Hindawi
Structural Control and Health Monitoring
Volume 2024, Article ID 2967927, 22 pages
https://doi.org/10.1155/2024/2967927

Research Article

Intelligent Detection of Surface Defects in High-Speed Railway Ballastless Track Based on Self-Attention and Transfer Learning

Wenlong Ye,[1,2,3] Juanjuan Ren,[1,2,3] Chen Li,[1,2,3] Wengao Liu,[1,2,3] Zeyong Zhang,[1,2,3] and Chunfang Lu[1,4]

[1]*School of Civil Engineering, Southwest Jiaotong University, Chengdu, China*
[2]*State Key Laboratory of Rail Transit Vehicle System, Southwest Jiaotong University, Chengdu, China*
[3]*MOE Key Laboratory of High-Speed Railway Engineering, Southwest Jiaotong University, Chengdu, China*
[4]*China Railway Society, Beijing, China*

Correspondence should be addressed to Juanjuan Ren; jj.ren@swjtu.edu.cn

Received 29 January 2024; Revised 18 April 2024; Accepted 24 April 2024; Published 18 May 2024

Academic Editor: Hoon Sohn

The detection of ballastless track surface (BTS) defects is a prerequisite for ensuring the safe operation of high-speed railways. Traditional convolutional neural networks fail to fully exploit contextual information and lack global pixel representations. The extensive stacking of convolutions leads deep learning models to play a black-box detection role, lacking interpretability. Due to the current lack of sufficient high-quality surface data for ballastless tracks, it is a severe constraint on the accurate identification of the substructure state in high-speed railways. This paper proposes an intelligent detection method for BTS defects named TrackNet based on self-attention and transfer learning. The method enhances the fusion ability of global features of BTS defects using multihead self-attention. The model's dependence on extensive defect data is reduced by transferring knowledge from large-scale publicly available datasets. Experimental results demonstrate that compared to advanced Swin Transformer model results, the TrackNet model achieves improvements in average accuracy and $F1$-score by 5.15% and 5.16%, respectively, on limited test data. The TrackNet model visualizes the decision regions of the model in identifying BTS defects, revealing the black-box recognition mechanism of deep learning models. This research performs engineering applications and provides valuable insights for the multiclass recognition of BTS defects in high-speed railways.

1. Introduction

As China's expansive high-speed railway lines, encompassing eight vertical and eight horizontal high-speed railway networks, have continuously developed, by the end of 2023, the total operational length of high-speed railways is about 45,000 km, securing a leading position globally. Currently, the operating speed of China's high-speed trains has attained a peak of 350 km/h, establishing them as the fastest globally [1]. The ballastless track is a critical infrastructure supporting the safe and stable operation of high-speed trains, mainly consisting of rail, fastener, track slab, self-compacting concrete, concrete roadbed, etc., as shown in Figure 1 [2]. With the extension of service time, under the influence of factors such as high-frequency train loads, temperature stress, and weathering, field research has found that ballastless tracks may exhibit surface defects, such as concrete cracks and fastener broken, as shown in Figure 2. Concrete cracks provide a pathway for rainwater to corrode the steel reinforcement and other structural elements of the ballastless track, thereby reducing the structural load-bearing capacity and significantly affecting the durability and service performance of the structure. Fasteners are critical components to ensure a reliable connection between the rail and the track slab. Once a fracture occurs, it can intensify the dynamic response between the wheel and the track and even lead to train derailment, potentially leading to safety accidents. Therefore, to ensure operational safety and high-speed railways, it is particularly important to understand the surface state of ballastless tracks and adopt efficient methods for detecting BTS defects.

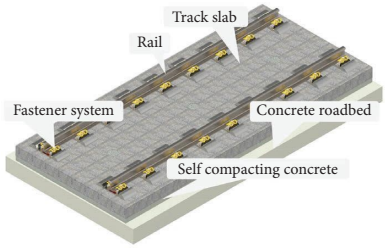


Figure 1: CRTS III ballastless track structure.

Figure 2: Ballastless track surface defects: (a) fastener broken [3]; (b) concrete crack.

Visual changes in railway infrastructure defects directly indicate the health status of the structure to inspection personnel [4]. Traditional surface defect detection primarily depends on manual inspection, and this method has limitations such as low efficiency, poor accuracy, and incompleteness. With the rapid development of information technology, a growing number of scholars are focusing on the automation of infrastructure surface detection using advanced computer visual information [5]. Computer visual inspection techniques are mainly categorized into image processing and deep learning-based detection techniques [6]. Common image processing techniques include edge extraction [7], region growing [8], threshold segmentation [9], and template matching [10]. The detection accuracy of these methods depends on manually adjusted feature parameters and is usually limited by the dataset. In the complex and variable night environment, the detection of BTS defects often lacks robustness.

In recent years, considerable advancements have been achieved in the realm of surface defect detection within infrastructure such as railways, pavements, and bridges, thanks to advanced deep learning technologies. Santur et al. [11] adopted laser cameras for railway surface inspection to reduce time loss and additional costs during railway maintenance. James et al. [12] proposed a multiphase rail surface defect detection technique based on deep learning, which improves detection performance by reducing false alarm rates. Wang et al. [13] presented a quantitative classification approach for track cracks utilizing deep learning networks, achieving the determination of track crack severity levels. Guo et al. [14, 15] proposed real-time railway component detection and pixel-level segmentation models, which have significant advantages in terms of accuracy and processing speed. Wu et al. [16, 17] used drone images for rail defect and track component detection and proposed the RBGNet and AOYOLO models based on hybrid deep learning. Cai et al. [18] introduced a few-shot learning model for ballastless track defect detection, effectively addressing the challenge posed by the scarcity of high-quality data necessary for training deep learning models. Ye et al. [19] developed a pixel-level segmentation-quantification method suitable for nighttime ballastless track crack detection, completing the precise measurement of track crack width in nighttime environments. Zhang et al. [20] proposed ShuttleNet for multiple distress detection on asphalt roads, which can learn and integrate context information at different resolution levels many times to enhance expression. Tran et al. [21] assessed the performance of five advanced target detection algorithms in bridge crack recognition and integrated the YOLOv7 model with the U-Net algorithm for bridge crack detection and segmentation. Their research studies significantly improved the capability of infrastructure surface defect detection, highlighting the robustness of deep learning models based on convolutional neural networks in defect detection. However, deep convolutional neural networks overlook the long-distance dependencies between pixels in the feature extraction and modeling process of defects, posing challenges to their global modeling capability [22].

The Transformer model employs a self-attention mechanism to encode and represent input sequences, offering higher computational efficiency compared to deep convolutional neural networks [23]. It is particularly adept at capturing dependencies of feature patterns at different positions within a global view, enabling long-distance context awareness. This architecture has been widely applied and achieved significant success in various fields, including natural language processing (e.g., ChatGPT) [24], autonomous driving [25], and audio-video recognition [26]. However, it is crucial to acknowledge that Transformer models frequently necessitate a substantial volume of training data samples. To reduce the model's dependence on extensive image data samples, transfer learning is an effective solution, enhancing the model's performance in target domains by transferring knowledge from large-scale public datasets [27]. Shamsabadi et al. [28] utilized transfer learning and the ViT model to detect cracks on asphalt and concrete surfaces, demonstrating strong robustness against various noise signals. Pan et al. [29] presented a novel transfer learning model founded on MobileNet, which addresses the issue of limited data samples and effectively detecting welding defects. Bunrit et al. [30] studied models pretrained on the ImageNet large dataset, which can be trained to classify construction material images in a shorter time through transfer learning and fine-tuning schemes. These studies provide new insights for the precise detection of limited BTS defects.

This study reviews the development of current infrastructure damage detection technologies and the application of artificial intelligence. It is found that deep convolutional neural networks have certain limitations in extracting long-distance dependencies between defect features, while Transformer models require a substantial amount of training samples. At the same time, these detection models mainly play the role of a black box, and

interpreting these models remains a challenging task [31]. Therefore, an interpretable track detection network (TrackNet) model based on the Swin Transformer and utilizing transfer learning is proposed for the detection of BTS defects. This paper primarily contributes in the following ways:

(1) The TrackNet model can load pretrained weights from extensive public datasets and perform parameter adjustments on a limited BTS defect dataset, which accelerates the model's convergence and reduces the dependence on extensive training samples.

(2) The TrackNet model utilizes the window multihead self-attention to enhance the global modeling capability of the detection model. This hierarchical structure can provide feature information about BTS defects at various scales.

(3) The TrackNet model introduces activation heatmaps and a nonlinear dimensionality reduction strategy, making the decision-making process of the model more interpretable and transparent. By reducing high-dimensional spatial semantic information to 2D or 3D visual spaces, a clearer expression is achieved in the paper.

To be more specific, the contents of the subsequent sections are as follows. Section 2 introduces the interpretable transfer learning method. Section 3 describes the process of constructing a dataset for surface defects of ballastless tracks at night. Section 4 details the experimental process, including the training environment and evaluation metrics. Section 5 discusses the test results and provides visualizations of the model's decision regions and dimensionality reduction identification results. Section 6 demonstrates the application effectiveness of the model in identifying on-site damage images. Finally, the conclusions and future work are given in Section 7.

2. Interpretable Transfer Learning Method

This section introduces an interpretable transfer learning method named TrackNet for detecting multitarget BTS defects. This method utilizes the ImageNet-1k data to obtain pretrained weights and optimizes training based on the Swin Transformer model, which accelerates the model's convergence speed and enhances its generalization ability [32]. Additionally, it incorporates the Grad-CAM (Gradient-weighted Class Activation Mapping) mechanism [33] to interpret the deep network model in the form of heatmaps. The t-SNE [34] nonlinear dimensionality reduction technique is also used to reduce the model's high-dimensional space data representation to 2D or 3D space, facilitating the visualization of the identification results.

2.1. Transfer Learning Model. Figure 3 illustrates the overall structure of the TrackNet model. The Swin Transformer model is trained on the publicly available large dataset ImageNet-1k to obtain pretrained weights. Subsequently, through transfer learning, the model undergoes further fine-tuning on the track defect dataset. Firstly, the sizes of the BTS defect images are uniformly adjusted to $256 \times 256 \times 3$ and fed into the entire network structure. The images undergo Patch Partitions, specifically, a 4×4 2D convolutional operation with a stride of 4 within the channel dimension. This process yields a feature map, denoted as F_0, with dimensions of $64 \times 64 \times 48$. Subsequently, F_0 is utilized in the construction of four stages to extract defect features at different scales. This hierarchical structure is similar to the network structures of VGG [35] and Resnet [36], which are capable of capturing damage information at multiple scales. Shallow-layer structures handle more data, while top-layer structures handle less data, yet with richer semantic information.

The feature map passes through a Linear Embedding layer in the stage 1 block, where a linear transformation alters the image shape to $64 \times 64 \times 96$ ($C = 96$). Then, the feature map goes through a transformer block, maintaining its size after passing through the self-attention computation. To reduce the quadratic complexity of the token count and facilitate the transmission of information between adjacent windows, the transformer block employs two types of MSA (multihead self-attention) mechanisms. Specifically, it contains the W-MSA (Windows MSA) and SW-MSA (Shifted Windows MSA), which also elucidates the rationale behind the presence of an even number of transformer blocks at different stages (i.e., 2, 2, 6, 2). The computation of the entire consecutive sequence of transformer blocks is performed as follows:

$$\begin{aligned} \hat{\mathbf{z}}^{l} &= \mathrm{W-MSA}\left(\mathrm{LN}\left(\mathbf{z}^{l-1}\right)\right) + \mathbf{z}^{l-1}, \\ \mathbf{z}^{l} &= \mathrm{MLP}\left(\mathrm{LN}\left(\hat{\mathbf{z}}^{l}\right)\right) + \hat{\mathbf{z}}^{l}, \\ \hat{\mathbf{z}}^{l+1} &= \mathrm{SW-MSA}\left(\mathrm{LN}\left(\mathbf{z}^{l}\right)\right) + \mathbf{z}^{l}, \\ \mathbf{z}^{l+1} &= \mathrm{MLP}\left(\mathrm{LN}\left(\hat{\mathbf{z}}^{l+1}\right)\right) + \hat{\mathbf{z}}^{l+1}, \end{aligned} \tag{1}$$

where $\hat{\mathbf{z}}^{l}$ and $\hat{\mathbf{z}}^{l+1}$ represent the output features of the W-MSA and SW-MSA, respectively, and $\mathbf{z}^{l}$ and $\mathbf{z}^{l+1}$ represent the output features of the MLP.

Each feature map generates a total of $H/M \times W/M$ windows, each with a size of $M \times M$. The W-MSA operates within each window, where each patch in the window is considered a token, and its features are regarded as the concatenation of the original pixel values. Each token generates three vectors Q, K, and V through three trainable transformation matrices W_q, W_k, and W_v. To facilitate the model to concern itself with information from various subspace locations, Q, K, and V vectors are linearly projected into a low-dimensional space h times according to (2) and (3), to perform self-attention operations. After obtaining the self-attention outputs from different subspaces, they are concatenated and projected back to a higher dimension through a linear projection to obtain the final output, thereby enhancing the semantic expression of the defect features.

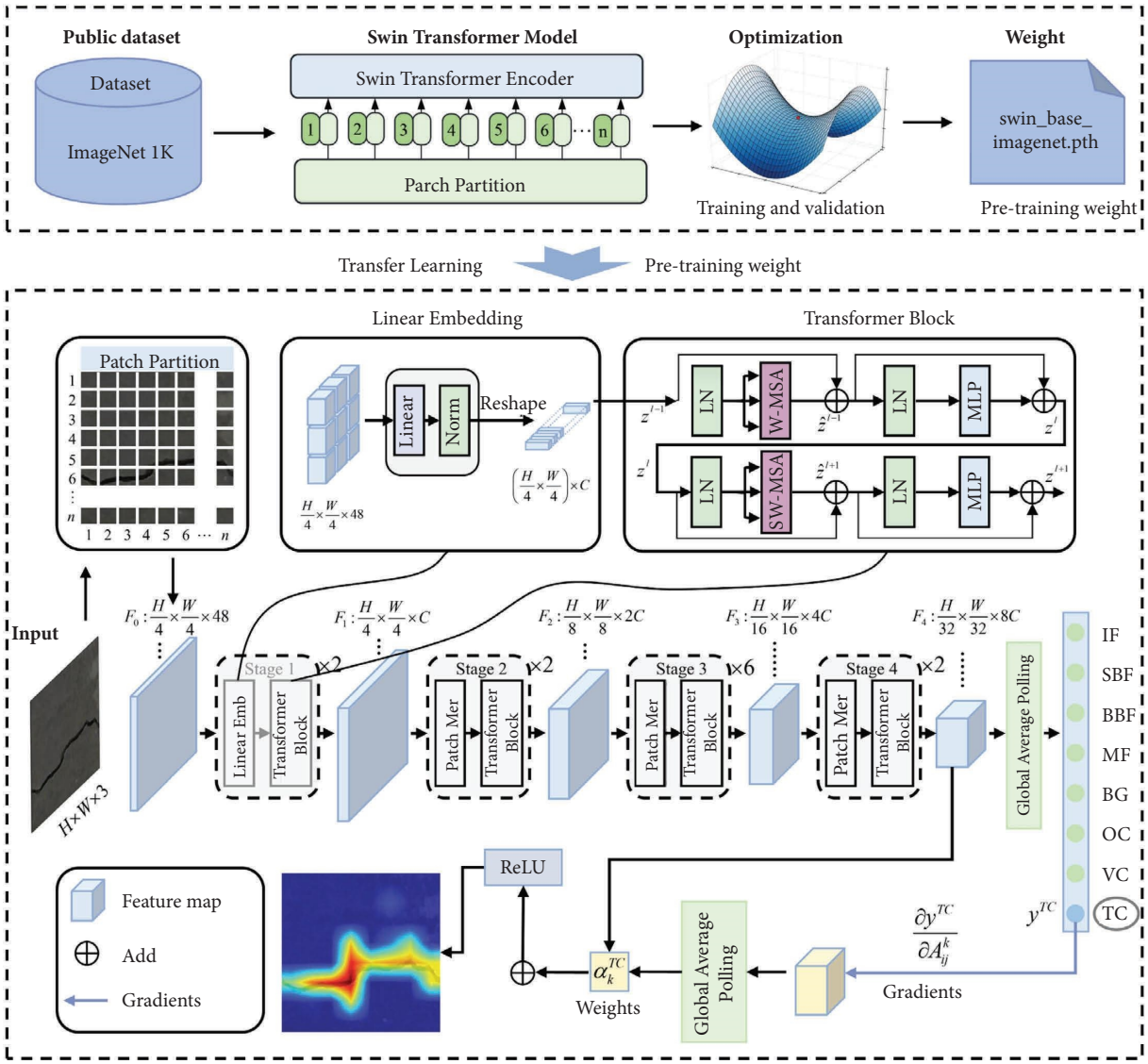


FIGURE 3: Overall structure of the proposed TrackNet model.

$$\text{MultiHead}(Q, K, V) = \text{Concat}W^{o}, \tag{2}$$

$$\begin{aligned}\text{head}_{i} &= \text{Attention}(Q_i, K_i, V_i) \\ &= \text{softmax}\left(\frac{Q_i K_i^{T}}{\sqrt{d_k}}\right)V_i,\end{aligned} \tag{3}$$

where Q is query, K is key, and V represents the information extracted from each token.

To enhance the information connection between different windows, after completing W-MSA, the feature map undergoes SW-MSA computation. Taking an 8×8 feature map as an example, Figure 4 illustrates the shifted window partition. At the l layer, a standard window partitioning approach is utilized, with $M \times M$ $(M = 4)$ patches and 2×2 windows. Subsequently, self-attention calculations are performed within each window. At the $l+1$ layer, it is visible that the window partitioning has shifted, displacing $(M/2, M/2)$ pixels from the conventionally divided windows to create new windows. The self-attention computation crosses the previous window boundaries, providing connections between windows. This helps in transferring information between different windows and strengthens the global connections of the feature map.

In Stage 2, the feature map undergoes downsampling through the Patch Merging layer, resulting in a halved size and doubled channel number, ultimately forming a new feature map with dimensions of $32 \times 32 \times 192$. The key to this process is the reduction of the number of tokens to lower computational complexity while preserving contextual information. As illustrated in Figure 5, Patch Merging extracts pixels at the same position within each adjacent 2×2 region in the $H \times W \times C$ feature map and concatenates them in the channel dimension, yielding $H/2 \times W/2 \times 4C$ feature maps. These feature maps undergo a linear transformation in the channel dimension through the LayerNorm layer, reducing the depth of the feature

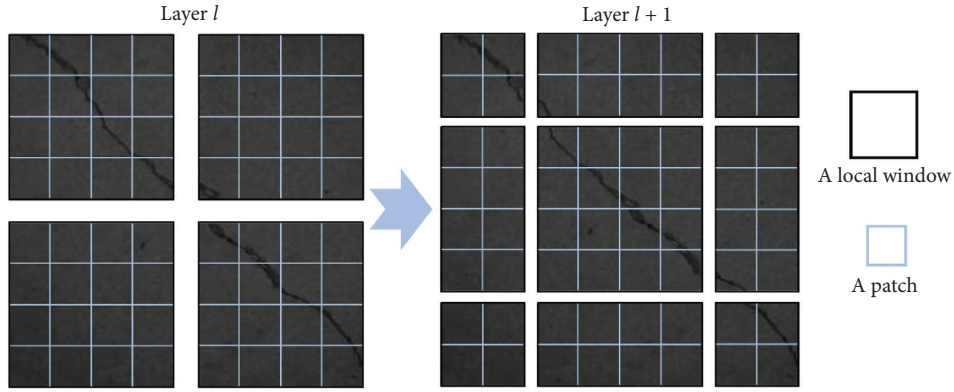


Figure 4: Shifted window partition.

maps from 4C to 2C. Subsequently, they enter the transformer block for multihead self-attention computation, extracting BTS defect features.

Similarly, after passing through Stage 3 and Stage 4, the feature map becomes $16 \times 16 \times 384$ and $8 \times 8 \times 768$, respectively, through the processing of the Patch Merging and the transformer block. As the stage blocks are stacked, the multiscale feature information of BTS defects is further extracted, enhancing the expression of deep semantic features. At the same time, as the number of stages decreases, the number of tokens requiring self-attention computation also correspondingly reduces. Finally, after global feature extraction, the feature map outputs the category of BTS defects. Through such processing, the model is capable of capturing the defect features at different levels and scales, achieving effective detection and classification of BTS defects.

2.2. Grad-CAM Module. To build a transparent and interpretable model and to reveal the working mechanism of the model in the decision-making process for BTS defect recognition, the Grad-CAM module has been integrated into the TrackNet model. Figure 6 demonstrates how the Grad-CAM module functions. In deep learning models, the output feature map of the last convolutional layer has the greatest impact on the recognition results. The TrackNet model assesses the significance of each neuron in the decision-making process regarding BTS defects by analyzing the gradient information that flows into the model's last convolutional layer. The last convolutional layer of the TrackNet network is the output layer of stage 4.

In Figure 6, with a crack image as the input and after the forward propagation through the TrackNet model, the original score for identifying the crack category is calculated through the image classification task. For neurons identified as belonging to the crack category, their gradients are set to 1, while for neurons of other categories, the gradients are set to 0. This gradient information is backpropagated to the targeted corrected feature map and then pointwise multiplied with the backpropagated gradients to obtain a decision heatmap.

Specifically, by backpropagating the predicted value y^c for category c (cracks), the partial derivative of the value y^c to the feature map A^k is calculated, obtaining the contribution of each element in the feature map A^k to the value y^c. The neuron importance weight α_k^c is computed as follows:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}, \tag{4}$$

where y^c represents the score predicted by the TrackNet model for category c before softmax activation, A_{ij}^k represents the data in A at the location of coordinate (i, j) on channel k, and Z is the multiplication of the width and height.

The gradient heatmap is obtained by a weighted combination and applying the ReLU activation function. The expression is given in the following equation:

$$L_{\text{map}}^c = \text{ReLU} \sum_k \alpha_k^c A^k. \tag{5}$$

2.3. t-SNE Module. The t-SNE module [34] is added to the TrackNet model to reduce 768-dimensional semantic features to a 2D or 3D space for visualizing identification results. The core idea of t-SNE is to transform the Euclidean distance between data points into a probability distribution that represents their similarity, which is then mapped to a low-dimensional space. Specifically, data points that are far apart in high-dimensional space remain distant after being mapped to the low-dimensional space, and vice versa. In high-dimensional space, a Gaussian distribution is used to calculate the similarity between points. p_{ij} is defined as follows:

$$p_{\text{ij}} = \frac{\exp\left(-\left\|x_i - x_j\right\|^2 / 2\sigma^2\right)}{\sum_{k \neq l} \exp\left(-\left\|x_k - x_l\right\|^2 / 2\sigma^2\right)}, \tag{6}$$

where p_{ij} represents the similarities between x_i and x_j in high-dimensional space and σ means the variance of the Gaussian.

To avoid crowding of data points in the low-dimensional space, a heavy-tailed distribution called the Student-t distribution is used instead of the Gaussian distribution to compute the similarity between points. The similarity in the low-dimensional is given by

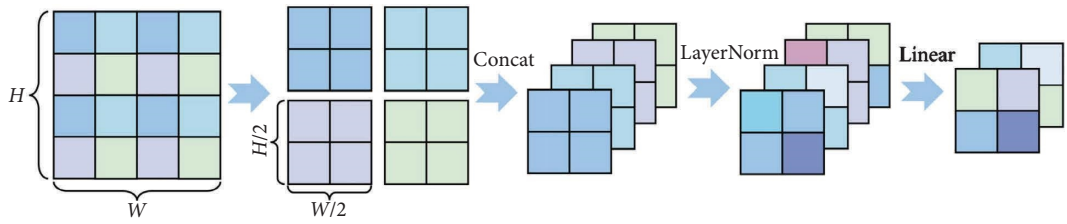


Figure 5: Illustration of Patch Merging.

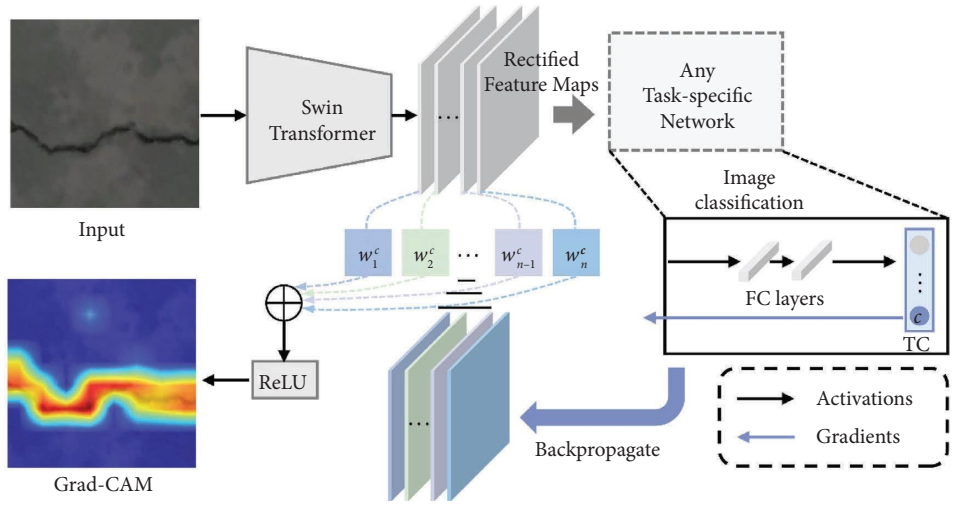


Figure 6: Illustration of the Grad-CAM module.

$$q_{ij} = \frac{\left(1 + \left\|y_i - y_j\right\|^2\right)^{-1}}{\sum_{k \neq l}\left(1 + \left\|y_k - y_l\right\|^2\right)^{-1}}, \tag{7}$$

where q_{ij} represents the similarities between y_i and y_j in the low-dimensional space.

To ensure the convergence of the probability distributions before and after dimensionality reduction, the KL (Kullback–Leibler) divergence is employed to formulate the loss function for dimensionality reduction. The discrepancy between the probability distributions in the low- and high-dimensional spaces is optimized by minimizing the KL divergence. Here, the loss C is defined as

$$C = \mathrm{KL}(P \| Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}, \tag{8}$$

where C denotes the loss function for dimensionality reduction and P and Q represent the joint probability distribution in high- and low-dimensional spaces, respectively.

3. Ballastless Track Surface Defect Dataset

This study focuses on the BTS defects, primarily involving concrete cracks and fastener defects. The ability to detect these defects effectively during nighttime relies on the image quality. This section introduces the data preparation and analysis.

3.1. Image Preparation. The concrete crack data used in this study are derived from two sources: firstly, crack images of the ballastless at the HSR site are captured using a high-definition camera and crack images of the test platform at the Southwest Jiaotong University (SWJTU) are captured using a drone; secondly, the study refers to publicly available concrete crack datasets [6, 37]. Fastener data are collected from the images on the experimental platform at SWJTU through a ballastless track inspection vehicle [38]. To achieve precise detection of various BTS defects, this study selected 2,400 images from the mentioned datasets. The dataset is stratified into a training set of 1,680 images and a test set of 720 images in a 7 : 3 ratio, ensuring balanced data distribution to improve the model's performance [39]. The identified target categories comprise eight classes, namely, background (BG), transverse crack (TC), vertical crack (VC), oblique crack (OC), intact fastener (IF), single-broken fastener (SBF), both-broken fastener (BBF), and missing fastener (MF). As shown in Figure 7, the type of crack is determined based on the angle of the crack [40]. Specifically, a crack appearing in the range of region I is defined as a transverse crack, a crack appearing in the range of region II is defined as a vertical crack, and a crack appearing in the range of region III is defined as an oblique crack.

As Chinese high-speed trains operate during the day, maintenance work on ballastless tracks can only be conducted at night. To better align the reference image data with nighttime conditions, these images undergo further darkening processing. Specifically, the images consist of pixel

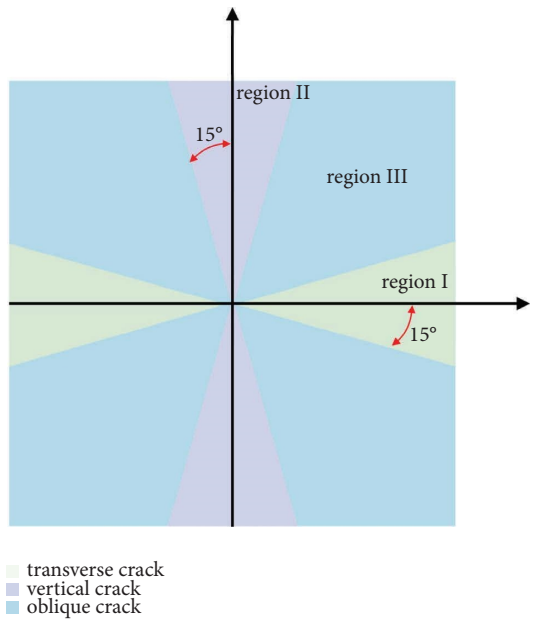


Figure 7: Crack classification rules.

matrices for the RGB (Red, Green, Blue) channels, with each pixel matrix containing 256×256 pixel values. The pixel values span from 0 to 255, with 0 denoting black and 255 signifying white. The higher the pixel, the brighter the image. In Figure 8, this study achieves the darkening of the images by batch-processing all images and implementing a coefficient reduction on pixel values through a program.

3.2. Image Analysis. This section analyzes the BTS defect images after the image darkening process. Table 1 displays images of eight categories of BTS defects in the dataset, and the pixel values of the images form a normal distribution, indicating good image quality after the darkening process. In the histograms, the x-axis represents the pixel values, with smaller values indicating darker regions and larger values indicating brighter regions. The pixel values are roughly distributed between 50 and 70 across the histograms of BTS defect images for the eight categories. Additionally, after the image darkening process, some pixel values for fastener defects are 0, which is related to the fact that fasteners themselves have a black color.

To further validate whether the image darkening process affects the features of BTS defects, Table 1 presents the gradient diagrams of BTS images. In the gradient diagrams, it is clear that the original gradient features along the edges of concrete cracks and fractured fasteners are retained, indicating no loss of defect features. This further confirms the effectiveness of the constructed multiclass dataset, ensuring the reliability of fine-tuning model parameters in the transfer learning process for identifying BTS defects.

4. Experimental Design

This section introduces the model training configuration, the training optimization process, and the evaluation metrics used for the recognition of BTS defects.

4.1. Training Configuration. To mitigate the impact of different devices on training and testing results, and to ensure a fair hardware environment for experimental study, all algorithms are trained and tested on an identical computing system. The computer is equipped with an Intel i7-11700 processor as the CPU and an NVIDIA GeForce RTX 3080 Ti GPU for accelerated computation. All models ran successfully in the Python 3.8 version and PyTorch 1.9 environment.

When configuring the learning strategy, the batch size is established at 16, and the transfer learning model undergoes training for a duration of 100 epochs. The training optimization employs the SGD optimizer with initial parameters, including a learning rate of 0.1, a momentum value of 0.9, and a weight decay of 0.0001. A learning rate adjustment strategy is implemented, involving stepwise reduction of the learning rate to 1/10 of its initial value at specific epochs, coupled with a linear warm-up of the learning rate. Figure 9 illustrates the detailed learning rate strategy.

4.2. Model Training. The training procedure of the TrackNet model primarily involves two crucial phases: forward propagation (FP) and backward propagation (BP). During the FP phase, the model utilizes training image data and weight parameters to compute the predicted output. The BP phase introduces a loss function to assess the disparity between actual labels and predicted values. Subsequently, an optimizer is employed to adjust the model's weight parameters. The FP and BP phases are executed in a loop within the specified number of epochs.

In this study, Label Smoothing Loss is employed as the loss function. This loss function first smoothes the labels to prevent the network from being overly confident, thus avoiding overfitting and improving the model's generalization capability [41]. According to (9), the labels are smoothed by introducing a small hyperparameter α, making the label distribution more uniform. Ultimately, the loss is represented by the cross-entropy expectation between the actual and predicted values, as in (10).

$$y_k^{ls} = y_k(1-\alpha) + \frac{\alpha}{K}, \tag{9}$$

$$\text{Loss} = H(y,p) = \sum_{k=1}^{K} -y_k^{ls}\log(p_k), \tag{10}$$

where y_k is the original actual label, y_k^{ls} is the actual label after smoothing, p_k is the predicted value, and K indicates the BTS defect category.

To enhance result accuracy while reducing computational time, this study employed the approach of transfer learning. Figure 10 illustrates the training processes of various algorithms. As the epoch increases, the model's loss gradually decreases and stabilizes. The TrackNet model, based on transfer learning, loaded pretrained weights on the ImageNet-1K dataset during training. From Figure 10, it can be observed that after 100 training epochs, the loss of the TrackNet model stabilizes around 1.0. In contrast, the Swin Transformer model exhibits slower training speed, with

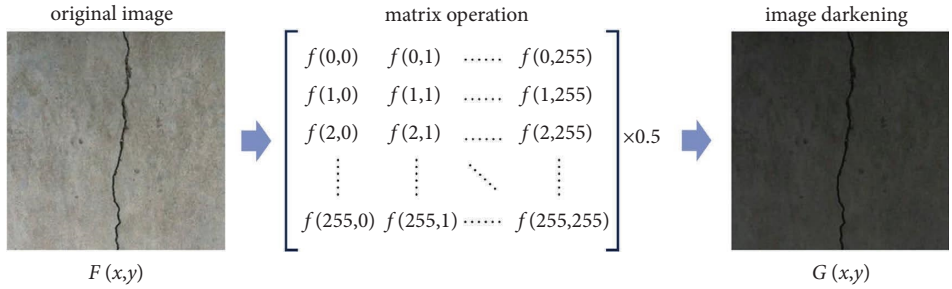


Figure 8: Image darkening of BTS defects.

a smaller rate and magnitude of loss reduction. By the time the training epoch reaches 200, the loss gradually stabilizes around 1.6. The TrackNet model not only demonstrates a faster rate of loss reduction but also achieves a smaller final loss value. This suggests that the transfer learning model has achieved a more significant performance improvement within the same training period.

4.3. Evaluation Metrics. This section introduces the basic concepts required for calculating these evaluation metrics. TP is the number of correctly predicted instances of the target defect, categorized into eight target classes in this study. FP is the number of instances where other defect categories are incorrectly predicted as the target category. TN is the number of instances where other defect categories (i.e., categories outside the target defect category) are correctly identified as other respective defect categories. FN is the number of instances where the target defect category is incorrectly identified as other defect categories.

Accuracy is a commonly used overall performance metric for models, as shown in the following equation.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}. \tag{11}$$

Precision, as demonstrated in (12), is defined as the proportion of true positive samples to the total number of samples predicted as the positive class by the model.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{12}$$

Recall, as depicted in (13), is calculated as the fraction of samples predicted as true positives relative to the total number of actual positive class samples.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{13}$$

The $F1$-score, represented in (14), is defined as the harmonic mean of precision and recall.

$$\text{F1} - \text{score} = \frac{2\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{14}$$

5. Results and Discussion

This section presents the outcomes achieved by various algorithms on the limited test set and provides a comprehensive evaluation of the test results using a confusion matrix, PR curve, and ROC curve. Additionally, this section visualizes the decision regions of the models in identifying BTS defects and the data representations in low-dimensional space.

5.1. Test Results. The confusion matrix serves as a tool for assessing a model's detection performance by summarizing its predictions across diverse categories and presenting the outcomes in matrix format. Figure 11 illustrates the confusion matrices of different models in the limited test results, serving as an assessment of the models' generalization capabilities. In the confusion matrix, each column represents the true labels of the eight types of BTS defects, while each row represents the model's predictions for defect categories. Taking the example of transverse crack recognition in Figure 11(a), the value 77 indicates the number of correctly predicted samples of transverse cracks. The value 9 represents the number of transverse cracks mistakenly classified as background samples. The value 4 represents the number of transverse cracks mistakenly classified as oblique crack samples. The more concentrated the predicted values on the diagonal of the confusion matrix, the better the performance of the model in detecting BTS defects. From the calculation in Figure 11(b), the detection accuracy of the TrackNet model is obtained as 99.17%, which is an improvement of 5.15% compared to the Swin Transformer model results.

Tables 2 and 3, respectively, present the evaluation metrics of the Swin Transformer and the proposed TrackNet model on a limited test set. This limited test set consists of 90 images, covering various types of defects such as background, transverse crack, vertical crack, oblique crack, intact fastener, single-broken fastener, both-broken fastener, and missing fastener. Table 2 reveals that the average precision, recall, and $F1$-score are 94.63%, 94.31%, and 94.29% for detecting various types of BTS defects, respectively. Specifically, Swin Transformer demonstrates a more significant improvement in the detection accuracy of concrete cracks compared to fastener defects. In contrast, Table 3 reveals that the average precision, recall, and $F1$-score of the proposed TrackNet model are 99.20%, 99.17%, and 99.16%, respectively. When compared to the detection outcomes of the Swin Transformer algorithm, the results of the TrackNet model show an improvement of 4.82%, 5.15%, and 5.16%, respectively. The assessment outcomes suggest that the TrackNet model exhibits superior performance on the

Table 1: Image analysis of ballastless track surface defects.

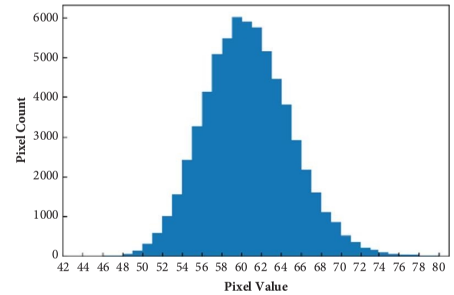
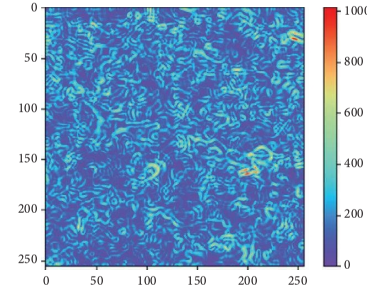
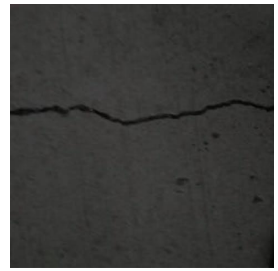
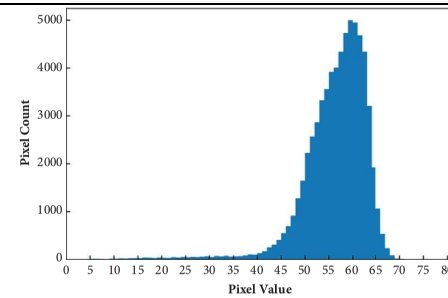
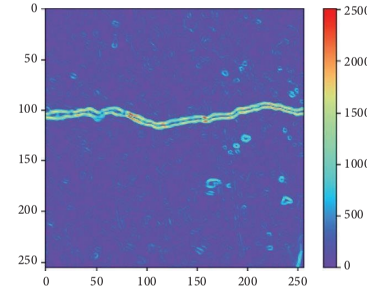
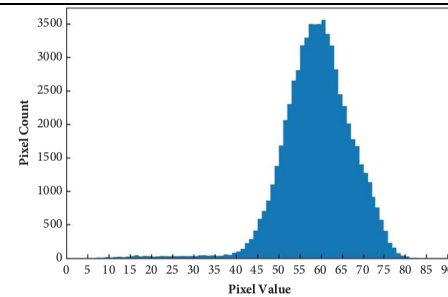
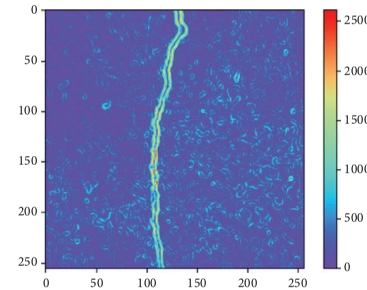
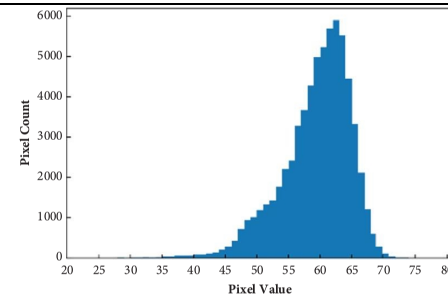
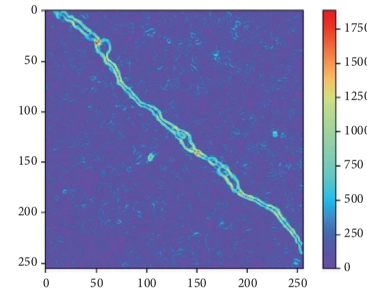
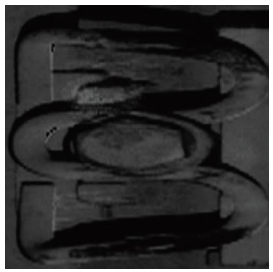
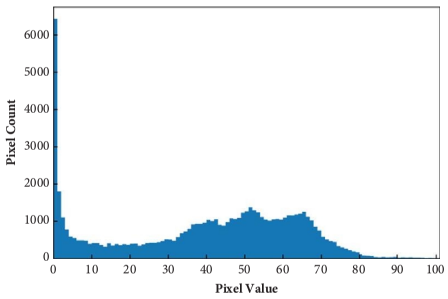
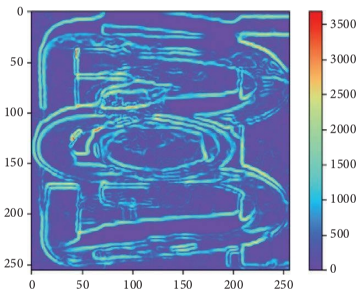
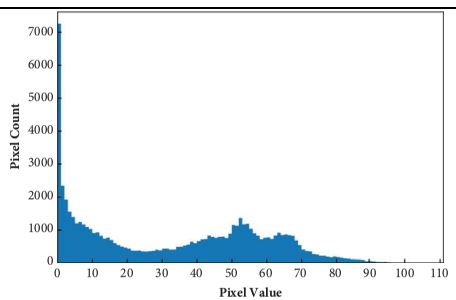
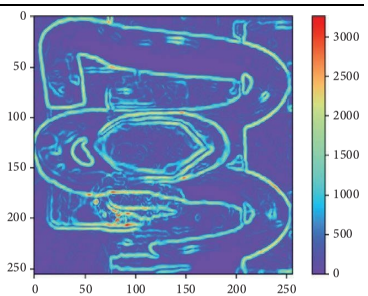
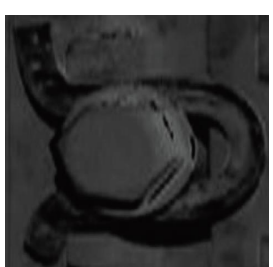
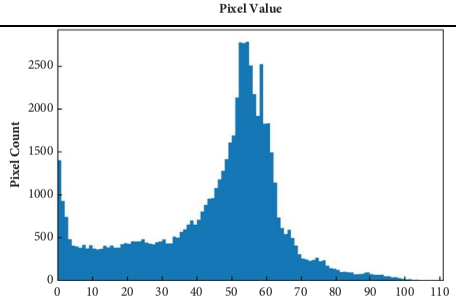
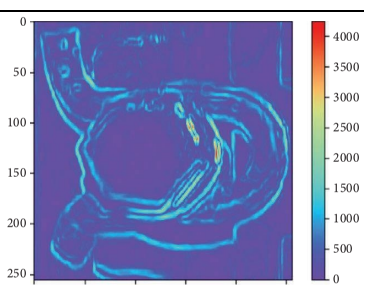
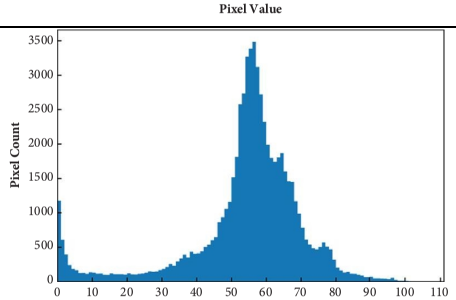
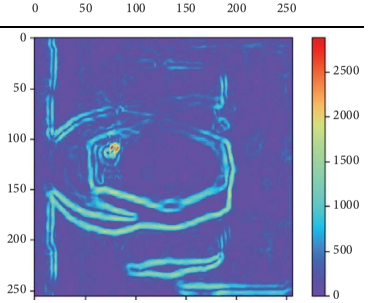
Track defects		Histogram	Gradient diagram
Background (BG)			
Transverse crack (TC)			
Vertical crack (VC)			
Oblique crack (OC)			

Table 1: Continued.

Track defects		Histogram	Gradient diagram
Intact fastener (IF)			
Single-broken fastener (SBF)			
Both-broken fastener (BBF)			
Missing fastener (MF)			

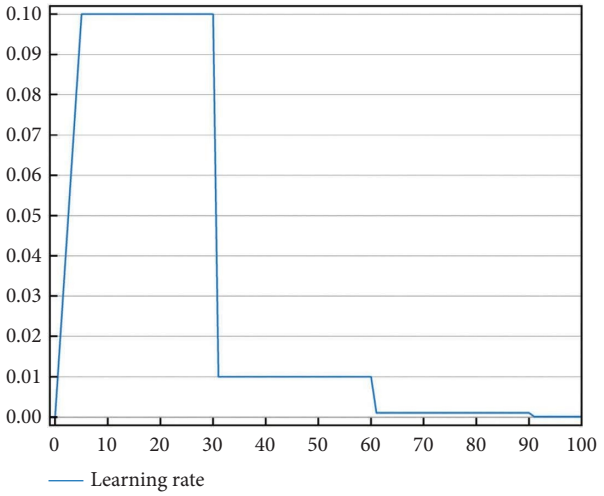


Figure 9: Learning rate strategy.

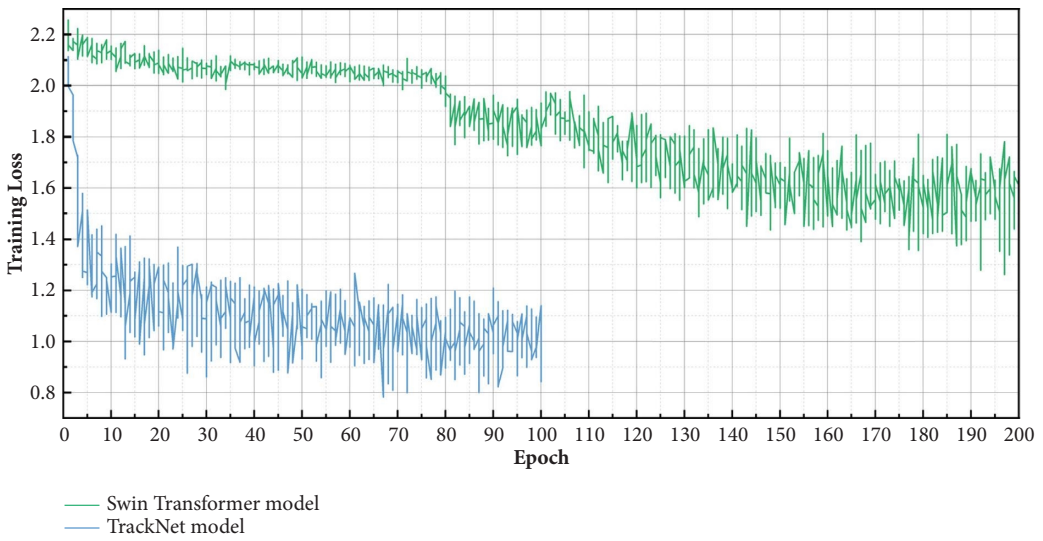
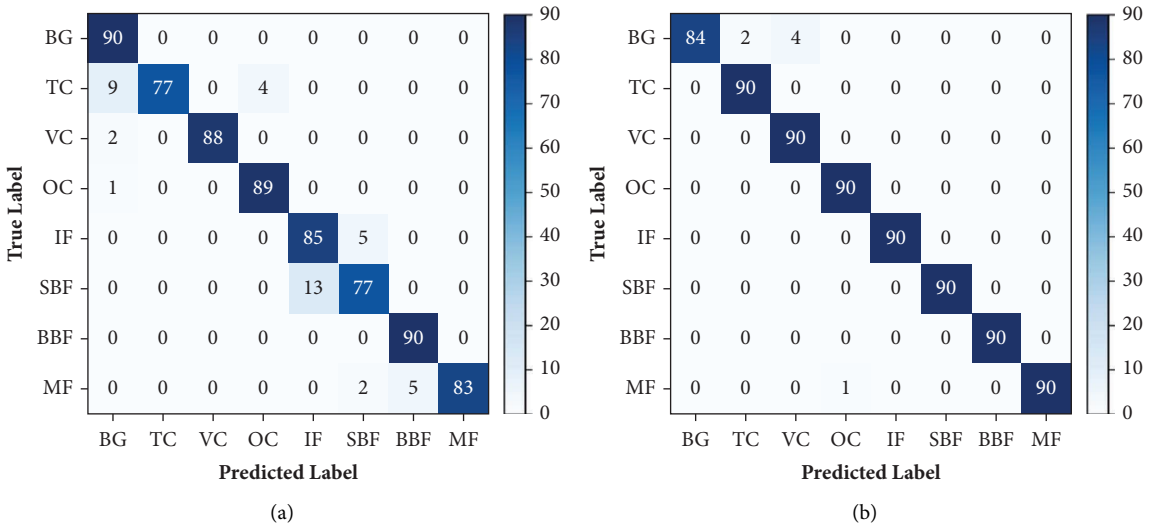


Figure 10: Loss variation curves of various algorithms.



(a) (b)

Figure 11: Confusion matrix of different models in limited test results: (a) Swin Transformer; (b) TrackNet model.

limited test set, particularly in significantly enhancing the detection accuracy of fastener defects.

To enhance the recognition performance of the model for surface defects on ballastless tracks in nighttime conditions, the study conducts a darkening process on the image pixel values in Section 3.1. The section adds the performance of the TrackNet model in recognizing ballastless track defects under different pixel reduction coefficients. In the comparative experiments, the pixel coefficients were set to 0.25, 0.50, and 0.75, where a smaller pixel coefficient indicates a darker image of ballastless track defects. From Table 4, it can be observed that as the pixel coefficient decreases, the recognition performance of the TrackNet model weakens, to some extent indicating that the recognition performance of deep learning models is still closely related to the lighting environment.

The PR and ROC curves serve as tools for assessing the models' performance in BTS defect detection across various thresholds [42]. In the PR curve, recall is plotted on the x-axis, and precision is plotted on the y-axis. The curve closer to the upper right corner indicates better model performance. The Area under the PR curve (AUC_{PR}) serves as an evaluation metric for model performance, where a higher AUC_{PR} value signifies better performance. In the ROC curve, the x-axis is the FPR, and the y-axis is the TPR. Similar to the PR curve, a curve that approaches the upper left corner signifies better model performance. The Area under the ROC curve (AUC_{ROC}) is another evaluation metric for model performance, where a larger AUC_{ROC} indicates superior performance.

Figure 12 displays the PR curves of the Swin Transformer and TrackNet model on the limited test set. The average AUC_{PR} value for the Swin Transformer model to identify all BTS defects is 0.981. The average AUC_{PR} of the proposed TrackNet model is 0.999, which is 1.83% higher than that of the Swin Transformer model. From Figure 12(a), it is evident that the Swin Transformer model has the maximum area under the curve for detecting vertical cracks, suggesting superior performance in detecting vertical cracks. Figure 12(b) illustrates that the TrackNet model performs exceptionally well in detecting all types of BTS defects without the background category.

Figure 13 illustrates the ROC curves of the Swin Transformer and TrackNet model on the limited test set. The Swin Transformer model has an AUC_{ROC} value of 0.989, while the TrackNet model achieves an AUC_{ROC} value of 0.999, representing a 1.01% improvement. From Figures 13(a) and 13(b), it can be observed that the result patterns are similar to the trends seen in the PR curves, emphasizing the outstanding performance of the TrackNet model in detecting BTS defects.

5.2. Result Visualization. The section visualises the model decision-making process and the identification results of the dimensionality reduction. Interpretability studies are essential for researchers to gain a deeper understanding of the decision-making processes of complex models. Therefore, Grad-CAM is introduced to the TrackNet model. This approach leverages gradient information that enters the final convolutional layer of the model to discern the significance of each neuron in the decision-making process. The study prefers to obtain decision heatmaps at different stages. From Figure 14, it can be observed that the decision heatmaps output by the first three stages are not very effective. This is mainly because the shallow convolutional layers usually learn low-level features of the image, such as edges and textures, which may not be sufficiently distinctive for the final classification decision. In contrast, the last convolutional layer typically learns more advanced feature representations, which are more conducive to the model making correct decisions.

Figure 15 shows the decision heatmap of the last convolutional layer output of the different network models. The red regions indicate the important feature areas that the models focus on during defect identification decisions. Specifically, in the recognition process of concrete cracks, the Swin Transformer model focuses on areas mainly distributed at the edges of crack defect features. In contrast, the TrackNet model concentrates its attention on the crack

Table 2: Results of the Swin Transformer model.

Type	Precision (%)	Recall (%)	$F1$-score (%)	Support
BG	88.24	100.00	93.75	90
TC	100.00	85.56	92.22	90
VC	100.00	97.78	98.88	90
OC	95.70	98.89	97.27	90
IF	86.73	94.44	90.42	90
SBF	91.67	85.56	88.51	90
BBF	94.74	100.00	97.30	90
MF	100.00	92.22	95.95	90
Macro avg	94.63	94.31	94.29	720

Table 3: Results of the TrackNet model.

Type	Precision (%)	Recall (%)	$F1$-score (%)	Support
BG	100.00	93.33	96.55	90
TC	97.83	100.00	98.90	90
VC	95.74	100.00	97.82	90
OC	100.00	100.00	100.00	90
IF	100.00	100.00	100.00	90
SBF	100.00	100.00	100.00	90
BBF	100.00	100.00	100.00	90
MF	100.00	100.00	100.00	90
Macro avg	99.20	99.17	99.16	720

Table 4: Results of the TrackNet model with different pixel reduction coefficients.

Coefficient	Precision (%)	Recall (%)	$F1$-score (%)	Support
0.25	97.45	97.22	97.34	720
0.50	99.20	99.17	99.16	720
0.75	99.59	99.58	99.59	720

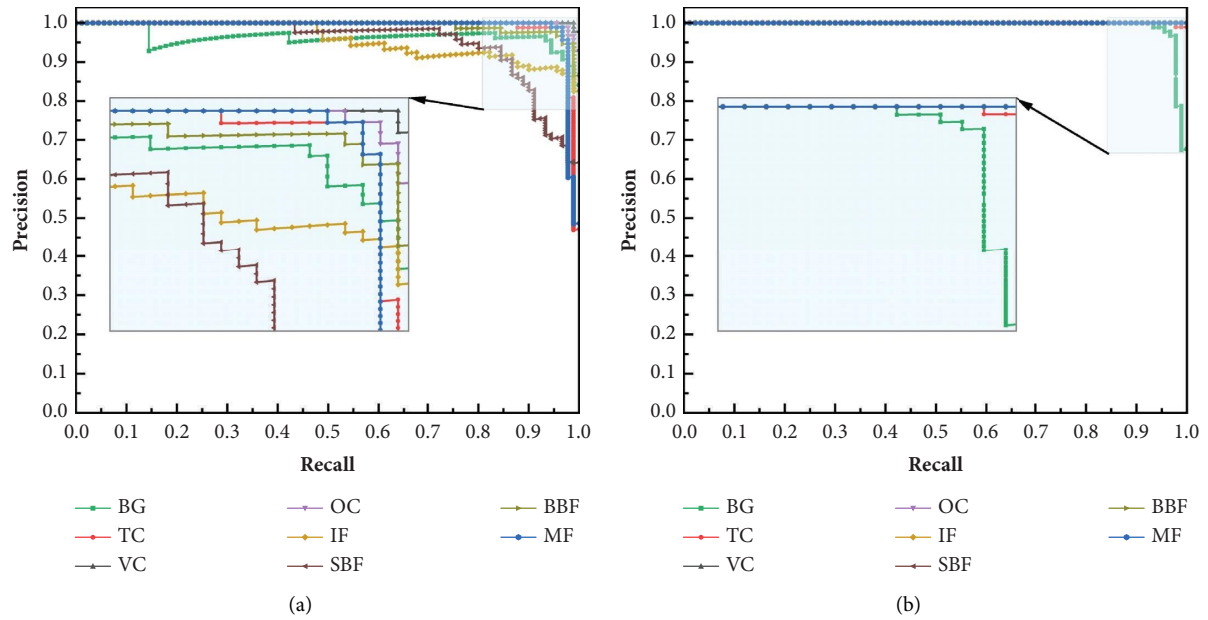


Figure 12: PR curve of different models in limited test results: (a) Swin Transformer model; (b) TrackNet model.

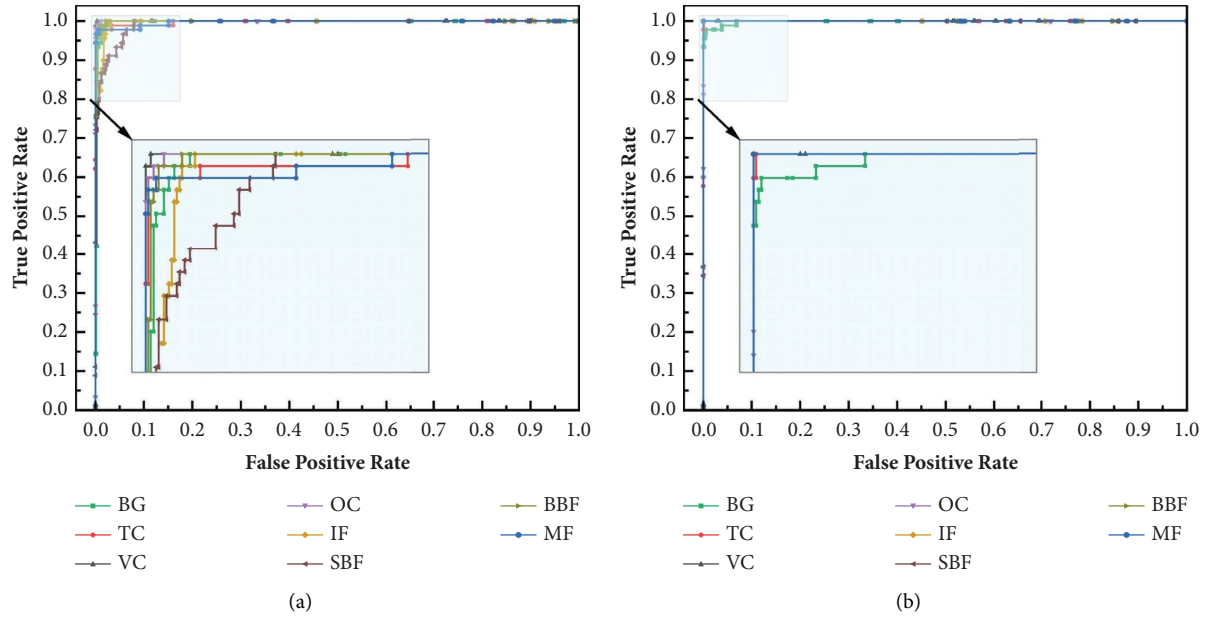


Figure 13: ROC curves of different models in limited test results: (a) Swin Transformer model; (b) TrackNet model.

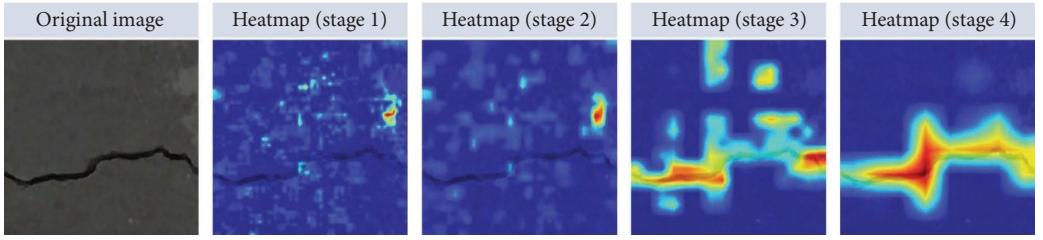


Figure 14: Decision-making heatmaps obtained at different stages.

feature regions. In the recognition process of fastener defects, both models have more widespread attention areas, without a relatively concentrated attention region. This is due to the fastener features occupying a more extensive image area. By comparing the visual interpretation effects of decision regions for different models, it can be observed that the TrackNet model highlights more distinct important features.

The results produced by deep learning models typically present high-dimensional semantic information, making it challenging for individuals to comprehend intuitively. To address this issue, the TrackNet model incorporates t-SNE nonlinear dimensionality reduction technology. This technique can map high-dimensional semantic recognition information to a low-dimensional space. It also preserves the similarity relationships among data points in the high-dimensional space, providing researchers with a more intuitive understanding.

Figure 16 illustrates the recognition results of the Swin Transformer and the TrackNet models on the limited test set, where high-dimensional information is mapped to a 2D space using nonlinear dimensionality reduction technology. The results of the Swin Transformer model in the 2D space are displayed in Figure 16(a). Different crack defects with similar features are clustered in the close region. However, due to the similar features and model recognition errors between intact fasteners and single-broken fastener defects, there is an overlap in the clustered regions. Figure 16(b) presents the results of the TrackNet model in 2D space, where different types of BTS defects are distinctly clustered in their respective independent areas. Compared to the results of the Swin Transformer model, the TrackNet model's outcomes after dimensionality reduction more distinctly separate various defect categories. This indicates that the TrackNet model possesses superior detection performance, allowing for a clear identification of various defect types in the original high-dimensional semantic information.

Figure 17 displays the recognition results of the Swin Transformer and the TrackNet models, using nonlinear dimensionality reduction technology to map high-dimensional information to 3D space. Compared to the data in 2D space, the results in 3D space present information on the distribution of data in three-dimensional directions, with a more pronounced visual impact. The results in 3D space provide more information, contributing to a more comprehensive understanding of the model's ability to recognize different defect categories. The overall distribution of results in 3D space is similar to that in 2D space.

The results of the Swin Transformer model in Figure 17(a) still exhibit spatial overlap when the model recognizes intact fasteners and single-broken fasteners. In Figure 17(b), the recognition results of the TrackNet model show that various types of BTS defects can still be clustered into different regions. It is noteworthy that the overall distribution of crack defects along the z-axis is higher than that of fastener defects in the results of both the Swin Transformer and the TrackNet models. The model's perception of different defect features can be further inferred by observing the relative positions of cracks and fastener defects in the results of dimensionality reduction in 3D space.

5.3. Result Discussion. To verify whether the selection of datasets and training weights affects the detection performance of the model, the section discusses the detection effects of the TrackNet model when loading different pretraining weights. Comparative experiments are conducted using the NEU-DET and ImageNet-1k datasets for pretraining. The NEU-DET dataset contains images of 6 types of surface defects on hot-rolled steel strips, totaling 1800 images [39]. On the other hand, the ImageNet-1k dataset contains 1000 categories, with approximately 1.3 million images.

Table 5 shows the evaluation metrics of the TrackNet model using different pretraining weights for identifying different surface defects on ballastless tracks. The results show that the TrackNet model pretrained with the ImageNet-1k dataset performs excellently in terms of identification with precision, recall, and $F1$-score reaching 99.20%, 99.17%, and 99.16%, respectively. In comparison, the performance of the TrackNet model pretrained with the NEU-DET dataset is slightly inferior.

Transferring pretraining weights from different datasets may result in different computational costs. In this study, pretraining weights obtained from the ImageNet-1k dataset can be acquired by referencing open-source code repositories, thus incurring no additional computational cost [32]. However, if one chooses to pretrain weights using the NEU-DET dataset, an additional 1.77 hours would be required for pretraining on that dataset before training on the BTS defect dataset can commence.

The proposed TrackNet achieves near-perfect results in Section 5.1, especially in OC, IF, SBF, BBF, and MF, where the evaluation metrics reach 100%. To enhance the credibility of the results, the study repeatedly splits the training and test sets. The section conducts an additional three random splits of the training and testing sets while maintaining the same parameters and environment for training and testing. Tables 6–8 present the evaluation metrics of the TrackNet model on the three resplit test datasets. In classification tasks, precision and recall are often conflicting indicators. The $F1$-score can comprehensively consider these two indicators, providing a more comprehensive assessment of the TrackNet model's performance. From the additional experimental results in this study, it can be seen that the $F1$-score of the TrackNet model in identifying OC, IF, SBF, and BBF defects did not reach 100% in every test. However, it is undeniable that the $F1$-score of the model for MF defects remained at 100%. To further enhance the credibility of the TrackNet model's test results, this study further analyzes the similarity between MF defect images. Nine randomly selected MF defect images from the training and testing sets, as shown in Figure 18, demonstrate the pixel distribution characteristics of these images using histograms. It can be observed from the Figure 18 that the pixel values of the MF defect images are mainly distributed around 50, indicating a certain degree of similarity in this type of damage. In future research, the authors will collect a more extensive range of images of ballastless track defects from railway test sites or laboratories to train the model and improve its generalization ability.

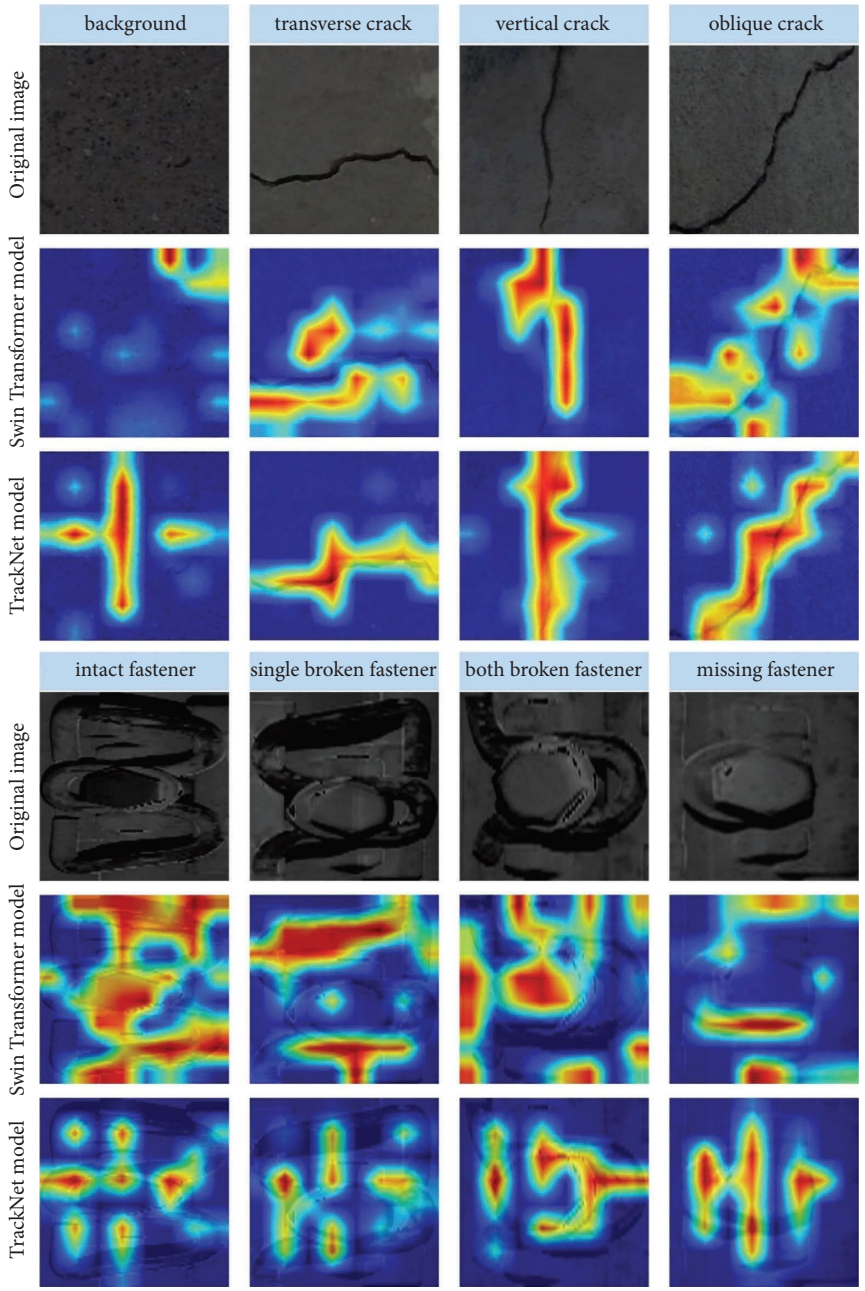


Figure 15: Heatmaps of different models in recognizing defects.

5.4. Comparison with Other Algorithms. To explore the superiority of the TrackNet model in detecting BTS defects, comparative experimental studies are conducted in this section. EfficientNet [43], Vision Transformer [44], and Swin Transformer [32] are all advanced intelligent detection models and are used for comparative research. Figure 19 shows the results of different detection algorithms on a limited test set of BTS defects. Among all the detection results, the Vision Transformer model has the lowest average accuracy, precision, recall, and $F1$-score. This is because models based on self-attention mechanisms often require a large number of training samples. The proposed TrackNet model effectively overcomes this challenge, achieving high performance on the limited dataset of BTS defects through transfer learning and multihead self-attention mechanisms. Compared to the results of the EfficientNet, Vision Transformer, and Swin Transformer models, the accuracy of the TrackNet model has increased by 0.15%, 11.39%, and 5.15%, respectively.

6. Engineering Applications

The TrackNet model with a global attention mechanism obtained through transfer learning demonstrates outstanding advantages in detection performance. In this section, engineering application validation is conducted using actual nighttime environmental images.

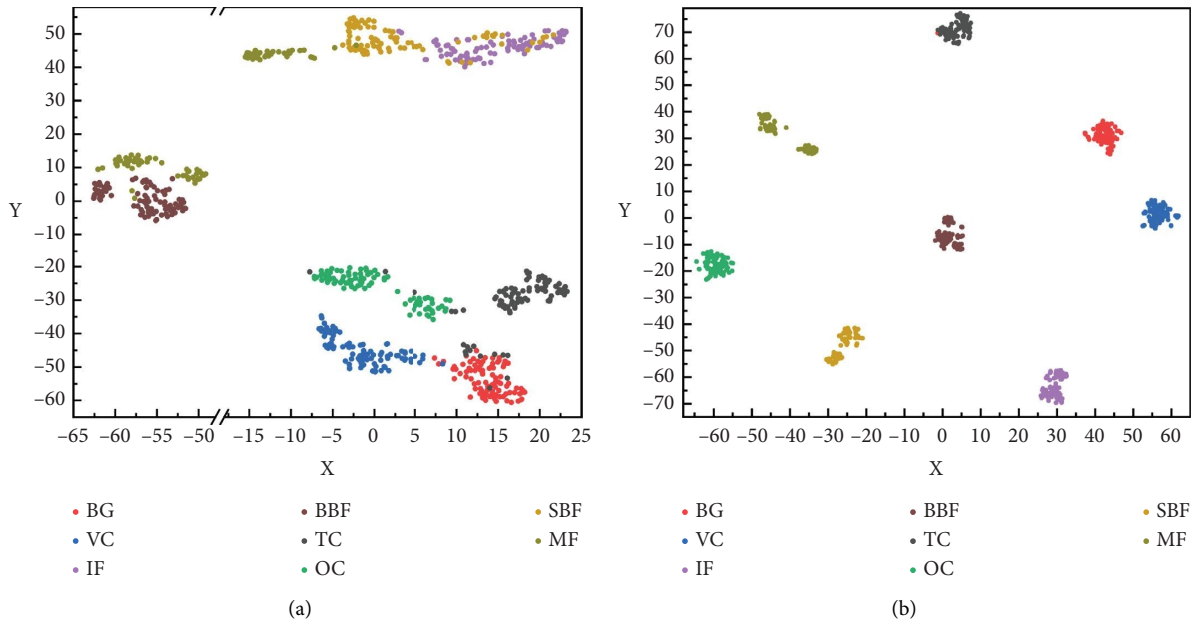


Figure 16: Results of dimensionality reduction in 2D space: (a) Swin Transformer model; (b) TrackNet model.

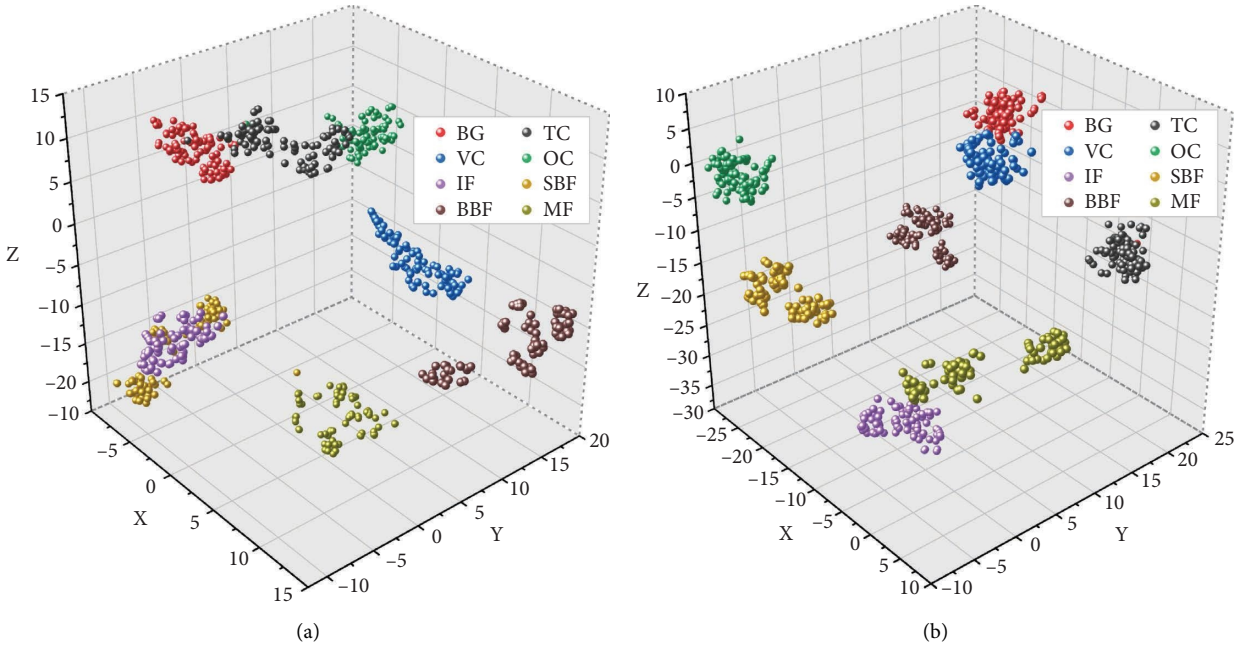


Figure 17: Results of dimensionality reduction in 3D space: (a) Swin Transformer model; (b) TrackNet model.

Table 5: Evaluation metrics of the TrackNet model with different pretraining weights on the test set.

Weight	Precision (%)	Recall (%)	*F*1-score (%)	Support
NEU-DET	95.00	94.86	94.93	720
ImageNet-1k	99.20	99.17	99.16	720

To further validate the scientific validity and effectiveness of the proposed TrackNet model, this study selected nighttime environmental images collected from a certain high-speed railway line in China and the full-scale test platform of ballastless track at SWJTU for applied research. Figure 20 illustrates the full-scale ballastless track structure

Table 6: Test results of split dataset I.

Type	Precision (%)	Recall (%)	$F1$-score (%)	Support
BG	100.00	88.89	94.12	90
TC	98.90	100.00	99.45	90
VC	91.75	98.89	95.19	90
OC	98.90	100.00	99.45	90
IF	100.00	97.78	98.88	90
SBF	97.83	100.00	98.90	90
BBF	98.90	100.00	99.45	90
MF	100.00	100.00	100.00	90
Macro avg	98.29	98.19	98.18	720

Table 7: Test results of split dataset II.

Type	Precision (%)	Recall (%)	$F1$-score (%)	Support
BG	100.00	88.89	94.12	90
TC	95.74	100.00	97.83	90
VC	93.68	98.89	96.22	90
OC	100.00	99.45	99.72	90
IF	100.00	100.00	100.00	90
SBF	100.00	100.00	100.00	90
BBF	100.00	100.00	100.00	90
MF	100.00	100.00	100.00	90
Macro avg	98.68	98.40	98.49	720

Table 8: Test results of split dataset III.

Type	Precision (%)	Recall (%)	$F1$-score (%)	Support
BG	100.00	95.56	97.73	90
TC	97.83	100.00	98.90	90
VC	96.74	98.89	97.80	90
OC	100.00	100.00	100.00	90
IF	100.00	98.89	99.44	90
SBF	98.90	100.00	99.45	90
BBF	100.00	100.00	100.00	90
MF	100.00	100.00	100.00	90
Macro avg	99.18	99.17	99.16	720

model established indoors at SWJTU. The entire ballastless track model is approximately 20 m long, consisting of two structures: CRTS I-type slab ballastless track and CRTS III-type slab ballastless track. The surface of the ballastless track exhibits predefined cracks and fastener defects.

This study utilizes a high-definition camera to capture images of surface cracks and fastener defects on ballastless tracks during nighttime and conduct applied validation. The upper part of Figure 21 demonstrates concrete crack images on the track slab and roadbed collected by the collaborative team during nighttime maintenance hours on a certain high-speed railway line. In the images, it is visible that actual nighttime capture scenes exhibit significant interference from shadows and illuminated areas. These interfering factors pose a considerable challenge to the detection of surface defects on ballastless tracks. The proposed TrackNet, leveraging its outstanding global feature extraction capability, successfully visualized the concrete crack regions.

The lower part of Figure 21 showcases images of fastener defects collected by our team during nighttime at the SWJTU high-speed ballastless track full-scale test platform. Through visual results, it can be observed that the attention of the TrackNet model is concentrated within a small area of the fasteners, validating the inference in Section 5.2. When defect areas occupy a large portion of the image, the model's attention tends to be more widespread. However, in engineering applications, defect areas typically occupy only a small portion of the image, leading the model's attention to be more focused. The application validation of the TrackNet model demonstrates its effectiveness in detecting surface defects on ballastless tracks, providing crucial insights for the identification of surface defects on high-speed railway ballastless tracks.

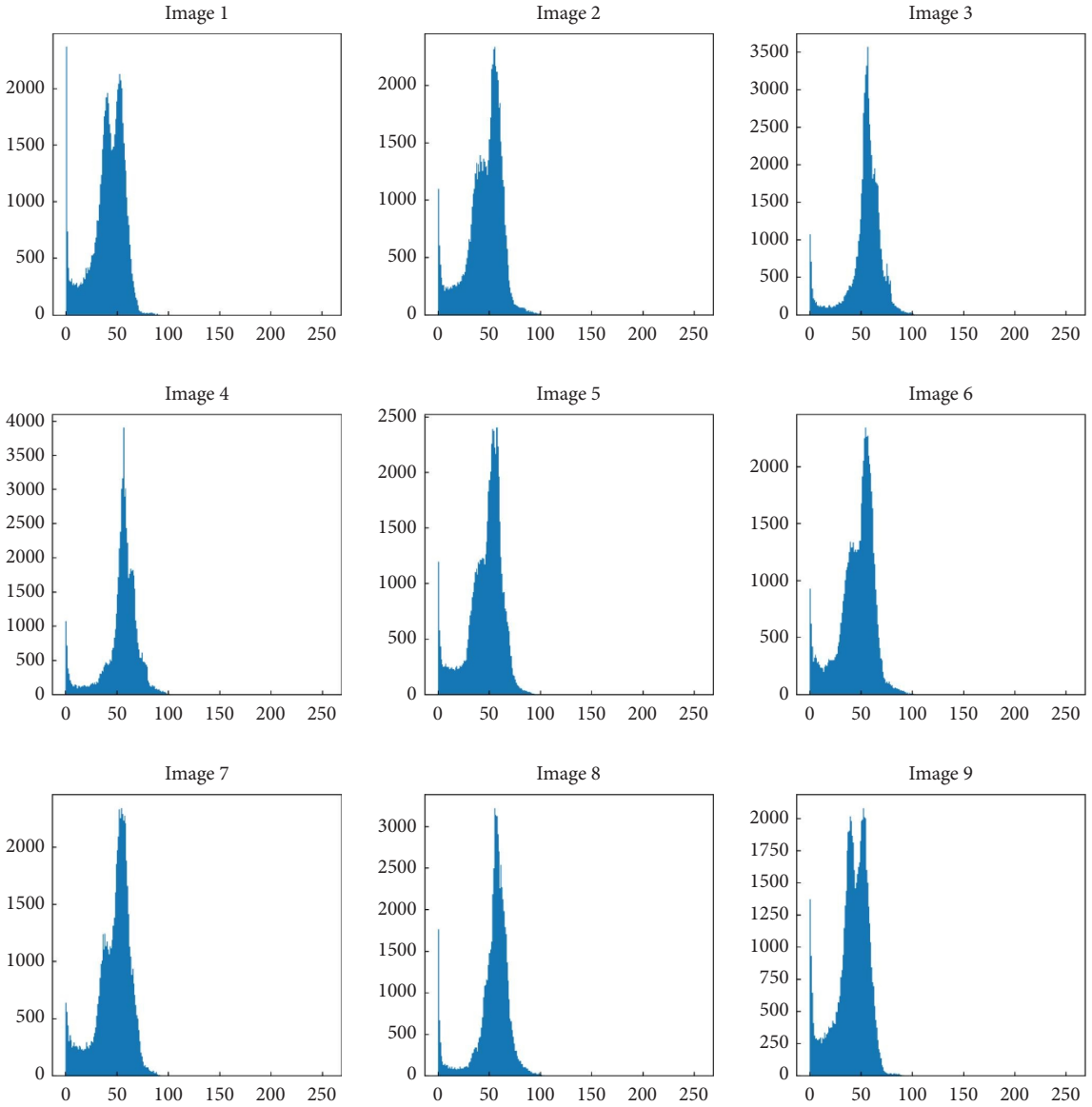


Figure 18: Analysis of similarity in images of BTS defects (with MF defects as an example).

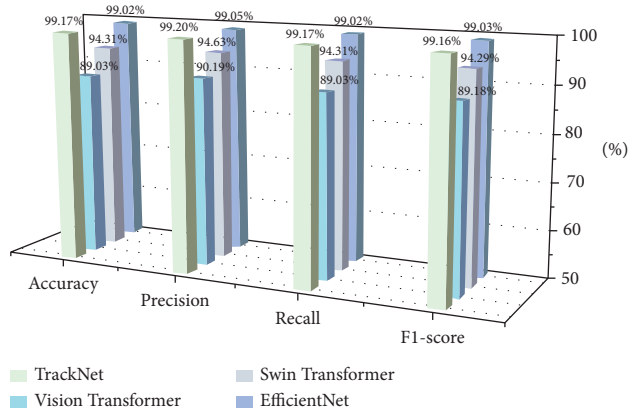


Figure 19: Comparison results of different detection algorithms.

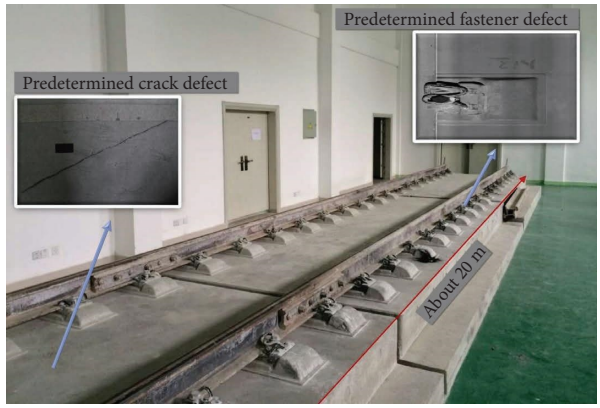


Figure 20: Ballastless track full-scale test platform at SWJTU.

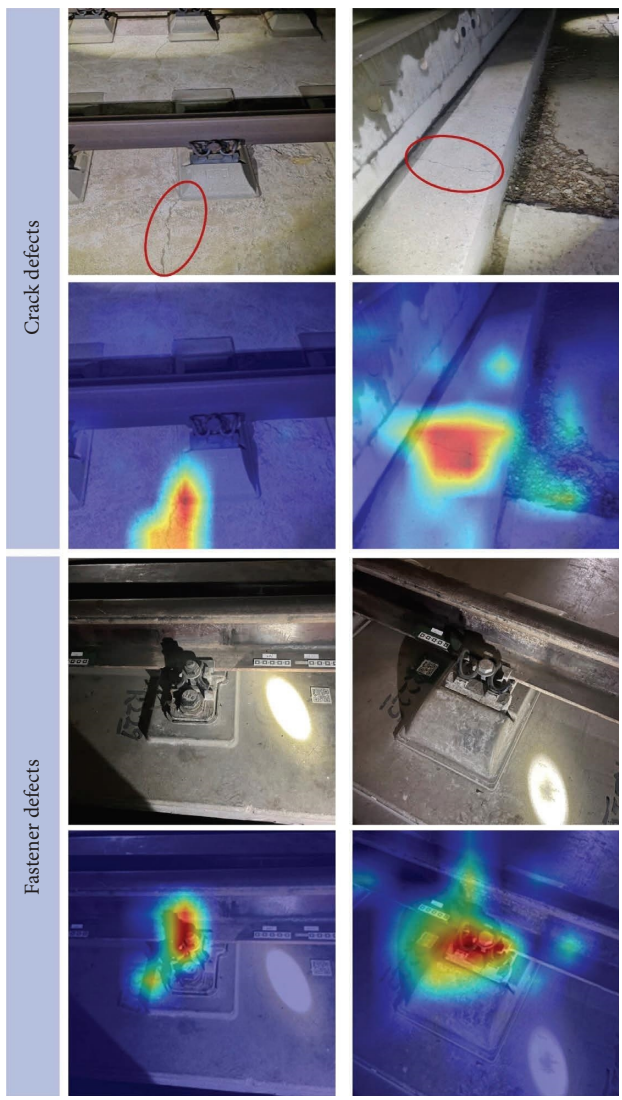


Figure 21: Results of engineering applications.

7. Conclusion

This study aims to detect limited surface defects of ballastless tracks through transfer learning and conduct interpretability research on model decisions and high-dimensional information. Based on experiments and evaluations conducted on a constructed limited nighttime defect dataset, the following conclusions were drawn:

(1) In the context of railway nighttime detection, this study uniformly pixel-reduced three-channel color images to construct a multiclass defect dataset for nighttime BTS defects. An interpretable intelligent detection method for BTS defects based on transfer learning is proposed. This method can effectively alleviate the shortcomings of deep learning in recognizing small-sample infrastructure defects at night by loading pretrained weights obtained on publicly available large datasets.

(2) The proposed TrackNet model in the paper enhances the extraction of global features of BTS defects in limited test images covering eight categories of BTS defects. Its average accuracy, precision, recall, and $F1$-score are 99.17%, 99.20%, 99.17%, and 99.16%, respectively. When compared to the detection outcomes of the Swin Transformer algorithm, it shows an improvement of 5.15%, 4.82%, 5.15%, and 5.16%, respectively. This indicates that the transfer learning model exhibits better detection performance on a limited test set.

(3) When performing transfer learning, it is advisable to prioritize pretraining weights from large-scale datasets available in open-source code repositories. This approach can improve the model's detection performance while avoiding additional computational costs.

(4) In terms of interpreting model decisions, this study visualizes the decision regions of the TrackNet model in recognizing BTS defects using heatmaps, revealing the black-box recognition mechanism of deep learning models. Additionally, the nonlinear dimensionality reduction technique shows the model's recognition results mapped from high-dimensional space to 2D or 3D space, aiding in the understanding of abstract data distributions.

The authors plan to develop an intelligent detection device for ballastless tracks in the future, but this is a long-term research project. The next specific work is mainly divided into two aspects. Firstly, due to the limited collection of track defect data in this study, which may affect the detection accuracy to some extent, the authors will collect more extensive images of ballastless track defects from railway test sites or laboratories to train the model and improve its generalization ability. Secondly, the authors will use the TrackNet model as the backbone network and combine it with object detection algorithms to achieve defect position localization. Finally, the model will be deployed in the intelligent detection vehicle for ballastless track defects.

Data Availability

The data are not publicly available due to project requirements.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Wenlong Ye was responsible for original draft preparation, review and editing, methodology, and data curation. Juanjuan Ren was responsible for funding acquisition. Chen Li, Wengao Liu, and Zeyong Zhang were responsible for data curation. Chunfang Lu was responsible for supervision.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (no. 52278461), the National Key Research and Development Program of China (no. 2021YFF0502100), the Sichuan Province Youth Science and Technology Innovation Team (no. 2022JDTD0015), and the Research and Development Program of China Railway Design Corporation (no. 2023A0223804).

References

[1] C. Lu, "A discussion on technologies for improving the operational speed of high-speed railway networks," *Transportation Safety and Environment*, vol. 1, pp. 22–36, 2019.

[2] W. Ye, J. Ren, P. Zhang, Q. Zhang, and L. Li, "Review of integrated full life cycle data management and application of the slab tracks," *Intelligent Transportation Infrastructure*, vol. 1, pp. 1–14, 2022.

[3] X. Dai, *3D Detection System Integration and Recognition Algorithm for High-Speed Railway Fastener Defects*, Southwest Jiaotong University, Chengdu, China, 2018.

[4] Y. J. Cha, W. Choi, G. Suh, S. Mahmoudkhani, and O. Büyüköztürk, "Autonomous structural visual inspection using region-based deep learning for detecting multiple damage types," *Computer-Aided Civil and Infrastructure Engineering*, vol. 33, no. 9, pp. 731–747, 2018.

[5] Y. Wu, Y. Qin, Y. Qian, and F. Guo, "Automatic detection of arbitrarily oriented fastener defect in high-speed railway," *Automation in Construction*, vol. 131, Article ID 103913, 2021.

[6] W. Ye, S. Deng, J. Ren, X. Xu, K. Zhang, and W. Du, "Deep learning-based fast detection of apparent concrete crack in slab tracks with dilated convolution," *Construction and Building Materials*, vol. 329, Article ID 127157, 2022.

[7] M. Karakose, O. Yaman, M. Baygin, K. Murat, and E. Akin, "A new computer vision based method for rail track detection and fault diagnosis in railways," *International Journal of Mechanical Engineering and Robotics Research*, vol. 6, pp. 22–27, 2017.

[8] H. Oliveira and P. L. Correia, "Road surface crack detection: improved segmentation with pixel-based refinement," in *Proceedings of the 2017 25th European Signal Processing Conference (EUSIPCO)*, IEEE, Kos island, Greece, September 2017.

[9] A. Akagic, E. Buza, S. Omanovic, and A. Karabegovic, "Pavement crack detection using Otsu thresholding for image

segmentation," in *Proceedings of the 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, IEEE, Opatija, Croatia, May 2018.

[10] H. Ma, Y. Min, C. Yin et al., "A real time detection method of track fasteners missing of railway based on machine vision," *International Journal of Performability Engineering*, vol. 14, pp. 1190–1200, 2018.

[11] Y. Santur, M. Karaköse, and E. Akin, "A new rail inspection method based on deep learning using laser cameras," in *Proceedings of the 2017 International Artificial Intelligence and Data Processing Symposium (IDAP)*, IEEE, Malatya, Turkey, September 2017.

[12] A. James, W. Jie, Y. Xulei et al., "Tracknet-a deep learning based fault detection for railway track inspection," in *Proceedings of the 2018 International Conference on Intelligent Rail Transportation (ICIRT)*, IEEE, Singapore, December 2018.

[13] W. Wang, W. Hu, W. Wang et al., "Automated crack severity level detection and classification for ballastless track slab using deep convolutional neural network," *Automation in Construction*, vol. 124, pp. 103484–103517, 2021.

[14] F. Guo, Y. Qian, and Y. Shi, "Real-time railroad track components inspection based on the improved YOLOv4 framework," *Automation in Construction*, vol. 125, pp. 103596–103615, 2021.

[15] F. Guo, Y. Qian, Y. Wu, Z. Leng, and H. Yu, "Automatic railroad track components inspection using real-time instance segmentation," *Computer-Aided Civil and Infrastructure Engineering*, vol. 36, no. 3, pp. 362–377, 2021.

[16] Y. Wu, Y. Qin, Y. Qian, F. Guo, Z. Wang, and L. Jia, "Hybrid deep learning architecture for rail surface segmentation and surface defect detection," *Computer-Aided Civil and Infrastructure Engineering*, vol. 37, no. 2, pp. 227–244, 2022.

[17] Y. Wu, P. Chen, Y. Qin, Y. Qian, F. Xu, and L. Jia, "Automatic railroad track components inspection using hybrid deep learning framework," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–15, 2023.

[18] X. Cai, X. Tang, S. Pan et al., "Intelligent recognition of defects in high-speed railway slab track with limited dataset," *Computer-Aided Civil and Infrastructure Engineering*, 2023.

[19] W. Ye, J. Ren, A. A. Zhang, and C. Lu, "Automatic pixel-level crack detection with multi-scale feature fusion for slab tracks," *Computer-Aided Civil and Infrastructure Engineering*, vol. 38, no. 18, pp. 2648–2665, 2023.

[20] A. A. Zhang, K. C. Wang, Y. Liu et al., "Intelligent pixel-level detection of multiple distresses and surface design features on asphalt pavements," *Computer-Aided Civil and Infrastructure Engineering*, vol. 37, no. 13, pp. 1654–1673, 2022.

[21] T. S. Tran, S. D. Nguyen, H. J. Lee, and V. P. Tran, "Advanced crack detection and segmentation on bridge decks using deep learning," *Construction and Building Materials*, vol. 400, Article ID 132839, 2023.

[22] E. Zhang, L. Shao, and Y. Wang, "Unifying transformer and convolution for dam crack detection," *Automation in Construction*, vol. 147, Article ID 104712, 2023.

[23] A. Vaswani, N. Shazeer, N. Parmar et al., "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[24] P. P. Ray, "ChatGPT: a comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope," *Internet of Things and Cyber-Physical Systems*, 2023.

[25] Y. Bie and H. Tan, "Image recognition in autonomous driving based on improved Swin transformer," in *Proceedings of the 2022 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*, IEEE, Dalian, China, June 2022.

[26] Q. Song, B. Sun, and S. Li, "Multimodal sparse transformer network for audio-visual speech recognition," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 12, pp. 10028–10038, 2023.

[27] F. Zhuang, Z. Qi, K. Duan et al., "A comprehensive survey on transfer learning," *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, 2021.

[28] E. Asadi Shamsabadi, C. Xu, A. S. Rao, T. Nguyen, T. Ngo, and D. Dias-da-Costa, "Vision transformer-based autonomous crack detection on asphalt and concrete surfaces," *Automation in Construction*, vol. 140, Article ID 104316, 2022.

[29] H. Pan, Z. Pang, Y. Wang, Y. Wang, and L. Chen, "A new image recognition and classification method combining transfer learning algorithm and mobilenet model for welding defects," *IEEE Access*, vol. 8, pp. 119951–119960, 2020.

[30] S. Bunrit, N. Kerdprasop, and K. Kerdprasop, "Improving the representation of cnn based features by autoencoder for a task of construction material image classification," *Journal of Advances in Information Technology*, vol. 11, no. 4, pp. 192–199, 2020.

[31] F. Liu and L. Wang, "UNet-based model for crack detection integrating visual explanations," *Construction and Building Materials*, vol. 322, Article ID 126265, 2022.

[32] Z. Liu, Y. Lin, Y. Cao et al., "Swin transformer: hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Montreal, BC, Canada, October 2021.

[33] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision*, Cambridge, MA, USA, June 2017.

[34] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, 2008.

[35] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *ICLR*, vol. 2014, pp. 1–14, 2015.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, June 2016.

[37] Ç.F. Özgenel and A. G. Sorguç, "Performance comparison of pretrained convolutional neural networks on crack detection in buildings, ISARC," in *Proceedings of the International Symposium on Automation and Robotics in Construction*, IAARC Publications, Berlin, Germany, July 2018.

[38] Y. Zhan, X. Dai, E. Yang, and K. C. Wang, "Convolutional neural network for detecting railway fastener defects using a developed 3D laser system," *International Journal of Reality Therapy*, vol. 9, no. 5, pp. 424–444, 2021.

[39] Y. He, K. Song, Q. Meng, and Y. Yan, "An end-to-end steel surface defect detection approach via fusing multiple hierarchical features," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 4, pp. 1493–1504, 2020.

[40] R. Fu, M. Cao, D. Novák, X. Qian, and N. F. Alkayem, "Extended efficient convolutional neural network for concrete crack detection with illustrated merits," *Automation in Construction*, vol. 156, Article ID 105098, 2023.

[41] R. Müller, S. Kornblith, and G. E. Hinton, "When does label smoothing help?" *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[42] S. Zhou, C. Canchila, and W. Song, "Deep learning-based crack segmentation for civil infrastructure: data types, architectures, and benchmarked performance," *Automation in Construction*, vol. 146, Article ID 104678, 2023.

[43] M. Tan and Q. Le, "Efficientnet: rethinking model scaling for convolutional neural networks," *International Conference on Machine Learning*, vol. 6105, 2019.

[44] A. Dosovitskiy, L. Beyer, A. Kolesnikov et al., "An image is worth 16x16 words: transformers for image recognition at scale," *International Conference on Learning Representations*, 2021.