algorithms

MDPI

Article

The Fine Feature Extraction and Attention Re-Embedding Model Based on the Swin Transformer for Pavement Damage Classification

Shizheng Zhang [1,*], Kunpeng Wang [1], Zhihao Liu [1], Min Huang [1] and Sheng Huang [2]

[1] Software Engineering College, Zhengzhou University of Light Industry, 136 Science Avenue, Zhengzhou 450000, China; 332316040988@zzuli.edu.cn (K.W.); 332213020801@zzuli.edu.cn (Z.L.); huangmin@zzuli.edu.cn (M.H.)
[2] School of Big Data and Software Engineering, Chongqing University, 174 Shazheng Street, Chongqing 400044, China; huangsheng@cqu.edu.cn
* Correspondence: zshizheng@zzuli.edu.cn

Abstract: The accurate detection and classification of pavement damage are critical for ensuring timely maintenance and extending the service life of road infrastructure. In this study, we propose a novel pavement damage recognition model based on the Swin Transformer architecture, specifically designed to address the challenges inherent in pavement imagery, such as low damage visibility, varying illumination conditions, and highly similar surface textures. Unlike the original Swin Transformer, the proposed model incorporates two key components: a fine feature extraction module and a multi-head self-attention re-embedding module. These additions enhance the model's ability to capture subtle and complex damage patterns. Experimental evaluations demonstrate that the proposed model achieves a 2.07% improvement in classification accuracy and a 0.97% increase in F1 score compared to the baseline while maintaining comparable computational complexity. Overall, the model significantly outperforms the baseline Swin Transformer in pavement damage detection and classification, highlighting its practical applicability.

Keywords: pavement damage detection; feature extraction optimization; multi-head self-attention re-embedding; Swin Transformer

check for updates

Academic Editor: Igor Aizenberg

Received: 14 May 2025
Revised: 11 June 2025
Accepted: 13 June 2025
Published: 18 June 2025

Citation: Zhang, S.; Wang, K.; Liu, Z.; Huang, M.; Huang, S. The Fine Feature Extraction and Attention Re-Embedding Model Based on the Swin Transformer for Pavement Damage Classification. *Algorithms* **2025**, *18*, 369. https://doi.org/10.3390/a18060369

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Pavement damage detection and classification are complex and challenging tasks in road maintenance engineering. With the widespread use of asphalt pavements and the construction of highways, this task has become increasingly important, significantly impacting the safe driving of motorists. Early pavement damage detection and classification primarily relied on manual inspection by professional personnel, who used empirical knowledge to conduct on-site measurements and visual inspections. This manual approach was inefficient and required substantial human effort, with results largely influenced by subjective judgment, causing inconsistencies in detection quality and elevated labor expenses. As a result, it was impractical for large-scale pavement detection [1]. With the development of image technology, the detection process evolved from fully manual to semi-automated. This semi-automated method utilizes high-speed downward-facing digital cameras installed on dedicated data collection vehicles to acquire pavement imagery [2]. For example, France's GERPHO road detection system [3] uses a 35 mm film GPS locator on a data collection vehicle to gather images of road surfaces. Professional staff then use indoor equipment to assess road damage and establish a database, thus moving away

from traditional manual methods and reducing the waste of human and traffic resources in maintenance inspections. Similarly, Japan's Komatsu road detection system [4] and Sweden's PAVUE detection vehicle [5] can perform semi-automated detection. While these methods save time by eliminating the need for on-site inspections, they still require human resources to analyze the collected digital images. This process is time-consuming, relies on the expertise of professional inspectors, consumes significant human resources, and remains inefficient, making it unsuitable for large-scale pavement detection.

With advancements in computer technology and its widespread adoption, digital image processing has emerged as an efficient method for pavement damage detection and classification, saving both human and material resources. The current development trend focuses on automating and mechanizing this process by employing machine learning, deep learning, and pattern recognition techniques to enhance its adaptability and robustness. Early intelligent detection technologies primarily relied on image processing techniques, manual feature extraction, and conventional classifiers to transform, process, and analyze digital images for pavement damage classification. Typically, image preprocessing is performed using conventional image processing techniques, followed by the application of dedicated feature extraction algorithms and classifiers to identify the presence of cracks and categorize their types, thereby achieving pavement damage classification detection. For instance, Ja et al. presented a method utilizing moment invariants combined with neural networks to process pavement images [6], demonstrating the practicality of this combination in distinguishing various crack types. This technique involves calculating the moment invariants corresponding to different pavement crack forms to extract relevant features, which are then input into a neural network for classification. Likewise, Nejad et al. proposed an expert system that integrates wavelet transform with a Radon neural network for pavement distress classification [7]. This system enhances the performance of scale-invariant feature extraction by computing the wavelet modulus, followed by transformation through the Radon neural network, with the resulting feature peaks and parameters used for training and testing. Lv et al. developed a model based on the Mask R-CNN which was effectively applied to road crack detection and hazard identification [8]. Liu et al. introduced a lightweight GAN architecture [9], aimed at achieving high computational efficiency with low processing cost for automated pavement damage detection. This design integrates several efficient modules, including Squeeze and Expand (SE), Multi-Scale Convolution (MC), and depthwise separable convolution (DSC). The SE module adaptively modulates channel-wise weights to enhance feature representation, the MC module extracts image features at various scales, and the DSC reduces both parameter count and computational load, thus improving overall network efficiency. This architecture ensures high detection accuracy while significantly lowering complexity and resource demands, making it well-suited for deployment in real-world, resource-constrained environments. Xu et al. proposed an innovative neural network architecture named the Selective Feature Fusion and Irregular Perception Network for pavement crack detection [10], which selectively fuses multi-level features, regulates the transmission of salient information at different stages, and effectively models irregular crack structures. Nonetheless, these methods still exhibit certain limitations, such as high reliance on domain expertise, the need for separate optimization of feature extraction and classification processes, complex preprocessing steps, and relatively limited generalization ability.

With its successful development and growing adoption, deep learning technology has been increasingly applied to pavement damage detection and classification. Due to the unique properties of pavement damage images such as high resolution, a low proportion of damaged areas, and the limitations in enabling end-to-end training. Sheng Huang et al. proposed a series of straightforward yet effective end-to-end deep learning

approaches named WSPLINs [11]. This framework initially segments pavement images into patches at multiple scales using different acquisition strategies, and subsequently employs a patch label inference network (PLIN) to predict labels for these patches. In another study, Zhun et al. [12] developed an automated pavement crack detection technique based on structured prediction utilizing a CNN [13–16]. This is a supervised algorithm based on deep learning, where the CNN learns and trains on raw pavement crack images, modeling crack detection as a multi-class label problem. This approach achieves effective pavement damage crack classification and detection without requiring additional preprocessing steps, showing certain performance improvements. While these methods have shown some effectiveness in applying deep learning to pavement damage detection and classification, they still have certain shortcomings. For instance, most of these methods involve deep learning models wherein pavement damage detection and classification are treated as simple object detection or image classification problems, with relatively little consideration for the subtle features of pavement damage images [17].

Pavement images often exhibit high similarity between normal and damaged areas. Specially, normal images and damaged images differ only slightly in the damaged areas, which occupy a small proportion of the entire pavement image. In addition, different types of cracks (such as alligator cracks, transverse cracks, and longitudinal cracks) have high similarity, making it difficulty to categorize the damage accurately. To overcome the mentioned problems, this paper proposes a novel method named fine feature extraction and attention re-embedding (FFEAR) based on the Swin Transformer for pavement damage detection and classification. Overall, FFEAR surpasses the original Swin Transformer and offers a specialized approach for pavement damage detection and classification.

Firstly, we developed a specialized model, FFEAR, based on the CQU-BPDD pavement damage dataset. Building on the Swin Transformer architecture, this model is specifically optimized for pavement damage detection and classification. To tackle challenges unique to pavement images such as low damage visibility, high similarity between crack types, and uneven lighting we introduced a multi-head self-attention feature re-embedding mechanism. This mechanism enhances feature extraction, improving the model's ability to capture critical damage characteristics and achieve higher classification accuracy.

Secondly, the fine feature extraction module named FFE integrates depthwise separable convolutions to reprocess features obtained from the multi-head self-attention mechanism. This allows the model to capture small-scale features that may have been previously overlooked. The re-extracted features are subsequently re-embedded, thereby strengthening the model's representational capacity. As a result, the model becomes more effective at distinguishing and categorizing various feature types, particularly subtle crack patterns, leading to enhanced recognition and classification performance.

Finally, extensive experiments show that the FFEAR model significantly outperforms the baseline Swin Transformer model. The FFEAR model exhibits a notable 2.07% increase in accuracy and a 0.97% enhancement in F1 score based on the CQU-BPDD dataset.

These results underscore the model's superior performance in accurately detecting and classifying pavement damage. The introduction of depthwise separable convolutions and multi-head self-attention mechanisms has proven effective in enhancing the model's ability to capture subtle features, making it a highly capable and specialized model for pavement damage detection and classification tasks.

2. Related Works

2.1. Pavement Distress Classification

Recently, the swift progress of intelligent transportation systems and advancement in autonomous driving technologies have drawn significant attention to the research on

pavement damage detection and classification. Pavement damage not only affects driving safety but also accelerates vehicle wear and tear. Traditional manual inspection approaches are both time-consuming and labor-intensive, rendering them impractical for large-scale applications. As a result, automated detection techniques utilizing computer vision and deep learning have emerged as prominent research directions. The integration of deep learning, particularly CNNs, has notably enhanced the accuracy of pavement damage classification. For instance, Maeda et al. [18] developed a detection method based on deep CNNs, using large-scale datasets to train models capable of automatically identifying various damage types, including cracks and potholes. Benedetto et al. [19] investigated the use of U-Net-based semantic segmentation networks for pavement damage detection. Utilizing fully convolutional networks (FCNs [20–22]), they achieved pixel-level classification, enabling not only the detection of damage but also the precise localization of the damaged areas. With the development of technologies like multimodal data fusion, Alessandro et al. [23] proposed a multi-sensor fusion approach for road damage detection, which integrates RGB images and data from multispectral imaging sensors. By leveraging depth information from multispectral sensors to supplement incomplete visual data, they achieved notable improvements in both detection accuracy and model robustness. This multi-sensor fusion approach enhances the model's adaptability to complex environments while ensuring more accurate and reliable detection outcomes. Gopalakrishnan et al. [24] explored the combination of thermal imaging with traditional visible light images to detect sub-surface pavement damage. They discovered that fusing multispectral data allows for the earlier detection of potential structural issues in the pavement. In the realm of supervised learning and multiple-instance learning, Tang et al. [25] proposed an ultra-weakly supervised visual Transformer named PicT for pavement damage classification based on the Swin Transformer [26]. They developed a new pavement image classification Transformer (PicT) for this task. To better leverage the discriminative information at the patch level in pavement images, they proposed a patch label teacher model. In addition to computer vision-based detection approaches, Abubeker et al. [27] proposed a linear viscoelastic (LVE) method for simulating pavement responses under moving loads. This method is based on the multi-layered elastic theory (MLET) combined with the elastic–viscoelastic correspondence principle. By employing a numerical inverse Laplace transform, the model enables the computation of time-dependent stress–strain responses and incorporates a time-collocation interpolation scheme to improve computational efficiency. Gkyrtis et al. [28] developed a comprehensive mechanistic framework that integrates both elastic and viscoelastic theories for the in situ performance evaluation of asphalt pavements. The study emphasized the impact of material modeling assumptions (elastic vs. viscoelastic) on strain prediction and subsequent maintenance decision-making. Based on a combination of Falling Weight Deflectometer (FWD) and Ground-Penetrating Radar (GPR) data, along with core sample testing, the authors constructed master curves using the Huet–Sayegh rheological model and performed dynamic response simulations using the ViscoRoute software. These approaches collectively offer various solutions to the problem of pavement damage classification from different perspectives. The model generates pseudo-labels for image patches dynamically in each iteration, enabling the weakly supervised learning of discriminative patch features through patch label inference. These approaches collectively offer various solutions to the problem of pavement damage classification from different perspectives.

2.2. Swin Transformer

Transformers were initially developed for applications in machine translation and natural language processing [29,30]. The powerful self-attention mechanism of Transformers has been well-received across various tasks. As the technology has evolved, it has gradually

been applied to the field of computer vision. For example, Chen et al. [31] proposed the DePatch, which adaptively partitions images into patches of varying positions and scales in a data-driven way, replacing the use of fixed, predefined patches. This approach effectively retains semantic information within the patches for subsequent detection and classification tasks. Zheng et al. [32] propose a novel end-to-end network, named GRDATFusion, which utilizes an attention residual module to process critical image details and a Transformer module to capture global information and model long-term dependencies. However, there are significant differences between visual images and text. To address these challenges, the Vision Transformer (ViT) [33] emerged, bridging this gap by treating images as a series of patches. The ViT demonstrated that Transformer architectures could be applied to images with performance comparable to CNNs in image recognition tasks. The Swin Transformer [26] introduced a hierarchical Transformer structure with a shifting window scheme, limiting self-attention calculations to non-overlapping local windows. Compared to the global attention mechanism of the ViT, this significantly improves computational efficiency. Additionally, the Swin Transformer enables cross-window interactions, which facilitate improved contextual information capture and enhance flexibility, generalization capability, and robustness. Despite these advancements, Transformers are not entirely suited for pavement damage detection and classification tasks due to the unique characteristics of pavement damage images compared to other image classification scenarios.

2.3. Feature Re-Embedding

The feature re-embedding model proposed by Tang et al. [34], is an enhanced version of the Transformer architecture. The model introduces a re-embedding region Transformer (RRT) designed for online re-embedding of instance features, enabling the capture of fine-grained local details and the establishment of inter-region connections. In contrast to prior approaches that rely on pre-trained feature extractors or complex instance aggregation designs, RRT specifically focuses on the real-time re-embedding of instance features. Inspired by the feature re-embedding idea of the RRT model, this paper proposes an attention feature re-embedding module, which targets feature extraction and re-embedding for road damage images and makes full use of the damage features in the road damage images to achieve targeted detection and classification.

3. Methods

Pavement damage detection and classification tasks differ significantly from other image classification tasks due to their unique characteristics, such as large pixel variations, uneven illumination, and the presence of small damage areas within the overall image. To tackle these challenges, this study presents FFE and MHAR, built upon the Swin Transformer architecture.

The overall architecture of FFEAR is illustrated in Figure 1, while FFE and MHAR are shown in Figures 2 and 3. FFE and MHAR are designed to work together both are essential components of the system. Section 3.1 will provide a detailed description of the overall architecture and workflow of FFEAR. Section 3.2 will focus on FFE, and Section 3.3 will elaborate on MHAR. The experimental results confirm that the proposed FFEAR model effectively mitigates the challenge posed by the low damage ratio in pavement images, outperforming the original Swin Transformer in overall performance.

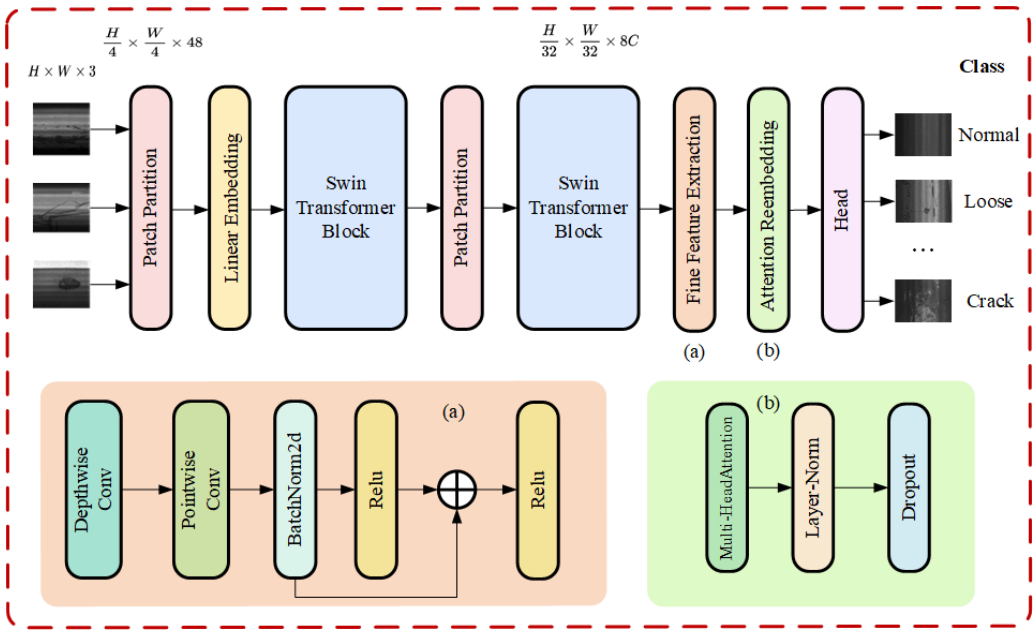


Figure 1. Overall architecture of FFEAR-Swin Transformer. The part (**a**) is FFE. The part (**b**) is MHAR.

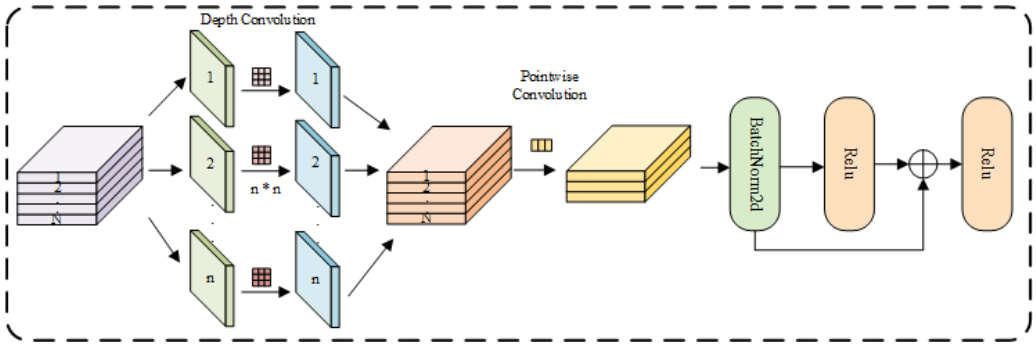


Figure 2. Architecture of FFE.

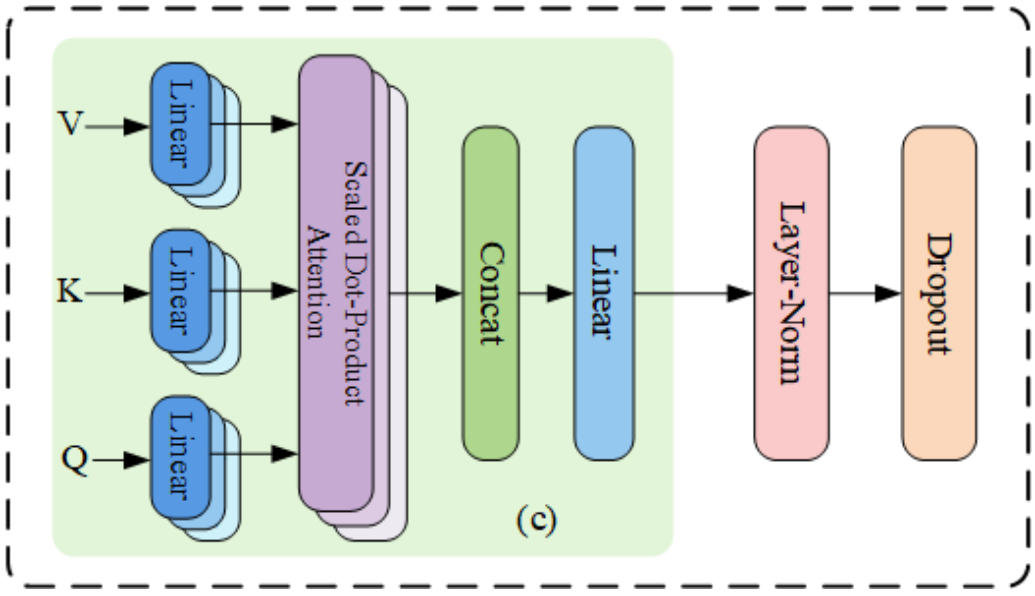


Figure 3. Architecture of the MHAR. The part (c) is the multi-head self-attention mechanism.

3.1. FFEAR-Swin Transformer

The FFEAR-Swin Transformer is an enhanced variant of the original Swin Transformer. After passing through four original Swin Transformer blocks, the FFEAR module is in-

troduced for feature-based attention re-filtering and embedding. Finally, the processed features are passed to the classification head layer. This approach maintains the original input dimensions, effectively avoiding additional model complexity while still achieving excellent classification results, including higher success rates and precision.

As illustrated in Figure 1, when a pavement damage image is input into the model, it is first processed by the patch partition module. This module divides the input image of size HWC into non-overlapping patches of equal size, where H represents the height of the original image, W represents the width, and C denotes the number of channels (typically $C = 3$ for RGB images). If the input is a grayscale image or another type, it needs to be transformed into an RGB image with three channels before processing. After the patch partition module, an image of size $H \times W \times 3$ is segmented into $N \times (P \times P \times 3)$ patches, where each $P \times P \times 3$ patch is treated as a token. Here, N denotes the number of tokens, $P = 4$ represents the patch size, and 3 is the number of channels. Specifically, each image of size $H \times W \times 3$ is divided into $N = \frac{H}{4} \times \frac{W}{4}$ tokens, with each token having a size of $P \times P \times C = 4 \times 4 \times 3 = 48$ dimensions. In the second step, a fully connected layer (linear embedding) projects these N tokens of 48 dimensions to an arbitrary dimension C, resulting in a linear embedding of size $N \times C$. This data is then fed into the first Swin Transformer block with self-attention. During the first input stage, the number of tokens remains unchanged at $N = \frac{H}{4} \times \frac{W}{4}$. The first Swin Transformer block and the linear embedding layer are considered as Stage 1.

When tokens are fed into a Swin Transformer block, they first pass through a Layer-Norm layer before proceeding to the window-based multi-head self-attention (W-MSA) module. This module divides the image into non-overlapping windows and performs self-attention calculations within each window independently, significantly reducing computational complexity and improving model efficiency. The process can be expressed using Equation (1).

$$\hat{x}^l = WMSA\left(LN\left(x^{l-1}\right)\right) + x^{l-1} \tag{1}$$

In the equation, x^{l-1} denotes the output feature from the previous $l-1$ layer. $LN(\cdot)$ represents Layer Normalization, which standardizes the input feature distribution. $WMSA(\cdot)$ denotes the window-based multi-head self-attention mechanism, where self-attention is computed within non-overlapping local windows. The intermediate output $hatx^l$ refers to the feature map after the attention module in the current layer.

After the W-MSA stage, a residual connection layer is applied, followed by another Layer-Norm layer, then the tokens proceed through the MLP layer and undergo another residual connection. This completes the first Swin Transformer block.The process can be expressed using Equation (2).

$$x^l = MLP(LN(\hat{x}^l)) + \hat{x}^l \tag{2}$$

In the equation, $MLP(\cdot)$ denotes a multi-layer perceptron block consisting of fully connected layers with GELU activation and Dropout regularization. x^l is the final output of the current layer, serving as the input to the next layer.

However, using the W-MSA module means that self-attention is computed only within each independent, non-overlapping window, potentially missing direct information exchange between adjacent windows. To address this issue, the second Swin Transformer block employs the shifted window-based multi-head self-attention (SW-MSA) module, which introduces shifted W-MSA. In this module, the windows are shifted, allowing for information exchange between adjacent windows and enhancing inter-window information interaction. The remaining steps are similar to those in the W-MSA-based Swin Transformer block, including Layer-Norm, MLP, and residual connections. It is important to note that

W-MSA and SW-MSA are used in pairs, not individually, as each Swin Transformer block alternates between the two. The process can be expressed using Equations (3) and (4).

$$\hat{x}^{l+1} = SWMSA\left(LN\left(x^{l}\right)\right) + x^{l} \tag{3}$$

$$x^{l+1} = MLP(LN(\hat{x}^{l+1})) + \hat{x}^{l+1} \tag{4}$$

In the equations, the $SWMSA(\cdot)$ operation is similar to $WMSA(\cdot)$ but incorporates a shifted window strategy to enable cross-window information exchange. The output x^{l+1} corresponds to the feature representation after the shifted self-attention module, while x^{l+1} represents the final output of the current block after applying Layer Normalization, the MLP module, and a residual connection—serving as the input to the subsequent block.

As the network depth increases, the number of tokens is progressively reduced through the patch merging layers, with each layer reducing the token count to one-fourth of its previous value while doubling the number of channels. Each patch merging layer consolidates each 22 neighboring pixel block into a single patch, creating four feature maps from the original input. These four feature maps are then concatenated along the depth dimension, followed by a Layer Norm layer. Finally, a fully connected layer performs a linear transformation along the depth of the feature maps, increasing the depth from C to $2C$. The first patch merging layer and the second Swin Transformer block together constitute Stage 2. This process is repeated to form Stage 3 and Stage 4. After passing through four stages, the original input image is transformed into tokens of size $\frac{H}{32} \times \frac{W}{32} \times 8C$.

Subsequently, these tokens are processed by MHAR, which is specifically designed to address the unique characteristics of pavement damage images. This self-attention mechanism enhances the extraction of features that might be overlooked due to the low damage ratio, thus making better use of the information contained in the images. The number of attention heads in this multi-head attention mechanism is consistent with the number of heads used in the previous stages, and the embedding dimensions match those of Stage 4. In MHAR, we also incorporate the FFE block to capture localized, subtle features from the input tokens. This approach significantly improves the model's ability to represent features. Afterward, multi-head self-attention is employed to compute attention across multiple heads. This process is described by Equations (5) and (6).

$$x^{l+1} = Dropout(LN(MHA(\hat{x}^{l+1}))) \tag{5}$$

$$x^{l+1} = ReLU(BN(PC(DC(x^{l+1}))) + x^{l+1}) \tag{6}$$

In these equations, $Dropout(\cdot)$ serves as a regularization technique to prevent overfitting by randomly discarding a subset of features during training. $LN(\cdot)$ denotes Layer Normalization, which stabilizes the training process. $MHA(\cdot)$ refers to multi-head self-attention, which is used to capture global dependencies within the input features. $DC(\cdot)$ represents depthwise convolution, which applies convolution independently to each input channel, thereby reducing computational cost. $PC(\cdot)$ denotes Pointwise Convolution (1×1 convolution), used for channel mixing or dimensionality adjustment. $BN(\cdot)$ refers to Batch Normalization, which helps stabilize and accelerate gradient propagation. $ReLU(\cdot)$ is a nonlinear activation function that enhances the model's representational capacity. The variable x^{l+1} denotes the input feature passed from the previous layer and also represents the output after sequential processing through $Dropout(\cdot)$, $LN(\cdot)$, $MHA(\cdot)$, $DC(\cdot)$, $PC(\cdot)$, $BN(\cdot)$, $ReLU(\cdot)$, along with residual connections, which is then propagated to the next layer.

Following these operations, all tokens are fed into the Swin Transformer framework, undergoing Layer Norm and global average pooling. Finally, they are passed through

the head layer to achieve the final classification. This completes the overall process of the FFEAR-Swin Transformer. The following Sections 3.2 and 3.3 will provide a detailed description of FFE and MHAR.

3.2. Fine Feature Extraction Module

FFE is specifically designed to tackle the challenge posed by the low damage ratio in pavement images. This module employs depthwise separable convolution techniques to handle each channel separately, followed by linear combination of the depthwise convolution outputs to mix information from different channels. This approach enhances the feature representation capability. Figure 2 illustrates the structure of FFE, and the expression formula is given in Equation (6). This module primarily consists of depthwise separable convolution, complemented by a BatchNorm2d layer for normalization and a ReLU activation function for nonlinear activation. Additionally, residual connections are incorporated to prevent information loss.

When the input tokens enter the FFE module, they first undergo depthwise separable convolution with a kernel size of 3×3, a stride of 2, and padding of 1. This convolution maintains the same dimensionality as the original input to avoid information loss due to dimensional changes. After performing convolution separately on each channel, the outputs are linearly combined to mix information across different channels, enhancing inter-channel information transfer. Subsequently, Batch Normalization and ReLU activation are applied to improve model stability and feature representation capabilities. A residual connection is then added to prevent information loss. Finally, another ReLU activation is performed to obtain the final processed result, which is then passed to the multi-head self-attention module for subsequent classification tasks.

3.3. Multi-Head Self-Attention Re-Embedding Module

The feature re-embedding model (RRT) [34] proposed by Tang et al. has demonstrated significant performance improvements for image classification tasks. Tang et al.'s feature re-embedding model integrates additional information, such as positional encoding, into the features processed by the original model, and then performs a second embedding of the features based on the positional encoding. Our MHAR leverages the self-attention mechanism to achieve feature embedding in the same dimensional space. It retains the original feature information while the re-embedded features are subjected to secondary extraction through depthwise separable convolution, thereby enhancing the representation capability of the features.

Pavement damage images have unique characteristics, such as high similarity between normal and damaged images—with only minor differences at the damaged areas—and the relatively small proportion of these damaged areas within the entire image, as well as the high similarity among different types of cracks (e.g., crocodile cracks, transverse cracks, and longitudinal cracks), and our multi-head self-attention re-embedding is particularly well-suited to addressing these challenges. We input the tokens processed by the Swin Transformer framework into our re-embedding layer, leveraging the strong performance of the Swin Transformer while specifically addressing the characteristics of pavement damage images with our re-embedding technique. This approach enhances feature representation without significantly increasing the model's complexity, efficiently and conveniently accomplishing the task of pavement damage image classification.

As illustrated in Figure 3, MHAR consists of the multi-head attention calculation module, a linear embedding layer, and a dropout layer. The input tokens first pass through the fine feature extraction module, where they undergo depthwise separable convolution with a 33 kernel, a stride of 1, and padding of 1. After this processing, the tokens are

fed into the attention module for multi-head self-attention calculation. The calculation formula is given in Equation (7). In this process, Q, K, and V represent the three vectors: Query, Key, and Value, respectively. Attention is computed by aggregating the Query with the Key to determine the relevance, and then the appropriate Value is selected based on this relevance. This results in the allocation of attention weights and generates the final output. The output is then normalized using a Layer Norm layer to stabilize the training process. Finally, a dropout layer is applied to randomly discard some of the data to prevent overfitting, thereby enhancing the model's robustness and generalization capability.

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{7}$$

4. Experiments

In this section, the whole experimental process will be introduced in detail, including the configuration of experimental environment, datasets, evaluation metrics, experimental results, and ablation experiments.

4.1. Experimental Environment

All experiments were performed on a 64-bit Ubuntu 20.04.5 LTS operating system. The hardware configuration comprised an Intel® Core™ i5-13600KF CPU and an NVIDIA RTX 4080 GPU with 16 GB of memory. The model was implemented in Python using the open-source PyTorch framework, with CUDA 12.2 utilized to accelerate training.

4.2. Datasets and Evaluation Metrics

Before the work presented in [25], most models for pavement distress classification were evaluated using proprietary datasets. However, due to the limited size and inconsistent quality of such private datasets, we utilized the publicly available CQU-BPDD dataset. The dataset comprises 60,056 asphalt pavement images collected by professional inspection vehicles across multiple locations in South China at various times. It classifies pavement conditions into eight categories: transverse cracks, large cracks, crocodile cracks, crack filling, longitudinal cracks, wrinkles, repairs, and normal. Representative samples from the CQU-BPDD dataset are presented in Figure 4. These categories cover common types of pavement damage and include characteristics such as relatively small proportions of distressed areas, some of which may be in low-light conditions and difficult to identify. The CQU-BPDD dataset was initially divided into a training set and a test set. The training set includes 10,137 images, comprising 5137 abnormal pavement images (encompassing all damage categories) and 5000 normal images. The test set contains 49,919 images, with 11,589 labeled as abnormal and 38,330 as normal.

In addition to employing the CQU-BPDD dataset as the primary experimental dataset, to further evaluate the generalizability and robustness of the proposed FFEAR model, the well-established Crack500 pavement crack dataset was included in the ablation study. As a benchmark dataset in pavement damage detection, Crack500 comprises 3363 diverse crack images. It is divided into training and testing subsets at a 7:3 ratio, with 2354 images used for training and 1005 reserved for evaluation. This division ensures a balanced approach to assessing model robustness across varying image samples.

For comparative analysis with other pavement damage classification models, the CQU-BPMDD dataset was additionally employed. This dataset comprises 9851 disease images and 29,143 normal images, covering categories such as longitudinal cracks, transverse cracks, repairs, looseness, potholes, massive cracks, and wave crowding. Unlike CQU-BPDD, all images in CQU-BPMDD were captured with exposure compensation, and most depict small-scale cracks, making them more representative of real-world scenarios and

posing greater challenges for detection. Representative samples from the CQU-BPMDD dataset are presented in Figure 5.

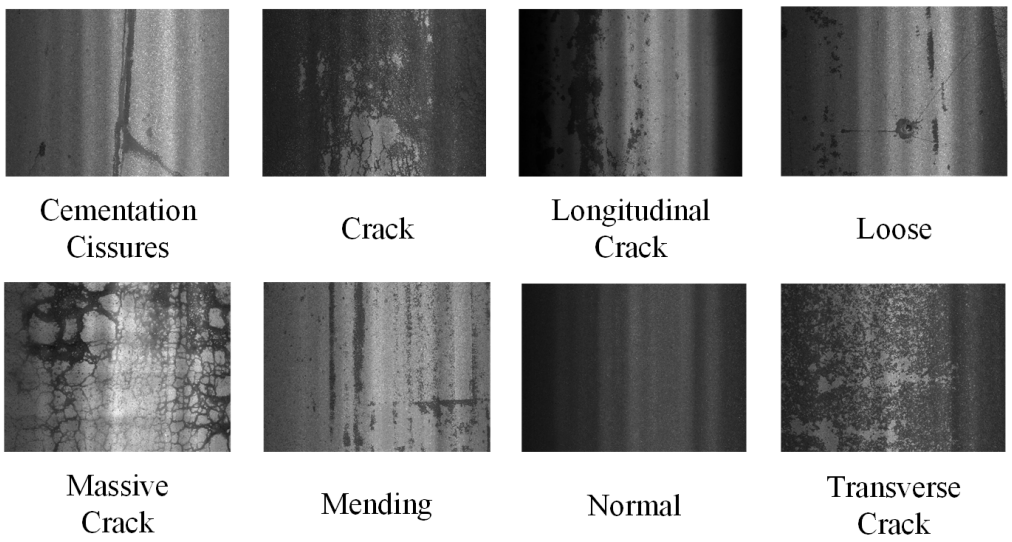


Figure 4. Eight sample images from CQU-BPDD dataset.

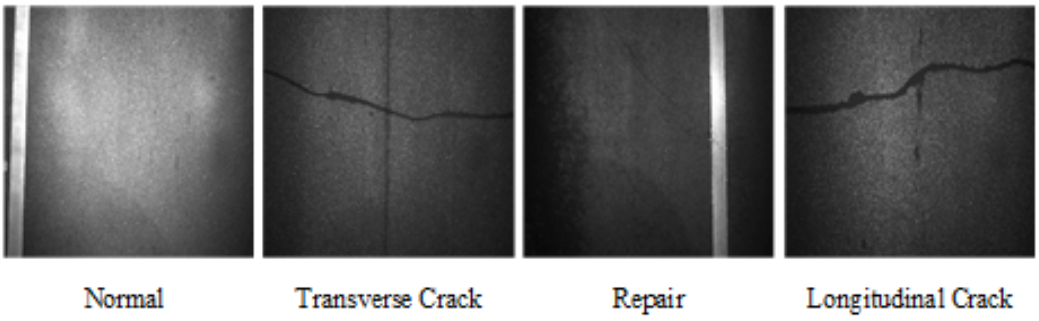


Figure 5. Four sample images from CQU-BPMDD dataset.

The design of the fundamental parameters for all experiments is shown in Table 1. Accuracy, F1 score, and precision, commonly used metrics in image classification tasks, were employed as evaluation indicators.

Table 1. Experimental parameter settings.

Parameter	Value
Learning rate	0.0001
Weight decay	0.0005
Batch size	32
Epoch	50

Accuracy represents the proportion of correctly classified samples out of the total number of samples, which is a comprehensive indicator to measure the accuracy of the model on all categories. It is described by Equation (8). Precision represents the proportion of samples truly belonging to a certain class among all samples predicted by the model to be of that class. Precision focuses on the correctness of the prediction results and how many of the predicted positive samples are truly positive, and it is described by Equation (9). F1 score integrates precision and recall and measures the model's performance through their

harmonic mean, as shown in Equation (10).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

$$F1 = 2 \times \frac{\frac{TP}{TP+FP} \times \frac{TP}{TP+FN}}{\frac{TP}{TP+FP} + \frac{TP}{TP+FN}} \tag{10}$$

4.3. Experimental Analysis

A summary of comparative experiments on the CQU-BPDD dataset is provided in Table 2. The models AlexNet [35], VGG16 [36], MobileNetV2 [37], MobileNetV3 [38], FFVT [39], TransFG [40], Vision Mamba [41], MambaOut [42], Vision Transformer, and Swin Transformer were comprehensively compared. The experimental results indicate that, compared to Swin-T-S (baseline), FFEAR achieves a 2.07% improvement in accuracy and a 0.97% increase in F1 score, with similar precision. To demonstrate the superior performance of FFEAR in pavement damage classification, we selected professional pavement breakage classification models of PicT [25], DMTC [43], and DDACDN [44] to compare with FFEAR. Overall, FFEAR demonstrates strong performance in the road damage detection and classification task.

Table 2. Comparison of experimental data on CQU-BPDD dataset.

Model	Accuracy	Precision	F1 Score
MobileNetV3	82.31%	79.91%	79.90%
MobileNetV2	83.40%	79.55%	79.71%
AlexNet	73.43%	75.16%	73.48%
VGG16	77.23%	84.63%	79.45%
Swin-T-S	87.43%	**88.65**%	87.50%
Swin-T-B	87.88%	88.34%	87.85%
ViT-B-16	86.08%	85.96%	85.76%
ViT-B-32	84.39%	82.37%	82.99%
Vision Mamba	87.12%	88.21%	87.33%
MambaOut	87.37%	88.61%	87.63%
FFVT	81.98%	81.85%	81.26%
TransFG	80.88%	80.87%	80.22%
DMTC	84.32%	83.92%	82.68%
DDACDN	83.62%	83.73%	83.57%
PicT	88.58%	88.14%	87.85%
FFEAR	**89.24**%	88.36%	**88.35**%

As shown clearly in Table 2, FFEAR performs better than other classical classification models on the main evaluation indicators. Due to the special characteristics of road surface damage images, there are some drawbacks in directly applying the classical classification model to this task, so the performance is not as good as that of FFEAR. Firstly, we introduce FFE to solve the difficulties of the low damage ratio and high similarity of damage types in road surface damage images and use depthwise separable convolution for secondary feature extraction to enhance the feature expression ability. Secondly, MHAR is introduced to address issues such as uneven lighting and background noise interference. The features extracted twice by the fine feature extraction module are re-embedded using multi-head attention, which not only strengthens the representation of limited features but also mitigates the challenges posed by background noise. These improvements make FFEAR achieve better performance in road surface damage classification tasks. FFEAR exhibits a 2.07%

improvement in accuracy and a 0.97% increase in F1 score, with similar precision to Swin-T. Compared with the ViT, FFEAR achieves a 3.67% improvement in accuracy, a 3.02% increase in F1 score, and a 2.8% increase in precision.

To further highlight the strong performance of FFEAR in pavement damage classification, comparisons were conducted with several specialized pavement damage classification models. The experimental results, presented in Figure 6, demonstrate that FFEAR outperforms comparable models across multiple evaluation metrics. By incorporating FFE and MHAR, FFEAR effectively leverages limited features to achieve superior classification accuracy.



Figure 6. Comparison of FFEAR and pavement damage classification models on CQU-BPMDD dataset.

In order to verify the robustness and generalization ability of FFEAR, we conducted generalization experiments with other pavement damage classification models on the CQU-BPMDD dataset. The experimental results are shown in Table 3. Compared with the baseline on CQU-BPMDD dataset, FFEAR achieved a 0.92% improvement in accuracy, a 1.93% improvement in precision, and a 2.4% improvement in F1 score. These results also prove that FFEAR is superior to other similar models.

Table 3. Results of generalization experiments on CQU-BPMDD dateset.

Model	Accuracy	Precision	F1 Score
Swin-T	85.63%	84.72%	83.91%
DMTC	84.32%	83.92%	82.68%
DDACDN	81.62%	81.37%	80.93%
FFEAR	**89.24**%	88.36%	**88.35**%

4.4. Visualization Analysis

In the experimental process, we utilized Grad-CAM [45] to generate visual heatmaps, and Figure 7 provides an intuitive explanation of the decision-making process of both the FFEAR and Swin Transformer models. The generated heatmaps clearly demonstrate the specific regions of the images that these two models focus on when making classification predictions. For the road damage classification task, Grad-CAM produced heatmaps for different categories, visually displaying the areas of the image that the models attend to for each category, thus offering deeper insights into the models' decision-making rationale and feature extraction methods. As demonstrated in Figure 7, it is evident that the proposed FFEAR model exhibits superior accuracy in positioning and feature extraction when applied to road surface damage images with a low damage ratio. The FFEAR model is capable of identifying damage areas with greater precision, producing finer and more detailed extraction compared to the Swin-T model. This enhanced capability enables FFEAR to capture even smaller, subtler damage regions, providing a more accurate representation of road surface conditions.

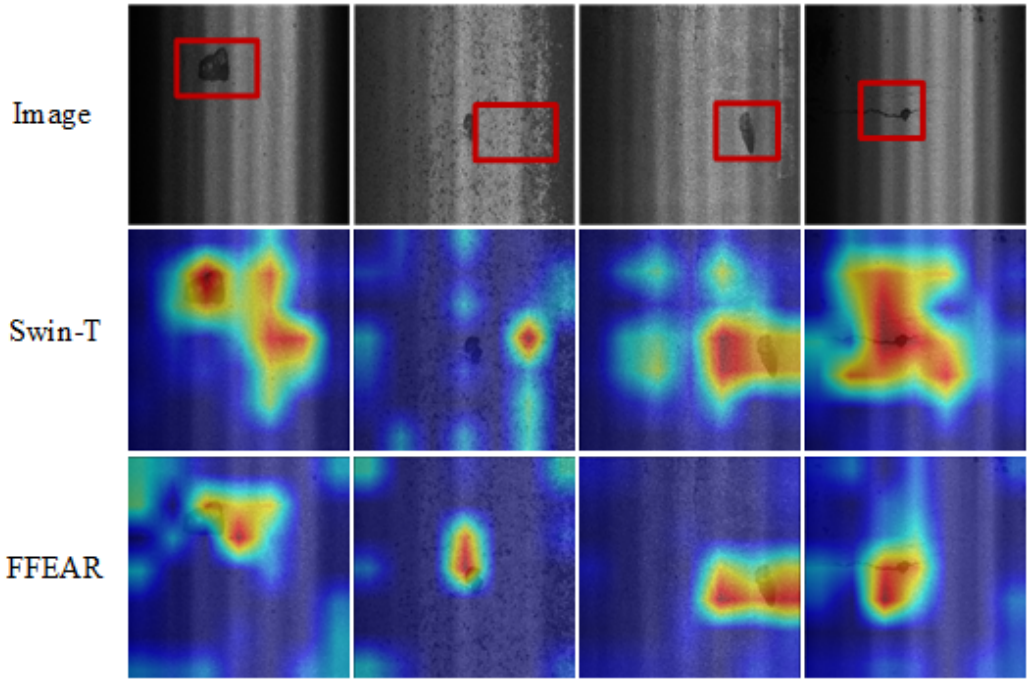


Figure 7. Visual analysis of heat maps. The first line of pictures is the pavement damage image, where the damage is marked by a red box. The second line is a visual analytical heat map of Swin-T. The third line is a visual analytical heat map of FFEAR.

4.5. Ablation Analysis

FFE and MHAR in the proposed FFEAR model are specifically designed to tackle the distinct challenges of pavement damage images, including low damage ratios, uneven illumination, similar crack patterns, and high visual similarity among samples. The corre-

sponding ablation study results are provided in Table 4, with a histogram illustrating these results shown in Figure 8.

Table 4. Ablation experimental results based on CQU-BPDD dataset and Crack500 dataset.

Dataset	Model	Accuracy	Precision	F1 Score	Param (M)
CQU-BPDD	Swin-T	87.43%	**88.65**%	87.50%	48
	MHAR	87.11%	85.51%	85.61%	58
	FFE	88.14%	88.30%	87.66%	48
	FFEAR	**89.24**%	88.36%	**88.35**%	58
Crack500	Swin-T	76.82%	76.64%	76.37%	48
	MHAR	76.52%	76.77%	76.42%	58
	FFE	77.24%	77.10%	76.50%	48
	FFEAR	**77.71**%	**77.65**%	**76.80**%	58

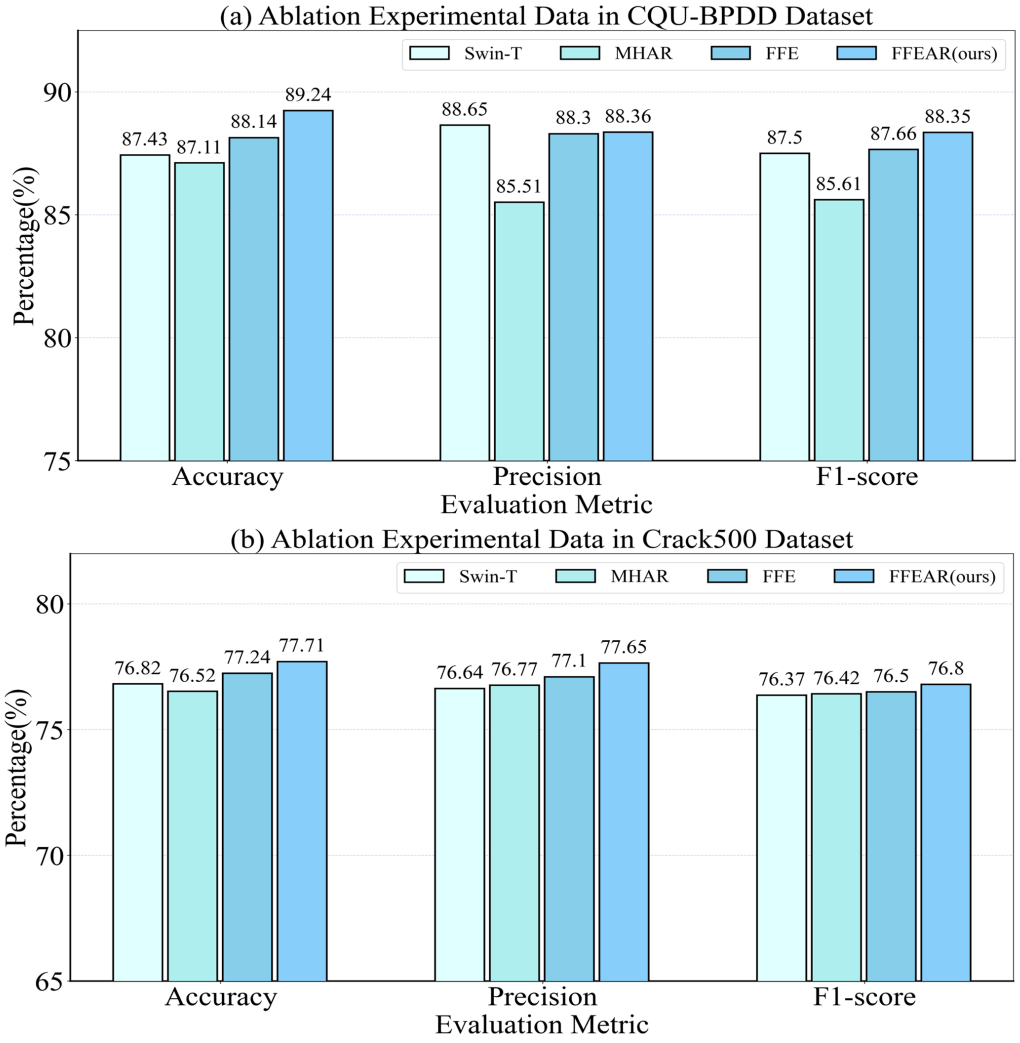


Figure 8. Histogram of ablation experiment; (**a**) is the results on the CQU-BPDD dataset, and (**b**) is the results on Crack500 dataset.

The experimental results indicate that using MHAR in isolation yields limited improvement in evaluation metrics. This limitation is attributed to the module's focus on re-embedding rather than re-extracting fine features from the pavement damage images, relying primarily on original convolutional outputs. As a result, MHAR remains par-

tial, reducing the model's representational capacity and limiting its focus on intricate damage details.

When employing only FFE, greater emphasis is placed on small-scale damage features, enhancing feature expression and maximizing their representation. The experimental results indicate that the proposed method achieves improvements across all evaluation metrics.

When combining the fine feature extraction and multi-head self-attention re-embedding modules, the model shows an accuracy increase of 2.07% and an F1 score improvement of 0.97% in the experimental results, achieving superior performance compared to using either module individually.

In terms of performance and efficiency comparison, the baseline Swin-T model has a parameter count of 48 M. After the MHAR module and the FFE module are incorporated separately, the parameter counts increase to 58 M and 48 M, respectively. When the two modules are integrated into the full FFEAR model, the total number of parameters reaches 58 M. Although the structural enhancements introduce a moderate increase in parameter count, the computational complexity remains at $O(n^2)$, without adding additional computational burden. This controlled increase in parameters leads to a consistent improvement in classification performance. The experimental results demonstrate that the FFEAR model outperforms the baseline in both accuracy and F1 score, confirming the effectiveness and practicality of the proposed design.

5. Conclusions

This paper presents a specialized solution to address the unique challenges of pavement damage images, including low damage visibility, variable lighting conditions, high crack-type similarity, and similar pavement textures. Building on the Swin Transformer, we introduce a fine feature extraction module and a multi-head self-attention re-embedding module. These enhancements enable secondary extraction of fine features and recalculation of self-attention, thus significantly improving feature representation while minimizing information loss. Extensive experiments demonstrate that FFEAR outperforms the Swin Transformer, making it better suited for pavement damage detection and classification tasks. However, due to the Transformer's inherent $O(n^2)$ complexity with respect to the number of image patches and the additional overhead introduced by depthwise separable convolutions and re-embedding operations, future work will focus on designing lightweight structures and attention-efficient mechanisms to reduce computational cost while maintaining performance, making the model more suitable for real-time and resource-constrained applications.

Author Contributions: Project administration, Writing—review and editing, Supervision, S.Z.; Methodology, Writing—original draft, Writing—review and editing, K.W.; methodology, Z.L.; supervision, M.H.; visualization, S.H. All authors have read and agreed to the published version of the manuscript.

Funding: The work is partially supported by the Henan Provincial Department of Science and Technology Research Project (Grant No. 242102210107; No. 252102211070; No. 252102210127) and the Key Scientific Research Projects of Higher Education Institutions in Henan Province (24B520038), in part by the Mass Innovation Space Incubation Project under Grant 2023ZCKJ216, and in part by the Key Research and Development Program of Shaanxi (Program No. 2024GX-YBXM-545).

Data Availability Statement: The datasets analyzed during the current study are available from the corresponding author on reasonable request. Codes from the current study may be obtained from the corresponding authors upon reasonable request.

Conflicts of Interest: We promise that there will be no conflicts of interest between the authors.

Abbreviations

The following abbreviations are used in this manuscript:

FFE	Fine feature extraction
MHAR	Multi-head self-attention re-embedding
R-CNN	Region-based convolutional neural network
GAN	Generative Adversarial Network
SE	Squeeze and Expand
MC	Multi-Scale Convolution
DSC	Depthwise separable convolution
WSPLINs	Weakly supervised patch label inference networks
PLIN	Patch label inference network
FCN	Fully convolutional network
LVE	Linear viscoelastic
MLET	Multi-layered elastic theory
FWD	Falling Weight Deflectometer
GPR	Ground-Penetrating Radar
ViT	Vision Transformer
Swin-T	Swin Transformer
RRT	Re-embedding region Transformer
W-MSA	Window-based multi-head self-attention mechanism
SW-MSA	Shifted window-based multi-head self-attention mechanism
VGG16	Visual Geometry Group 16-layer network
FFVT	Feature Fusion Vision Transformer
TransFG	Transformer Architecture for Fine-Grained Recognition
DMTC	Dense Multiscale Feature Learning Transformer
DDACDN	Deep Domain Adaptation for Pavement Crack Detection
PicT	Pavement image classification Transformer

References

1. Doshi, K.; Yilmaz, Y. Road damage detection using deep ensemble learning. In Proceedings of the 2020 IEEE International Conference on Big Data (Big Data), Atlanta, GA, USA, 10–13 December 2020; pp. 5540–5544.
2. Valipour, P.S.; Golroo, A.; Kheirati, A.; Fahmani, M.; Amani, M.J. Automatic pavement distress severity detection using deep learning. *Road Mater. Pavement Des.* **2023**, *25*, 1830–1846. [CrossRef]
3. Pan, N.; Liu, H.; Wu, D.; Liu, C.; Du, Y. Spatiotemporal matching method for tracking pavement distress using high-frequency detection data. *Comput.-Aided Civ. Infrastruct. Eng.* **2023**, *38*, 2257–2278. [CrossRef]
4. Fahad, M.; Nagy, R.; Guangpin, L.; Rosta, S. Pavement Crack Monitoring: Literature Review. *Iraqi J. Civ. Eng.* **2023**, *16*, 76–89. [CrossRef]
5. Zhao, Y.; Zhang, W.; Yang, Y.; Sun, H.; Wang, L. An efficient pavement distress detection scheme through drone–ground vehicle coordination. *Transp. Res. Part A Policy Pract.* **2024**, *180*, 103949. [CrossRef]
6. Chou, J.C.; O'Neill, W.A.; Cheng, H.D. Pavement distress classification using neural networks. *IEEE Int. Conf. Syst. Man Cybern.* **1994**, *1*, 397–401.
7. Nejad, F.M.; Zakeri, H. An expert system based on wavelet transform and radon neural network for pavement distress classification. *Expert Syst. Appl.* **2011**, *38*, 7088–7101. [CrossRef]
8. Lv, Z.; Cheng, C.; Lv, H. Automatic identification of pavement cracks in public roads using an optimized deep convolutional neural network model. *Philos. Trans. R. Soc. A* **2023**, *381*, 20220169. [CrossRef]
9. Liu, Z.; Pan, S.; Gao, Z.; Chen, N.; Li, F.; Wang, L.; Hou, Y. Automatic intelligent recognition of pavement distresses with limited dataset using generative adversarial networks. *Autom. Constr.* **2023**, *146*, 104674. [CrossRef]
10. Cheng, X.; He, T.; Shi, F.; Zhao, M.; Liu, X.; Chen, S. Selective feature fusion and irregular-aware network for pavement crack detection. *IEEE Trans. Intell. Transp. Syst.* **2024**, *25*, 3445–3456. [CrossRef]
11. Huang, S.; Tang, W.; Huang, G.; Huangfu, L.; Yang, D. Weakly supervised patch label inference networks for efficient pavement distress detection and recognition in the wild. *IEEE Trans. Intell. Transp. Syst.* **2023**, *24*, 5216–5228. [CrossRef]
12. Fan, Z.; Wu, Y.; Lu, J.; Li, W. Automatic pavement crack detection based on structured prediction with the convolutional neural network. *arXiv* **2018**, arXiv:1802.02208.

13. Sirhan, M.; Bekhor, S.; Sidess, A. Multilabel CNN model for asphalt distress classification. *J. Comput. Civ. Eng.* **2024**, *38*, 04023040. [CrossRef]
14. Zhu, J.; Zhong, J.; Ma, T.; Huang, X.; Zhang, W.; Zhou, Y. Pavement distress detection using convolutional neural networks with images captured via UAV. *Autom. Constr.* **2022**, *133*, 103991. [CrossRef]
15. Liang, J.; Gu, X.; Jiang, D.; Zhang, Q. CNN-based network with multi-scale context feature and attention mechanism for automatic pavement crack segmentation. *Autom. Constr.* **2024**, *164*, 105482. [CrossRef]
16. Li, P.; Zhou, B.; Wang, C.; Hu, G.; Yan, Y.; Guo, R.; Xia, H. CNN-based pavement defects detection using grey and depth images. *Autom. Constr.* **2024**, *158*, 105192. [CrossRef]
17. Tang, W.; Huang, S.; Zhao, Q.; Li, R.; Huangfu, L. An iteratively optimized patch label inference network for automatic pavement distress detection. *IEEE Trans. Intell. Transp. Syst.* **2021**, *23*, 8652–8661. [CrossRef]
18. Maeda, H.; Sekimoto, Y.; Seto, T.; Kashiyama, T.; Omata, H. Road damage detection using deep neural networks with images captured through a smartphone. *arXiv* **2018**, arXiv:1801.09454.
19. Di Benedetto, A.; Fiani, M.; Gujski, L.M. U-Net-based CNN architecture for road crack segmentation. *Infrastructures* **2023**, *8*, 90. [CrossRef]
20. Neto, O.P.V.; Luz, L.O.; Silva, P.A.R.; de Oliveira Bicalho, J.G.; Ruella, E.V.C.; Nacif, J.A.; Ferreira, R.S. The Impact of Information Flow Control on FCN Circuit Design. In Proceedings of the 2024 IEEE 24th International Conference on Nanotechnology (NANO), Gijón, Spain, 8–11 July 2024; pp. 448–453.
21. Noori, H.; Sarkar, R. Airport Pavement Distress Analysis. *Iran. J. Sci. Technol. Trans. Civ. Eng.* **2024**, *48*, 1171–1190. [CrossRef]
22. Xiong, B.; Hong, R.; Liu, R.; Wang, J.; Zhang, J.; Li, W.; Lv, S.; Ge, D. FCT-Net: A dual-encoding-path network fusing atrous spatial pyramid pooling and transformer for pavement crack detection. *Eng. Appl. Artif. Intell.* **2024**, *137*, 109190. [CrossRef]
23. Mei, A.; Zampetti, E.; Di Mascio, P.; Fontinovo, G.; Papa, P.; D'Andrea, A. ROADS—Rover for Bituminous Pavement Distress Survey: An Unmanned Ground Vehicle (UGV) Prototype for Pavement Distress Evaluation. *Sensors* **2022**, *22*, 3414. [CrossRef]
24. Gopalakrishnan, K.; Khaitan, S.K.; Choudhary, A.; Agrawal, A. Deep Convolutional Neural Networks with transfer learning for computer vision-based data-driven pavement distress detection. *Constr. Build. Mater.* **2017**, *157*, 322–330. [CrossRef]
25. Tang, W.; Huang, S.; Zhang, X.; Huangfu, L. PicT: A Slim Weakly Supervised Vision Transformer for Pavement Distress Classification. In Proceedings of the 30th ACM International Conference on Multimedia, Lisboa, Portugal, 10–14 October 2022; pp. 3076–3084.
26. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 10012–10022.
27. Ahmed, A.; Erlingsson, S. Viscoelastic Response Modelling of a Pavement under Moving Load. *Transp. Res. Procedia* **2016**, *14*, 748–757. [CrossRef]
28. Gkyrtis, K.; Loizos, A.; Plati, C. A mechanistic framework for field response assessment of asphalt pavements. *Int. J. Pavement Res. Technol.* **2021**, *14*, 174–185. [CrossRef]
29. Devlin, J. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
30. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 1–15.
31. Chen, Z.; Zhu, Y.; Zhao, C.; Hu, G.; Zeng, W.; Wang, J.; Tang, M. DPT: Deformable patch-based transformer for visual recognition. In Proceedings of the 29th ACM International Conference on Multimedia, Virtual, 20–24 October 2021; pp. 2899–2907.
32. Zheng, J.; Jeon, S.; Yang, X. GRDATFusion: A gradient residual dense and attention transformer infrared and visible image fusion network for smart city security systems in cloud and fog computing. *Expert Syst.* **2024**, *42*, e13685. [CrossRef]
33. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
34. Tang, W.; Zhou, F.; Huang, S.; Zhu, X.; Zhang, Y.; Liu, B. Feature Re-Embedding: Towards Foundation Model-Level Performance in Computational Pathology. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 11343–11352.
35. Eldem, H.; Ülker, E.; Işıklı, O.Y. AlexNet architecture variations with transfer learning for classification of wound images. *Eng. Sci. Technol. Int. J.* **2023**, *45*, 101490. [CrossRef]
36. Jiang, Y.; Pang, D.; Li, C.; Yu, Y.; Cao, Y. Two-step deep learning approach for pavement crack damage detection and segmentation. *Int. J. Pavement Eng.* **2023**, *24*, 2065488. [CrossRef]
37. Kumar, B.A.; Bansal, M. Pothole Detection of Road Pavement by Modified MobileNetV2 for Transfer Learning. In *International Conference on Soft Computing for Problem-Solving*; Springer Nature: Singapore, 2023; pp. 515–531.
38. Li, B.; Xu, J.; Lian, Y.; Sun, F.; Zhou, J.; Luo, J. Improved MobileNet V3-Based Identification Method for Road Adhesion Coefficient. *Sensors* **2024**, *24*, 5613. [CrossRef] [PubMed]
39. Wang, J.; Yu, X.; Gao, Y. Feature fusion vision transformer for fine-grained visual categorization. *arXiv* **2021**, arXiv:2107.02341.

40. He, J.; Chen, J.N.; Liu, S.; Kortylewski, A.; Yang, C.; Bai, Y.; Wang, C. TransFG: A transformer architecture for fine-grained recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, Online, 22 February–1 March 2022; Volume 36.
41. Liu, X.; Zhang, C.; Zhang, L. Vision mamba: A comprehensive survey and taxonomy. *arXiv* **2024**, arXiv:2405.04404.
42. Yu, W.; Wang, X. Mambaout: Do we really need mamba for vision? *arXiv* **2024**, arXiv:2405.07992.
43. Xu, C.; Zhang, Q.; Mei, L.; Shen, S.; Ye, Z.; Li, D.; Yang, W.; Zhou, X. Dense multiscale feature learning transformer embedding cross-shaped attention for road damage detection. *Electronics* **2023**, *12*, 898. [CrossRef]
44. Liu, H.; Yang, C.; Li, A.; Huang, S.; Feng, X.; Ruan, Z.; Ge, Y. Deep domain adaptation for pavement crack detection. *IEEE Trans. Intell. Transp. Syst.* **2022**, *24*, 1669–1681. [CrossRef]
45. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 618–626.

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