electronics

Article

Research on Rail Surface Defect Detection Based on Improved CenterNet

Yizhou Mao [1], Shubin Zheng [2,*], Liming Li [1,*], Renjie Shi [1] and Xiaoxue An [3]

[1] School of Urban Railway Transportation, Shanghai University of Engineering Science, Shanghai 201620, China; m405122210@sues.edu.cn (Y.M.); m405122207@sues.edu.cn (R.S.)
[2] Higher Vocational and Technical College, Shanghai University of Engineering Science, Shanghai 200437, China
[3] Engineering Training Center, Shanghai University of Engineering Science, Shanghai 201620, China; anxiaoxue@sues.edu.cn
* Correspondence: shubin.zheng@sues.edu.cn (S.Z.); liliming@sues.edu.cn (L.L.)

Abstract: Rail surface defect detection is vital for railway safety. Traditional methods falter with varying defect sizes and complex backgrounds, while two-stage deep learning models, though accurate, lack real-time capabilities. To overcome these challenges, we propose an enhanced one-stage detection model based on CenterNet. We replace ResNet with ResNeXt and implement a multi-branch structure for better low-level feature extraction. Additionally, we integrate SKNet attention mechanism with the C2f structure from YOLOv8, improving the model's focus on critical image regions and enhancing the detection of minor defects. We also introduce an elliptical Gaussian kernel for size regression loss, better representing the aspect ratio of rail defects. This approach enhances detection accuracy and speeds up training. Our model achieves a mean accuracy (*mAP*) of 0.952 on the rail defects dataset, outperforming other models with a 6.6% improvement over the original and a 35.5% increase in training speed. These results demonstrate the efficiency and reliability of our method for rail defect detection.

Keywords: rail defect detection; CenterNet; ResNeXt; SKNet attention mechanism; elliptical Gaussian kernel

check for updates

Citation: Mao, Y.; Zheng, S.; Li, L.; Shi, R.; An, X. Research on Rail Surface Defect Detection Based on Improved CenterNet. *Electronics* **2024**, *13*, 3580. https://doi.org/10.3390/electronics13173580

Received: 5 August 2024
Revised: 3 September 2024
Accepted: 7 September 2024
Published: 9 September 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

As an important part of the global public transportation infrastructure, the railroad network is essential in driving economic progress, enhancing regional interconnectivity, and supporting the movement of goods and people. When railroad operation time increases, irreversible damage is caused to the rail surface, which is inevitably affected by factors such as manufacturing process, wheel-rail contact stresses, and natural weathering, leading to the development of surface defects including seams, scars, abrasions, and spalling as shown in Figure 1 [1]. Wear in this picture above is caused by the gradual wearing away of metallic material from the surface of the rail as a result of prolonged contact with the wheels, and may cause train bumps. Depression is a depressed area caused by localized loss of material or indentations on the rail surface. Common defects in joint include looseness, misalignment, and poor joint can cause shock and noise when the train is moving, affecting passenger comfort and exacerbating track damage. Corrugation and abrasion may cause the rail surface to become rough, increasing friction with the wheels, which accelerates the wear process. Consequently, to uphold the smooth and efficient operation of the rail services, timely and accurate detection of defects on the surface of the railroad track has become an urgent problem in the railroad industry, which not only has significant practical application value, but also has far-reaching research significance.

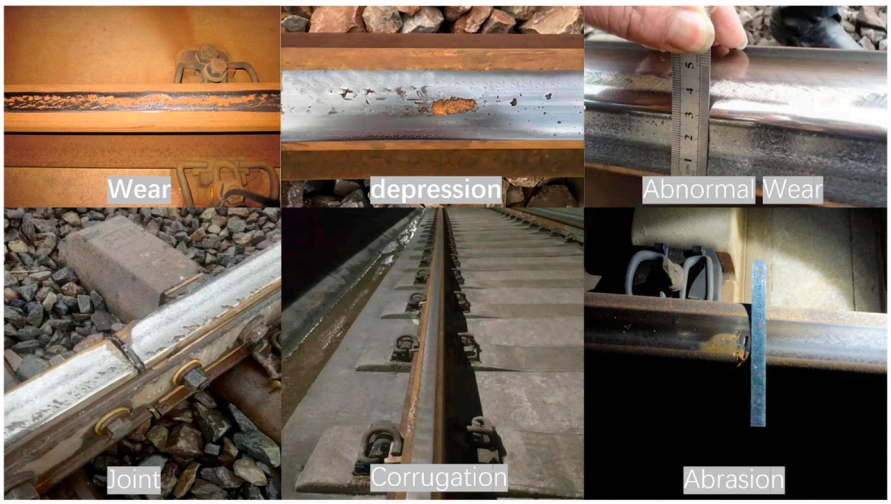


Figure 1. Examples of rail defects.

The early detection of rail surface defects primarily depends on manual inspection. Although this approach is cost-effective, it is often slow and inefficient, making it challenging to detect defects in their initial stages, potentially leading to more severe issues. With the advancement of technologies such as ultrasonic testing, magnetic particle inspection, and eddy current testing, detection capabilities have greatly improved, gradually replacing traditional manual inspection methods. M. Sun et al. [2] introduced a detection technique utilizing photoacoustic signals, which utilizes the photoacoustic effect for the detection of rail surface defects in response to the inadequacy of ultrasonic signals in detecting surface microcracks. Xiong, L et al. [3] studied and explored the use of ultrasonic guided wave technology for non-destructive testing of rail bottoms, which is a physical testing method that specializes in areas that are difficult to detect with conventional techniques. Han S-W et al. [4] proposed electromagnetic ultrasonic testing (UT) using Electromagnetic Acoustic Transducers (EMATs) is a non-contact ultrasonic inspection technology, which excites and detects ultrasonic waves on the metal surface through electromagnetic effect, and is particularly suitable for rail testing in high temperature environments. The above methods, although effective, still face limitations in terms of accessibility, real-time monitoring and effective coverage of long distances of railroad track. Rough or irregularly shaped rail surfaces may affect the coupling and detection results of ultrasonic inspection and ultrasonic inspection may not be sufficiently sensitive to defects located in the deeper layers of the material [5]; Magnetic Particle Inspection (MPI) is usually only able to qualitatively show the presence of defects, but it is difficult to determine the exact dimensions and depth of the defects [6]; Eddy current inspection is also difficult to achieve uniformity for complex or irregularly shaped workpieces [7].

With the increasing requirements for safety and reliability of railroad systems, traditional manual inspection methods can no longer meet the demand for efficient and accurate inspection. The existing body of work in vision-based systems for spotting rail imperfections exhibits significant strides and potential for utilization. Machine vision systems utilize cutting-edge image processing techniques for the automated recognition and pinpointing of defects on rail tracks [8].

Despite the significant progress of machine vision in rail defect detection, there are still some challenges. For example, in practical applications, complex ambient light changes, dirt and occlusions on the rail surface can affect the detection results. Moreover, conventional methods usually rely on manually set parameters such as thresholds, filter sizes, shape of structural elements, etc. These parameters need to be tuned for specific sizes of defects. These parameters need to be tuned for a specific size of defect, so when the defect size varies, the originally set parameters may no longer be applicable, leading to performance degradation. In addition, the operation of high-speed trains requires the system to have

extremely high real-time and stability, which puts higher requirements on the efficiency of the algorithms and hardware performance. To solve these problems, future research directions include developing more robust image processing algorithms, optimizing the computational efficiency of deep learning models, and designing smarter inspection system architectures. Overall, the research of machine vision technology in the detection of rail surface defects continues to deepen, providing strong technical support for the safety and security of railroad transportation.

Based on the above, the technical approaches for rail surface defect detection have all obtained good results, but at the same time there are problems and challenges in certain aspects. CenterNet [9] is a deep learning based target detection algorithm, which achieves target detection through a Fully Convolutional Network (FCN) [10], which makes it highly computationally efficient while maintaining a high level of accuracy. Key point estimation is applied to locate the center point of the target and regress to other attributes, which makes it flexible in dealing with defects of different sizes and shapes, and has a greater advantage for surface defects on rails with different sizes and types. The main contributions of this paper based on this paper are:

1. A novel fusion feature extraction module based on CenterNet is proposed for rail surface defect detection network.
2. Replacing the original backbone ResNet with an efficient feature extraction backbone ResNeXt, and changing the low-level features layer into a multi-branching layer increases the model's ability to capture low-level features, providing a richer feature representation for the accurate identification of rail surface defects.
3. The SKNet attention mechanism, which incorporates the C2f structure of the YOLOv8 network, is introduced so that the model can focus more on the key areas in the image, improving the detection accuracy of small defects and weak features.
4. Given the specific shape characteristics of rail defects, we change the traditional circular Gaussian kernel into an elliptical Gaussian kernel that further takes into account the aspect ratio of the GTbbox, which strengthens the dimensional regression loss of the network's loss values and improves the model training speed of the network.

2. Related Works

The application of machine vision technology in the detection of rail surface defects is mainly reflected in the traditional machine learning detection methods and deep learning detection methods.

2.1. Machine Learning Methods

Traditional machine learning inspection methods usually rely on manual feature extraction and definition and apply classical machine learning algorithms and subjectively classify various types of features into specific categories to identify and classify defects on rail surfaces. For example, support vector machines or decision trees [11] can be chosen for classification problems, and local binary [12] or histogram gradient can be chosen for extracting texture features of images. Zhang, C. et al. [13] introduced a detection approach for identifying rail surface flaws combining XGBoost and multi-scale feature extraction. The study captures more comprehensive defect information by performing multi-scale feature extraction on the original image and mapping the features to different resolution levels. Zhu W F et al. [14] proposed an automated rail surface defect detection method based on sparse representation. The method utilizes the theory of sparse representation to extract effective sparse features by representing the defect image as a linear combination of dictionaries. Deng, F et al. [15] explored the application of multiple kernel learning and image fusion techniques in the detection of rail surface defects. By combining multiple kernel functions, the multi-kernel learning technique can effectively capture the different characteristics of the data and achieve more accurate classification. Li, Q et al. [16] constructed a real-time visual inspection solution specifically designed to overcome the challenges of uneven light distribution and inconsistent reflections on the rail surface, and thus effectively identify

localized surface defects on the rail. Tian, S et al. [17] adapted and improved the Sobel algorithm to address its lack of sensitivity in recognizing defects in the X and Y directions, and the improved algorithm succeeded in enhancing the accuracy of the detection by achieving a 10% increase in efficiency.

2.2. Deep Learning Methods

As deep learning advances swiftly, the neural networks that employ convolutional structures excel in feature extraction and data handling, making them a superior choice for identifying defects on railway surfaces, surpassing traditional machine learning methods. Defect detection includes one-phase detection and two-phase detection.

The two-stage detection method has a large advantage in detection accuracy, but the inference speed is relatively slow. Among them, Choi, J.-Y et al. [18] proposed a rail surface defect detection method based on Faster R-CNN with integrated attention mechanism. The study improved the sensitivity and detection accuracy of the model to the fine defect region by adding the attention module to the Faster-RCNN model. Aydın I et al. [19] proposed a two-stage deep learning network for the classification and detection of rail surface defects. The first phase of the detection process identifies suspect defect areas via a Region Proposal Network (RPN), and the subsequent phase employs a convolutional neural network (CNN)-assisted classifier for the precise classification and exact localization of these areas.

One-stage detection accomplishes both candidate region (object proposals) generation and target classification and localization in a single forward propagation of the network, which is suitable for scenarios that require real-time processing, but at the expense of some accuracy in exchange for speed. The one-phase detection approach produces region candidates and simultaneously achieves the classification and exact positioning of targets in one network iteration, relying on the sacrifice of some accuracy in exchange for inference speed, suitable for scenarios requiring real-time processing, the most classic such as SSD [20]. Zheng Z et al. [21] Introduced a high-performance rail defect detection approach leveraging the YOLOv3 framework, further enhanced by transfer learning to boost detection capabilities. YOLOv3 optimizes the feature pyramid network (FPN) [22] to substantially decrease computational demands while sustaining top-notch detection precision. Zhang C et al. [23] introduces an enhanced YOLOX model coupled with an image enhancement technique for the detection of rail surface defects, addressing the limitations of existing detection models and overcoming challenges posed by poor illumination and sparse defect data in collected images. Zhu X et al. [24] constructed the TPH-YOLOv5 detection model, which utilizes the CBAM module [25] for the detection of tiny targets and employs a transformer predictor head instead of the traditional predictor head to optimize the detection of dense targets in complex environments.

3. Results

The core idea of CenterNet is to consider a target as its center point and to determine the target's boundaries by predicting the location, size and offset of the target's center point through a single-stage network. The entire process does not require complex candidate frame generation or post-processing steps, which enables CenterNet to achieve fast, concise, and accurate real-time target detection. However, the original CenterNet is still insufficient in dealing with complex and subtle defects on the rail surface. To this end, we have made several improvements to CenterNet with the aim of enhancing its performance in the detection of rail surface defects. First, we modified the backbone network part, and we proposed an improved multi-branch ResNeXt network structure to replace the original ResNet network; second, after the backbone network part of the feature fusion, we introduced a fused multi-scale attention mechanism, C2f_SKNet module, to further enhance the expression of important features; then, in order to better deal with defects at different scales, we introduce FPN into the network to enhance the detection of multi-scale defects; finally, we replace the previous circular Gaussian kernel with an elliptical Gaussian

kernel in the heat map generation part of the network head to better match the aspect ratio of the defects, which improves the accuracy and robustness of the defects detection, and also accelerates the training speed of the network at the same time. Our network's refined structure is presented in Figure 2.

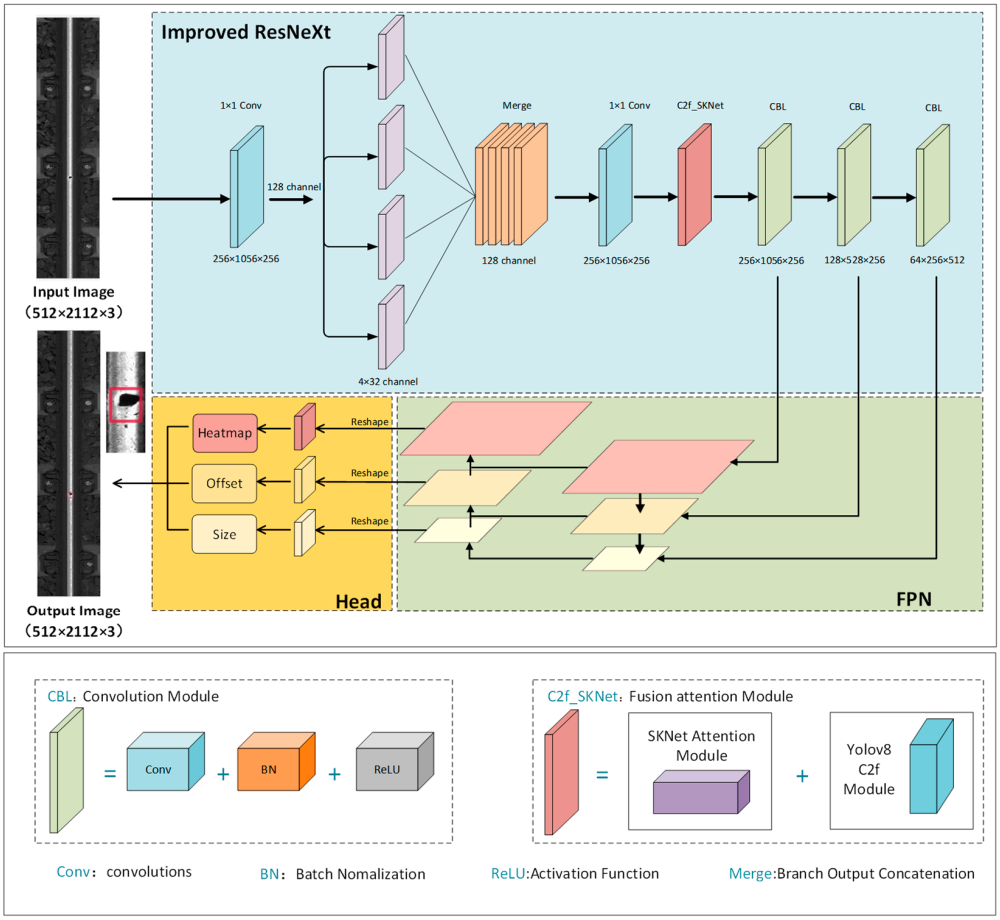


Figure 2. Diagram of the improved network structure.

3.1. ResNeXt_b

ResNeXt [26] is an innovative convolutional neural network architecture based on ResNet, which improves the efficiency and performance of the network by introducing the concept of group convolution to extend the previous residual module. However, grouped convolution may lose some information in low-level feature extraction because it limits the perceptual field of each convolutional kernel to process only some of the channels of data features are usually missing to capture are the basic visual elements in the image, such as edges, corners, textures, etc. For the rail scar detection task, low-level features are very important, and the original ResNeXt network may result in a lack of features making the detection ineffective [27]. Since scars usually exhibit edge abnormalities or texture changes. Therefore, we propose a modified ResNeXt_B network architecture as in Figure 3. The architecture is based on the classical ResNeXt and enriches it for low-level feature extraction in images by changing to four parallel grouped convolutional structures at the input feature layer, followed by the outputs of the four branches being merged into a 128-channel feature map in order to integrate the features of the different branches for enhanced feature representation.

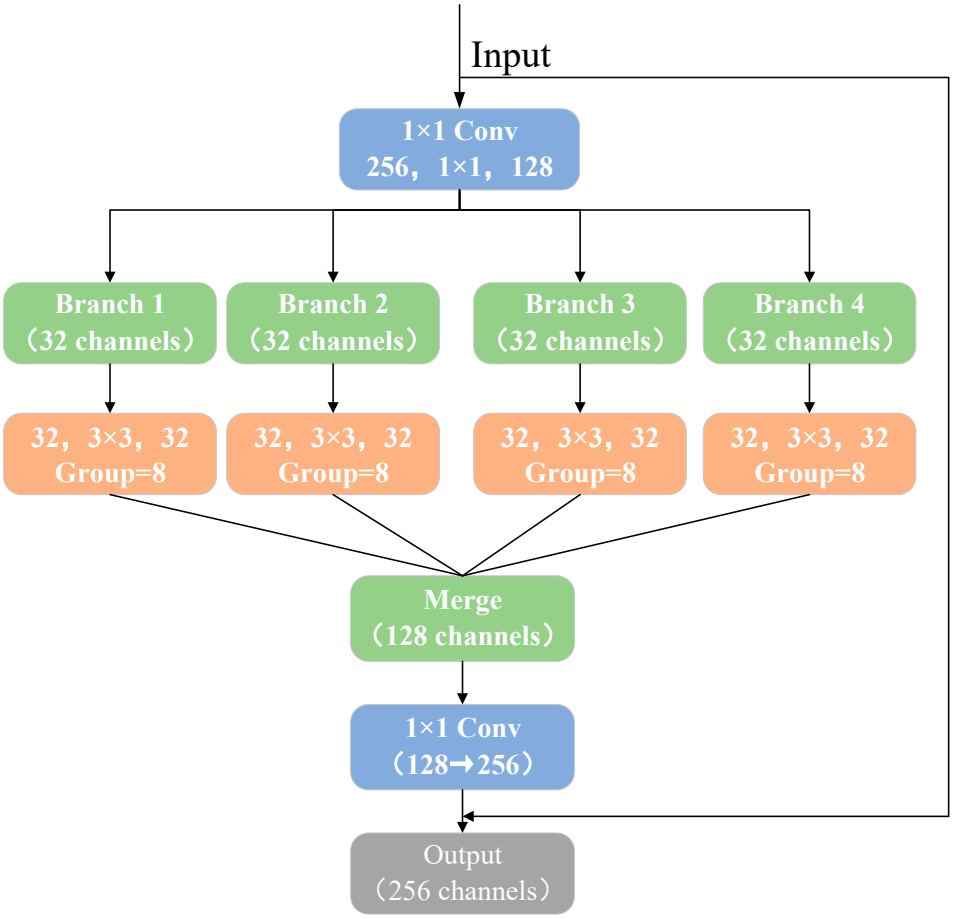


Figure 3. ResNeXt_B Network Architecture Diagram.

3.2. C2f_SKNet Module

Based on the characteristics of different scale sizes of rail surface defect species, we introduced the SKNet [28] attention module, which is able to adaptively select different convolution kernels to better capture features at different scales in the image. The structure of the network module is shown in Figure 4.

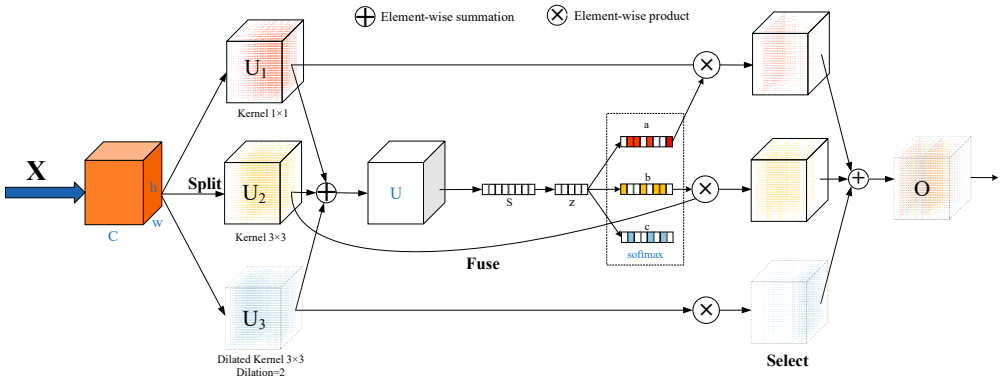


Figure 4. Structure of SKNet module.

SKNet is implemented by three operations Split, Fuse and Select, the Split operation is a multi-scale feature extraction technique. In this operation, the input feature map is convolved through multiple convolution kernels of different sizes and each convolution operation forms a feature branch. These differences in convolutional kernel sizes allow them to capture features at different scales, and multi-branch SKNet [29] can adapt to features at multiple scales, so

we selected three convolutional kernel sizes, 1×1, 3×3, and 5×5, for the Split operation. And since it is a small target detection, we introduce dilation convolution [30] into the 5×5 convolutional kernel, which is also called cavity convolution, a method to extend the sensory field of convolutional kernel. It works by inserting blank 0-value pixels between each pixel point of the convolution kernel, thus indirectly increasing the coverage of the convolution kernel without significantly increasing the amount of computation or the number of parameters. It is done by inserting gaps between every two neighboring elements of the convolution kernel. These voids do not contain actual weight values, they just make the coverage area of the convolution kernel larger. For the convolution kernel we designed with a 3×3 convolution kernel, using a null convolution with an expansion rate of 2 is equivalent to inserting 1 gap between every two neighboring elements. Due to these inserted gaps, the convolutional kernel's receptive field is equivalent to a 5×5 convolutional kernel, which can capture a wider range of image contextual information rather than being limited to the region covered by the original convolutional kernel size. Despite the increase in the receptive field, the actual computed convolution operation still only considers the original element positions of the convolution kernel. As a result, the number of parameters and the amount of computation do not increase with the expansion of the receptive field, and the results of the comparison between the standard convolution kernel and the expanded convolution kernel are shown in Figure 5.

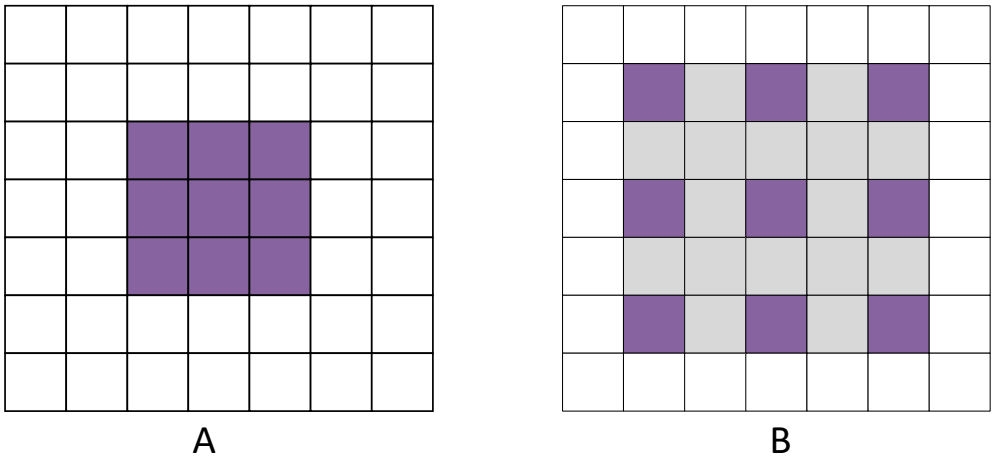


Figure 5. Comparison of standard convolution and dilated convolution. (**A**) Standard convolution; (**B**) dilated convolution with dilation ratio of 2.

The Split operation is followed by the Fuse operation to integrate the global information of the spliced features, and the weights for different scales are generated through the fully connected layer and feature stacking operations. Then finally these weights are applied to different feature branches through Select operation to get a weighted combined feature map, such weighted combination makes the network able to flexibly adjust its sensory field, which can be adapted to different sizes of targets. Based on SKNet's ability to enhance feature expression through multi-scale convolution and dynamic selection mechanism, we also fused the cross-stage part of the connectivity structure from the C2f module of the YOLOv8 network at the back of the network structure to turn it into a more adaptable attention model—C2f_SKNet, which enhances the feature expression ability of the model through shared feature map and multiscale feature fusion to enhance the feature expression capability of the model to provide efficient feature fusion strategies while maintaining the integrity of feature information [31]. The final network structure is shown in Figure 6.

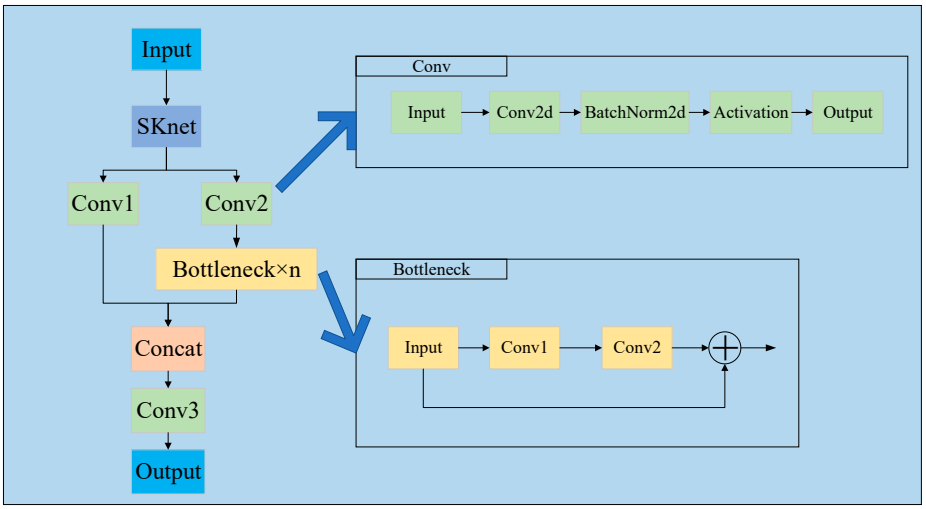


Figure 6. C2F_SKNet Attention Mechanism Network Architecture Diagram.

The overall structure includes, SKNet module, two-branch convolutional layer, Bottleneck module, and feature splicing layer. Among them, the Conv structure is the 2D convolutional module of YOLOv8, which uses a cross-stage partially connected structure to keep some features unchanged, and the other part is processed by the Bottleneck module and then feature fusion, which retains the rich contextual information, and significantly improves the accuracy of small target detection and detection in complex backgrounds.

3.3. Elliptic Gaussian Kernel

In CenterNet, the anchor-free based design generates a heat map by applying a Gaussian kernel at the center of the target, where the center point of the target has the highest value and decreases with distance, each target generates a Gaussian distribution at its center, and then all the Gaussian distributions are superimposed on a shared heat map. The value of the 2D Gaussian kernel at the image coordinates (x, y) is calculated as in Equation (1):

$$G(x,y) = \exp\left(-\frac{(x-x_c)^2 + (y-y_c)^2}{2\sigma^2}\right) \tag{1}$$

where (x_c, y_c) are the coordinates of the center point of the target in the target in the heat map, σ is the standard deviation of the Gaussian kernel, which determines the degree of expansion of the Gaussian distribution. In the inference phase, the model outputs a predictive heat map, and by looking for peaks in the heat map, the location and confidence level of the presence of the target can be determined, but relying on the centroid alone for size regression also poses a training challenge. The fact that only the information from the center point is used may cause the model to be less efficient in dealing with the size and scale of the target, which in turn increases the difficulty of training.

Improving training efficiency usually takes two approaches: increasing the learning rate and reducing the data augmentation, where too much of the former may result in training dispersion and instability, and the latter may cause the model to focus more on basic feature learning and lead to overfitting. Therefore, we draw on the theoretical ideas of [32] and change the original circular Gaussian kernel into an elliptical Gaussian kernel. This improvement further considers the aspect ratio of the target on the basis of the center point, which makes the model reflect the shape and size of the target more accurately in the localization regression, and it also conforms to the characteristics of the defects of the surface of the rail in the shape of a variety of sizes. We obtain the length-short axis ratio of the elliptical Gaussian kernel by labeling the width as well as the height of the target, and list a new elliptical Gaussian kernel ground calculation formula as in Equations (2)–(4):

$$G(x,y) = \exp(-(\frac{(x-x_c)^2}{2{\sigma_x}^2} + \frac{(y-y_c)^2}{2{\sigma_y}^2})) \tag{2}$$

$$\sigma_x = \frac{width}{6} \tag{3}$$

$$\sigma_y = \frac{height}{6} \tag{4}$$

where σ_x and σ_y are the long and short axes of the elliptical Gaussian kernel, dynamically adjusted according to the width and height of the target, the heat map can be more in line with the characteristics of the rail surface defects of unequal proportions of the length and width of the circular Gaussian kernel heat map and the elliptical Gaussian kernel heat map comparison effect is shown in Figure 7.

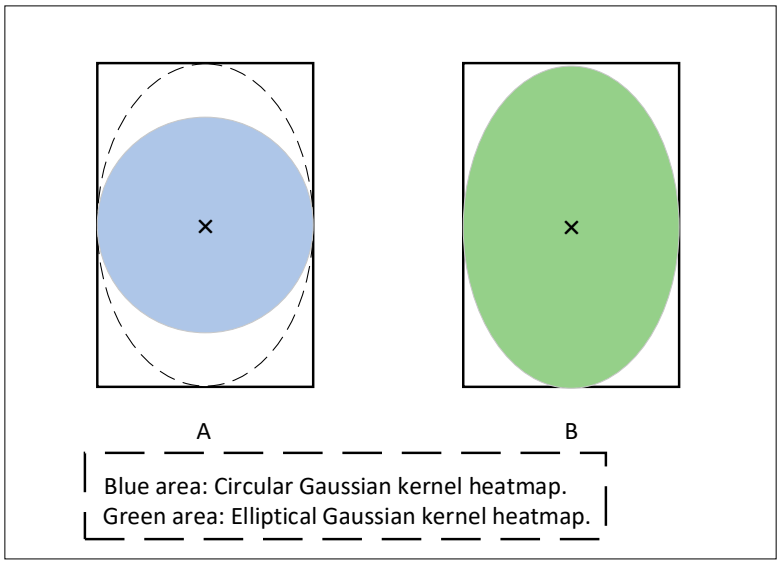


Figure 7. Comparison of circular Gaussian kernel heatmap and elliptical Gaussian kernel heatmap effects.

From the Figure 7, we can see that for the same target labeling box, the heat map generated by the elliptical Gaussian kernel occupies a larger distribution density, which makes the model predict a larger region of interest and improves the model's detection accuracy and training efficiency.

4. Experiments and Results

4.1. Experimental Environment

The test environment for this experiment is Windows 11 operating system, the hardware configuration system is Inte ® Core i7-14650HX CPU, NVIDIA GeForce RTX 4060 graphics card and 16 GB RAM. The deep learning configuration uses CUDA11.7, CUDNN8.9 and the framework platform is pytorch version 2.2.1.

4.2. Image Acquisition

The rail surface image acquisition system described in this paper primarily consists of a linear image acquisition unit and a rail detection beam, as depicted in Figure 8. This system includes an industrial-grade high-speed linear CCD camera and a non-visible light source. To achieve synchronized image capture and spatial isometric sampling with the two CCD linear cameras, a high-precision speed sensor is mounted on the wheelset. The key parameters of the image capture unit are provided in Table 1.

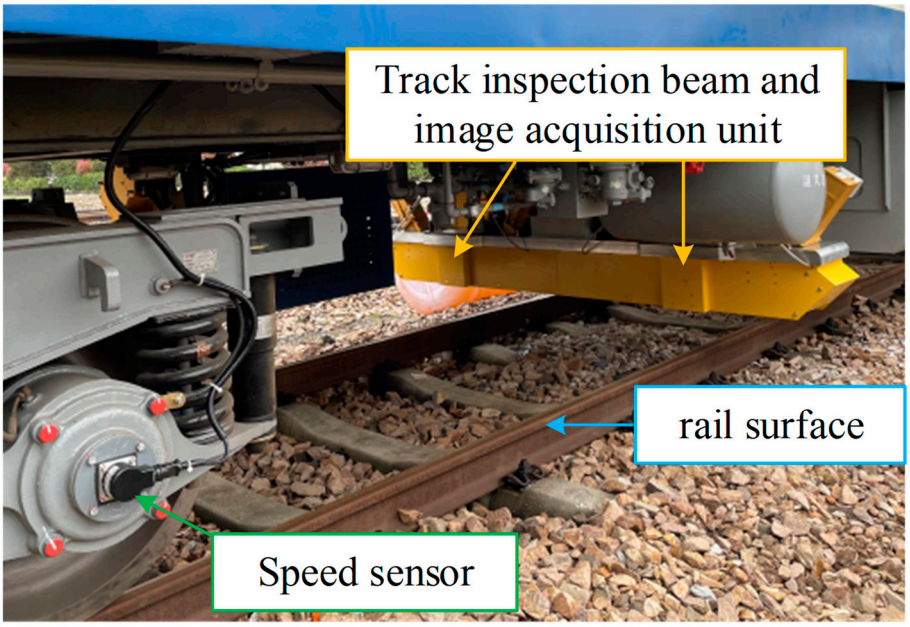


Figure 8. Image acquisition system.

Table 1. Parameters of the image acquisition unit.

Parameter	Value
Camera lens	8 mm
Line array camera resolution	2048 pixels
Data output	GigE
Laser power	15 W
Laser center wavelength	808 nm
Divergence angle of light source	75°
Camera field of view angle	67°
Sunlight interference resistance	Resistant to diffuse sunlight

The maximum inspection speed of the railroad inspection vehicle is about 80 km/h. The size of each captured image is 512 mm × 2112 mm, and a single image acquisition unit can capture 25 images per second.

4.3. Rail Surface Defect Dataset

This study uses a customized rail surface defect dataset designed to detect and classify different types of defects on the rail surface. The dataset consists of 321 images of rail surface defects, including four common types of rail surface defects: abrasions, scars, dents, and seams. Each image clearly shows a specific type of defect for model training and testing.

4.4. Evaluation Indicators

To validate the improvement of the model, we applied precision, recall, and mean accuracy (*mAP*) for comprehensive evaluation. They are calculated as in Equations (5)–(8):

$$Precision = \frac{TP}{TP + FP} \times 100\% \tag{5}$$

where *TP* denotes the number of true positive samples that were correctly predicted as positive, and *FP* denotes the number of true negative samples that were incorrectly predicted as positive. Accuracy reflects how accurately the model recognizes positive samples, i.e., how many of the results that the model predicts as positive samples are correct.

$$Recall = \frac{TP}{TP + FN} \times 100\% \tag{6}$$

where FN denotes the number of samples that are true positive samples but incorrectly predicted as negative samples. Recall reflects the model's ability to cover positive samples, i.e., how many of all positive samples are correctly identified.

$$AP = \frac{1}{11} \sum_{r \in \{0, 0.1 \cdots, 1\}} Precision(r) \tag{7}$$

For each category, the area under the precision-recall curve is calculated as the average precision (AP) for that category. A common calculation is to divide the recall into points and then average the maximum precision for each point.

$$mAP = \frac{1}{N} \sum_{i=1}^{N} AP_i \tag{8}$$

The APs of all categories are averaged to obtain mAP, where N is the total number of categories. mAP in the rail surface defect detection task is obtained experimentally to measure the overall detection performance of the model on various defect types.

We assess the computational complexity of the model using FLOPs, which indicate the number of floating-point operations required during a single forward pass. The size of the model is measured by the number of parameters; fewer parameters result in a smaller model size and lower storage and computation requirements. In addition, we also use FPS (frames per second) to measure the actual running speed of the model as an important indicator, and the average training time (Training time) to measure the model training efficiency.

4.5. Ablation Experiment

In order to verify the influence of each component on the performance of our proposed model for detecting surface defects on rails, we conducted ablation experiments. The setup and results of the experiment are shown in Table 2.

Table 2. Ablation test result. (↑ means higher index is better, ↓ means lower index is better, √ means the current module is added.)

Experiment	ResNeXt_b	C2f_SKNet	EGK	*mAP*_50 ↑	Recall ↑	Training Time ↓
1	-	-	-	0.893	0.725	0.2925
2	√	-	-	0.927	0.751	0.3147
3	-	-	√	0.915	0.724	0.1852
4	√	√	-	0.938	0.779	0.3654
5	√	√	√	0.952	0.803	0.1894

Experiment 1 is the unimproved CenterNet network model without any components, and the mAP and Recall reach 0.893 and 0.725, respectively, while the training time is longer at 0.2925 s/L. We replace the ResNet backbone network of the original model with a multi-branch ResNeXt network in Experiment 2, which improves the performance of the network by 3.8% and enriches the extraction of low-level features by 2.5%, due to the introduction of group convolution. network performance as well as enriching the extraction of low-level features, the mAP is improved by 3.8% and Recall by 2.5%. In Experiment 3, we improved the prediction ability of the model as well as the training efficiency by replacing the original circular Gaussian heat map with an elliptical Gaussian heat map with a larger region of interest, with a 2.4% improvement in mAP and a 36.6% faster Training time than before. In order to verify the enhancement of the fused multi-scale attention on the feature extraction of the network model, we added the C2f_SKNet module to Experiment 2, which

resulted in better *mAP* and Recall, but the training time of Experiment 4 became slower at the same time, due to the substantial increase in the number of parameters as well as a more complex number of network layers. Therefore, after we used all the improvements together in Experiment 5, the model performance is further improved with 4.9% improvement in *mAP*, 10.7% improvement in Recall, and 35.2% speedup in Training time.

4.6. Performance Comparison Experiment

In order to verify the effectiveness of our proposed method in the detection of rail surface defects, we compared it with several mainstream target detection networks, and a detailed comparison of the experimental results is shown in Table 3, and a graph comparing the effect of network performance is shown in Figure 9, the four color boxes in Figure 9 are the corresponding detection boxes for each of the four types of defects. Among them, Faster-RCNN [33], Cascade-RCNN [34] and TridentNet [35] belong to the two-stage networks with a large number of parameters to improve precision at the expense of processing speed. Faster R-CNN has high detection accuracy. It is optimized for region extraction and feature processing, and is a benchmark for many high-precision detection tasks, and is well suited for complex backgrounds and multi-scale defects such as rail defects. Cascade R-CNN is an improved version of its predecessor, which further improves detection accuracy through a multi-stage cascade structure, especially for difficult objects. TridentNet introduces a multi-branch structure to better handle objects of different scales. a multi-branch structure to better handle objects of different scales, which is similar to our improvement and can be compared with the advantages and disadvantages by experimental results. But they are still not as good as the networks in this paper in terms of precision and recall. Our network is 9.17%, 9.17%, and 5.43% better than the previous three in terms of precision, and 41.62%, 27,30%, and 21.12% better than the previous three in terms of recall. SSD is one of the most classical one-stage networks, it is suitable for real-time inspection tasks for evaluating the balance between speed and accuracy of new models. Although the network parameters as well as the processing speed are small, the number of parameters is 33.78% smaller compared to our network, and the FPS is improved by 55.63% compared to our network, the detection accuracy and recall are far inferior to our network. FCOS is a single-stage detector without anchor frames, closest to CenterNet, the initial network of this paper, which utilizes a fully convolutional network to directly predict the location and class of an object, simplifying the detection process and improving speed. YOLOv8 is the more effective YOLO architecture algorithm, which combines a number of advanced detection techniques. However, due to the lack of strategies designed for small objects, he is not as effective as our network, outperforming it by 1.93% and 5.24% in terms of precision and recall.

Table 3. Comparision of Detection Performance of Different Networks. (↑ means higher index is better, ↓ means lower index is better.)

Model	*mAP*-50 ↑	*mAP* ↑	*Recall* ↑	Flops/G ↓	Params/M ↓	FPS img/s ↑
Faster-RCNN	0.872	0.531	0.567	215.04	99.639	13.7
FCOS	0.864	0.615	0.671	90.882	32.502	31.5
SSD	0.787	0.442	0.510	21.576	24.288	75.3
YOLOv8	0.934	0.694	0.763	29.36	30.06	40.93
Cascade-RCNN	0.872	0.575	0.631	946.176	77.411	12.9
TridentNet	0.903	0.592	0.663	799.744	33.438	9.8
Ours	0.952	0.763	0.803	90.86	32.498	33.4

We also performed a scatter plot depiction as in Figure 10 based on the relationship between each network Params and *mAP*, respectively, to demonstrate the comprehensive performance of each model. From the figure, it can be visualized that our network significantly improves the detection accuracy while keeping the complexity low, highlighting the advantages of the comprehensive performance.

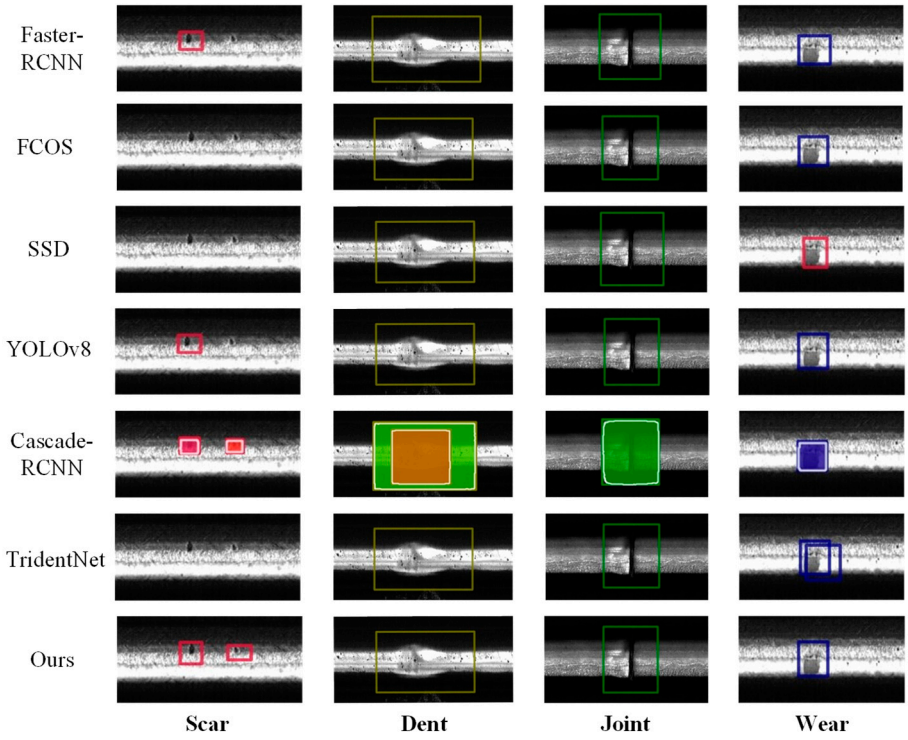


Figure 9. Comparison chart of the effect of detection.

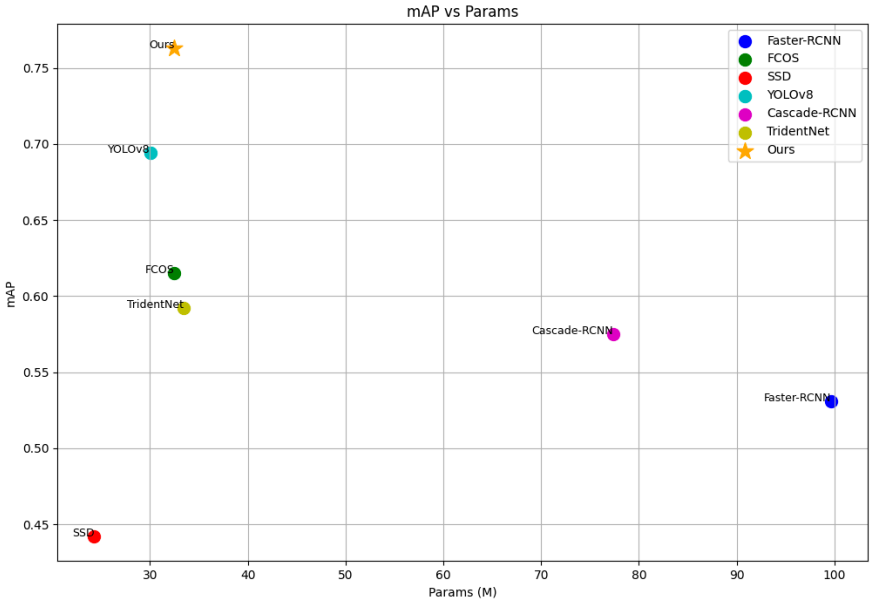


Figure 10. Scatter Plot of the Relationship between Network Parameters and *mAP* Metrics.

5. Conclusions

In this paper, we propose a new model based on CenterNet that integrates a multi-branch ResNeXt backbone network with a fused multi-scale attention mechanism, specifically targeting the challenges of detecting small targets and handling a diverse range of defects on the rail surface. Our model achieves the highest mean Average Precision (*mAP*) of 0.763, outperforming several mainstream target detection models. In addition, the Recall rate of our model reaches 0.803, significantly higher than other models, which indicates its superior performance in detecting small and less prominent defects.

In terms of efficiency, our model maintains a relatively low computational complexity with 90.86 GFLOPs and 32.498 million parameters, while achieving a processing speed

of 33.4 images per second. This makes it not only faster than two-stage networks like Faster R-CNN but also competitive with other single-stage detectors, as demonstrated. This balance between accuracy and efficiency highlights the effectiveness of integrating the SKNet attention mechanism with the C2f module from YOLOv8, as detailed in the ablation experiment in Table 2.

By comparing the various indices from our experiments, we demonstrate that our model achieves high-accuracy rail surface defect detection without significantly sacrificing speed and computational resources. Specific improvements include a 6.6% increase in *mAP* and a 35.5% reduction in training time compared to the unimproved model. The model's detection performance on real datasets as shown in Figure 10 reflects its lower leakage rate and reliable target detection, making it more suitable for deployment in real-time applications where both accuracy and efficiency are critical. These results confirm that our approach provides a robust and efficient solution for real-time rail defect detection.

Author Contributions: Conceptualization, Y.M. and S.Z.; methodology, Y.M.; software, Y.M.; validation, Y.M., L.L. and R.S.; formal analysis, X.A.; investigation, Y.M.; resources, S.Z.; writing—original draft preparation, Y.M.; writing—review and editing, L.L.; visualization, R.S. and X.A.; supervision, Y.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (Grant No. 51975347).

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Cannon, D.F.; Edel, K.O.; Grassie, S.L.; Sawley, K. Rail defects: An overview. *Fatigue Fract. Eng. Mater. Struct.* **2003**, *26*, 865–886. [CrossRef]
2. Sun, M.; Lin, X.; Wu, Z.; Liu, Y.; Shen, Y.; Feng, N. Non-destructive photoacoustic detecting method for high-speed rail surface defects. In Proceedings of the 2014 IEEE International Instrumentation and Measurement Technology Conference (I2MTC) Proceedings, Montevideo, Uruguay, 12–15 May 2014; pp. 896–900.
3. Xiong, L.; Jing, G.; Wang, J.; Liu, X.; Zhang, Y. Detection of Rail Defects Using NDT Methods. *Sensors* **2023**, *23*, 4627. [CrossRef] [PubMed]
4. Han, S.-W.; Cho, S.-H.; Jang, G.-W.; Park, J.-H. Non-contact inspection of rail surface and internal defects based on electromagnetic ultrasonic transducers. *J. Intell. Mater. Syst. Struct.* **2016**, *27*, 427–434. [CrossRef]
5. Felice, M.V.; Fan, Z. Sizing of flaws using ultrasonic bulk wave testing: A review. *Ultrasonics* **2018**, *88*, 26–42. [CrossRef] [PubMed]
6. Wu, Q.; Dong, K.; Qin, X.; Hu, Z.; Xiong, X. Magnetic particle inspection: Status, advances, and challenges-demands for automatic non-destructive testing. *NDT E Int.* **2023**, *143*, 103030. [CrossRef]
7. Rajamäki, J.; Vippola, M.; Nurmikolu, A.; Viitala, T. Limitations of eddy current inspection in railway rail evaluation. *Proc. Inst. Mech. Eng. Part F J. Rail Rapid Transit* **2018**, *232*, 121–129. [CrossRef]
8. Kumar, A.; Harsha, S.P. A Systematic Literature Review of Defect Detection in Railways Using Machine Vision-Based Inspection Methods. *Int. J. Transp. Sci. Technol.* **2024**. [CrossRef]
9. Duan, K.; Bai, S.; Xie, L.; Qi, H.; Huang, Q.; Tian, Q. CenterNet: Keypoint Triplets for Object Detection. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6568–6577.
10. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
11. Smadja, D.; Touboul, D.; Cohen, A.; Doveh, E.; Santhiago, M.R.; Mello, G.R.; Krueger, R.R.; Colin, J. Detection of subclinical keratoconus using an automated decision tree classification. *Arch. Ophthalmol.* **2013**, *156*, 237–246.e1. [CrossRef] [PubMed]
12. Ojala, T.; Pietikainen, M.; Harwood, D. Performance evaluation of texture measures with classification based on Kullback discrimination of distributions. In Proceedings of the 12th International Conference on Pattern Recognition, Jerusalem, Israel, 9–13 October 1994; pp. 582–585.
13. Zhang, C.; Zhao, Q.; Shen, T.; Sun, B. Rail Defect Detection Method Based on Improved XGBoost. In Proceedings of the International Conference on Computer Engineering and Networks, Haikou, China, 4–7 November 2022; Springer Nature: Singapore, 2022; pp. 911–920.
14. Zhu, W.; Xiang, Y.; Zhang, H.; Cheng, Y.; Fan, G.; Zhang, H. Research on ultrasonic sparse DC-TFM imaging method of rail defects. *Measurement* **2022**, *200*, 111690. [CrossRef]

15. Deng, F.; Li, S.-Q.; Zhang, X.-R.; Zhao, L.; Huang, J.-B.; Zhou, C. An Intelligence Method for Recognizing Multiple Defects in Rail. *Sensors* **2021**, *21*, 8108. [CrossRef] [PubMed]
16. Li, Q.; Ren, S. A Real-Time Visual Inspection System for Discrete Surface Defects of Rail Heads. *IEEE Trans. Instrum. Meas.* **2012**, *61*, 2189–2199. [CrossRef]
17. Shi, T.; Kong, J.-Y.; Wang, X.-D.; Liu, Z.; Zheng, G. Improved Sobel algorithm for defect detection of rail surfaces with enhanced efficiency and accuracy. *J. Cent. South Univ.* **2016**, *23*, 2867–2875. [CrossRef]
18. Choi, J.-Y.; Han, J.-M. Deep Learning (Fast R-CNN)-Based Evaluation of Rail Surface Defects. *Appl. Sci.* **2024**, *14*, 1874. [CrossRef]
19. Aydın, I.; Akın, E. Two-Stage Rail Defect Classification Based on Fuzzy Measure and Convolutional Neural Networks. In Proceedings of the International Conference on Intelligent and Fuzzy Systems, Izmir, Turkey, 19–21 July 2022; Springer International Publishing: Cham, Switzerland, 2022; pp. 769–776.
20. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016, Proceedings of the 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016*; Proceedings, Part I 14; Springer International Publishing: Cham, Switzerland, 2016; pp. 21–37.
21. Zheng, Z.; Qi, H.; Zhuang, L.; Zhang, Z. Automated rail surface crack analytics using deep data-driven models and transfer learning. *Sustain. Cities Soc.* **2021**, *70*, 102898. [CrossRef]
22. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
23. Zhang, C.; Xu, D.; Zhang, L.; Deng, W. Rail Surface Defect Detection Based on Image Enhancement and Improved YOLOX. *Electronics* **2023**, *12*, 2672. [CrossRef]
24. Zhu, X.; Lyu, S.; Wang, X.; Zhao, Q. TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 2778–2788.
25. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
26. Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; He, K. Aggregated residual transformations for deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1492–1500.
27. Du, J.; Zhang, R.; Gao, R.; Nan, L.; Bao, Y. RSDNet: A New Multiscale Rail Surface Defect Detection Model. *Sensors* **2024**, *24*, 3579. [CrossRef] [PubMed]
28. Li, X.; Wang, W.; Hu, X.; Yang, J. Selective kernel networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 510–519.
29. Hu, B.; Liu, J.; Xu, Y.; Huo, T. An Integrated Bearing Fault Diagnosis Method Based on Multibranch SKNet and Enhanced Inception-ResNet-v2. *Shock Vib.* **2024**, *2024*, 9071328. [CrossRef]
30. Yu, F.; Koltun, V. Multi-scale context aggregation by dilated convolutions. *arXiv* **2015**, arXiv:1511.07122.
31. Liu, H.; Zhou, K.; Zhang, Y.; Zhang, Y. ETSR-YOLO: An improved multi-scale traffic sign detection algorithm based on YOLOv5. *PLoS ONE* **2023**, *18*, e0295807. [CrossRef] [PubMed]
32. Liu, Z.; Zheng, T.; Xu, G.; Yang, Z.; Liu, H.; Cai, D. Training-time-friendly network for real-time object detection. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 11685–11692.
33. Faster, R. Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *9199*, 2969239–2969250.
34. Cai, Z.; Vasconcelos, N. Cascade R-CNN: Delving Into High Quality Object Detection. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6154–6162.
35. Li, Y.; Chen, Y.; Wang, N.; Zhang, Z. Scale-Aware Trident Networks for Object Detection. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6053–6062.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.