Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.060661

ARTICLE

MSCM-Net: Rail Surface Defect Detection Based on a Multi-Scale Cross-Modal Network

Xin Wen*, Xiao Zheng and Yu He

School of Software Engineering, Shenyang University of Technology, Shenyang, 110870, China

*Corresponding Author: Xin Wen. Email: wen_xin@sut.edu.cn

Received: 07 November 2024; Accepted: 21 January 2025; Published: 06 March 2025

ABSTRACT: Detecting surface defects on unused rails is crucial for evaluating rail quality and durability to ensure the safety of rail transportation. However, existing detection methods often struggle with challenges such as complex defect morphology, texture similarity, and fuzzy edges, leading to poor accuracy and missed detections. In order to resolve these problems, we propose MSCM-Net (Multi-Scale Cross-Modal Network), a multiscale cross-modal framework focused on detecting rail surface defects. MSCM-Net introduces an attention mechanism to dynamically weight the fusion of RGB and depth maps, effectively capturing and enhancing features at different scales for each modality. To further enrich feature representation and improve edge detection in blurred areas, we propose a multi-scale void fusion module that integrates multi-scale feature information. To improve cross-modal feature fusion, we develop a cross-enhanced fusion module that transfers fused features between layers to incorporate interlayer information. We also introduce a multimodal feature integration module, which merges modality-specific features from separate decoders into a shared decoder, enhancing detection by leveraging richer complementary information. Finally, we validate MSCM-Net on the NEU RSDDS-AUG RGB-depth dataset, comparing it against 12 leading methods, and the results show that MSCM-Net achieves superior performance on all metrics.

KEYWORDS: Surface defect detection; multiscale framework; cross-modal fusion; edge detection

1 Introduction

Railways are a critical component of railway infrastructure, directly influencing the safety and reliability of transportation. Repeated train loads and environmental factors can lead to defects in in-service tracks, compromising normal operations. Thus, defect detection in in-service tracks is of vital importance. However, unused tracks also warrant attention. Although these tracks have not yet been put into operation and have not been subjected to complex external forces, potential defects may already exist due to manufacturing, transportation, and storage processes. Once these tracks are operational, such defects can rapidly propagate under the combined effects of train loads and environmental factors, evolving into more severe quality issues and even posing safety risks. Therefore, early detection and remediation of defects in unused tracks can eliminate potential risks before they are put into use, ensuring safety and reliability during the initial service period. This study provides critical support for the lifecycle health management of railway tracks and lays a foundation for further advancements in defect detection technologies for in-service tracks. The inspection process for unused rails, as shown in Fig. 1, can help mitigate potential risks before they are put into service.

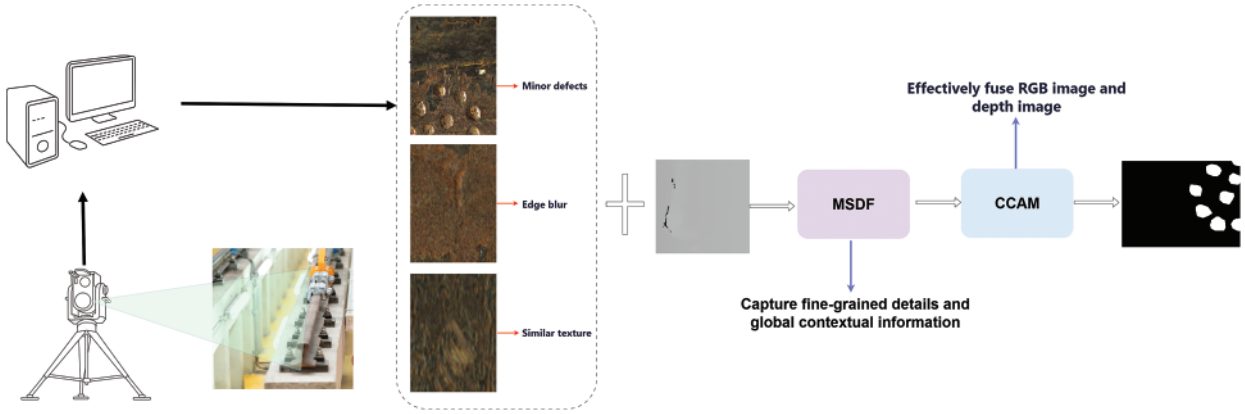

Figure 1: Defect detection process for unused rails

Currently, rail inspection methods primarily include physical techniques such as ultrasonic testing, eddy current testing, and magnetic flux leakage detection. However, these methods typically focus only on detecting internal rail defects and are often limited by slow inspection speeds and sensitivity to the condition of the detectors. In contrast, machine vision-based inspection methods have gained popularity among researchers due to their non-contact nature, high speed, and low cost. These methods are gradually replacing manual inspection for surface defect detection, and there are now many high-performance approaches available for rail surface defect detection.

The rapid progress in deep learning and enhanced computational power have enabled convolutional neural networks (CNNs) to transform rail defect detection. For example, Gibert et al. [1] introduced a deep learning approach that combines a multi-task learning framework with multiple detectors, achieving notable improvements in detection efficiency and accuracy, which underscores the potential of deep learning in railway maintenance. Zhang et al. [2] designed the Multiple Contextual Information Segmentation Network (MCnet), which utilizes dense blocks, pyramid pooling modules, and multi-information integration mechanisms to greatly enhance the segmentation accuracy of surface defects on unserviceable rails, effectively leveraging contextual information in complex scenarios. Zhou et al. [3] introduced a Dense Attention-guided Cascade Network (DACNet) that enhances the detection capabilities for significant defects on rail surfaces, particularly excelling in the identification of small targets through the application of an attention mechanism. Ni et al. [4] developed an IoU-guided centroid estimation network focused on improving the detection accuracy of surface defects on rails, ensuring efficient detection results by accurately pinpointing defective regions. Collectively, these studies illustrate diverse technical approaches and innovative strategies that contribute substantially to the advancement of rail defect detection.

Existing studies primarily focus on general rail defect detection in standard railway systems, where the effectiveness of detection methods is limited when addressing complex challenges such as large areas of shadowing, irregular boundaries, and small target regions. In response to these challenges, we categorized the dataset based on several defect types that affect detection accuracy. The dataset contains a total of 1852 images, which are classified into five categories: 130 images of edge-effect defects, 614 images of single-target defects, 937 images of multiple defects, 71 images of minor defects, and 110 images of texture-similar defects, as shown in Fig. 2.

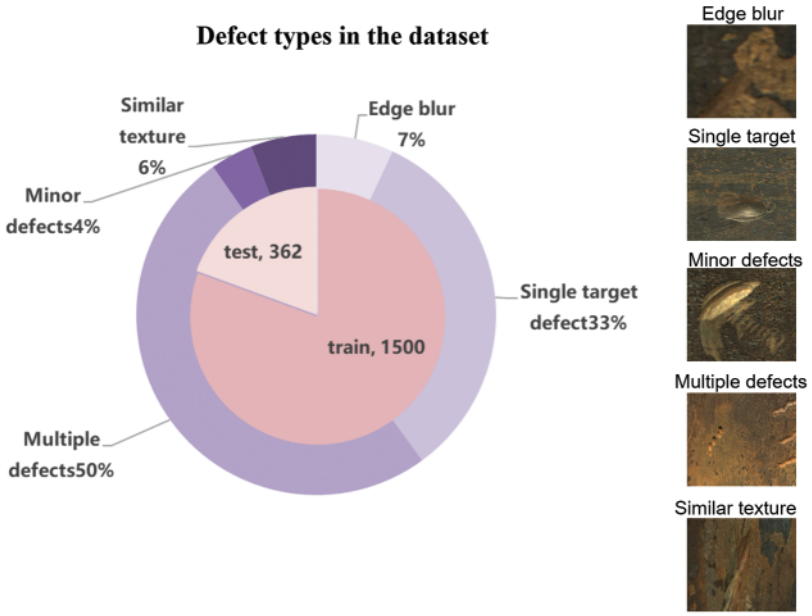

Figure 2: Defect types in the dataset

Because of the resemblance in color and texture, detecting defects using a single modality often fails to differentiate those with textures similar to the background. Therefore, the introduction of depth images is essential, as their pixel values provide depth information of objects within the camera coordinate system, enabling better differentiation between the similar foreground and background textures found on rails. However, deep learning-based detection still encounters significant technical challenges, such as optimizing and improving algorithm performance, enhancing accuracy, and reducing model size.

Fig. 3 illustrates several representative CNN-based RGB-SOD (RGB Surface Object Detection) methods. The detection results reveal issues such as missed defect detection and blurred edges, particularly in cases of irregular boundaries, similar foreground and background, and minor defects, resulting in imprecise outcomes. To overcome these challenges, we propose a new network. The main contributions of this paper are as follows:

1) Our proposed module utilizes multi-scale dilated convolution in conjunction with a channel-spatial attention mechanism, enabling the model to capture both local and global features across a range of receptive fields. Through the incorporation of dilated convolutions at different rates, the MSDF (Multi-Scale Dilation Fusion Module) module strengthens the model's capability to capture fine textures and edge details, thereby improving feature representation.

2) We construct a multi-modal fusion module (TSFM: Temporal Scale Fusion Module) that effectively combines detailed information from RGB images with the geometric structure from depth images, achieving more precise feature extraction. This design demonstrates strong adaptability in multi-modal tasks and significantly improves the model's performance in recognizing targets within complex scenes.

3) We develop a decoder structure capable of adaptively allocating weights to different features during the stepwise recovery of image details. This design effectively addresses challenges such as object deformation, texture similarity, and noise interference while preserving high-level semantic information, substantially increasing the model's capacity to identify boundaries and reassemble details in complicated situations, thereby improving robustness and generalization.

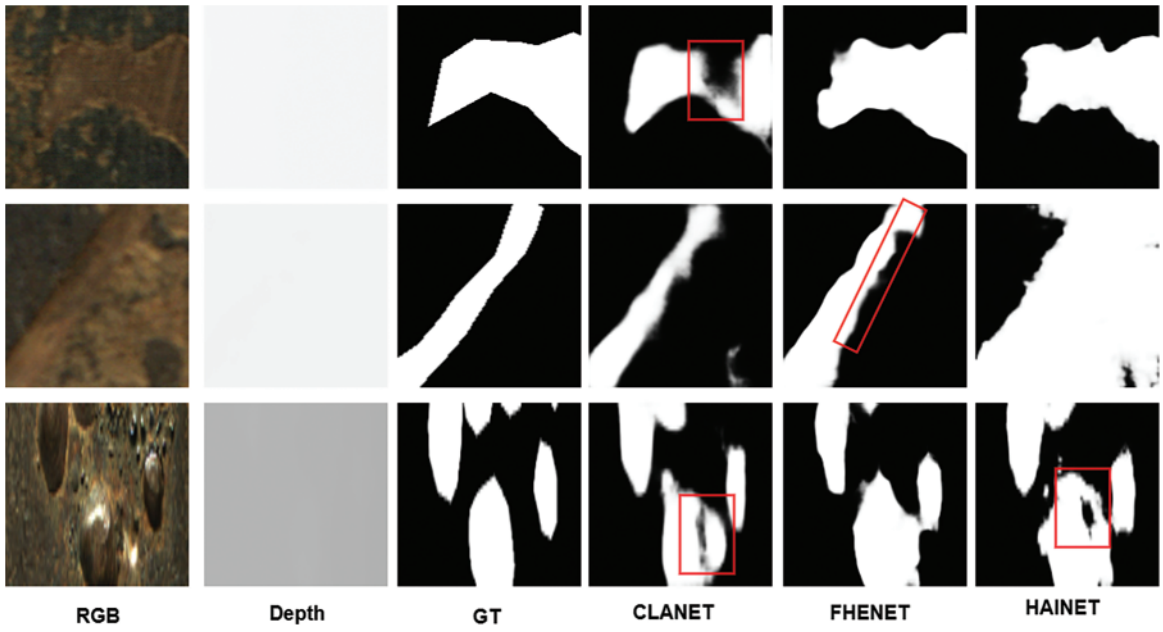

Figure 3: Saliency maps produced by three leading methods across three challenging scenarios

2 Related Works

2.1 Rail Surface Defect Detection

Detection methods for rail surface defects primarily include manual inspection, eddy current testing, magnetic flux leakage testing, ultrasonic testing, and machine vision-based inspection. Although traditional manual inspection is simple to operate, it is inefficient and poses safety risks. Ultrasonic testing technology, despite its broad detection range, has its accuracy limited by the contact-based nature of the measurement [5–7]. Eddy current testing [8,9] allows for non-destructive evaluation but operates at a slower speed and is susceptible to environmental interference. While magnetic flux leakage testing [10,11] can enable quantitative analysis, it is greatly affected by external magnetic fields. With the rapid advancement of computer vision, machine vision-based detection methods have gained popularity for their non-contact and high-efficiency advantages. In 2016, Faghih-Roohi et al. [12] took the lead in applying deep convolutional neural networks to this field, developing an end-to-end framework for automatically identifying defects like cracks and spalling. In 2018, Mercy et al. [13] systematically evaluated machine learning algorithms, including decision trees and random forests, for predicting rail defects, demonstrating that intelligent algorithms have, in certain scenarios, surpassed traditional detection methods. In 2019, Yu et al. [14] introduced a novel multi-scale detection approach that splits the detection process into coarse detection and fine analysis phases, improving system efficiency while preserving recognition accuracy. By 2020, research in this field entered a phase of rapid development. The network model proposed by Dong et al. [15] incorporates pyramid feature fusion and global context attention mechanisms, improving accuracy in surface defect detection and greatly enhancing the model's adaptability to various detection scenarios. He et al. [16] focused on the challenge of feature extraction, designing a multi-level feature fusion mechanism based on ResNet combined with an RPN (Region Proposal Network) network to construct a complete end-to-end detection system, thereby enhancing the algorithm's ability to recognize different defect types. Qin et al. [17] proposed an innovative multi-branch Deepfake detection algorithm that effectively utilizes fine-grained feature analysis. By incorporating a Feature Localization Module (FLM) and a Global Attention Module (GAM), this approach

significantly enhances the accuracy of identifying manipulated features. This design, based on multi-branch fine-grained feature extraction and reinforced attention mechanisms, also provides new insights for railway defect detection in complex environments. Meanwhile, Cao et al. [18] proposed the PSMFNet model, which uses partially separated convolutions and integrates multi-scale features to achieve impressive image super-resolution performance at a relatively low computational cost, demonstrating the advantages of multi-scale feature extraction. These methods offer valuable references for advancing techniques in related fields. Song et al. [19] proposed MFANet, a framework designed for cross-granularity few-shot defect segmentation, effectively leveraging coarse-grained data to improve fine-grained defect identification. In 2022, Wang et al. [20] introduced a collaborative attention network leveraging RGB-D images, effectively improving defect detection by integrating RGB images and depth information and achieving outstanding results on multiple public datasets. Considering the practical implementation of future models, these two papers [21,22] have provided me with new ideas. The content on sensor fusion methods, cost-effective alternatives, and fine-grained feature extraction holds significant reference value for rail defect detection.

These studies have advanced rail inspection technology toward greater intelligence and automation, laying the groundwork for further research in areas such as real-time performance optimization and system integration.

2.2 Traditional RGB-D Salient Object Detection

In early RGB-D salient object detection techniques, manually designed features were primarily used to identify target areas by analyzing properties such as contrast and spatial distribution in depth maps and RGB images. In 2012, Niu et al. [23] pioneered stereo saliency detection, leveraging disparity maps to extract depth cues from stereo images and established the STEREO dataset. In the same year, Lang et al. [24] applied a Gaussian mixture model to simulate the saliency distribution induced by depth, highlighting the critical role of depth information in saliency detection. As research progressed, more complex saliency detection models were developed. Desingh et al. [25] improved saliency detection accuracy in indoor environments by considering the 3D structural features of objects. Cheng et al. [26] combined color, depth contrast, and spatial deviation to propose a multi-cue depth-enhanced saliency detection method and created the DES dataset to validate its effectiveness. Feng et al. [27] introduced the Local Background Enclosure (LBE) method, enhancing saliency detection by integrating depth information with RGB data. This approach computes the depth variation between each pixel and its surrounding background to produce a saliency map, thereby better distinguishing foreground from background. However, this approach is highly dependent on depth map quality, has a high computational complexity, and its simple linear fusion limits the expressive relationship between RGB and depth information.

2.3 Deep Learning RGB-D Saliency Object Detection

With deep learning advancing rapidly, RGB-D saliency detection has evolved. CNNs enhance feature extraction, enabling models to learn complex RGB-depth interactions and overcome the limits of traditional features. In 2017, Qu et al. [28] pioneered the use of deep learning in RGB-D SOD, employing CNNs for multi-level feature extraction and generating saliency maps via Laplacian propagation. Subsequent models and fusion strategies further enhanced detection performance. Chen et al. [29] designed a progressive synergistic perception fusion network that improves detection results by fusing information across modalities and levels. In 2019, Piao et al. [30] developed a depth-induced multi-scale recurrent attention network to improve multimodal interaction and fusion via a recurrent mechanism.

To better handle multimodal information, Liu et al. [31] proposed the S2MA network, which optimizes multimodal fusion by applying an adaptive weighted attention mechanism to RGB and depth features

extracted by a dual-stream encoder, addressing issues of information loss and modal distribution differences found in traditional methods. Ji et al. [32] introduced an innovative collaborative learning framework, which jointly learns edge detection, saliency detection, and depth estimation tasks, effectively enhancing the mutual support between various features. Fan et al. [33] developed a Depth Depuration Unit (DDU) to automatically filter low-quality depth maps and enable cross-modal feature learning. UC-Net [34] first employed a conditional variational autoencoder (CVAE) to capture human-annotated uncertainty and produce diverse saliency maps, while introducing a depth correction network to denoise depth maps. Li et al. [35] proposed a hierarchical saliency detection network to manage feature interactions, while Zhai et al. [36] were the first to apply branched feature extraction and a cascade refinement mechanism to RGB-D saliency detection, helping to reduce noise in lower-level features. Niu et al. [37] proposed an improved YOLO model that integrates deep residual networks and densely connected networks, greatly boosting the precision and efficiency of object detection in complex situations. Finally, Song et al. [38] designed a modality-aware decoder that dynamically learns and leverages relationships between RGB and depth information during decoding, thereby achieving more efficient RGB-D fusion.

3 Methodology

This chapter presents MSCM-Net, a multi-scale cross-modal network that fuses RGB and depth images, aimed at efficiently detecting rail defects. The neural network architecture consists of three stages: feature extraction, cross-modal fusion, and defect localization and segmentation. Specifically, the feature extraction backbone network effectively captures the color features and depth features of rail defects by deeply mining information from different network layers. Subsequently, the fusion of cross-modal information aims to strengthen the distinction between defects and the background, enhancing the accuracy of defect identification and achieving precise prediction and localization of rail defects. Fig. 4 visually illustrates the overall architecture of MSCM-Net.

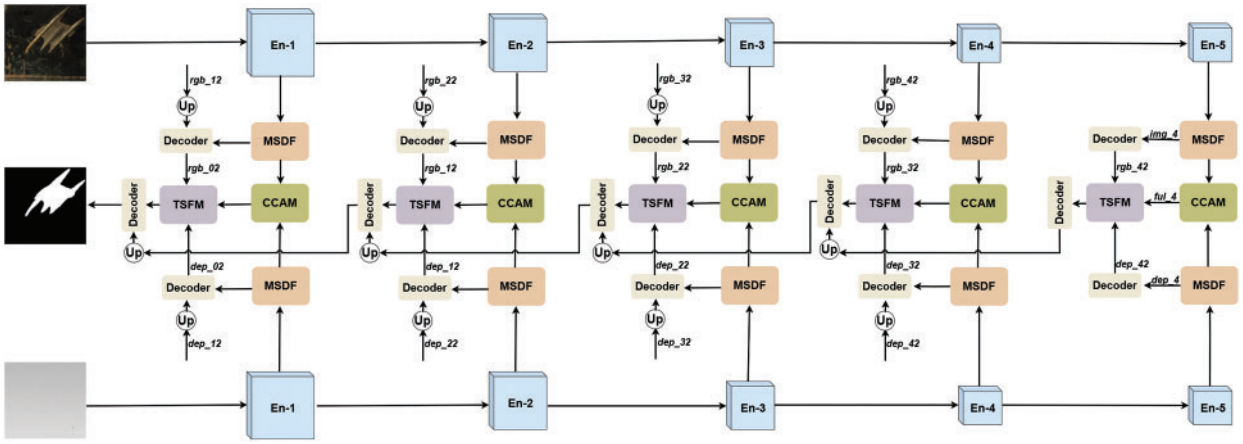

Figure 4: Framework of MSCM-Net

3.1 Feature Extraction

As in Fig. 4, the network uses a dual-stream model with a pretrained ResNet-34 encoder from ImageNet. The primary function of the dual-stream encoder is to extract features from RGB and depth images, resulting in five different levels of feature representations. Each feature extraction module is denoted as En(i), where i = 1, 2, 3, 4, 5 corresponds to different feature extraction layers.

3.2 Cross-Modal Information Fusion

As outlined in this paper, the proposed cross-modal fusion module incorporates three fundamental components to ensure seamless integration: the Multi-Scale Dilation Fusion Module (MSDF), the Criss-Cross Attention Module (CCAM), and the Temporal Scale Fusion Module (TSFM). In the upcoming part of this section, a thorough overview of each module will be presented.

1) Multi-Scale Dilation Fusion Module (MSDF):

In the task of rail defect detection, defects often exhibit characteristics such as irregular boundaries and small target areas. Traditional CNNs have limitations in multi-scale feature extraction and the reinforcement of key features. To address this issue, our MSDF module combines multi-scale dilated convolutions and attention mechanisms. This adaptive structure makes feature extraction more flexible and context-sensitive, making it better suited for detecting defects with irregular boundaries and small targets.

As shown in Fig. 5, the Multi-Scale Dilation Fusion Module (MSDF) consists of a multi-scale dilation convolution component and a channel-space attention fusion component. First, the MSDF module employs five parallel convolutional branches to achieve multi-scale feature extraction. The first branch uses a 1×1 convolutional kernel to directly extract fine-grained local features. The second to fourth branches utilize 3×3 convolutional kernels with dilation rates set to 6, 12, and 18, respectively, thereby gradually increasing the receptive field to capture contextual information over varying ranges. The fifth branch extracts global features through global average pooling, enhancing the model's understanding of the overall layout. This multi-scale convolutional structure ensures that the model can flexibly capture multi-scale feature information while maintaining spatial resolution, which is particularly crucial for handling the diverse morphology of rail defects.

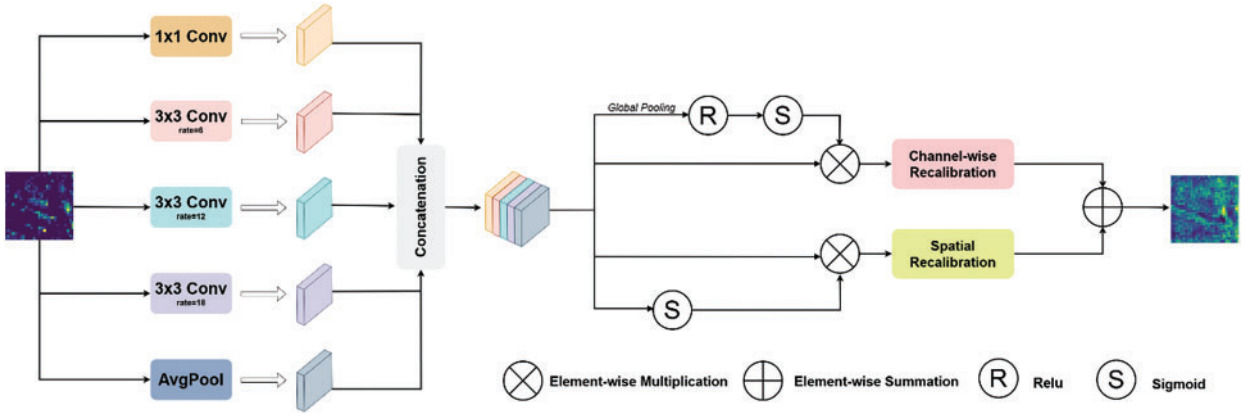

Figure 5: MSDF model diagram

The specific formulas are as follows:

$$\begin{aligned} F_{multi} = \; & Concat\,(\mathrm{ReLU}\,(\mathrm{BN}\,(Conv_{1\times1}\,(F_{in}))), \\ & \mathrm{ReLU}\,(BN\,(Conv_{3\times3,d=6}\,(F_{in}))), \\ & \mathrm{ReLU}\,(BN\,(Conv_{3\times3,d=12}\,(F_{in}))), \\ & \mathrm{ReLU}\,(BN\,(Conv_{3\times3,d=18}\,(F_{in}))), \\ & \mathrm{ReLU}\,(BN\,(Conv_{1\times1}\,(GlobalAvgPool\,(F_{in}))))) \end{aligned} \tag{1}$$

Here, $GlobalAvgPool(\cdot)$ represents global average pooling, and $Concat(\cdot)$ denotes the tensor concatenation operation.

In addition, the MSDF module utilizes both channel and spatial attention mechanisms to adaptively modify the feature weights based on the data context. Channel attention assesses the significance of each channel by applying global pooling followed by a 1D convolution, which boosts the response of essential channels. Meanwhile, spatial attention calculates the weights for each spatial location using 1×1 convolution, highlighting critical regions. These attention mechanisms effectively select important features while suppressing redundant information, consequently, this boosts the model's performance in detecting relevant patterns. Finally, the outputs from the multi-scale dilation convolutions and the attention mechanisms are fused and integrated through a 1×1 convolution for dimensionality reduction, resulting in the final output feature map. This module design equips the model with a stronger ability to capture features and achieve higher recognition accuracy when handling complex rail defect detection tasks. The specific formulas are as follows:

$$\begin{aligned} F_{channel} &= F_{multi} \otimes \text{Interpolate}\left(\text{ReLU}\left(Conv_{1\times1}\left(AvgPool\left(F_{multi}\right)\right)\right), \text{size} = (H, W)\right) \\ F_{spatial} &= F_{\text{in}} \otimes Sigmoid\left(Conv_{1\times1}\left(F_{multi}\right)\right) \\ F_{fusion} &= F_{channel} \oplus F_{spatial} \end{aligned} \tag{2}$$

Here, $\otimes$ denotes element-wise multiplication, $\oplus$ represents element-wise addition, and $AvgPool(\cdot)$ indicates average pooling.

2) Criss-Cross Attention Module:

Depth and RGB images each provide unique benefits in the fusion process. To better leverage these advantages and improve fusion quality, this paper introduces the enhanced CCAM module. While the original CAAF module [39] could fuse autocorrelation features from RGB and depth images, its utilization of depth information was limited, as illustrated in Fig. 6. Compared to traditional methods, our modules focus on addressing the issues of context-awareness and computational complexity in multimodal information fusion. CCAM strengthens pixel-level detail representation by capturing contextual information both horizontally and vertically. Compared to traditional networks that use dense connections to capture global context, CCAM's cyclic attention mechanism significantly reduces computational complexity and memory usage, making it particularly suitable for the efficiency demands of rail defect detection.

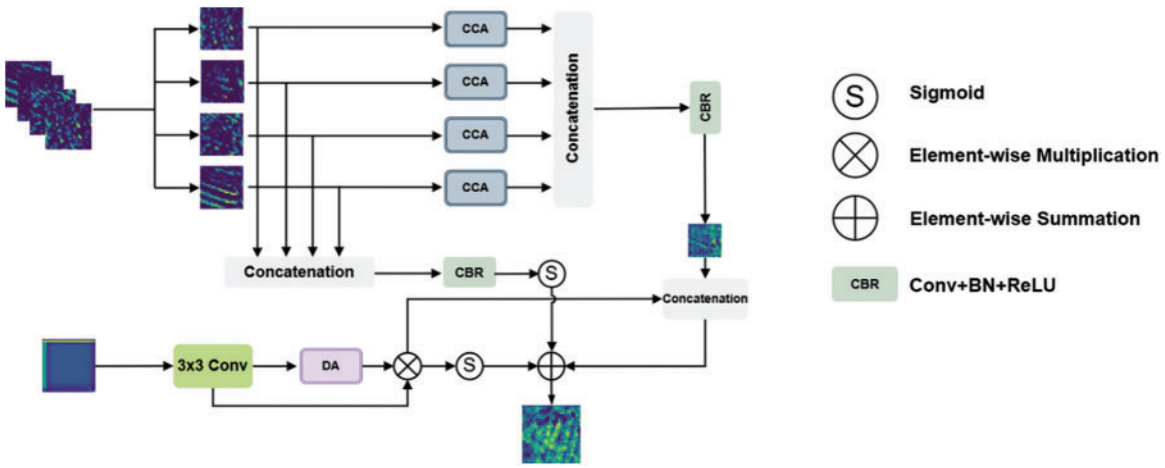

Figure 6: CCAM model diagram

The core of the CCAM module lies in enhancing contextual awareness through attention mechanisms and multi-scale feature extraction. Among these, the CrissCross Attention (CCA) mechanism serves as one of the key components, as illustrated in Fig. 7, and can be represented by the following formula:

$$\begin{aligned} \text{Attention}_H &= V_H \cdot \text{softmax}\left(Q_H K_H^T \oplus \text{INF}\right) \\ \text{Attention}_W &= V_W \cdot \text{softmax}\left(Q_W K_W^T\right) \\ \text{CCA}\left(F_{in}\right) &= \gamma \cdot \left(\text{Attention}_H \oplus \text{Attention}_W\right) \oplus F_{in} \end{aligned} \tag{3}$$

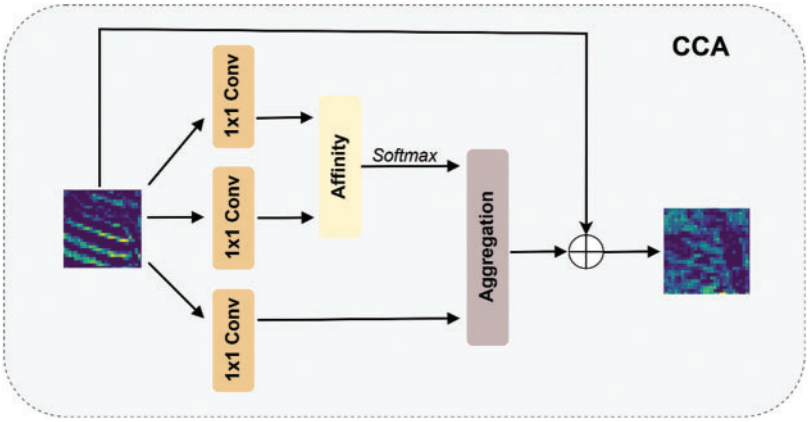

Figure 7: CCA model diagram

Here, γ is a learnable parameter, and INF is a negative infinity mask. Through this mechanism, the model effectively captures long-range dependencies, enhances feature representation capabilities, and excels particularly in multimodal fusion scenarios.

The CCAM module also includes an RGB processing branch and a depth processing branch. In the RGB branch, we apply the CCA mechanism:

$$F_{rgb\_weighted} = F_{rgb} \otimes \text{CCA}\left(F_{rgb}\right) \tag{4}$$

In the depth processing branch, we adopt a lightweight design:

$$F_{depth\_weighted} = Conv\left(F_{depth}\right) \otimes \sigma\left(Conv\left(F_{depth}\right)\right) \tag{5}$$

Here, σ represents the sigmoid function.

Finally, we designed an adaptive fusion strategy that combines RGB and depth features:

$$\text{out} = Concat\left(F_{rgb}, F_{depth}\right) \oplus w_{depth} \cdot F_{depth} \oplus w_{early} \cdot F_{early\_rgb} \tag{6}$$

Here, w_{depth} and w_{early} are learnable weight parameters.

Experimental results indicate that the CCAM module significantly enhances the performance of rail defect detection. Compared to traditional methods, our approach shows a notable improvement in detection accuracy, particularly excelling in the recognition of minor defects. This is largely attributed to the module's multi-scale feature extraction capability and global contextual awareness. The fusion of multimodal information provides a more comprehensive representation of defects, enabling the model to simultaneously detect both surface and deep defects.

3) Temporal Scale Fusion Module:

To fully leverage the unique information provided by RGB images and depth images, as well as the complementary characteristics of their fusion, we specifically designed a dual-modal fusion module (TSFM) for effective rail defect detection. Unlike simple element-wise addition or multiplication-based fusion methods, this module leverages specific multi-scale feature extraction and temporal attention mechanisms to achieve deeper information interaction. The core components of the TSFM include a Multi-Scale Feature Extractor (MSFE) and a Temporal Fusion Attention Module (TFAM). As illustrated in Fig. 8, these components are simply integrated and forwarded to the subsequent decoder.

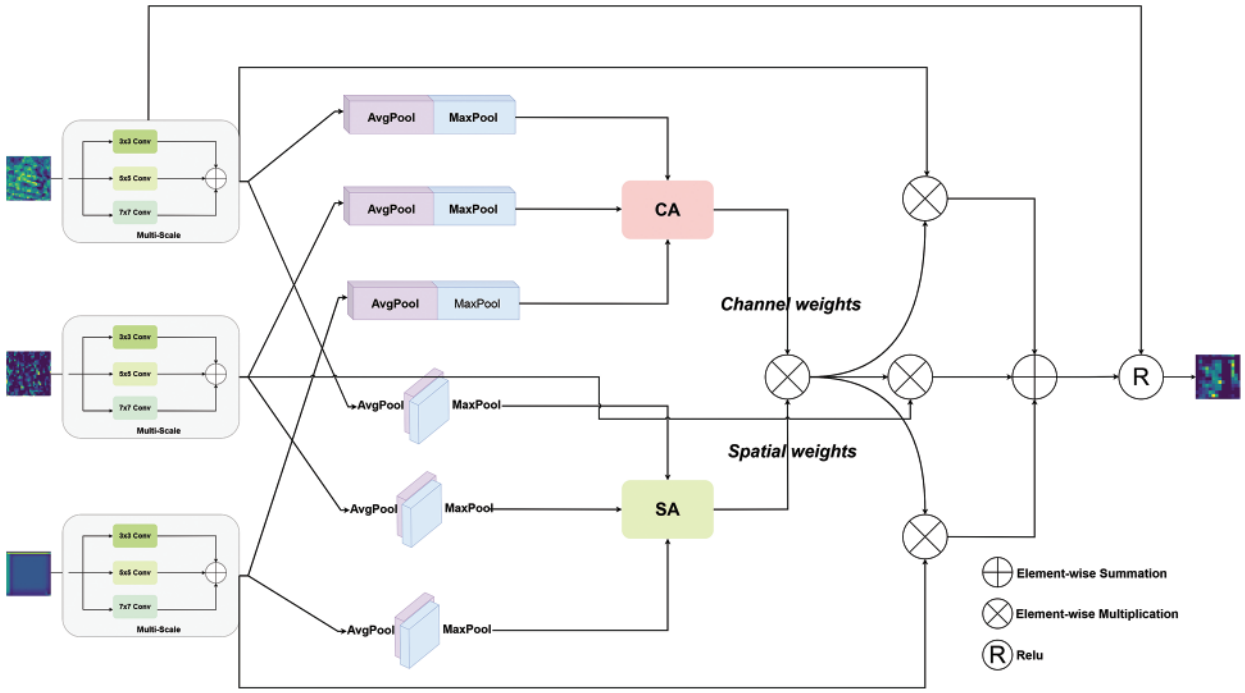

Figure 8: TSFM model diagram

After the preceding calculations, we obtain the specific modal features F^r and F^d from the output of the MSDF module, as well as the cross-modal fusion features F^{ful} from the output of the CCAM module. In the next step, these three features are integrated into the decoder.

The Multi-Scale Feature Extractor (MSFE) employs a parallel multi-scale convolutional structure, incorporating convolutional kernels of sizes 3 × 3, 5 × 5, and 7 × 7. This design, tailored to the multi-scale characteristics of rail defects, not only extracts features through different receptive fields but also avoids information redundancy by aggregating features through addition, ensuring efficient and fine-grained feature fusion. Compared to traditional methods, MSFE significantly enhances the model's sensitivity to defects of varying scales.

$$\begin{aligned} F_1 &= Conv_{3\times3}\left(F_{in}\right) \\ F_2 &= Conv_{5\times5}\left(F_{in}\right) \\ F_3 &= Conv_{7\times7}\left(F_{in}\right) \\ F_{out} &= \mathrm{ReLU}\left(F_1 \oplus F_2 \oplus F_3\right) \end{aligned} \tag{7}$$

The TFAM module innovatively integrates channel attention and spatial attention mechanisms to optimize feature representation, enabling the efficient integration of RGB images, depth images, and fused image features. Adaptive 1D convolutional kernels are employed in channel attention, with their sizes set according to the input channel count, thereby improving the model's computational efficiency and adaptability. The spatial attention, on the other hand, captures local contextual information through a 7×7 convolution operation. This dual attention mechanism enables selective enhancement along both the feature channel and spatial dimensions, significantly improving the effectiveness of feature extraction.

Additionally, TFAM is capable of simultaneously processing RGB images, depth images, and the feature maps generated from their fusion. By calculating and integrating the attention weights of feature maps from different sources, TFAM effectively leverages the complementary information among these features, thereby enhancing the model's ability to represent rail defect characteristics. The specific formulas are as follows:

$$\begin{aligned} P_{avg} &= Concat\left(AvgPool\left(F_{out}^{r}\right), AvgPool\left(F_{out}^{d}\right), AvgPool\left(F_{out}^{ful}\right)\right) \\ P_{max} &= Concat\left(MaxPool\left(F_{out}^{r}\right), MaxPool\left(F_{out}^{d}\right), MaxPool\left(F_{out}^{ful}\right)\right) \\ A_c &= \sigma\left(Conv1D\left(\left[P_{avg}, P_{max}\right]\right)\right) \\ \sigma &= \text{Softmax along temporal dimension} \end{aligned} \tag{8}$$

The formula presented demonstrates the channel attention mechanism, where $F_{out}^{r}, F_{out}^{d}, F_{out}^{ful}$ are the input feature maps from three different categories, and $Conv1D$ indicates the 1D convolution process, with the kernel size determined by the quantity of input channels. The spatial attention mechanism operates in a similar manner.

After strengthening the features using channel and spatial attention mechanisms, an element-wise multiplication is performed with the multi-scale convolutional features. This allows the model to achieve fine-grained feature fusion while preserving spatial and channel information. Following the feature fusion, batch normalization is applied to standardize the feature distribution, improving the model's convergence speed and stability. By introducing non-linearity, the subsequent ReLU activation function strengthens the model's ability to express complex patterns. To prevent overfitting and improve generalization, we conclude with a Dropout layer set at a 0.5 dropout rate, the final fused feature map achieves comprehensive multimodal information integration.

$$\begin{aligned} A_c &= \text{ChannelAttention}\left(F_{out}^{ful}, F_{out}^{r}, F_{out}^{d}\right) \\ A_s &= \text{SpatialAttention}\left(F_{out}^{ful}, F_{out}^{r}, F_{out}^{d}\right) \\ F_{fuse} &= A_c \otimes A_s \otimes \left[F_{out}^{ful}, F_{out}^{r}, F_{out}^{d}\right] \\ F'_{fuse} &= \text{ReLU}\left(F_{fuse} \oplus F_{out}^{ful}\right) \\ F_{out} &= Dropout\left(\text{ReLU}\left(BN\left(F'_{fuse}\right)\right)\right) \end{aligned} \tag{9}$$

A hybrid loss function was employed as a supervisory mechanism in the network training process. This hybrid loss function is composed of binary cross-entropy (BCE) loss and Intersection over Union (IoU) loss. The BCE component drives feature prediction for RGB and depth maps, while the IoU component delivers stage-wise supervision in the final decoder, enhancing the network's ability to capture essential features.

$$\begin{aligned} R_d^5 &= d\left(F_{funsion5}^{\ r}\right) \\ D_d^5 &= d\left(F_{funsion5}^{\ d}\right) \\ \mathcal{L}_{total} &= \mathcal{L}_R + \mathcal{L}_D + \mathcal{L}_{FD} \end{aligned}$$

$$\mathcal{L}_R = \ell_{bce}\left(R_d^5, G\right) + \ell_{iou}\left(R_d^5, G\right)$$
$$\mathcal{L}_D = \ell_{bce}\left(D_d^5, G\right) + \ell_{iou}\left(D_d^5, G\right)$$
$$\mathcal{L}_{FD} = \sum_{m=1}^{5}\left(\ell_{bce}\left(F_{out}, G\right) + \ell_{iou}\left(F_{out}, G\right)\right) \tag{10}$$

The final-layer single-modal features, En_5^r and En_5^d, are processed by the MSDF module, producing output features $F_{funsion5}^{\ r}$ and $F_{funsion5}^{\ d}$, which are directly fed into the decoder to obtain R_d^5 and D_d^5. Meanwhile, F_{out} represents the multi-modal feature obtained through the TSFM module. Where $\ell_{bce}(\cdot)$ represents the binary cross-entropy loss function, $\ell_{iou}(\cdot)$ represents the intersection over union loss function, and GT represents the ground truth.

4 Experimental Results

We trained and tested the model using the NEURSDDS-AUG dataset [20], implementing it with the PyTorch library. The training was conducted on an NVIDIA 3080 GPU with 40 GB of memory. The backbone network is ResNet-34. Due to the differing number of channels in RGB and depth images, the depth images were adjusted to three channels before being input into the backbone network. The model was optimized with the Adam optimizer, beginning with a learning rate of 10^{-4} that was reduced by ten every 60 epochs. Input resolution for RGB and depth images was 256 × 256, with data augmentation techniques like flipping, rotation, and cropping used for better diversity and generalization. Training spanned 200 epochs with a batch size of 6, lasting about 16 h.

We compared MSCM-Net with 12 state-of-the-art methods, including S2AM, BBS, CoNet, HAI, CLA, DRER [40], FHE [41] and CAVER [42], which employ deep learning approaches, and ACSD, CDCP, DCMC, and DF, which are based on handcrafted features. Six commonly used evaluation metrics were employed to assess the performance of the proposed framework. These metrics include Structure measure (Sm), Maximum enhanced-alignment measure (maxE), Mean Absolute Error (MAE), Maximum F-measure (maxF), Precision-Recall (P-R) curve, and the F-measure curve varying with thresholds. The experimental results demonstrate that MSCM-Net achieves significant improvements across all these six metrics.

Quantitative outcomes for the four evaluation metrics of the test set are listed in Table 1. The proposed MSCM-Net consistently outperforms all comparison methods based on these metrics.

Table 1: Quantitative comparison and evaluation of RGB-D SOD algorithms

Models	*Sm*↑	*maxE* ↑	*maxF*↑	*MAE*↓
DCMC	0.484	0.595	0.498	0.287
ACSD	0.556	0.670	0.575	0.360
DF	0.564	0.713	0.636	0.241
CDCP	0.574	0.694	0.591	0.236
HAI	0.718	0.829	0.803	0.171
S2MA	0.775	0.864	0.817	0.141
CONET	0.786	0.878	0.834	0.101
BBS	0.828	0.909	0.867	0.073
CLANet	0.835	0.920	0.878	0.069
FHENet	0.836	0.926	0.881	0.064
DRERNet	0.844	0.933	0.891	0.059

(Continued)

Table 1 (continued)

Models	*Sm*↑	*maxE* ↑	*maxF*↑	*MAE*↓
CAVER	0.838	0.918	0.884	0.069
Ours	**0.849**	**0.934**	**0.893**	**0.057**

Fig. 9 provides a visual comparison of the detection results for railway surface defects.

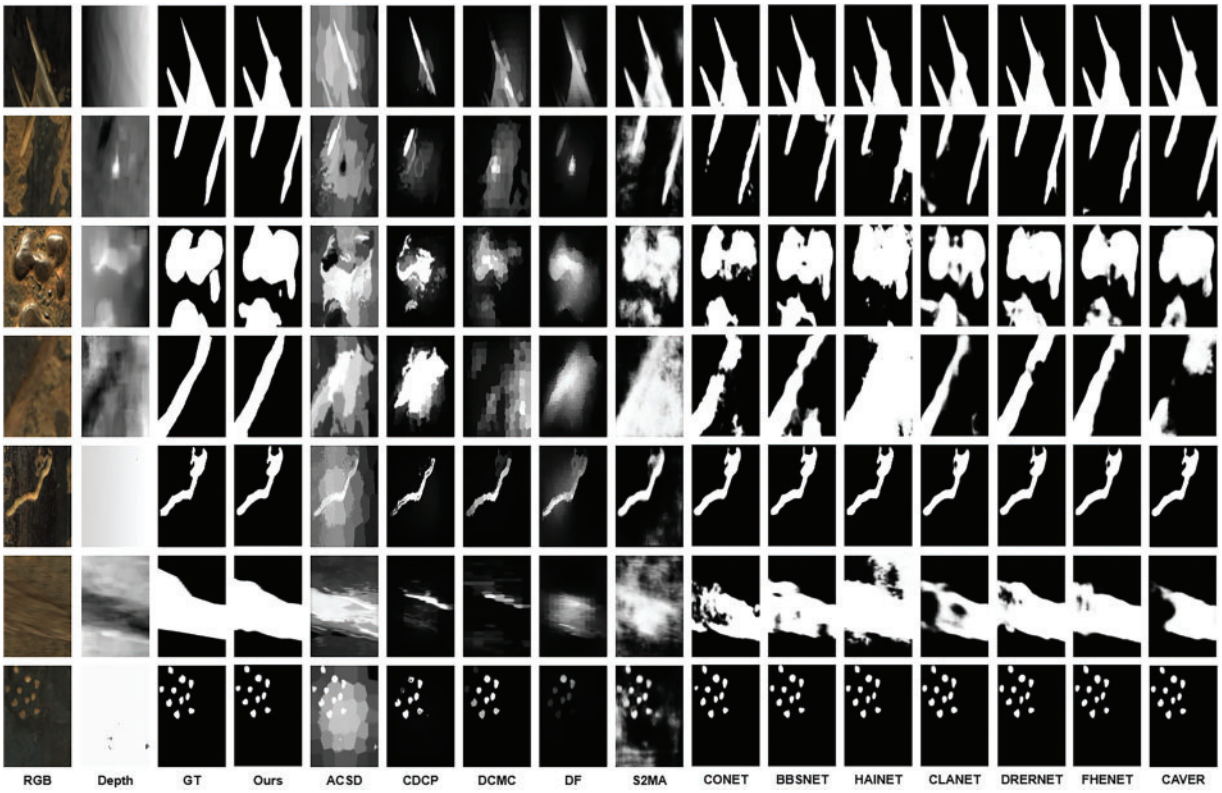

Figure 9: Comparison of multi-model experimental results

Recall is plotted on the x-axis and precision on the y-axis in the PR curve. Precision is defined by the formula TP/(TP+FP), and recall is calculated as TP/(TP+FN), where TP, FP, and FN represent true positives, false positives, and false negatives, respectively. Fig. 10 illustrates both the PR curve and the threshold F-measure curve, highlighting comparisons with alternative methods.

As shown in Fig. 9, comparative analysis of the prediction results of MSCM-Net and several typical methods reveals significant shortcomings in handling complex defect boundaries, particularly with irregular shapes, where details are often missed. Additionally, many methods struggle to accurately locate defect areas when the defects are similar to background textures, leading to decreased detection performance. In contrast, MSCM-Net excels in surface defect detection, especially on unserviced rail tracks. Thanks to the introduction of the Multi-Scale Feature Aggregation module (MSDF), Cross-Channel Attention module (CCAM), and Two-Stream Feature Fusion module (TSFM), MSCM-Net can precisely refine complex defect boundaries while effectively distinguishing similar textures, ensuring accurate defect location identification.

Furthermore, the network effectively reduces noise during the fusion and prediction processes, enhancing the clarity and stability of the detection results.

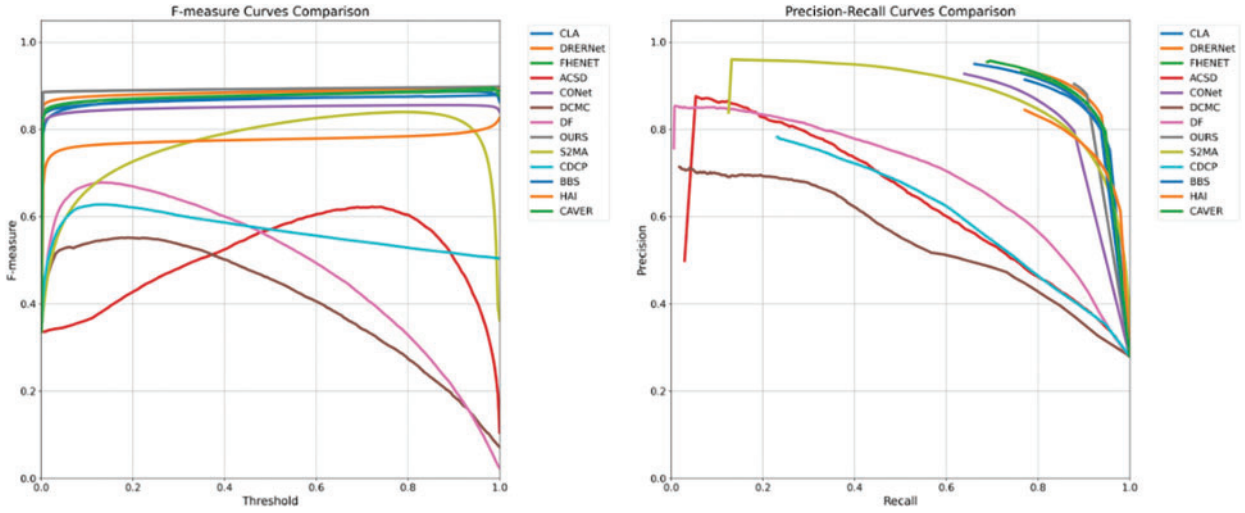

Figure 10: PR curve graph and threshold F-measure curve graph

4.1 Performance on Other Datasets

To validate the effectiveness of MSCM-Net, we evaluated the proposed network on three public RGB-D SOD datasets to demonstrate its generalization ability. We selected 1485 image pairs from NJU2K, 700 image pairs from NLPR as the training set, and the remaining image pairs, which are from STERE, NJU2K, and NLPR, were used for testing.

NJU2K [43]: Contains 1985 stereo image pairs, covering a variety of lighting conditions and scene types.

NLPR [44]: Includes 1000 stereo image pairs, encompassing both indoor and outdoor environments.

STERE [23]: The first stereo image SOD dataset consists of 1000 pairs of stereo images from the internet, covering a range of real-world scenes.

As shown in Table 2, MSCM-Net achieved top-three performance in the generalization tests on other publicly available datasets, demonstrating strong competitiveness. We observed that these datasets exhibit differences in image quality, capturing conditions, and defect feature distributions compared to the primary test dataset, which may have influenced the performance. Nevertheless, MSCM-Net maintained stable results, indicating its robust capability. In the future, we plan to enhance the model's generalization ability by incorporating more diverse training data or leveraging transfer learning techniques.

Table 2: Quantitative comparison on three representative large-scale benchmark datasets (↑Higher is Better, ↓Lower is Better)

Models	NJU2K				NLPR				STERE			
	***Sm*↑**	***maxE*↑**	***maxF*↑**	***MAE*↓**	***Sm*↑**	***maxE*↑**	***maxF*↑**	***MAE*↓**	***Sm*↑**	***maxE*↑**	***maxF*↑**	***MAE*↓**
S2MA	0.894	0.930	0.889	0.053	0.915	0.953	0.902	0.030	0.890	0.932	0.882	0.051
BBS	0.921	0.949	0.920	0.035	0.930	0.961	0.918	0.023	0.908	0.942	0.903	0.041
HAI	0.912	0.944	0.915	0.038	0.921	0.960	0.915	0.024	0.907	0.944	0.906	0.040
SPNet	0.924	0.957	0.927	0.029	0.928	0.962	0.918	0.021	0.907	0.949	0.906	0.037

(Continued)

Table 2 (continued)

Models	NJU2K				NLPR				STERE			
	***Sm*↑**	***maxE*↑**	***maxF*↑**	***MAE*↓**	***Sm*↑**	***maxE*↑**	***maxF*↑**	***MAE*↓**	***Sm*↑**	***maxE*↑**	***maxF*↑**	***MAE*↓**
CLANet	0.911	0.923	0.900	0.040	0.932	0.959	0.894	0.021	—	—	—	—
DRER	0.906	0.943	0.907	0.038	0.915	0.953	0.901	0.024	0.895	0.943	0.891	0.042
CAVER	0.926	0.958	0.928	0.030	0.934	0.970	0.928	0.021	0.918	0.955	0.916	0.033
Ours	**0.911**	**0.949**	**0.914**	**0.033**	**0.924**	**0.963**	**0.915**	**0.020**	**0.903**	**0.948**	**0.900**	**0.035**

4.2 Ablation Experiments

We carried out a series of ablation experiments to validate the effectiveness of the core components of MSCM-Net. We used ResNet-34 as the backbone network and progressively added the Multi-Scale Dilated Fusion module (MSDF), the Criss-Cross Attention Integration Module (CCAM), and the Triple Temporal Multi-Scale Fusion Module (TSFM). The experiments were conducted on the NEU RSDDS-AUG dataset. Table 3 presents the performance comparison of different model configurations.

Table 3: Ablation experiments for each module of MSCM-Net

Models	NEU RSDDS-AUG			
	***Sm*↑**	***maxE*↑**	***maxF*↑**	***MAE*↓**
Backbone	0.830	0.923	0.877	0.069
Backbone+MSDF	0.835	0.926	0.880	0.065
Backbone+CCAM	0.839	0.928	0.886	0.063
Backbone+TSFM	0.839	0.928	0.888	0.062
Backbone+MSDF+CCAM+TSFM	0.849	0.934	0.893	0.057

MSDF: The addition of MSDF significantly improved metrics such as Maximum F-measure and Structure Measure (Sm), demonstrating its effectiveness in capturing multi-scale features.

CCAM: Incorporating CCAM resulted in notable improvements in feature alignment and cross-modal information fusion, particularly enhancing the complementarity and alignment accuracy between different modalities.

TSFM: The TSFM module improved structure and alignment measures, playing a crucial role in fusing the three image features and optimizing the information flow, further boosting the overall performance of MSCM-Net.

The ablation study results demonstrate that the proposed Multi-Scale Dilated Fusion module (MSDF), CrissCross Attention Augmented module (CCAM), and Triple Temporal Multi-Scale Fusion module (TSFM) all contribute positively to the overall performance of MSCM-Net. The significant improvements across various evaluation metrics validate the rationality and necessity of these components' design. Nonetheless, future research could explore the interactions between these modules and their performance on different datasets.

5 Conclusion

To fulfill the need for surface defect detection in unused rails, this paper proposes a detection method utilizing deep learning and RGB-D fusion. By employing a multi-scale atrous convolution module, the

method effectively merges information gathered from multiple receptive fields, significantly enhancing the model's performance in extracting complex textures and edge details. Additionally, the multi-modal feature fusion module exploits the complementarity between RGB and depth information, enabling comprehensive extraction of fine-grained details and geometric structures. The improvement boosts the model's performance in detecting defects, especially in conditions involving blurry boundaries and similar texture patterns. According to experimental outcomes, the proposed technique exceeds the performance of traditional unimodal and several multimodal approaches in terms of accuracy and robustness, which effectively enhances the model's generalization and stability. Future research can further optimize the network structure to achieve model lightweighting and improved real-time performance, thereby increasing its applicability in practical industrial inspection scenarios. Moreover, this method shows good scalability in the field of surface defect detection, with the potential to be applied to more complex surface inspection tasks, providing broader technical support for intelligent detection systems.

Acknowledgement: The authors would like to express their gratitude to the members of the research group for their support.

Funding Statement: This research was funded by the National Natural Science Foundation of China (grant number 62306186), the Technology Plan Joint Foundation of Liaoning Province (grant number 2023-MSLH-246), and the Technology Plan Joint Foundation of Liaoning Province (grant number 2023-BSBA-238).

Author Contributions: Study conception and design: Xin Wen, Xiao Zheng, Yu He; data collection, analysis and interpretation of results: Xin Wen, Xiao Zheng, Yu He; draft manuscript preparation: Xin Wen, Xiao Zheng, Yu He. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The experimental data supporting the study's conclusions may be obtained from the corresponding author upon request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Gibert X, Patel VM, Chellappa R. Deep multi-task learning for railway track inspection. arXiv:1509.05267. 2015 Sep 17.
2. Zhang D, Song K, Xu J, He Y, Niu M, Yan Y. MCnet: multiple context information segmentation network of no-service rail surface defects. IEEE Trans Instrum Meas. 2021;70:1–9. doi:10.1109/TIM.2021.3127641.
3. Zhou X, Fang H, Liu Z, Zheng B, Sun Y, Zhang J, et al. Dense Attention-guided cascaded network for salient object detection of strip steel surface defects. IEEE Trans Instrum Meas. 2022;71:1–14. doi:10.1109/TIM.2021.3132082.
4. Ni X, Ma Z, Liu J, Shi B, Liu H. Attention network for rail surface defect detection via consistency of intersection-over-union (IoU)-guided center-point estimation. IEEE Trans Ind Inform. 2022 Mar;18(3):1694–1705. doi:10.1109/TII.2021.3085848.
5. Mariani S, Nguyen T, Zhu X, Lanza Di, Scalea F. Field test performance of noncontact ultrasonic rail inspection system. J Transp Eng, Part A: Systems. 2017 May;143(5):04017007. doi:10.1061/JTEPBS.0000026.
6. Sabato A, Niezrecki C. Feasibility of digital image correlation for railroad tie inspection and ballast support assessment. Measurement. 2017 Jun;103(3):93–105. doi:10.1016/j.measurement.2017.02.024.
7. Ramatlo DA, Wilke DN, Loveday PW. Development of an optimal piezoelectric transducer to excite guided waves in a rail web. NDT E Int. 2018 Apr;95:72–81. doi:10.1016/j.ndteint.2018.02.002.
8. Kwon S-G, Lee T-G, Park S-J, Park J-W, Seo J-M. Natural rail surface defect inspection and analysis using 16-channel eddy current system. Appl Sci. 2021 Aug;11(17):8107. doi:10.3390/app11178107.

9. Alvarenga TA, Carvalho AL, Honorio LM, Cerqueira AS, Filho LMA, Nobrega RA. Detection and classification system for rail surface defects based on eddy current. Sensors. 2021 Nov;21(23):7937. doi:10.3390/s21237937.
10. Jia Y, Lu Y, Xiong L, Zhang Y, Wang P, Zhou H. A filtering method for suppressing the lift-off interference in magnetic flux leakage detection of rail head surface defect. Appl Sci. 2022 Feb;12(3):1740. doi:10.3390/app12031740.
11. Long Y, Huang S, Peng L, Wang W, Wang S, Zhao W. Internal and external defects discrimination of pipelines using composite magnetic flux leakage detection. In: 2021 IEEE International Instrumentation and Measurement Technology Conference (I2MTC); 2021 May; Glasgow, UK: IEEE; p. 1–6.
12. Faghih-Roohi S, Hajizadeh S, Nunez A, Babuska R, De Schutter B. Deep convolutional neural networks for detection of rail surface defects. In: 2016 International Joint Conference on Neural Networks (IJCNN); 2016 Jul; Vancouver, BC, Canada: IEEE.
13. Mercy KG, Srinivasa Rao SK. A framework for rail surface defect prediction using machine learning algorithms. In: 2018 International Conference on Inventive Research in Computing Applications (ICIRCA); 2018 Jul; Coimbatore: IEEE; p. 972–7.
14. Yu H, Li Q, Tan Y, Gan J, Wang J, Geng YA, et al. A coarse-to-fine model for rail surface defect detection. IEEE Trans Instrum Meas. 2019 Mar;68(3):656–66. doi:10.1109/TIM.2018.2853958.
15. Dong H, Song K, He Y, Xu J, Yan Y, Meng Q. PGA-Net: pyramid feature fusion and global context attention network for automated surface defect detection. IEEE Trans Ind Inf. 2020 Dec;16(12):7448–58. doi:10.1109/TII.2019.2958826.
16. He Y, Song K, Meng Q, Yan Y. An end-to-end steel surface defect detection approach via fusing multiple hierarchical features. IEEE Trans Instrum Meas. 2020 Apr;69(4):1493–504. doi:10.1109/TIM.2019.2915404.
17. Qin W, Lu T, Zhang L, Peng S, Wan D. Multi-branch deepfake detection algorithm based on fine-grained features. Comput Mater Contin. 2023;77(1):467–90. doi:10.32604/cmc.2023.042417.
18. Cao S, Liang J, Cao Y, Huang J, Yang Z. PSMFNet: lightweight partial separation and multiscale fusion network for image super-resolution. Comput Mater Contin. 2024;81(1):1491–509. doi:10.32604/cmc.2024.049314.
19. Song K, Feng H, Cao T, Cui W, Yan Y. MFANet: multifeature aggregation network for cross-granularity few-shot seamless steel tubes surface defect segmentation. IEEE Trans Ind Inf. 2024 Jul;20(7):9725–35. doi:10.1109/TII.2024.3383513.
20. Wang J, Song K-C, Zhang D, Niu M, Yan Y. Collaborative learning attention network based on RGB image and depth image for surface defect inspection of no-service rail. IEEE/ASME Trans Mechatronics. 2022 Dec;27(6):4874–84. doi:10.1109/TMECH.2022.3167412.
21. Abro GEM, Ali ZA, Rajput S. Innovations in 3D object detection: a comprehensive review of methods, sensor fusion, and future directions. IECE Trans Sens Commun Control. 2024 Oct;1(1):3–29. doi:10.62762/TSCC.
22. Qi P, Chai L, Ye X. Unsupervised industrial anomaly detection based on feature mask generation and reverse distillation. Chin J Inf Fusion. 2024 Sep;1(2):160–74. doi:10.62762/CJIF.
23. Niu Y, Geng Y, Li X, Liu F. Leveraging stereopsis for saliency analysis. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition; 2012 Jun; Providence, RI: IEEE; p. 454–61.
24. Lang C, Nguyen TV, Katti H, Yadati K, Kankanhalli M, Yan S. Depth matters: influence of depth cues on visual saliency. In: Fitzgibbon A, Lazebnik S, Perona P, Sato Y, Schmid C, editors. Computer vision–ECCV 2012. Berlin/Heidelberg: Springer Berlin Heidelberg; 2012. vol. 7573. p. 101–15.
25. Desingh K, Madhava Krishna K, Rajan D, Jawahar C. Depth really matters: improving visual salient region detection with depth. In: Procedings of the British Machine Vision Conference 2013; 2013; Bristol: British Machine Vision Association; p. 98.1–11.
26. Cheng Y, Fu H, Wei X, Xiao J, Cao X. Depth enhanced saliency detection method. In: Proceedings of International Conference on Internet Multimedia Computing and Service; 2014 Jul; Xiamen, China: ACM; p. 23–7.
27. Feng D, Barnes N, You S, McCarthy C. Local background enclosure for RGB-D salient object detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun; Las Vegas, NV, USA: IEEE; p. 2343–50.
28. Qu L, He S, Zhang J, Tian J, Tang Y, Yang Q. RGBD salient object detection via deep fusion. IEEE Trans Image Process. 2017 May;26(5):2274–85. doi:10.1109/TIP.2017.2682981.

29. Chen H, Li Y. Progressively complementarity-aware fusion network for RGB-D salient object detection. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun; Salt Lake City, UT: IEEE; p. 3051–60.
30. Piao Y, Ji W, Li J, Zhang M, Lu H. Depth-induced multi-scale recurrent attention network for saliency detection. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV); 2019 Oct; Seoul, Republic of Korea: IEEE; p. 7253–62.
31. Liu N, Zhang N, Han J. Learning selective self-mutual attention for RGB-D saliency detection. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun; Seattle, WA, USA: IEEE; p. 13753–62.
32. Ji W, Li J, Zhang M, Piao Y, Lu H. Accurate RGB-D salient object detection via collaborative learning. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020; Springer International Publishing; 2020. p. 52–69.
33. Fan D-P, Lin Z, Zhang Z, Zhu M, Cheng M-M. Rethinking RGB-D salient object detection: models, data sets, and large-scale benchmarks. IEEE Trans Neural Netw Learning Syst. 2021 May;32(5):2075–89. doi:10.1109/TNNLS.2020.2996406.
34. Zhang J, Fan DP, Dai Y, Anwar S, Saleh FS, Zhang T, et al. UC-Net: uncertainty inspired RGB-D saliency detection via conditional variational autoencoders. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun; Seattle, WA, USA: IEEE; p. 8579–88.
35. Li G, Liu Z, Lin W, Ling H. Hierarchical alternate interaction network for RGB-D salient object detection. IEEE Trans Image Process. 2021;30:3528–42. doi:10.1109/TIP.2021.3062689.
36. Fan D-P, Zhai Y, Borji A, Yang J, Shao L. BBS-Net: RGB-D salient object detection with a bifurcated backbone strategy network. In: European Conference on Computer Vision; 2020; Cham: Springer International Publishing; p. 275–92.
37. Niu J, Hu Q, Niu Y, Zhang T, Kumar Jha S. Real-time recognition and location of indoor objects. Comput Mater Contin. 2021;68(2):2221–9. doi:10.32604/cmc.2021.017073.
38. Song M, Song W, Yang G, Chen C. Improving RGB-D salient object detection via modality-aware decoder. IEEE Trans Image Process. 2022;31:6124–38. doi:10.1109/TIP.2022.3205747.
39. Yan Y, Jia X, Song K, Cui W, Zhao Y, Liu C, et al. Specificity autocorrelation integration network for surface defect detection of no-service rail. Opt Lasers Eng. 2024 Jan;172(8):107862. doi:10.1016/j.optlaseng.2023.107862.
40. Wu J, Zhou W, Qiu W, Yu L. Depth repeated-enhancement RGB network for rail surface defect inspection. IEEE Signal Process Lett. 2022;29:2053–7. doi:10.1109/LSP.2022.3211199.
41. Zhou W, Hong J. FHENet: lightweight feature hierarchical exploration network for real-time rail surface defect inspection in RGB-D images. IEEE Trans Instrum Meas. 2023;72:1–8. doi:10.1109/TIM.2023.3237830.
42. Pang Y, Zhao X, Zhang L, Lu H. CAVER: cross-modal view-mixed transformer for bi-modal salient object detection. IEEE Trans Image Process. 2023;32:892–904. doi:10.1109/TIP.2023.3234702.
43. Ju R, Ge L, Geng W, Ren T, Wu G. Depth saliency based on anisotropic center-surround difference. In: 2014 IEEE International Conference on Image Processing (ICIP); 2014 Oct; Paris, France: IEEE; p. 1115–9.
44. Peng H, Li B, Xiong W, Hu W, Ji R. RGBD salient object detection: a benchmark and algorithms. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T, editors. Computer vision–ECCV 2014. Cham: Springer International Publishing; 2014. p. 92–109.