

$$\rho_y = \left(\frac{b_{c_y}^{gt} - b_{c_y}}{c_h}\right)^2 \tag{10}$$

$$\gamma = 2 - \Lambda \tag{11}$$

As β approaches 0, the contribution of distance cost significantly decreases. On the other hand, as β approaches 45°, the distance cost Δ contributes more. With increasing angle, γ is defined as a distance value that prioritizes time.

3) SHAPE COST

The shape cost Ω is defined as:

$$\Omega = \sum_{t=w,h} \left(1 - e^{-\omega_t}\right)^{\theta} \tag{12}$$

where:

$$\omega_w = \frac{|w - w^{gt}|}{\max(w, w^{gt})}, \omega_h = \frac{|h - h^{gt}|}{\max(h, h^{gt})} \tag{13}$$

The value of θ defines the importance of the shape cost, and it is unique for each dataset. When θ is set to 1, it optimizes the shape and restricts its free movement of the shape. To avoid excessive shape loss that could affect the movement of the predicted bounding box, the value is set to 1 in this paper.

4) IoU LOSS

IoU loss is the intersection ratio union of the real box and the prediction box. The formula is as follows:

$$IoU = \frac{|B \cap B^{GT}|}{|B \cup B^{GT}|} \tag{14}$$

Combining equations (8) and (12), the final definition of the SIoU loss function is as follows:

$$L_{SIoU} = 1 - I_{IoU} + \frac{\Delta + \Omega}{2} \tag{15}$$

The original IoU loss function is enhanced by introducing angle cost Δ and distance cost Ω. This modification reduces the probability of having penalty terms equal to zero, resulting in a more stable convergence process. Additionally, it improves the accuracy of inference.

D. GLOBAL ATTENTION MECHANISM ATTENTION (GAM ATTENTION)

In recent years, various attention mechanisms have emerged, and their performance has continuously improved. Squeeze-and-Excitation Networks (SENet) [21], Convolutional block attention module (CBAM) [22], Bottleneck attention module (BAM) [23], Triplet attention module (TAM) [24] While attention mechanisms have achieved good results, they often focus on two dimensions and may not fully utilize the information from all three dimensions. In this paper, we propose

the use of GAM attention to fully leverage information from all three dimensions, leading to improved performance.

The structure of the GAM is illustrated in Figure 5. This mechanism incorporates the sequential channel-spatial attention mechanism from CBAM (Convolutional Block Attention Module) and introduces redesigned modules. Given an input feature map $F1 \in R^{C*H*W}$, the intermediate states F2 and the output F3 are defined as follows:

$$F2 = Mc\,(F1) \otimes F1 \tag{16}$$

$$F3 = Ms\,(F2) \otimes F2 \tag{17}$$

where Mc and Ms are the channel and spatial attention maps, respectively; $\otimes$ denotes element-wise multiplication.

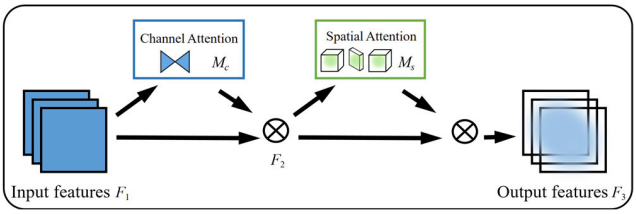

FIGURE 6. The overview of GAM.

The channel attention sub-module transforms the input feature map F1 by performing dimensionality reduction. It then passes through a two-layer perceptron and undergoes dimensionality restoration before being processed by a sigmoid function. The process is illustrated in Figure 6.

The spatial attention submodule primarily employs convolution operations for processing. Initially, the input is subjected to a 7-kernel size convolution operation to reduce the number of channels, thus reducing computational complexity. Subsequently, it undergoes another convolution operation to increase the number of channels, ensuring uniformity in channel dimensions. Finally, the output is further processed through a sigmoid function. The entire process is visually depicted in Figure 7.

IV. EXPERIMENTAL SETUP AND RESULTS ANALYSIS

Experimental Setup and Hardware Environment.

A. PREPARATION FOR THE EXPERIMENT

1) DATA COLLECTION

The image data utilized in this study was acquired through a novel rail inspection device developed by a company based in Beijing. This device was purposefully designed to capture images of CTRSII-type rail panels on high-speed railways during nighttime inspections. Equipped with a laser line scanning camera, the device enables image acquisition at a high resolution of 4096 × 4096 pixels.

2) DATA SET ESTABLISHMENT

The collected images underwent an initial screening process to identify the original images that contained cracks. Subsequently, the grayscale images, with a pixel resolution of

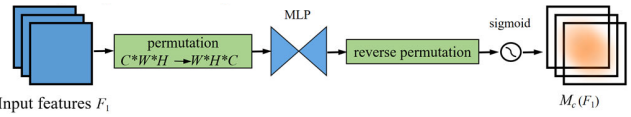

FIGURE 7. Channel attention submodule.

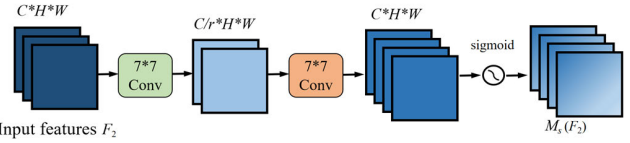

FIGURE 8. Spatial attention submodule.

TABLE 1. Experimental environment and configuration.

Project	Environment
System	Windows 10, 64
GPU	GTX3090
CPU	I7-13700
RAM size	32GB
Programming language	Python 3.8.13
Deep learning framework	Torch1.12.0+cu116

4096 × 4096, were subjected to segmentation and a secondary screening. This approach effectively improved the input quality by mitigating excessive feature loss caused by extensive image adjustments, while simultaneously enhancing detection speed and maintaining accuracy. A total of 663 rail panel images containing cracks were carefully selected and further divided into training, testing, and validation sets in an 8:1:1 ratio. Given the limited availability of training samples, data augmentation techniques were employed on the segmented images. These techniques included flipping, Gaussian filtering, and brightness adjustment, effectively expanding the training dataset to 2655 images. This augmentation strategy successfully addressed the data scarcity issue.

B. EVALUATION INDEX

To assess the performance of the improved model, this study primarily employed precision (P), recall (R), and mean Average Precision at IoU 0.5 (mAP_0.5) as the evaluation metrics. Precision is calculated as the ratio of correctly located cracks to the total number of detected cracks, while recall is the ratio of correctly located cracks to the total number of cracks in the samples. The calculation formulas are as follows:

$$P = \frac{T_P}{T_P + F_N} \times 100\% \tag{18}$$

$$R = \frac{T_P}{T_P + F_P} \times 100\% \tag{19}$$

$$A_{AP} = \int_0^1 P\,(R)\,dR \tag{20}$$

FIGURE 9. Intelligent rail inspection vehicle.

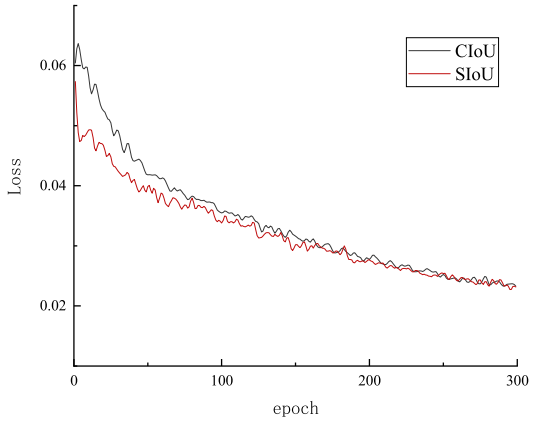

FIGURE 10. Function comparison.

where: T_P represents correctly predicted crack results, F_N represents undetected crack results, F_P represents falsely identified results as cracks.

C. RESULT ANALYSIS

1) LOSS FUNCTION COMPARISON

Under identical experimental conditions, the convergence of the enhanced model's loss function was validated. The training curves depicting the two loss functions in relation to the number of iterations are illustrated in the graph below.

The graph clearly demonstrates the convergence of the model's loss as the number of iterations increases. Notably, the loss function employed in this study exhibits a faster convergence rate. Consequently, the utilization of SIoU as the loss function in the improved model holds substantial significance in enhancing its performance.

2) MODEL COMPARISON

The curves in the graph show the mAP_0.5 and precision improvement for crack detection on railway sleepers before and after enhancing the model enhancement. The enhanced model in this study converges faster and exhibits an increase of 2.39% in crack detection precision

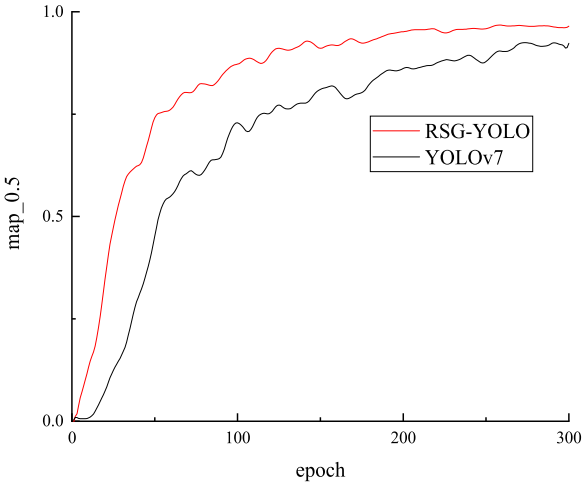

FIGURE 11. Model mAP_0.5 curve comparison.

TABLE 2. Comparison of different model results.

model	P	R	Map_0.5
YOLOv5	0.8668	0.8738	0.9050
YOLOv7	0.8928	0.8932	0.9158
YOLOx	0.8722	0.8828	0.9076
Faster-RCNN	0.8683	0.8824	0.8922
RSG-YOLO	0.9362	0.8955	0.9466

and a 3.08% improvement in mAP_0.5. Compared to the YOLOv7 network, the RSG-YOLO network demonstrates better performance in detecting cracks on railway sleepers.

To assess the detection algorithm's performance, the proposed method in this study was compared to widely adopted models, namely YOLOv5, YOLOx, and Faster-RCNN. The comparison was carried out using consistent initial conditions and training parameters to ensure a fair evaluation. The specific results are presented in the table below:

3) DETECTING RESULTS

To elucidate the learning effects of the GAM attention mechanism, we employed the Gradient-weighted Class Activation Mapping (Grad-CAM) [25] technique. Grad-CAM is a discriminative visual classification method used to pinpoint specific regions within images. This functionality enhances the explainability and visual interpretability of neural network models. Grad-CAM operates by utilizing a particular layer within an image classification model to generate a localization map. By applying global average pooling to the gradients of the convolutional layers, channel-wise weights are computed. These weights are then used to linearly combine the feature maps, ultimately generating a Class Activation Map (CAM) [26] Consequently, this approach enables the visualization of neural networks. By implementing this methodology, we effectively visualized the effects of the GAM, thus providing an improved means of assessing the model's performance.

The visualization results are presented in Figure 12. Subfigures b), d), and f) illustrate the detection outcomes using

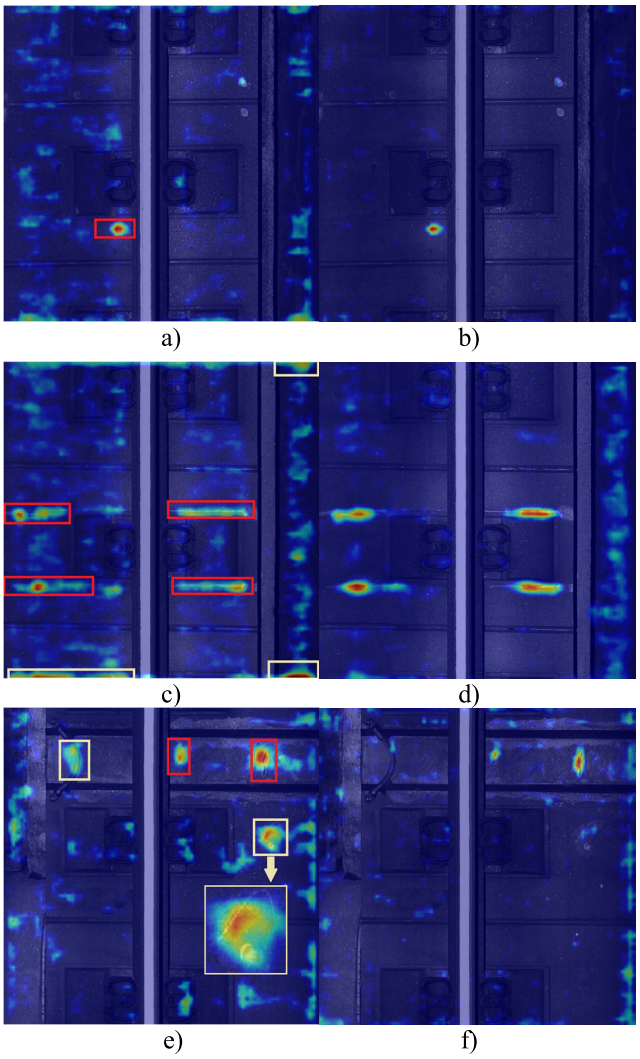

FIGURE 12. Visual interpretation of the effect of GAM by Grad-CAM on crack datasets.

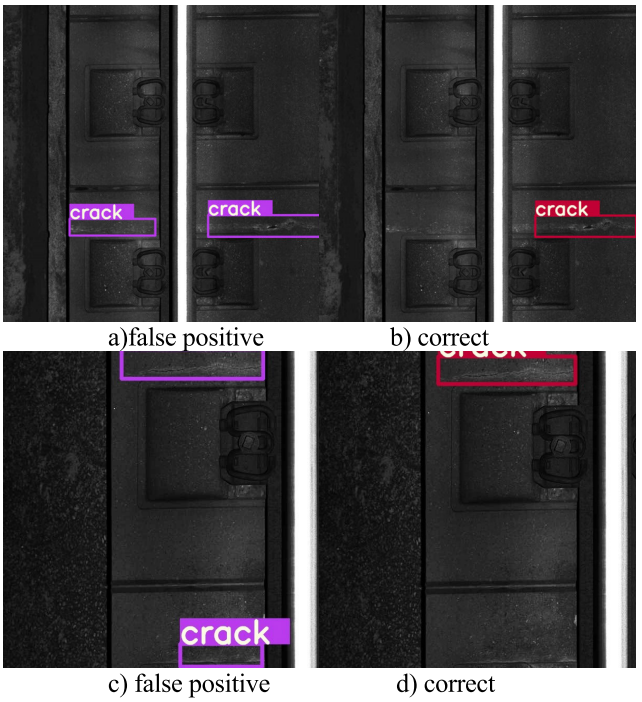

FIGURE 13. Comparison of test results: left (YOLOv7), right (RSG-YOLO).

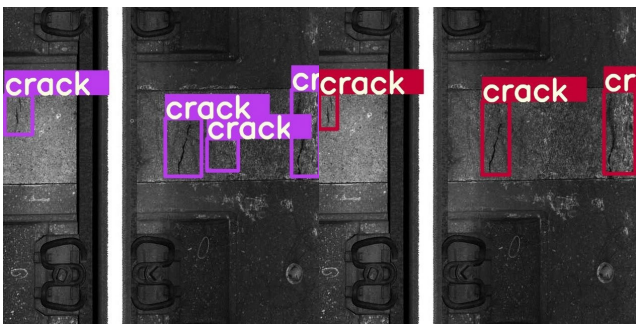

FIGURE 14. Result of survey: left (YOLOv7), right (RSG-YOLO).

the GAM, while subfigures a), c), and e) do not utilize the GAM. The red boxes demarcate actual crack regions, while the yellow boxes highlight areas mistaken for cracks. In subfigure a), the model exhibits significant attention to regions without cracks. In subfigure b), precise focus is directed exclusively to the area containing the crack. Subfigure c) reveals the model's tendency to emphasize regions at the image periphery. In contrast, subfigure d) accurately directs attention to the cracked region. In subfigure e), attention is drawn to the ground ropes and marked lines, resulting in misclassification. subfigure f) appropriately focuses on the crack region, leading to reduced attention on ropes and ground markings.

Select a few pictures of track plate cracks and utilize both the model proposed in this paper and the YOLOv7 model to predict and locate the cracks. The results are illustrated in the figure. As observed from the figure, the YOLOv7 model identifies the suspected crack location as a crack, while the model proposed in this paper accurately recognizes it, thereby reducing the false detection rate. It should be noted that the joint between the rail platform and the track slab bears similarities to the characteristics of cracks, leading to misclassification by the YOLOv7 model. However, the model presented in this paper effectively distinguishes itself, exhibiting superior positioning capabilities.

The cracks in the track slab are complex, and the detection performance of the YOLOv7 network is not satisfactory. It often misidentifies cables and other backgrounds as cracks, resulting in numerous false detections. In contrast, the model proposed in this paper demonstrates exceptional crack detection capabilities. As depicted in Figure 14, the YOLOv7 model misclassifies the background as a crack, whereas the model presented in this paper achieves precise crack localization. Moreover, the proposed model employs more reasonable positioning anchor frames, enabling effective crack detection.

V. CONCLUSION AND FUTURE WORK

The RSG-YOLO detection model is proposed in this paper to address the rail track crack detection problem. Its main contributions are as follows:

1. By incorporating the idea of reparameterized fusion feature pyramid and YOLOv7, the Neck component is improved to enhance the model's feature extraction capability and extract detailed information about track slab cracks.

2. The original CIoU loss function is optimized by using the SIoU loss function to enhance the model's localization ability for cracks, thereby improving the detection accuracy and reducing false detections and missed detections.

3. The GAM attention mechanism is added to the detection head to enhance the model's sensitivity to channel and spatial information, improving the localization ability for crack information and enhancing the overall performance of the model.

Experimental results demonstrate that the proposed rail track slab cracks detection method achieves higher precision and recall rates, accurately localizes rail track cracks, and significantly improves the detection effect of cracks in complex background. Compared to methods such as YOLOv5 and YOLOv7, this method achieves higher recognition accuracy, with a final precision of 93.6%, recall rate of 89.5%, and mAP_0.5 of 94.7%.

Future work includes segmenting and quantitatively analyzing crack images, determining damage levels based on maintenance rules for ballastless high-speed railway tracks, and studying crack expansion for tracking purposes. These efforts aim to provide assistance in maintaining rail track crack maintenance work.

REFERENCES

[1] H. Cho, H.-J. Yoon, and J.-Y. Jung, "Image-based crack detection using crack width transform (CWT) algorithm," *IEEE Access*, vol. 6, pp. 60100–60114, 2018, doi: 10.1109/ACCESS.2018.2875889.

[2] R. Ali, J. H. Chuah, M. S. A. Talip, N. Mokhtar, and M. A. Shoaib, "Structural crack detection using deep convolutional neural networks," *Autom. Construct.*, vol. 133, Jan. 2022, Art. no. 103989, doi: 10.1016/j.autcon.2021.103989.

[3] W. Song, G. Jia, D. Jia, and H. Zhu, "Automatic pavement crack detection and classification using multiscale feature attention network," *IEEE Access*, vol. 7, pp. 171001–171012, 2019, doi: 10.1109/ACCESS.2019.2956191.

[4] Z. Qu, J. Mei, L. Liu, and D.-Y. Zhou, "Crack detection of concrete pavement with cross-entropy loss function and improved VGG16 network model," *IEEE Access*, vol. 8, pp. 54564–54573, 2020, doi: 10.1109/ACCESS.2020.2981561.

[5] S. Meng, S. Kuang, Z. Ma, and Y. Wu, "MtlrNet: An effective deep multitask learning architecture for rail crack detection," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–10, Jun. 2022, doi: 10.1109/TIM.2022.3181940.

[6] Z.-W. Li, X.-Z. Liu, H.-Y. Lu, Y.-L. He, and Y. Zhou, "Surface crack detection in precasted slab track in high-speed rail via infrared thermography," *Materials*, vol. 13, no. 21, p. 4837, Oct. 2020, doi: 10.3390/ma13214837.

[7] G. Wu, X. Sun, L. Zhou, H. Zhang, and J. Pu, "Research on morphological wavelet operator for crack detection of asphalt pavement," in *Proc. IEEE Int. Conf. Inf. Autom. (ICIA)*, Ningbo, China, Aug. 2016, pp. 1573–1577, doi: 10.1109/ICInfA.2016.7832069.

[8] M. Salman, S. Mathavan, K. Kamal, and M. Rahman, "Pavement crack detection using the Gabor filter," in *Proc. 16th Int. IEEE Conf. Intell. Transp. Syst. (ITSC)*, The Hague, The Netherlands, Oct. 2013, pp. 2039–2044, doi: 10.1109/ITSC.2013.6728529.

[9] S. Chambon, P. Subirats, and J. Dumoulin, "Introduction of a wavelet transform based on 2D matched filter in a Markov random field for fine structure extraction: application on road crack detection," *Proc. SPIE*, vol. 7251, Feb. 2009, Art. no. 72510A, doi: 10.1117/12.805437.

[10] F.-C. Chen and M. R. Jahanshahi, "NB-CNN: Deep learning-based crack detection using convolutional neural network and Naïve Bayes data fusion," *IEEE Trans. Ind. Electron.*, vol. 65, no. 5, pp. 4392–4400, May 2018, doi: 10.1109/TIE.2017.2764844.

[11] L. O. Chua and T. Roska, "The CNN paradigm," *IEEE Trans. Circuits Syst. I, Fundam. Theory Appl.*, vol. 40, no. 3, pp. 147–156, Mar. 1993, doi: 10.1109/81.222795.

[12] R. Fan, M. J. Bocus, Y. Zhu, J. Jiao, L. Wang, F. Ma, S. Cheng, and M. Liu, "Road crack detection using deep convolutional neural network and adaptive thresholding," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Paris, France, Jun. 2019, pp. 474–479, doi: 10.1109/IVS.2019.8814000.

[13] X. Xiang, Z. Wang, and Y. Qiao, "An improved YOLOv5 crack detection method combined with transformer," *IEEE Sensors J.*, vol. 22, no. 14, pp. 14328–14335, Jul. 2022, doi: 10.1109/JSEN.2022.3181003.

[14] C. Li, L. Li, Y. Geng, H. Jiang, M. Cheng, B. Zhang, Z. Ke, X. Xu, and X. Chu, "YOLOv6 v3.0: A full-scale reloading," 2023, *arXiv:2301.05586*.

[15] C.-Y. Wang, A. Bochkovskiy, and H.-Y.-M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 7464–7475.

[16] X. Zhu, S. Lyu, X. Wang, and Q. Zhao, "TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 2778–2788.

[17] D. Zhou, J. Fang, X. Song, C. Guan, J. Yin, Y. Dai, and R. Yang, "IoU loss for 2D/3D object detection," in *Proc. Int. Conf. 3D Vis. (3DV)*, Quebec City, QC, Canada, Sep. 2019, pp. 85–94, doi: 10.1109/3DV.2019.00019.

[18] 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/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 658–666.

[19] 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.*, Apr. 2020, vol. 34, no. 7, pp. 12993–13000, doi: 10.1609/aaai.v34i07.6999.

[20] Z. Gevorgyan, "SIoU loss: More powerful learning for bounding box regression," May 2022, *arXiv:2205.12740*, doi: 10.48550/arXiv.2205.12740.

[21] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[22] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

[23] J. Park, S. Woo, J.-Y. Lee, and I. S. Kweon, "BAM: Bottleneck attention module," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2018, pp. 1–14.

[24] D. Misra, T. Nalamada, A. U. Arasanipalai, and Q. Hou, "Rotate to attend: Convolutional triplet attention module," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 3139–3148.

[25] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[26] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jan. 2022, pp. 2921–2929, Accessed: Jun. 2016. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2016/html/Zhou_Learning_Deep_Features_CVPR_2016_paper.html

TANGBO BAI received the Ph.D. degree in mechanical engineering from the China University of Petroleum, Beijing, China, in 2016.

He is currently an Associate Professor with the School of Mechanical-Electronic and Vehicle Engineering, Beijing University of Civil Engineering and Architecture, Beijing. His research interests include mechanical equipment condition monitoring and fault diagnosis, and rail infrastructure detection.

BAILE LV received the B.E. degree in control engineering from the School of Mechanical-Electronic and Vehicle Engineering, Beijing University of Civil Engineering and Architecture, Beijing, in 2021, where he is currently pursuing the M.D. degree in mechanical engineering.

His research interests include computer vision, image processing, and track slab damage detection.

YING WANG received the Ph.D. degree in geological resources and geological engineering from the China University of Petroleum, Beijing, China, in 2016.

She is currently a Senior Engineer with the Beijing Institute of Aerospace Control Devices, Beijing. She is also a part-time Senior Engineer with the Qingdao Marine Science and Technology Center, Qingdao. Her research interests include signal process and navigation control.

JIALIN GAO received the M.E. degree in control engineering from the School of Mechanical-Electronic and Vehicle Engineering, Beijing University of Civil Engineering and Architecture, Beijing, China, in 2023. He is currently pursuing the Ph.D. degree in vehicle utilization engineering with the School of Mechanical and Electronic Control Engineering, Beijing Jiaotong University, Beijing.

His research interests include computer vision and large language models.

JIAN WANG received the B.E. degree in control engineering from the School of Mechanical-Electronic and Vehicle Engineering, Beijing University of Civil Engineering and Architecture, Beijing, in 2021, where he is currently pursuing the M.D. degree in mechanical engineering.

His research interests include fault diagnosis and deep learning.