Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.062949

ARTICLE

Pyramid–MixNet: Integrate Attention into Encoder-Decoder Transformer Framework for Automatic Railway Surface Damage Segmentation

Hui Luo, Wenqing Li* and Wei Zeng

School of Information and Software Engineering, East China Jiaotong University, Nanchang, 330013, China

*Corresponding Author: Wenqing Li. Email: 2022068081000004@ecjtu.edu.cn

Received: 31 December 2024; Accepted: 18 April 2025; Published: 09 June 2025

ABSTRACT: Rail surface damage is a critical component of high-speed railway infrastructure, directly affecting train operational stability and safety. Existing methods face limitations in accuracy and speed for small-sample, multi-category, and multi-scale target segmentation tasks. To address these challenges, this paper proposes Pyramid-MixNet, an intelligent segmentation model for high-speed rail surface damage, leveraging dataset construction and expansion alongside a feature pyramid-based encoder-decoder network with multi-attention mechanisms. The encoding network integrates Spatial Reduction Masked Multi-Head Attention (SRMMHA) to enhance global feature extraction while reducing trainable parameters. The decoding network incorporates Mix-Attention (MA), enabling multi-scale structural understanding and cross-scale token group correlation learning. Experimental results demonstrate that the proposed method achieves 62.17% average segmentation accuracy, 80.28% Damage Dice Coefficient, and 56.83 FPS, meeting real-time detection requirements. The model's high accuracy and scene adaptability significantly improve the detection of small-scale and complex multi-scale rail damage, offering practical value for real-time monitoring in high-speed railway maintenance systems.

KEYWORDS: Pyramid vision transformer; encoder–decoder architecture; railway damage segmentation; masked multi-head attention; mix-attention

1 Introduction

The rapid expansion of global high-speed rail networks has created an urgent need for advanced damage detection systems to maintain operational safety and infrastructure integrity. While rail surfaces inevitably develop critical defects like cracks, pitting, and corrugations due to extreme operational stresses [1], current manual inspection methods remain labor-intensive, subjective, and incapable of real-time monitoring. This creates significant safety risks, as undetected damage may lead to catastrophic failures. The technical challenges are substantial: complex operating environments with variable lighting and weather conditions interfere with detection, while the diverse morphology of defects-ranging from microscopic cracks to extensive spalling—complicates automated segmentation. Furthermore, the scarcity of annotated training data for rare damage types and the inherent trade-off between model complexity and generalization performance continue to hinder practical solutions. These limitations highlight the urgent need for an intelligent inspection system that combines computational efficiency with robust defect recognition capabilities.

While digital image processing has advanced rail surface inspection, traditional methods remain limited. Classical computer vision techniques, though effective for plain-background railways, often require specialized feature engineering for broader infrastructure applications. Yuan et al. [2] used an improved

Ostu algorithm for damage segmentation, but it lacks generality due to scene-specific threshold adjustments. Cao et al. [3] combined laser sensors with an enhanced dynamic detection algorithm and digital railway surface alignment, suitable only for online inspections. Kundu et al. [4] deployed acoustic emission sensors and wavelet transform for damage localization, yet faced real-time performance issues, environmental interference, and blind spots. These methods are limited by their reliance on handcrafted features and their inability to adapt to the complex and dynamic environments of high-speed railways, resulting in poor generalization and high false positive rates.

Recent years have witnessed deep learning become the predominant approach for rail defect identification, owing to its superior feature representation and end-to-end learning capabilities. Notable innovations include: Guo et al. [5] enhanced Deeplabv3+ optimizing the accuracy-efficiency trade-off for surface defects, and Wu et al. [6] proposed RBGNet leveraging rail boundary-rail surface complementarity for improved detection robustness. Zhang et al. [7] demonstrated residual CNNs' effectiveness in vibration-based defect characterization, Si et al. [8] put up the Rail-STrans addresses small-defect segmentation through transformer-enhanced feature learning. While deep learning-based methods have made remarkable progress in railway damage detection and segmentation, they still face limitations including heavy reliance on large annotated datasets, high computational complexity, and insufficient generalization across diverse real-world scenarios.

To address the difficulty of segmenting discrete damage on high-speed railway surfaces, we present a unique approach that combines deep learning Transformer and semantic segmentation networks with a variety of attention strategies. This comprehensive technology not only offers immediate damage location and segmentation, enabling maintenance professionals to analyze and intervene more rapidly, but also when applied to rail surface damage segmentation, it improves detection speed and accuracy, reduces operational costs, and boosts the level of automation in the inspection process. The paper's primary contributions are:

1. A new end-to-end, encoder-decoder and effective convolutional network for railway rail damage segmentation is proposed, which achieves the precise location and detailed description of the important parts in rail damage.
2. The encoding network includes a progressive shrinking pyramid, a Spatial-Reduction Attention (SRA) and a Masked Multi-Head Attention (MMHA). These capabilities enable the fusion of features across diverse scales, effectively consolidating global information within the input sequence, which not only diminishes redundant computations but also enhances the model's flexibility and scalability.
3. The decoding network includes Mix-Attention (MA), which is able to capture the correlation between different scales simultaneously, which improves the computational speed and generalisation ability of the model.
4. By adopting transversal connections between encoder and decoder levels as feature queries for the attention module, we deviate from conventional jump connection methodologies, elegantly harmonizing the integration of high-level semantic information with low-level structural details and refining the model's capacity to cohesively understand and represent complex data patterns.
5. By using a hybrid loss model that combines the Dice and Focal loss functions, the significant issue of data imbalance is resolved, and segmentation accuracy and speed are increased.

2 Related Work

Semantic segmentation has evolved through distinct methodological phases. The breakthrough of Fully Convolutional Networks (FCNs) [9] established an end-to-end paradigm for pixel-wise prediction, replacing traditional patch-based approaches. Subsequent innovations like U-Net [10] and pyramid pooling modules [11] addressed scale variability but remained constrained by local receptive fields. The DeepLab

series [12,13] mitigated this via dilated convolutions, yet inherent limitations persisted in modeling long-range dependencies—a critical requirement for complex scenes such as railway infrastructure. While CNNs excel at hierarchical feature extraction, their inductive biases (e.g., translation equivariance) may hinder adaptability to irregular structures or occlusions common in real-world environments.

Originally developed for machine translation, Transformers have demonstrated remarkable success in computer vision through Vision Transformers (ViTs). By leveraging self-attention mechanisms, ViTs excel at modeling long-range dependencies and handling variable input sizes, outperforming traditional CNNs in various visual tasks. Liu et al. [14] introduced a Bridge Transformer (BrT) for 3D object detection, enhancing accuracy across visual and point cloud data. Guo et al. [15] proposed RailFormer, a Transformer-based network with overlapped patch merging and Criss-Cross attention, achieving state-of-the-art mIoU for rail surface defect detection on RSDD datasets. Chen et al. [16] developed RailSegVITNet, a vision transformer-based encoder-decoder model for rail track segmentation, maintains its lightweight architecture while achieving comparable or higher segmentation performance. Despite their remarkable performance in computer vision, Vision Transformers face significant challenges when applied to railway defect detection–a demanding real-time application with constrained resources.

The integration of CNN's local feature extraction with Transformer's global modeling capabilities has emerged as a transformative approach for semantic segmentation. Representative works demonstrate distinct architectural innovations: SETR [17] pioneers pure-Transformer segmentation via sequence-to-sequence prediction, while SegFormer [18] combines hierarchical Transformers with lightweight MLP decoders for efficiency. Pyramid Vision Transformer [19] and Twin-Svt [20] further optimize computational overhead through pyramid structures and spatially separable attention, respectively. Mask2Former [21] advances this trend with unified masked attention for panoptic segmentation. CNN-Transformer hybrid architectures synergize CNN's local feature extraction with Transformer's global contextual modeling, enabling both precise identification of micrometer-scale rail surface cracks and robust handling of complex scenarios such as ballast occlusion.

3 Implementation Details

3.1 Overall Architecture

As shown in Fig. 1, the overall structure of the proposed Pyramid-MixNet is a U-shaped hierarchical network with lateral links between the encoder and decoder. Specifically, given an input image of size $H * W * 3$, we first divide it into $\frac{HW}{4^2}$ patches, each of size $4 * 4 * 3$. Then, we send these flattened patches to a linear projection to obtain embedded patches of size $\frac{HW}{4^2} * C_1$. Afterwards, the embedded patches are passed through the Transformer encoder at layer L_1 along with the positional embedding, and the output is reshaped into a feature map F_1 of size $\frac{H}{4} * \frac{W}{4} * C_1$. Similarly, using the feature maps from the previous stage as inputs, the corresponding feature maps can be obtained at each stage: F_2, F_3, and F_4, which are relative to the input image in steps of 8, 16 and 32 pixels, respectively. The encoder uses an incremental compression technique to manage the feature map scale via the patch embedding layer, which allows for flexible scaling of the feature maps at every stage. For feature reconstruction, the proposed decoder also consists of four stages, each step generating fine features D_{4-i+1} by performing a hybrid attention sequence, where the features of the query x_q^i are equal to the respective lateral encoder feature maps. The features of key and value x_{kv}^i are given by encoder and decoder level mixing. It is worth noting that our decoder reflects the dimensions of the encoder-level output. The decoder features are upsampled using bilinear interpolation to match the

height and width of D_i. Finally, the spliced features are subjected to MLP processing to predict a semantic segmentation map of $\frac{H}{4} * \frac{W}{4} * C_1$.

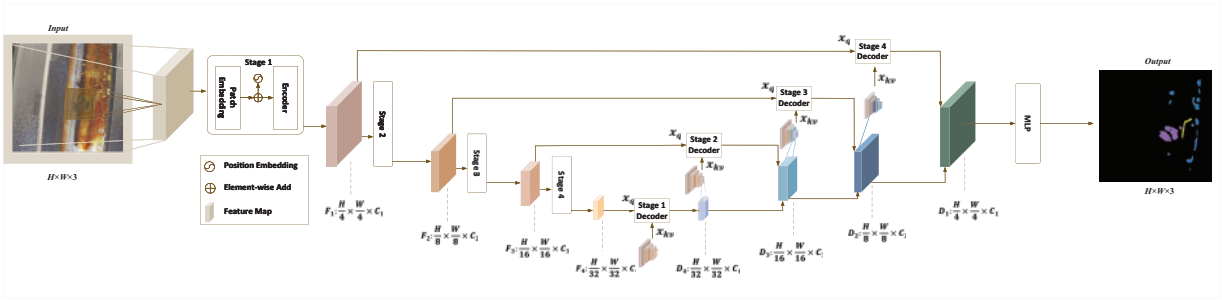

Figure 1: Pyramid-MixNet overall architecture

3.2 Pyramid Transformer Encoder

Pyramid Vision Transformer (PVT) [19] uses an asymptotic shrinkage approach to adjust the scale of the feature maps through patch embedding layers. Here, P_i is used to denote the patch size of stage i, and C_i is the number of channels output from stage i. At the beginning of stage i, the input feature map $F_{i-1} \in \mathbb{R}^{H_{i-1} \times W_{i-1} \times C_{i-1}}$ is first uniformly partitioned into $\frac{H_{i-1}W_{i-1}}{P_i^2}$ blocks, and then each block is flattened and projected into the C_i dimensional embedding. After linear projection, the shape of the embedded block can be viewed as $\frac{H_{i-1}}{P_i} \times \frac{W_{i-1}}{P_i} \times C_i$, where the height and width are P_i times smaller than the input. In this way, as shown in Fig. 2, the scaling of the feature mapping can be flexibly adjusted at each stage so that a feature pyramid can be constructed for the Transformer.

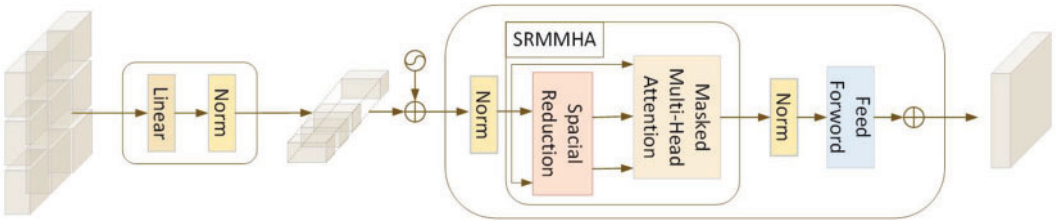

Figure 2: Pyramid-MixNet encoder architecture

3.3 Masked Multi-Head Attention

The Transformer encoder for stage i is comprised of L_i encoder layers, each incorporating both an attention layer and a feedforward layer. Given that the Pyramid Vision Transformer (PVT) necessitates processing high-resolution feature maps, we propose the integration of a Masked Multi-Head Attention (MMHA) layer to supplant the conventional Multi-Head Attention (MHA) layer within the encoder architecture. The network connectivity in the figure below is the Masked Saled Dot-Product Attention layer. SRMMHA takes as input a query Q, a key K and a value V and outputs a fine-grained feature. The difference is that SRMMA reduces the spatial scale of K and V prior to the attention operation. As shown in Fig. 3 below, this significantly reduces the memory overhead. The SRMMHA for each stage is represented as follows.

$$\text{SRMMA}\left(Q,K,V\right)=\text{Concat}\left(A_0,\ldots,A_H\right)W^o, \tag{1}$$

$$A_h=\text{Attention}\left(\text{Q}W_h^Q,\text{SR}\left(K\right)W_h^K,\text{SR}\left(V\right)W_h^V\right), \tag{2}$$

where Concat(·) is the join operation, $h=\{1,2,\ldots,H\}$ is the header index, and d_{model} is the size of each query, key and value, $d_k=d_v=\frac{d_model}{H}$, $A_h\in\mathbb{R}^{d_{model}\times d_k}$, $W_h^Q\in\mathbb{R}^{d_{model}\times d_k}$, $W_h^K\in\mathbb{R}^{d_{model}\times d_k}$, $W_h^V\in\mathbb{R}^{d_{model}\times d_k}$. It can be seen that the outputs of all the heads are concatenated and linearly projected through the weight matrix $W_h^O\in\mathbb{R}^{\text{H}d_v\times d_model}$ obtained from learning to form the final output of the SRMMHA module. Residual concatenation is applied from the inputs to the outputs of the SRMMHA module, followed by frame-by-frame layer normalisation. SR(·) is a dimensionality reduction operation on the spatial dimension of the input sequence (i.e., K or V), which can be written as:

$$\text{SR}\left(x\right)=\text{Concat}\left(\text{Reshape}(x,R_i)W^S\right). \tag{3}$$

$x\in\mathbb{R}^{H_iW_i\times C_i}$ denotes an input sequence, R_i denotes the shrinkage rate of the layer of interest at stage i. $\text{Reshape}(x,R_i)$ denotes the reshaping of the input sequence x into a sequence of size $\frac{H_iW_i}{R_i^2}\times(R_i^2C_i)$. $W^S\in\mathbb{R}^{(R_i^2C_i)\times C_i}$ is an operation that downscales the input sequence down to a linear projection of C_i. Norm(·) is layer normalisation and Attention(·) is denoted as:

$$\text{Attention}\left(Q_h,K_h,V_h\right)=\text{Softmax}\left(M+\frac{\text{Q}_h^{K_{hT}}}{\sqrt{d_k}}\right)V_h, \tag{4}$$

where $Q_h\in\mathbb{R}^{L\times d_k}$, $K_h\in\mathbb{R}^{L\times d_k}$, $V_h\in\mathbb{R}^{L\times d_k}$, $M\in\mathbb{R}^{L,L}$ is used to mask similarities including future frames and ensure causality. Since the latter operation is a Softmax function, masking is performed by adding $-\infty$. After the masking process, each row of the sequence similarity matrix is transformed into a probability distribution through the application of the Softmax activation function, ensuring that the values sum to one and lie within the range of 0 and 1. Ultimately, a novel representation is derived through the computation of the dot product of the normalized similarity matrix with V_h. With these formulations, SRMMHA is less computationally expensive and it is capable of handling larger input feature maps even when resources are limited.

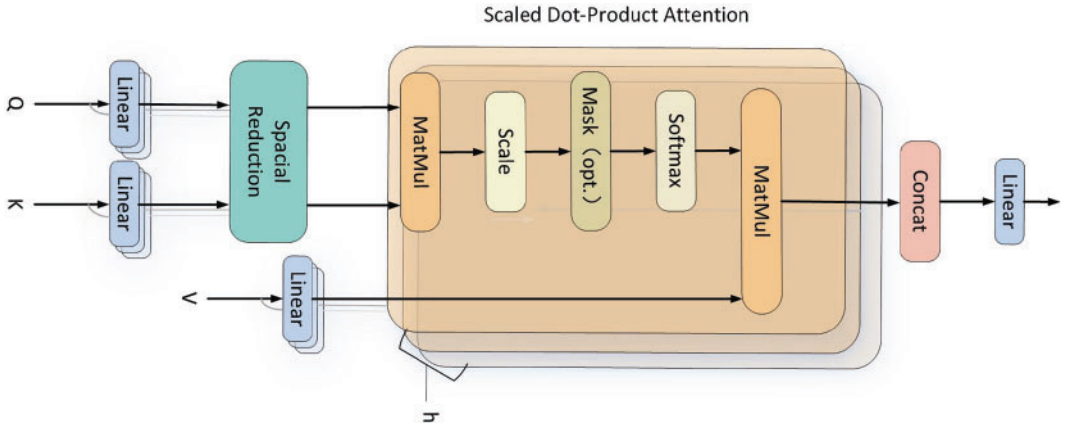

Figure 3: Spacial reduction masked multi-head attention

3.4 Mix-Attention

The traditional cross-attention approach uses two independent sets of features: x_q and x_{kv}, each originating from a different source (See Fig. 4). This technique, while useful in many situations, may neglect the potential benefits of combining information at many sizes or degrees of detail. In contrast, our proposed hybrid attention mechanism offers a more subtle approach, using a mix of x_{kv} features from several multi-scale stages. This method allows queries to seek matches not only at their current scale, but also across a range of contextual granularities. This capacity considerably improves the functional refining process by allowing for a more in-depth knowledge and exploitation of contextual information, enhancing the model's overall performance and flexibility [22].

The selected segmentation of the feature set F_i for decoder stage $i \in \{1, \ldots, N\}$ is of the following form:

$$F_i = \begin{cases} \{E_j\}_{j=1}^{N}, if\ i = 1 \\ \{E_j\}_{j=1}^{N-i+1} \cup \{D_j\}_{j=N-i+2}^{N}, otherwise \end{cases} \tag{5}$$

where in the initial stage of the decoder ($i = 1$), we choose all of the encoder's characteristics. In the subsequent decoder stage, the previously computed decoder output will be propagated by substituting the corresponding horizontal encoder features in F_i. To ensure that the spatial dimensions of the features in F_i are aligned, we use spatial approximation:

$$\hat{F}_j^i = \text{Linear}(C_j, C_j)(\hat{F}_j^i), \forall j \in \{1, ..., N-1\}, \tag{6}$$

where F_j^i represents the jth element in the feature set F_i, and pr_j indicates the pooling ratio corresponding to the alignment size with the minimum feature mapping F_N^i. To establish spatial consistency, these alignment-processed features are joined along the channel dimensions, yielding a new feature vector x_{kv} that contains both key and value information.

$$x_{kv}^i = \text{Concat}(\{\hat{F}_j^i\}_{j=1}^{N-1} \cup \{F_N^i\}). \tag{7}$$

The cross-attention module is replaced with a hybrid attention module using layer normalisation (LN) and feed-forward networks (FFN), as illustrated in Fig. 5, and the output of Decoder-Stage(i) is computed as follows:

$$A_i = \text{LN}(Mix - Attention.\left(LN(X_{kv}^i, X_q^i)\right) + LN(X_q^i)), \tag{8}$$

$$\text{Decoder-Stage(i)} = D_{N-i+1} = FFN(A_i) + A_i. \tag{9}$$

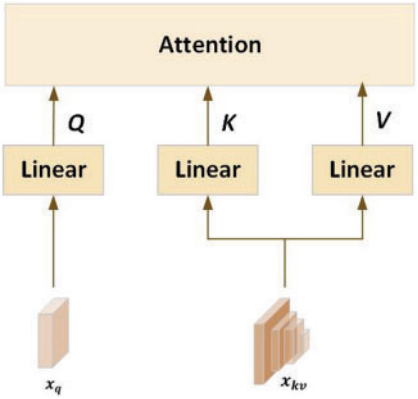

Figure 4: Mix attention

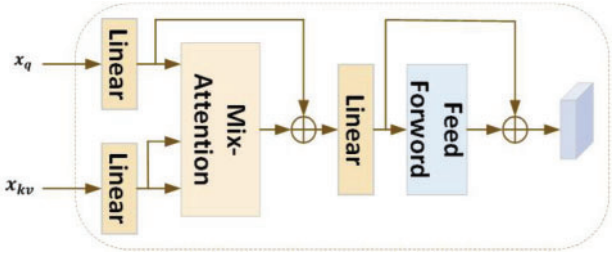

Figure 5: Pyramid-MixNet decoder architecture

3.5 Loss Function and Evaluation Metrics

For enhancing the efficiency of semantic segmentation on the dataset and successfully address the severe data imbalance problem, we use a hybrid loss model that incorporates the Dice and Focal loss functions. The Dice loss function is built around the definitions of precision and recall, and its formula is illustrated in (10). It is worth mentioning that Dice loss, a popular segmentation evaluation metric, and direct optimisation can considerably increase the model's performance. This method not only addresses the imbalance between categories, but also improves the model's capacity to identify samples from a limited number of categories, hence enhancing the overall segmentation result. The accuracy of Pyramid-MixNet model segmentation is maximum when θ takes the value of 1.

$$Dice(Precision, Recall) = 1 - (1 + \theta^2)\frac{Precision * Recall}{\theta^2 * Precision * Recall}. \tag{10}$$

Focal loss establishes a criterion for measuring the categorical focal loss between the true and predicted values, where σ takes the value of 0.25 and τ takes the value of 2.

$$Focal(GroundTruth, Predicted) = -GroundTruth * \sigma * (1 - Predicted)^{\tau} * log(Predicted). \tag{11}$$

The final loss function formulation is as follows:

$$\text{TotalLoss} = DiceLoss + FocalLoss. \tag{12}$$

4 Experiment and Result Analysis

4.1 Data Preparation

With the support of the relevant departments, our team obtained a total of 545 rail surface damage images by using a high-speed rail camera to acquire a specific cross-section of high-speed rail for on-site shooting of visible light images of the damage, and then constructed an initial dataset of rail surface damage images by segmenting the collected images, adjusting them to a fixed size, and then labelling them. From the 500 clear datasets, 400 images were selected and manually labelled at the pixel level, and then these images were subjected to data expansion in the form of flipping, scaling, panning and cropping. In this investigation, 3000 rail surface damage images were gathered. To ensure the effectiveness of model training and the accuracy of evaluation, these images were divided into three sets: training, test, and validation, in proportions of 60%, 20%, and 20%, respectively. The dataset covers only two main categories: defects and backgrounds, which account for 7.3 % and 92.7% of the dataset. In the samples taken from the data, there are several types of defects on the rail surface, such as scratches, dents, abrasions, breaks or surface corrosion, oxidation and rust, but we have united these types of defects into one type of surface damage. All data collection processes strictly adhere to national confidentiality laws and regulations. This dataset involves confidential information, thus its usage has restrictions and will not be shared public. Fig. 6 depicts part of the self-constructed dataset. The pre-processed images were labelled using EISeg software and the labels were divided into two categories: background and damage. Fig. 7 shows one image of the dataset and the image after being labelled.

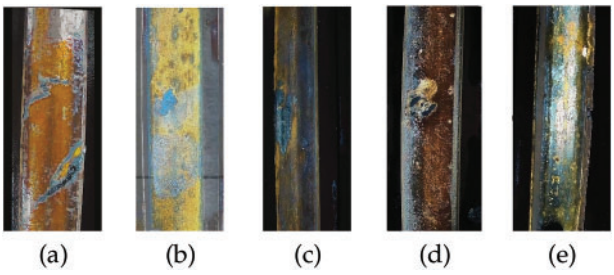

Figure 6: Partial presentation of the dataset created in this paper. (a) Scratches. (b) Paint loss. (c) Scuffs. (d) Holes. (e) Peeling

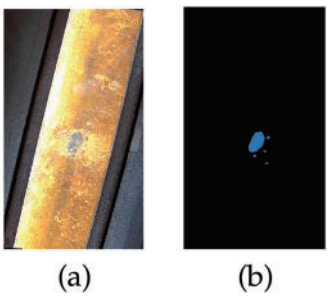

Figure 7: (a) A single image from the dataset. (b) Label image

4.2 Experiment Setup

The methodology described in this paper is based on the PyTorch 1.3 framework, which is developed in Python. All experiments and tests were conducted on the Windows 10 operating system on NVIDIA Tesla

P40 GPUs (with 24 GB of RAM and a single-precision performance of 12 teraflops) and the MMSegmentation platform. Throughout the training stage, the initial rate of learning was set at 0.001, and the model parameters were modified using the Adam optimisation algorithm, with a weight decay coefficient of 0.0005. In addition, a cosine recession approach was employed to adjust the learning rate. The input data was preprocessed uniformly before entering the model by scaling the image to 224 × 224 pixels.

4.3 Result Analysis

4.3.1 Ablation Experiments

In this paper, a complete series of ablation experiments were carried out to rigorously test the efficacy of the proposed algorithm. These trials demonstrated the critical role of essential modules such as MMHA (Masked Multi-Head Attention), Mix-Attention, and SRMMHA (Spatial Reduction Mixed Multi-Head Attention) in improving the algorithm's performance. The comprehensive trial results, thoroughly described in Table 1, provide light on the influence of each module and its synergistic combinations. The table displays a variety of network topologies along with their associated assessment metrics, which include Params Size, mAcc (mean accuracy), mDice (mean Dice coefficient), Background Accuracy, Damage Accuracy, Background Dice Coefficient, and Damage Dice Coefficient. As a baseline, the original Pyramid_unet architecture is characterized by an encoder with a Pyramid Vision Transformer-Medium backbone network that is seamlessly coupled with a U-Net-based code network.

Table 1: Ablation experiments show the effectiveness of various attention mechanisms and structures

Model	Params (M)	mAcc/%	mDice coefficient/%	Background Acc/%	Damage Acc/%	Background dice coefficient/%	Damage dice coefficient/%
U-Net	7.76	50.31	69.25	96.65	66.94	98.72	50.22
Pyramid_unet	51.96	53.37	74.85	97.03	68.41	98.89	61.28
MMHA	78.36	56.48	79.39	97.28	70.81	99.04	65.92
SRMMHA	70.55	58.81	85.19	98.21	76.88	99.25	69.38
MA	84.29	54.32	83.18	97.94	73.59	99.08	70.22
MMHA + MA	87.76	60.93	88.84	99.16	79.27	99.52	75.36
Pyramid-MixNet	80.93	62.17	89.05	99.32	83.41	99.68	80.28

The results of ablation experiments demonstrate the importance of MMHA (Masked Multi-Head Attention), SRMMHA (Spacial Reduction Masked Multi-Head Attention) and MA (Mix-Attention). By replacing the original Pyramid Vision Transformer with SRMMHA and MA in conjunction with the codec network in U-Net, it is possible to not only improve the generalisation ability of the model but also the mAcc and mDice coefficients.

(a) Effectiveness of MMHA

After using Masked Multi-Head Attention instead of Multi-Head Attention, the model mAcc is improved by 2.33% and the damage dice coefficient is improved from 61.28% to 65.92%. Experiments show that MMHA is able to capture more details, effectively improve feature extraction, and retain detailed information of the input image. This is mainly due to the fact that MMHA combines the mechanisms of Multi-Head Attention and Masked Attention, which is able to deal with missing values or information that needs to be ignored in the sequence data, which is essential for semantic segmentation tasks that need to accurately identify and locate the damage boundaries in an image.

(b) Effectiveness of SRMMMHA

After inserting the Spatial Reduction module into the decoder and coupling it with the Masked Multi-Head Attention, the model's mAcc increased dramatically to 58.81%. This strategy not only achieves great segmentation performance, but it also significantly decreases computational complexity and improves the model's robustness to changes in input images. In particular, the technique indicates some advantages when dealing with multi-scale data, demonstrating its potential and relevance in actual applications.

(c) Effectiveness of Mix-Attention

Mix-Attention is an innovative variant of the attention mechanism that achieves multi-dimensional attention to the input data by combining multi-head attention and other types of self-attention mechanisms, thereby enhancing the expressive and comprehension capabilities of the model and significantly improving its performance. By adding Mix-Attention to the decoder, the Damage Dice Coefficient is improved to 70.22%. When used with MMHA, the performance of the model was better and the mAcc reached 60.93%, providing new ideas and methods for model optimisation and improvement.

(d) Pyramid-MixNet Performance

The Pyramid-MixNet design improves performance by combining three important modules: Multi-column Mixed-Head Attention (MMHA), Spatial Reduction (SR), and Multi-column Attention (MA). The synergistic combination of SR, MMHA, and MA greatly improves the model's performance. Specifically, integrating SR with MMHA and MA is critical to improving the system's overall performance with 62.17% mAcc and 80.28% Damage Dice Coefficient, which results in more accurate segmentation of rail damage, better generalisation capability and robustness.

4.3.2 Comparison Experiments

In order to objectively evaluate the practicality of Pyramid-MixNet, we conducted comparative experiments with five widely-used semantic segmentation models: the classical U-Net, SegNet, Mask R-CNN, PSPNet, and DeepLabv3+. These models represent distinct technical approaches and architectural designs in semantic segmentation: U-Net excels in medical image segmentation with its symmetric encoder-decoder structure and skip connections; SegNet achieves efficient feature reconstruction through its encoder-decoder architecture with pooling indices; Mask R-CNN combines object detection and instance segmentation for multi-task learning in complex scenarios; PSPNet captures multi-scale contextual information via pyramid pooling modules; while DeepLabv3+ enhances segmentation precision while maintaining high-resolution features through atrous convolution and encoder-decoder structures. Using identical datasets, we systematically compared these models' performance across multiple metrics including segmentation accuracy, computational efficiency, parameter count, and robustness under different scenarios. The comparative experiments not only validate the superiority of our proposed model but also provide valuable references for future research, particularly regarding application potential in complex environments and multi-task learning. The experimental results are presented in Table 2.

Table 2: Results of comparative experiments

Method	FPS	mAcc/%	mIoU	Damage dice coefficient/%
U-Net	40.49	56.37	62.94	73.28
SegNet	48.61	60.91	67.89	76.54
Mask R-CNN	58.77	59.46	64.26	75.78
PSPNet	64.82	61.73	66.38	77.65

(Continued)

Table 2 (continued)

Method	FPS	mAcc/%	mIoU	Damage dice coefficient/%
DeepLabv3+	52.88	60.52	65.93	79.33
Pyramid-MixNet	56.83	62.17	68.62	80.28

It can be seen that the FPS of Pyramid-MixNet, although not the highest, this speed still meets the real-time demand for surface damage detection on high-speed railways; the average accuracy is the highest, reaching 62.17%, with the best overall classification performance; the mIoU reaches 68.62, reflecting the model's best segmentation results on all categories; the Damage Dice Coefficient comparison with these networks is also the highest. The enhancement of the proposed model in mAcc and Damage Dice Coefficient, especially in small-scale damage detection, can effectively reduce the leakage rate and improve the reliability of the detection results. In summary, the proposed model is more practical in practical applications because it not only meets the speed requirement of real-time detection, but also provides a more reliable and efficient solution for surface damage detection of high-speed railway through higher accuracy and stronger scene adaptability.

In addition, the visualisation results obtained using different segmentation networks are shown in Fig. 8. From the segmentation results, we find that the two models, U-Net and SegNet, from the actual segmentation effect, the boundaries between the damaged regions are not clear and definite, and it is difficult to accurately define the scope of different damages. The Mask R-CNN model improves the boundary delineation to a certain extent, but its segmentation effect is still unsatisfactory for the defective regions that are in patches. PSPNet performs relatively well for the segmentation of the first three rail damage images, but once the specific element of the rail seam appears in the image, the rail seam is segmented as well, leading to inaccurate segmentation results. The prediction of DeepLabv3+ shows a better performance on the source image, suggesting that appropriate processing techniques may help improve the segmentation results of this model, but it also misses some small damages and over-segmentation. Our proposed Pyramid-MixNet is able to accurately focus on the global region of interest, presenting the damage to be close to the original shape.

At last, we conducted cross-dataset testing and diversity data enhancement experiments. Firstly, we tested the performance of the model on a publicly available railway surface damage dataset (RSDDs). The experimental results show that the proposed model achieves a mAcc of 60.12% and a Damage Dice Coefficient of 78.45% on Rail-Dataset, which is comparable to the performance on the original dataset, proving the model's adaptability in different environments. Secondly, we introduced diverse data enhancement techniques during the training process, including random light changes, noise addition and background replacement, to simulate railway surface damage under different environments and conditions. By comparing the experimental results before and after enhancement, we find that data enhancement significantly improves the model's performance on complex backgrounds and small-scale damage, with a 2.3% increase in mAcc and a 1.8% increase in Damage Dice Coefficient.

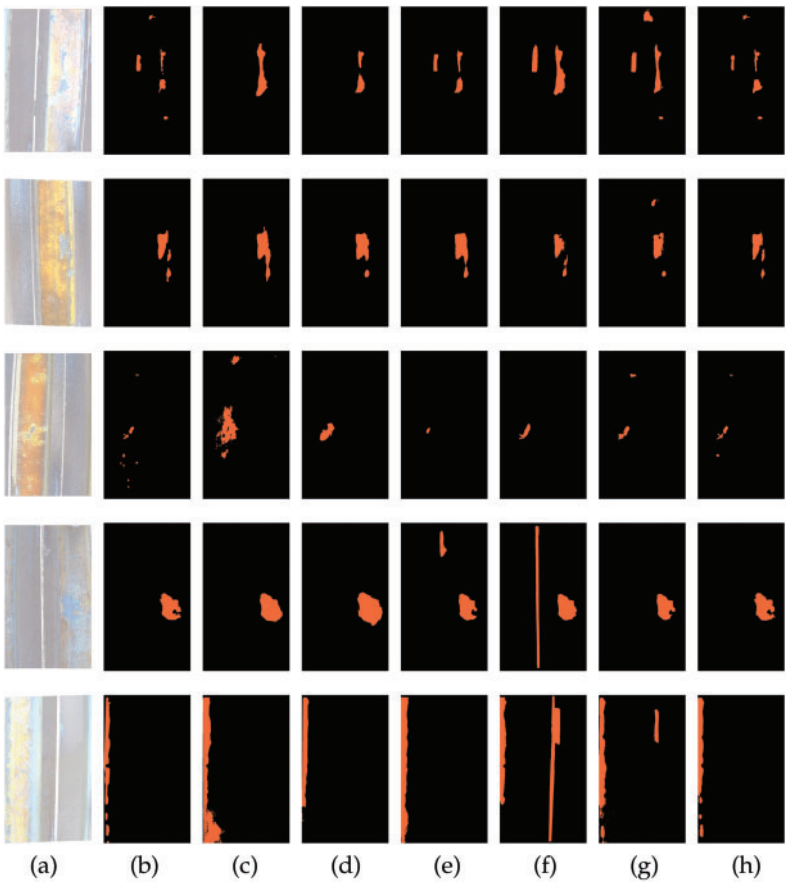

Figure 8: Visualized results of railway surface damage samples with different models. (a) Original damage image. (b) Ground-truth labeling. (c) U-Net. (d) SegNet. (e) Mask R-CNN. (f) PSPNet. (g) DeepLabv3+. (h) Pyramid-MixNet

5 Conclusion and Future Work

In this paper, we propose a method called Pyramid-MixNet, which combines Pyramid Vision Transformer (PVT) with U-Net segmentation network for the first time and incorporates multiple attention mechanisms. The integration of Spacial Reduction Masked Multi-Head Attention within the encoder effectively addresses the challenge of disregarded sequential data information, thereby enhancing the model's capability to process and understand complex sequence patterns. Complementarily, the introduction of Mix-Attention in the decoder facilitates a multi-dimensional focus on the input sequences, ensuring a more comprehensive extraction of pertinent features from diverse perspectives. Furthermore, the innovative application of lateral connection strategy within the encoder-decoder network architecture and the selection of self-made extended dataset as training samples play a pivotal role in accomplishing the segmentation task for surface defects on high-speed railways. By evaluating the accuracy and other indicators of the proposed method, the results show that the detection accuracy of our model reaches 62.17% while keeping the number of parameters. Compared with the mainstream segmentation detection algorithms nowadays, Pyramid-MixNet is more flexible and scalable, and has achieved satisfactory results in high-speed railway damage segmentation.

Nevertheless, our network is a preliminary implementation of automated rail surface damage detection and has several limitations, including a limited number of damage categories in the dataset, inability to

effectively detect covered rails, and slower inference time compared to lighter models. Future plans include expanding the damage image dataset to cover more complex scenarios, enhancing round segmentation capabilities, and exploring compression techniques to improve real-time performance and deployment flexibility.

Acknowledgement: Thanks to the Railway Authority for their support in taking pictures of the rails.

Funding Statement: This research was supported in part by the National Natural Science Foundation of China under Grant 6226070954 and Jiangxi Provincial Key R&D Programme under Grant 20244BBG73002.

Author Contributions: Hui Luo led the overall project execution. Wenqing Li conceived the main conceptual ideas and experimental arrangements, and conducted experiments and data analysis. Wei Zeng contributed to the final version of the manuscript through critical revisions that focused on interpretation of data. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: Data not available due to national confidentiality laws and regulations restrictions.This dataset involves confidential information, thus its usage has restrictions and will not be shared public. Please contact Email: 2022068081000004@ecjtu.edu.cn.

Ethics Approval: Not applicable.

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

References

1. Clark R. Rail flaw detection: overview and needs for future developments. NDT & E Int. 2004;37(2):111–8. doi:10.1016/j.ndteint.2003.06.002.
2. Yuan XC, Wu LS, Chen HW. Nanchang institute of technology; Nanchang university. Rail image segmentation based on Otsu's method. Opto-Electron Eng. 2016 1;24(7):1772–81.
3. Cao X, Xie W, Ahmed SM, Li CR. Defect detection method for rail surface based on line-structured light. Measurement. 2020;159(2):107771. doi:10.1016/j.measurement.2020.107771.
4. Kundu T, Datta AK, Topdar P, Sengupta S. Optimal location of acoustic emission sensors for detecting rail damage. Proc Inst Civ Eng Struct Build. 2022;177(3):254–63. doi:10.1680/jstbu.21.00074.
5. Guo F, Qian Y, Yu H. Automatic rail surface defect inspection using the pixelwise semantic segmentation model. IEEE Sens J. 2023;23(13):15010–8. doi:10.1109/JSEN.2023.3280117.
6. Wu Y, Qin Y, Qian Y, Guo F, Wang Z, Jia L. Hybrid deep learning architecture for rail surface segmentation and surface defect detection. Comput Aided Civ Infrastruct Eng. 2022;37(2):227–44. doi:10.1111/mice.12710.
7. Zhang Z, Che X, Song Y. An improved convolutional neural network for convenient rail damage detection. Front Energy Res. 2022;10:1007188. doi:10.3389/fenrg.2022.1007188.
8. Si C, Luo H, Han Y, Ma Z. Rail-STrans: a rail surface defect segmentation method based on improved swin transformer. Appl Sci. 2024;14(9):3629. doi:10.3390/app14093629.
9. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2015; Boston, MA, USA. p. 3431–40.
10. Sun Q, Xiao S, He W, Zuo X, Yu M. Rail surface defect detection using a U-Net convolutional neural network. In: Fourth International Conference on Signal Image Processing and Communication (ICSIPC 2024); 2024; Xi'an, China. Vol. 13253, p. 180–7.
11. Zhao H, Shi J, Qi X, Wang X, Jia J. Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2017; Honolulu, HI, USA. p. 2881–90.
12. Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL. Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans Pattern Anal Mach Intell. 2017;40(4):834–48. doi:10.1109/TPAMI.2017.2699184.

13. Liu C, Chen L, Schroff F, Adam H, Hua W, Yuille A, et al. Auto-DeepLab: hierarchical neural architecture search for semantic image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2019; Long Beach, CA, USA. p. 82–92.
14. Liu Y, Li Q, Yuan X, Li Y, Zhu Z. Rail base flaw detection and quantification based on the modal curvature method and the back propagation neural network. Eng Fail Anal. 2022;142:106792. doi:10.1016/j.engfailanal.2022.106792.
15. Guo F, Liu J, Qian Y, Xie Q. Rail surface defect detection using a transformer-based network. J Ind Inf Integr. 2024;38(6):100584. doi:10.1016/j.jii.2024.100584.
16. Chen Z, Yang J, Zhou F. RailSegVITNet: a lightweight VIT-based real-time track surface segmentation network for improving railroad safety. J King Saud Univ-Comput Inf Sci. 2024;36(1):101929. doi:10.1016/j.jksuci.2024.101929.
17. Zheng S, Lu J, Zhao H, Zhu X, Luo Z, Wang Y, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021; Nashville, TN, USA. p. 6881–90.
18. Xie E, Wang W, Yu Z, Anandkumar A, Alvarez JM, Luo P. SegFormer: simple and efficient design for semantic segmentation with transformers. Adv Neural Inf Process Sys. 2021;34:12077–90.
19. Wang W, Xie E, Li X, Fan D, Song K, Liang D, et al. Pyramid vision transformer: a versatile backbone for dense prediction without convolutions. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021; Montreal, QC, Canada. p. 548–58.
20. Chu X, Tian Z, Wang Y, Zhang B, Ren H, Wei X, et al. Twins: revisiting the design of spatial attention in vision transformers. Adv Neural Inf Process Sys. 2021;34:9355–66.
21. Cheng B, Misra I, Schwing AG, Kirillov A, Girdhar R. Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022; New Orleans, LA, USA. p. 1290–9.
22. Yeom SK, von Klitzing J. U-MixFormer: uNet-like transformer with mix-attention for efficient semantic segmentation. arXiv:231206272. 2023.