Responses to Reviewers' Comments

Manuscript ID 0090

Railway Wheel Tread Defect Recognition Method Using Improved Convolutional Neural Network Technology

We thank the respected reviewers for their interest in our manuscript and for their helpful comments that will definitely improve our manuscript and we have tried to do our best to respond to the points raised. The referees have brought up some good points and we appreciate the opportunity to clarify our research objectives and results. As indicated below, we have checked all the comments provided by the reviewers and have made necessary changes accordingly to their indications. For your convenience, the comments/questions are in black and our responses are in blue. The changes in the main text are also highlighted in blue. We hope that the new version of the paper passes your qualification criteria. However, we are open to further modifications if you think it is necessary. Thank you very much for your wonderful services.

Reviewer: 1

Comments:

This paper proposes a train wheel set tread defect recognition model based on attention feature fusion to address the under utilization of shallow features extracted from wheel set tread defects and the lack of defect samples during model training.

Question 1. My only minor comment is that, some figures (such as Fig. 1) are not of high resolution.

Response:

Thank you for your valuable comments. We agree that the resolution of the figures affects the readability of the article. We have made the necessary adjustments in the revised text.

Reviewer:2

Comments:

This paper develops a small target defect detection module and proposes a railway wheel tread defect recognition method based on an improved convolutional neural network. The selection has application value.

Question 1. Check the citation format of the references.

Response:

Thank you very much for pointing out these errors. We apologize for the reference formatting issues in the manuscript and have carefully revised them according to your suggestions.

Question 2. Check the reference format of the picture.

Response:

Thank you very much for pointing out these errors. We apologize for the reference formatting issues related to the images in the manuscript and have carefully revised them according to your suggestions.

Question 3. Fig.2 shows the situation that the text and figure span pages.

Response:

We apologize for the issue with the text and image of Fig. 2 spanning across pages in the manuscript, and have carefully revised it according to your suggestions..

Question 4. Explain the formula with "where" at the top.

回应:

We have carefully revised the text according to your suggestions.

Question 5. Attention segment spacing.

Response:

Thank you very much for pointing out these errors. We apologize for the spacing issues between paragraphs in the manuscript and have carefully revised them according to your suggestions.

Question 6. If you can, adjust the format using Latex.

Response:

Thank you very much for your comments, but we chose to use Word to revise and do a serious revision of the paper.

Question 7. Please describe the ablation experiments in detail.

Response:

Thank you very much for your valuable suggestions. Based on your recommendations, we have provided a more detailed description of the ablation experiments:

We clarified the dataset division and the control groups used during the experiments. Compared to the traditional YOLOV3 and YOLOV5, the network we designed demonstrates stronger recognition capability under the same environment and number of iterations. The proposed network structure reduces the difficulty of feature learning and eliminates information redundancy in unbalanced data.

Railway Wheel Tread Defect Recognition Method Using Improved Convolutional Neural Network Technology

He Jing[1][0000-0002-3650-3270], Zhipeng Ouyang[2] and Zhang Changfan[3][0000-0002-7439-1775]

[1] College of Electrical and Information Engineering, Hunan University of Technology, Zhuzhou 412000, Hunan, China
[2] College of Railway Transportation, Hunan University of Technology, Zhuzhou, Hunan 412000, China
[3] College of Railway Transportation, Hunan University of Technology, Zhuzhou, Hunan 412000, China
E-mail: hejing@263.net

Abstract. Wheel tread defect recognition is a crucial step to ensure the safety of the train wheel-rail system service. However, the diverse and complex nature of wheel tread defects, coupled with the presence of minor defect features, poses significant challenges to accurately identifying the defect by existing deep convolutional neural network. To address this issue, we develop a small target defect detection module and propose a railway wheel tread defect recognition method based on an improved convolutional neural network. First, a deformable convolutional deformable attention-enhanced bottleneck module is designed to achieve adaptive adjustment of the network receptive field in the backbone network. Secondly, an adaptive spatial and channel enhancement module is constructed to further improve the network's sensitivity and processing capabilities for different features. Thirdly, we design a new module called the spatial grouped attention fusion pyramid module, to enhance the extraction and fusion capabilities of multi-scale features through grouping and fusion spatial attention mechanisms, enabling effective extraction and discrimination of defect multi-layer semantic features. Finally, experiments are conducted on a tread defect dataset with an imbalance ratio of 10:1. Experimental results demonstrated the excellent performance of the proposed model on public datasets. The achieved average mAP@0.5 value jumps from 90.8% to 91.9%. Similarly, the observed average mAP@0.5:0.95 value boosts from 53.9% to 54.9%.

Keywords: Convolutional Neural Network, Defect Detection, Train Wheel Tread, Attention Mechanism, Feature Fusion.

1 introduction

With the advent of the new era of high-speed train movement, safety hazards in train operations have become increasingly prominent. Particularly, wheelsets, the critical supporting and traveling components of trains, often exhibit tread defects and damage due to wheel-rail rolling contact. If not addressed in time, these issues may further deteriorate, pose threat to the safety of trains operation. It has bring mounting difficulties for the conventional manual experience detection method to identify these defects. With the raise of image processing, technologies such as automatic defect recognition for wheelset tread images based on machine vision commence to emerge.

Since the proposal of the VGG [3] network model, deep learning models represented by Convolutional Neural Networks (CNNs) have been available for widely image recognition use due to their powerful feature extraction capabilities. CNNs automatically learn the features of defects from image samples to be identified, and the trained network can recognize and classify these defects. Currently, there are two models based on CNN: YOLO [5] and R-CNN [6]. In the YOLO scenario, the predication of location and category of objects in an image can be densely conducted, meaning that even small objects occupying a small number of pixels can be detected possibly. However, by utilizing a single regression method to directly predict the coordinates of bounding boxes and category labels, it may lead to deduction for the accuracy of bounding box localization, especially for small or overlapping objects, where the precision of localization is more likely to be affected. In contrast, R-CNN generates relatively fewer candidate regions through methods such as selective search, but this may lead to small objects being overlooked or an insufficient number of candidate boxes.

To effectively recognize features of different levels, some feature fusion modules with higher recognition rates for small defects had been proposed. CHEN [8] conduced an efficient feature extraction method called the ELAN module for tire defect feature extraction, which reduced the interference of redundant features while retaining important information, thereby improving the quality of feature representation. SHAO [9] et al. designed a spatial pyramid cross-connection module, where the parallel structure of the pooling part was changed to a serial structure, improving the network's inference speed without compromising accuracy. XU [10] et al., proposed an improved version of the bidirectional feature pyramid network to perform bidirectional feature fusion with HorNet, enhancing the detection performance of small-sized objects and thereby improving the detection accuracy of catenary components.

Additionally, it is difficult for small object detection with few features to extract beneficial semantic feature information during network training. Furthermore, a considerable amount of feature information for small objects is deleted after multiple down-sampling and pooling operations, making it challenging for the model to accurately locate and recognize small objects. Therefore, dozens of scholars dedicated to enhancing the feature information of small objects without increasing the model's complexity. DING [11] et al. used a new plug-and-play cross-fusion (CF) block to simultaneously aggregate features from different stages, optimizing the model's performance. ZHANG [12] et al. conducted an insulator defect feature enhancement module that emphasized target information and reduced the possibility of vanishing gradient problems. ZHAO [13] et al. proposed the C2fSE module, which used an attention mechanism to replace the backbone network's C2f module. The C2fSE module could obtain more image information without increasing the number of parameters and model size. To suppress the interference of invalid information in feature maps, the attention mechanism endowed the model with focusing capability, assigning greater weight to effective feature information and achieving the goal of suppressing unimportant noise interference information. Examples including BiFormer attention [14] and SimAM attention [15], which performed adaptive feature adjustment on the input data by learning data distribution and adaptive parameters, enhancing the model's expression capability and performance.

According to the above work, this paper proposes a train wheelset tread defect recognition model based on attention feature fusion to address the underutilization of shallow features extracted from wheelset tread defects and the lack of defect samples during model training. The main contributions of this paper are as follows:

(1) A novel deformation attention-enhanced bottleneck module is used to enhance the performance of the attention mechanism, especially when dealing with complex scenes and small objects, by introducing a dual-scale context module to capture more global and local information.

(2) A new adaptive spatial and channel enhancement module is proposed, which constructs a multi-level integrated feature representation method by jointly utilizing low-level detail information and high-level semantic information. This method can effectively extract and distinguish the multi-level semantic micro-features of defects.

(3) A pyramid split attention network space grouped attention fusion pyramid module is proposed, which can perform layered extraction of input features at multiple scales, while introducing multiple convolution kernels and grouped convolution structures to adapt to features of different scales.

2 Wheel Tread Defect Identification Model

The structure of the wheelset tread defect identification method is shown in Figure 1, including modules such as tread defect data collection, main feature extraction, and defect identification decision-making.

The defect identification process is as follows: First, the tread defect dataset is preprocessed to achieve conversion between images and markup languages used for storing and transmitting data. Subsequently, the images are sent to the backbone network to extract preliminary tread defect features before outputting two feature maps of different scales, with the feature map sizes down sampled to 2 and 4 times the input image size respectively. Next, the images extracted by the backbone network are imported into the deformable attention-enhanced bottleneck designed after convolution operations to obtain processed features. Then we apply the adaptive spatial and channel enhancement module after convolution operations to obtain processed feature maps. Again, the features are fed into the spatial grouping attention fusion pyramid module in this paper, and the results are introduced to the connection layer to export the target image features. Finally, the features are imported into the identification decision-making layer, utilizing the detection head to generate the result image and the detection results. We also employ the bounding box regression loss function to improve the recognition performance of imbalanced data.

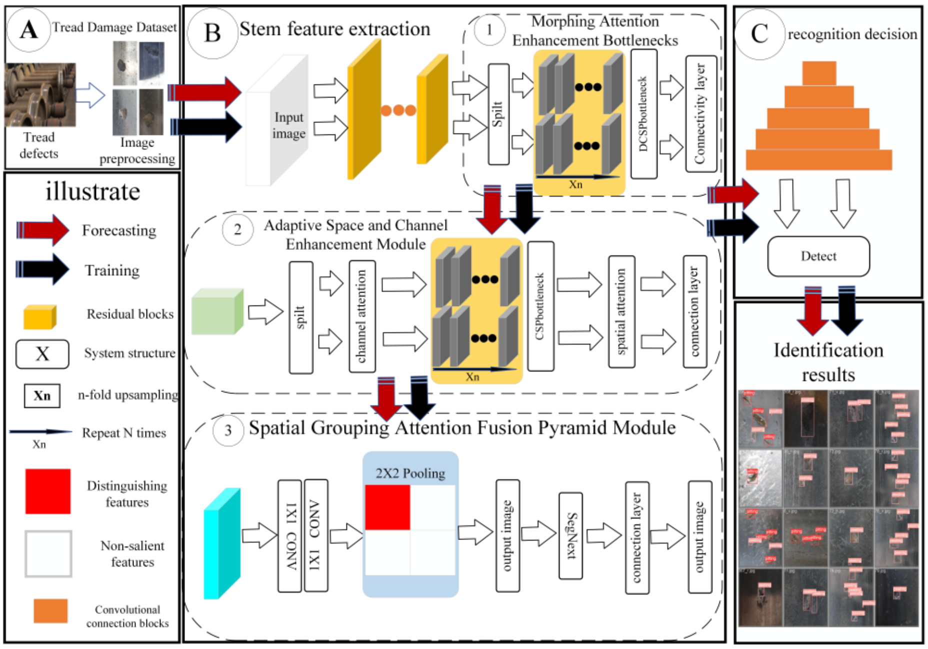


Fig. 1 Method for identifying defects in train wheelset tread

2.1 Deformable Attention-Enhanced Bottleneck Module

First, a convolution operation is applied on the input image. Next, a separation operation divides the image into two parts to facilitate parallel data processing and boost the model's efficiency. The bottleneck dual attention module incorporates a D attention module into the standard bottleneck module, allowing it to automatically identify and prioritize important features during feature extraction. This improves focus on significant features while suppressing irrelevant ones, thereby enhancing the model's performance and generalization capability. The D attention module is particularly effective for identifying tread defects. For further details, refer to Algorithm 1.

Algorithm 1 Bottleneck Convolution Algorithm

Step 1 Bottleneck Dual Attention

The input feature map undergoes a convolution operation, outputting the weighted feature map, and the convolution encoding process is shown in Figure 2.

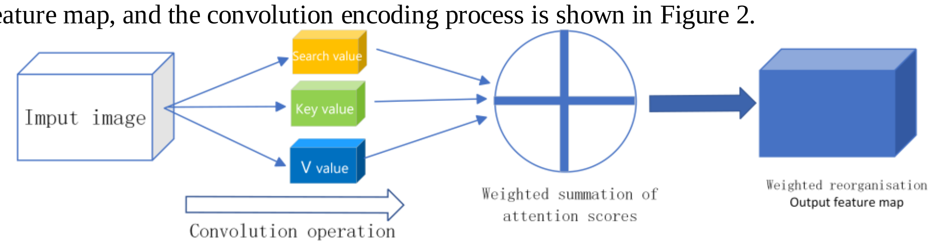


Fig. 2. Convolutional coding procedure

(1) Use convolution operations to calculate the offset at each position, so as to distort the position in the attention mechanism. The offset can be obtained from the query tensor *q* through convolution operations:

$$offset = conv\_offset(q) \tag{1}$$

where $offset$ is the offset tensor, and *q* is the query tensor.

(2). Use convolutional layers to project the input feature map to get query *q*, key *k*, and value *v*:

$$q = proj\_q(x), k = proj\_k(x), v = proj\_v(x) \tag{2}$$

where *q* is the query tensor, *k* is the key tensor, and *v* is the value tensor.

(3) Calculate relative position encoding, apply softmax normalization to get attention weights, and use these attention weights to perform weighted summation on the values to derive the output tensor:

$$attn = soft\max(q \bullet k^T) \tag{3}$$

$$y = proj\_drop(proj\_out(attn \bullet v)) \tag{4}$$

where *attn* is the attention weight tensor, and *y* is the output tensor.

Step 2 The introduction of position encoding enhances the model's ability to understand the spatial relationships between features. By encoding the position information of features, the module can manage the local and global relationships between features scientifically, improving the model's performance and robustness.

Step 3 Connect the feature blocks output by D attention to obtain the final output:

$$X_c = Concat(X_G, X_L) \tag{5}$$

where X_c is the attention weight tensor, X_G and X_L are the output tensors.

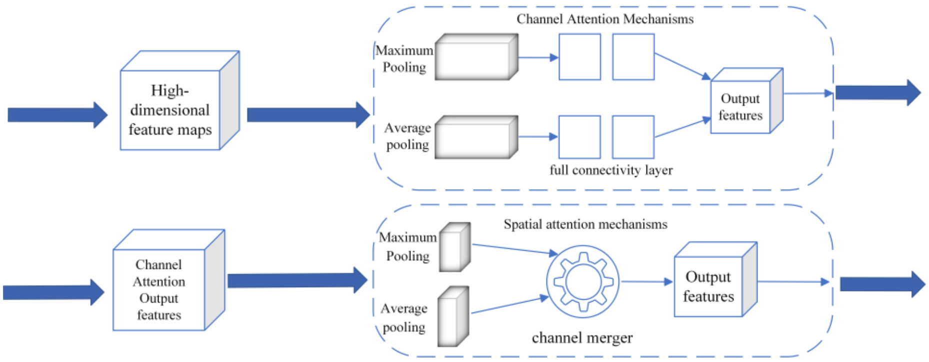


Fig. 3. Convolutional coding procedure

2.2 Adaptive Spatial and Channel Enhancement Module

Detecting small object poses captivating challenge in computer vision domains. Traditional object detection methods often struggle with issues such as inaccurate object localization, missed detections, and false detections when dealing with small objects. It is primarily due to weak feature information of small objects, which is difficult to accurately capture and represent.

We can integrate channel attention and spatial attention to improve the accuracy and robustness of object detection. Therefore, considering the detailed local information contained in low-level features and the rich global semantic information in high-level features, a new module called the Small Object Detection Block (SODB) is designed to fully integrate spatial and global information between different levels of features, as illustrated in Figure 5.

Firstly, a channel attention mechanism is employed, which weights the channel features through adaptive average pooling and convolution operations to enhance the response of important features. Subsequently, a spatial attention mechanism is applied, capturing the spatial relationships between different features by computing the mean and maximum values along the channel dimension to further optimize the feature representation.

The specific steps are as follows: Initially, perform convolutional separation on the image, followed by using the channel attention mechanism to learn the importance of each channel, whose weight will be adjusted in the feature map. This allows the network to better comprehend the importance of specific features for the task, improving the network's expressiveness and performance. Next, apply the spatial attention mechanism to adjust the weight of each pixel in the feature map, weighting according to the importance of each spatial location.

By adjusting the weights of each channel and spatial location in the feature map, the network enhances its ability to represent small target features, facilitating the capture of subtle details. The processed feature map is then fed into the bottleneck module for multiple training iterations. Finally, the outputs from the channels are concatenated to obtain the final result.

This paper introduces the improved SODB module as the Adaptive Spatial and Channel Enhancement Module after removing more redundant and noisy features. Compared to the original module, the Adaptive Spatial and Channel Enhancement Module enhances the feature representation of important positions through weighting, making the model more effective in utilizing spatial information and improving the model's effectiveness in detecting small targets.

The traditional structure and the Adaptive Spatial and Channel Enhancement Module proposed in this paper are shown in Figures 4 and 5.

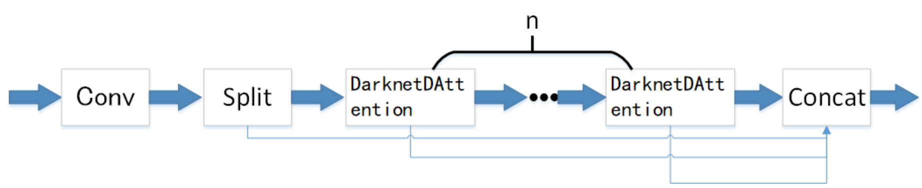


Fig. 4. Traditional module

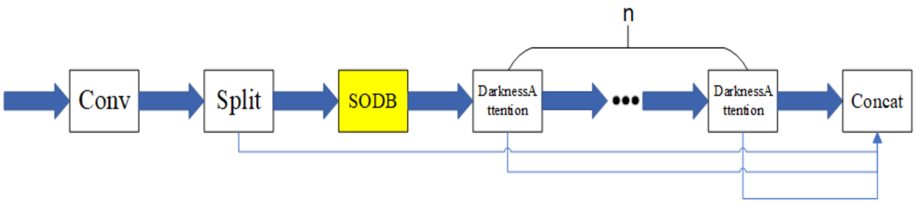


Fig. 5. Adaptive spatial and channel enhancement module

2.3 Spatial Grouping Attention Fusion Pyramid Module

In the context of wheelset treads, it is difficult to obtain sufficient samples, and there are many irrelevant stains, rust, and other redundant features in the process of tread image recognition. These irrelevant features propagate with the learning and training of the model, and their weights in the feature maps increase as the network layers deepen, thereby negatively impacting the model. To reduce the influence of useless and redundant features on small target defects, this paper proposes the Spatial Grouping Attention Fusion Pyramid Module (SGAFPM), as shown in Figure 6, to effectively eliminate background interference, filter important features, and refine semantic information.

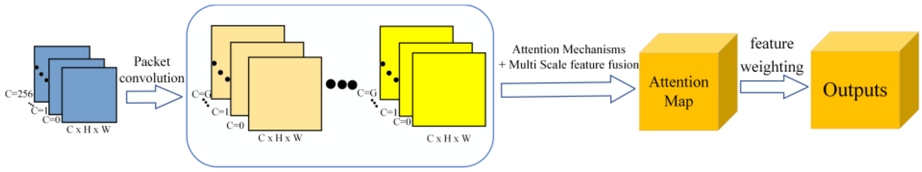


Fig. 6. Spatial Grouping Attention Fusion Pyramid Modul

The objective of the SGAFPM module is to introduce an attention mechanism to enhance the performance of semantic segmentation, while using feature fusion convolutional layers to extract features at different scales to improve the accuracy and robustness of semantic segmentation. Concurrently, the designed SEG module emphasizes the weighted fusion of the feature maps through the spatial attention mechanism to enhance the model's attention to features at different spatial positions, thereby improving the performance of the semantic segmentation model.

The SEG module comprises two key steps:

Step 1: Feature Grouping and Independent Convolutions

(1) Feature grouping and independent convolution operations. In this step, the input features are divided into multiple groups, where each group are independently convolved. Thus, the output features of all groups can be merged. We can express the feature as the following formula:

$$O_i = Concat(Conv(I_i)),\ for(i = 1, 2, ..., N) \tag{6}$$

where O_i is the output features of group , I_i represents the input features of group i, 3 expresses the number of groups the input features are divided into, Conv is the convolution operation, and concat represents the concatenation operation, which concatenates multiple feature maps along the channel dimension.

(2) Convolution operations in the horizontal and vertical directions to obtain feature information at different scales. Specifically, the grouped features are convolved in two directions, and then the results can be merged. The mathematical expression for this step is:

$$H = Conv\_Horizontal(M) \tag{7}$$

$$V = Conv\_vertical(M) \tag{8}$$

$$F = Concat(H, V) \tag{9}$$

where H is designated as the feature map in the horizontal direction, V represents the feature map in the vertical direction, and F is the merging of the horizontal and vertical feature maps.

Step 2: Attention Weighting

A simple attention mechanism is adopted to learn the importance of different regions in the feature map and weight the features according to their importance. The attention mechanism includes the following two steps:

(1) First, the attention map A is obtained through the grouped convolution operation. Then, the attention map is convolved in the horizontal and vertical directions, the results are also added to the original attention map to obtain the fused attention map FA, as described by the following formula:

$$F_A = A + Conv\_Fusion(A) \tag{10}$$

Where In the formula, A represents the attention map, and F_A represents the fused attention map.

$$F_F = M \odot F_A \tag{11}$$

Where In the formula, M is the original feature map, $\odot$ is the element-wise multiplication operation used for weighting the feature map, and F_F denotes the weighted feature map. Through this series of steps, the SEG module achieves the grouping, convolution, attention weighting, and fusion operations of features, thereby enhancing the model's ability to perceive features at different spatial locations, aiding the model in focusing more effectively on areas containing small targets and thus improving the performance of the semantic segmentation model.

3 Experiment results and analysis

This section first introduces the experimental dataset settings, parameters, and evaluation metrics. Then, the experimental results are presented and analyzed. The experiment comprises four parts: model training process, backbone feature extraction network comparison, attention fusion feature comparison, and model comparison evaluation.

3.1 Experimental Description

3.1.1 Experimental Dataset

A total of 232 defect samples are collected using the data acquisition system. Due to the difficulty in collecting real train wheelset data, the number of real defect samples is limited. To mitigate the risk of overfitting in the convolutional neural network model, this study applied data augmentation techniques to the defect samples, including horizontal mirroring, vertical flipping, and 180-degree rotation. Consequently, a total of 927 train wheel tread defect images are generated. To evaluate the performance of the model, the size of the images in the dataset is uniformly adjusted to 600×800, and the training and test set are divided in a 7:3 ratio. The augmented defect images are shown in Figure 2, with delamination and scratches corresponding to 262 and 562 images respectively, as illustrated in Figure 8. The final size of the training set is 371 images, and the size of the test set is 160 images.

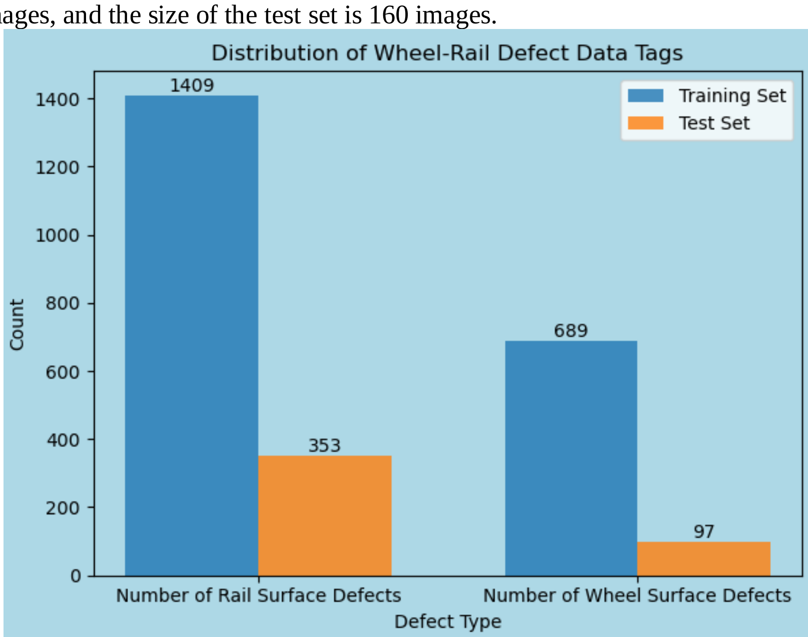


Fig. 7. Data label distribution diagram

3.1.2 Experimental Parameters

The network experiments are conducted on a Windows 10 operating terminal system, using Python 3.8 language, Intel(R) Xeon(R) Platinum 8269CY CPU @2.50GHz 2.49GHZ (2 processors), with NVIDIA Tesla T4 graphics configuration, trained and predicted under the Pytorch framework environment, and accelerated using CUDA 11.1. The optimizer used is Auto, the batch size is set to 16, the learning rate is initialized to 0.01, the number of iterations is 600, and an early stopping training strategy is adopted, stopping the training when the model shows overfitting leading to worse performance on the test set. Additionally, the network's Dropout rate is 0.25 to avoid feature overfitting, and batch normalization is used during network training to normalize the mean and variance of each layer's output, preventing gradient explosion, gradient vanishing, and network degradation. The training parameters in the network are shown in Table 1 below.

Table 1 Training related parameters

Parameters	Parameters Values
Image Size	600x800
epochs	600
batch size	16
Initial learning rate lr	0.01
Workers	128

3.1.3 Evaluation Metrics

This study aims to identify the status of collected wheelset tread defect images. The experiment primarily employs common evaluation metrics for assessment, specifically using Accuracy, Precision, Recall, and F1-score. The calculation formulas are provided below.

$$\begin{cases} Accuracy = \dfrac{TP+TN}{TP+TN+FP+TN} \times 100\% \\ Precision = \dfrac{TP}{TP+FP} \times 100\% \\ Recall = \dfrac{TP}{TP+FN} \times 100\% \\ F1-score = \dfrac{2 \times Precision \times Recall}{Precision + Recall} \times 100\% \end{cases} \tag{12}$$

Among them, TP (True Positive) and TN (True Negative) indicate that the predicted label is the same as the actual label; FP (False Positive) and FN (False Negative) indicate that the predicted label is different from the actual label.

3.2 Backbone Feature Extraction Network Comparison Experiment

When dealing with complex interference and class imbalance in tread defects, the importance of neural networks is conveyed in extracting visual features. To effectively describe the differences between different defect categories, a good feature extractor is needed, and the backbone network plays a key role in this process. Therefore, this paper introduces an improved neural network and compares it with the backbone feature extraction networks of YOLOV3[16] and YOLOV5[17] to evaluate the performance of the algorithms. The experimental results are presented in Table 2.

Table 2　Impact of different backbone networks on the dataset

Main Network	Map50/%	Map50-95/%	Model Size/MB
YOLOV3[16]	87.2	49.1	4.72
YOLOV5[17]	89.9	50.8	3.67
Method in this article	**91.9**	**54.9**	**6.39**

On the basis of results of Table 2, a comparison of the three backbone feature extraction networks shows that YOLOv3 is limited in feature extraction capability compared to other networks. The mainstream YOLOv5 uses CSP connections to make more full use of the information of input features, reducing information loss, which helps the network learn richer and more complex feature representations. It effectively addresses network degradation issues, significantly improving accuracy compared to YOLOv3. The model proposed in this paper improves Map50 and Map50-95 by 2.0% and 4.1% compared to YOLOv5, respectively due to the method has better feature extraction capabilities. Moreover, the introduction of a grouped attention fusion strategy in the residual block assigns more weight to important defect target areas and adaptively adjusts their feature weights, thereby effectively enhancing recognition performance.

3.3 Attention Fusion Feature Comparison Experiments

To demonstrate the effectiveness of the attention feature fusion strategy proposed in this paper, discussion experiments are conducted on the model. The contrast models in the experiments are: YOLOv8 without any fusion method, as a blank control network 1; YOLOv8 combined with a novel deformable attention-enhanced bottleneck module for feature extraction, as network 2; YOLOv8 combined with a spatial grouped attention fusion pyramid for feature extraction, as network 3; and YOLOv8 combined with an adaptive spatial and channel enhancement module, deformable attention-enhanced bottleneck, and spatial grouped attention fusion pyramid for feature extraction, as network 4. This is to evaluate the effectiveness of the fusion strategies in the proposed model. The experimental results are shown in Table 3.

Table 3　The effect of different fusion methods on algorithm performance

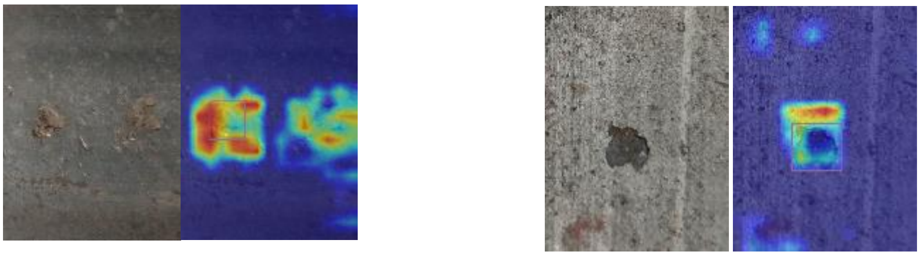
Fusion Method	Map50/%	Map50-95/%	Model Size/MB
Network 1	90.6	53.9	5.96
Network 2	91.2	54.9	6.11
Network 3	91.9	54.6	6.11
Network 4	91.4	54.9	6.19
Methods in this article	**91.9**	**54.9**	**6.39**

By comparing the results in Table 4, it can be seen that the recall rate of Network 1 without feature fusion is relatively the lowest, indicating that the introduction of positional encoding can improve the recognition ability of the algorithm. The accuracy of Network 3, which uses element fusion spatial grouping attention fusion pyramid, is higher than that of Network 1 without fusion, demonstrating that enhancing the performance of semantic segmentation while using feature fusion convolution layers to extract features at different scales is effective in improving the accuracy and robustness of semantic segmentation. By comparing Network 4 and Network 1, we find that, the use of the adaptive spatial and channel enhancement module to replace the original C2f module through the introduction of attention mechanisms and more complex feature fusion strategies, as well as more diverse module designs, can more effectively extract image features and achieve better performance in image processing tasks. The method proposed in this paper achieves the best results in all indicators, indicating that the network structure design based on the attention feature fusion strategy is reasonable and effective. At the same time, the design of adaptive spatial, channel enhancement modules, and spatial grouping attention fusion pyramid modules have all improved the accuracy of the model. The attention feature fusion strategy can specifically utilize detailed information in low-level feature maps and global information in high-level features while reducing noise interference in low-level feature maps.

2.4 Grad-CAM Heatmap

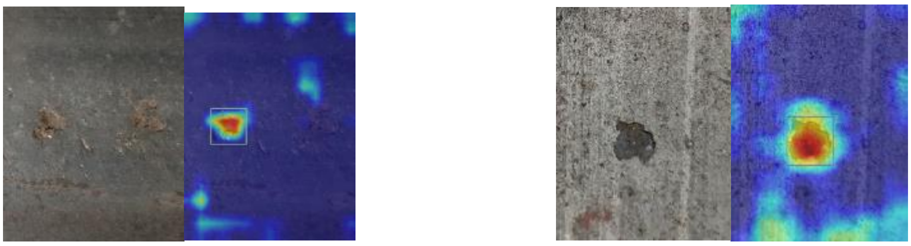
To further improve the interpretability of convolutional neural networks and verify the effectiveness of the residual attention module in mitigating noise interference, Grad-CAM (Gradient-weighted Class Activation Mapping) class activation visualization technology is used to visualize and compare the two defect types of tread scuffing and tread peeling. Based on the heatmap, the depth of its color reflects the model's attention to local areas of the image, with red areas being the regions of highest attention. The color distribution of the heatmap intuitively shows which areas of the image contribute significantly to category classification.

The weight files of the trained model in this paper are used to output the heatmap of the last convolutional layer in the backbone network using Grad-CAM technology. From the heatmap, as shown in Figure 9, it can be seen that the original unmodified model has a high degree of attention to non-defective regions with significant noise interference, shown in red, indicating that the model is severely affected by noise interference, which will affect the detection results. The method in this paper, as shown in Figure 10, greatly reduces the degree of attention to non-defective regions with noise interference and does not reduce the attention to real defect areas, which will help improve defect detection results, indicating that the three improvement modules have some effect in weakening noise interference.



(a)Depression Defect (b) Peeling Defect

Fig. 8. Heatmap before improvement



(a)Depression Defect (b) Peeling Defect

Fig. 9. Heatmap based on methods of this article

After enhancing the module of the model, we significantly reduce the impact of noise factors such as lighting and stains on detection performance. This improvement highlights the target to be detected within complex backgrounds, providing a solid foundation for the subsequent accurate detection of wheel defects.

2.5 Analysis of Model Generalization Ability

A To verify the proposed model's ability to accurately identify different types of defects, we conducted an additional experiment using a dataset from the Northeastern University surface defect database (NEU), created by Song et al. This dataset contains six typical surface defects of hot-rolled strip steel: inclusions (In), patches (Pa), cracks (Cr), pits (Ps), roll marks (RS), and scratches (Sc). Figure 10 illustrates sample images of these defects. The database consists of 1800 images, with 300 samples for each defect type, and each image has a resolution of 64×64 pixels. The dataset is well-balanced and offers a sufficient number of samples for training and testing.

In this experiment, we split the NEU dataset into a training set and a test set using a 7:3 ratio. This resulted in 1260 images in the training set and 540 images in the test set, providing a robust evaluation of the model’s generalization performance. The model structure was kept unchanged, with the only adjustment being the modification of the fully connected layer to accommodate six output neurons corresponding to the six defect classes in the NEU dataset.

Training and Experimental Setup

For training, we set the number of epochs to 600 to ensure thorough feature learning and parameter adjustment. To evaluate the effectiveness of our model, we conducted a comparative analysis with YOLOv3 and YOLOv5 under the same experimental conditions. The parameters for all models were kept consistent, and the same 7:3 dataset split was used across all experiments to ensure a fair comparison.

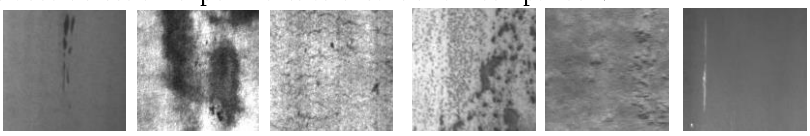


Fig. 10. NEU Data set sample image

Table 3 presents a comparison of our model's performance with YOLOv3 and YOLOv5 on the NEU dataset. The results indicate that our model demonstrates superior generalization ability in identifying various defect types, especially in scenarios involving complex backgrounds and low-contrast defects. The integration of the attention feature fusion module enables the model to effectively leverage shallow features, improving both detection accuracy and robustness.

Table 4 Model generalization performance experiment

Model	Map50/%	Map50-95/%	Model Size/MB
YOLOV3	61.5	27.6	4.78
YOLOV5	70.7	35.2	3.73
Methods in this article	76.0	42.1	6.34

It shows that compared with other algorithms, the performance of Map50 and Map50-95 have been significantly improved.

4 Conclusion

Aiming at the problem of sample category imbalance in rail surface defects, this paper proposes a method for dealing with the imbalance rail surface defect recognition problem based on an attention feature fusion network and verifies the effectiveness of the method on the imbalance defect dataset. The specific research content is summarized as follows:

(1) To address the challenge of detecting small defects, we propose a novel deformation attention-enhanced bottleneck module to enhance detection performance. This module incorporates a dual-scale context mechanism to capture both global and local information, specifically improving the detection of small targets.

(2) To mitigate the issue of underutilizing shallow semantic features of defects, we have refined the attention mechanism. We introduce an adaptive spatial and channel enhancement module that constructs a multi-level comprehensive representation by leveraging both spatial and channel attention. This approach excels in small target detection tasks by effectively extracting and distinguishing multi-level semantic features of defects.

(3) To address the problem of redundant features in the defect background, the network structure is improved. A spatial grouping attention fusion pyramid is proposed,

which can hierarchically extract input features at multiple scales, and multiple convolution kernels and grouped convolution structures are introduced to adapt to features of different scales.

Future research will focus on diversifying YOLO model types by drawing inspiration from the information processing mechanisms of the cerebral cortex and human visual cortex. By exploring different supervised learning algorithms and designing biologically plausible YOLO network models, we aim to achieve superior recognition results.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (52172403, 62173137).

References

1. Haonan Yang , Jing He , Zhigang Liu,Changfan Zhang, "LLD-MFCOS: A Multiscale Anchor-Free Detector Based on Label Localization Distillation for Wheelset Tread Defect Detection," in IEEE Transactions on Instrumentation and
2. Measurement,2024,73(5003815): 1-15.
3. Z. Huang, C. Zhang, L. Ge, Z. Chen, K. Lu and C. Wu, "Joining Spatial Deformable Convolution and a Dense Feature Pyramid for Surface Defect Detection," in IEEE Transactions on Instrumentation and Measurement, 2024,73(5012614): 1-14.
4. Z. Li, B. Li, S. G. Jahng and C. Jung, "Improved VGG Algorithm for Visual Prosthesis Image Recognition," in IEEE Access, 2024,12: 45727-45739.
5. Changfan Zhang， Yifu Xu， Zhenwen Sheng, et al. Deformable residual attention network for defect detection of train wheelset tread. The Visual Computert 2024,40(3), 1775–1785.
6. X. Qian, X. Wang, S. Yang and J. Lei, "LFF-YOLO: A YOLO Algorithm With Lightweight Feature Fusion Network for Multi-Scale Defect Detection," in IEEE
7. Access, 2022,10: 130339-130349
8. H. Dong, Y. Li and R. Liu, "A Detection Algorithm Based on Improved Cascade RCNN for UAV Aerial Images," 2023 IEEE 3rd International Conference on Electronic Technology, Communication and Information (ICETCI), Changchun, China, 2023, pp. 700-704, doi: 10.1109/ICETCI57876.2023.10176806.2023,
9. L. Yang et al., "An Improving Faster-RCNN With Multi-Attention ResNet for Small Target Detection in Intelligent Autonomous Transport With 6G," in IEEE Transactions on Intelligent Transportation Systems, vol. 24, no. 7, pp. 7717-7725, July 2023, doi: 10.1109/TITS.2022.3193909.
10. C. Peng, X. Li and Y. Wang, "TD-YOLOA: An Efficient YOLO Network With Attention Mechanism for Tire Defect Detection," in IEEE Transactions on Instrumentation and Measurement, vol. 72, pp. 1-11, 2023, Art no. 3529111, doi: 10.1109/TIM.2023.3312753.
11. S. J. Li, Y. X. Liu, M. Li and L. Ding, "DF-YOLO: Highly Accurate Transmission Line Foreign Object Detection Algorithm," in IEEE Access, vol. 11, pp. 108398-108406, 2023, doi: 10.1109/ACCESS.2023.3321385.

12. S. Xu et al., "A Locating Approach for Small-Sized Components of Railway Catenary Based on Improved YOLO With Asymmetrically Effective Decoupled Head," in IEEE Access, vol. 11, pp. 34870-34879, 2023, doi: 10.1109/ACCESS.2023.3264441.
13. Q. Ding et al., "CF-YOLO: Cross Fusion YOLO for Object Detection in Adverse Weather With a High-Quality Real Snow Dataset," in IEEE Transactions on Intelligent Transportation Systems, vol. 24, no. 10, pp. 10749-10759, Oct. 2023, doi: 10.1109/TITS.2023.3285035.
14. Q. Zhang, J. Zhang, Y. Li, C. Zhu and G. Wang, "IL-YOLO: An Efficient Detection Algorithm for Insulator Defects in Complex Backgrounds of Transmission Lines," in IEEE Access, vol. 12, pp. 14532-14546, 2024, doi: 10.1109/ACCESS.2024.3358205.
15. Z. Liangjun, N. Feng, X. Yubin, L. Gang, H. Zhongliang and Z. Yuanyang, "MSFAYOLO: A Multi-Scale SAR Ship Detection Algorithm Based on Fused Attention," in IEEE Access, vol. 12, pp. 24554-24568, 2024, doi: 10.1109/ACCESS.2024.3365777.
16. Y. Zhang, Z. Wu, X. Wang, W. Fu, J. Ma and G. Wang, "Improved YOLOv8 Insulator Fault Detection Algorithm Based on BiFormer," 2023 IEEE 5th International Conference on Power, Intelligent Computing and Systems (ICPICS), Shenyang, China, 2023, pp. 962-965, doi: 10.1109/ICPICS58376.2023.10235397.
17. L. -Q. Zhang, Z. -T. Liu and C. -S. Jiang, "An Improved SimAM Based CNN for Facial Expression Recognition," 2022 41st Chinese Control Conference (CCC), Hefei, China, 2022, pp. 6510-6514, doi: 10.23919/CCC55666.2022.9902045.
18. K. Kim, J. Kim, H. -G. Lee, J. Choi, J. Fan and J. Joung, "UAV Chasing Based on YOLOv3 and Object Tracker for Counter UAV Systems," in IEEE Access, vol. 11, pp. 34659-34673, 2023, doi: 10.1109/ACCESS.2023.3264603.
19. Z. Lu, L. Ding, Z. Wang, L. Dong and Z. Guo, "Road Condition Detection Based on Deep Learning YOLOv5 Network," 2023 IEEE 3rd International Conference on Electronic Technology, Communication and Information (ICETCI), Changchun, China, 2023, pp. 497-501, doi: 10.1109/ICETCI57876.2023.10176545.
20. Y. Liu, H. Yang and C. Wu, "Unveiling Patterns: A Study on Semi-Supervised Classification of Strip Surface Defects," in IEEE Access, vol. 11, pp. 119933-119946, 2023, doi: 10.1109/ACCESS.2023.3326843.