buildings

MDPI

Article

Pavement Crack Detection Based on the Improved Swin-Unet Model

Song Chen [1], Zhixuan Feng [2], Guangqing Xiao [1], Xilong Chen [1], Chuxiang Gao [2], Mingming Zhao [3] and Huayang Yu [2,*]

[1] Guangzhou Guangjian Construction Engineering Testing Center Co., Ltd., Guangzhou 510000, China; 13928707040@163.com (S.C.); xiaogq2022@126.com (G.X.); chenxilong@cngttc.cn (X.C.)
[2] School of Civil Engineering and Transportation, South China University of Technology, Guangzhou 510006, China; 202111081855@scut.edu.cn (Z.F.); 202232202225@mail.scut.edu.cn (C.G.)
[3] Yantai Donghua Material Science Co., Ltd., Yantai 264006, China; zhaomingming@dhchem.com
* Correspondence: huayangyu@scut.edu.cn

Abstract: Accurate pavement surface crack detection is crucial for analyzing pavement survey data and the development of maintenance strategies. On the basis of Swin-Unet, this study develops the improved Swin-Unet (iSwin-Unet) model with the developed skip attention module and the residual Swin Transformer block. Based on the channel attention mechanism, the pavement crack region can be better captured while the crack feature channels can be assigned more weights. Taking advantage of the developed residual Swin Transformer block, the encoder architecture can globally model the pavement crack feature. Meanwhile, the crack feature information can be efficiently exchanged. To verify the pavement crack detection performance of the proposed model, we compare the training performance and visualization results with the other three models, which are Swin-Unet, Swin Transformer, and Unet, respectively. Three public benchmarks (CFD, Crack500, and CrackSC) have been adopted for the purpose of training, validation, and testing. Based on the test results, it can be found that the developed iSwin-Unet achieves a significant increase in mF1 score, mPrecision, and mRecall compared to the existing models, thereby establishing its efficacy in pavement crack detection and underlining its significant advancements over current methodologies.

Keywords: deep learning; pavement crack detection; semantic segmentation; transformer

Citation: Chen, S.; Feng, Z.; Xiao, G.; Chen, X.; Gao, C.; Zhao, M.; Yu, H. Pavement Crack Detection Based on the Improved Swin-Unet Model. *Buildings* **2024**, *14*, 1442. https://doi.org/10.3390/buildings14051442

Received: 26 March 2024
Revised: 26 April 2024
Accepted: 9 May 2024
Published: 16 May 2024

1. Introduction

Surface cracks are among the most common pavement distresses [1,2]. The infiltration of moisture into these cracks can compromise the compaction of materials in the deeper layers of the pavement, resulting in evident reductions in the bearing capacity of the entire pavement structure. This not only has the potential to impair the functionality of the pavement surface, and shorten the overall service life, but also poses risks to the safety of vehicle operation. Therefore, fast and precise detection of surface cracks plays a crucial role in timely maintenance efforts, preventing the worsening of pavement conditions [3,4]. Conventional methods for detecting surface cracks heavily rely on manual inspection, known for being time-consuming, resource-intensive, and subjective. Recent technological advancements in image processing, automation, and artificial intelligence (AI) have significantly influenced the assessment and measurement of surface cracks on pavement. Consequently, transportation management agencies can prioritize and strategically plan road network maintenance to extend the pavement's service life [2,5,6].

Digital image processing techniques have been successfully applied to detect distress in transportation infrastructures [7]. This methodology involves two key steps: image acquisition and image detection, respectively. Pavement surface images are specifically captured by automatic inspection vehicles developed by various agencies. In these acquired images, cracks are primarily categorized into two types: linear cracks and block cracks [8,9].

Manually designed features, including grayscale, edges, filters, wavelets, etc., have been successfully employed for their detection. While these methods are easily applicable in the field for linear cracks, detecting complex cracks, with various shapes, varying widths, and disruptions such as oil spots, poses significant challenges. Notably, the performance of these methods is constrained, and issues like poor contrast around cracked pixels due to unfavorable imaging conditions further complicate crack detection. Thus, manually designed features prove inferior in extracting cracks from real-world inspection images [10].

With the rapid evolution of AI, convolutional neural networks (CNNs) offer an automated approach to discerning features of target objects in pavement surface crack detection. By employing layers such as convolution and pooling, CNNs can learn and classify crack features without the need for human intervention in designing. Many CNN-based crack detection algorithms have been developed by leveraging object detection or classification methods, such as locating pavement surface cracks in the acquired image. However, there are challenges in achieving pixel-level accuracy with these methods and thus the further evaluation of these cracks cannot be conducted. To achieve pixel-level crack detection, representative models such as fully convolutional networks (FCNs) [11], Unet [12], DeepLabv3 [13], etc. have been adopted to train and test on pavement image datasets. For instance, in [14], Dung et al. have utilized FCN to detect pavement surface cracks. Nevertheless, this approach overlooks the fact that cracks of varying widths may necessitate distinct sizes of context information. U-net and DeepLabv3 models have also been utilized for pavement surface crack detection, but they do not consider the relationships between neighbor pixels and fail to globally model crack features.

Recently, motivated by the remarkable success of transformers, researchers have endeavored to extend the application of transformers into different computer vision tasks. The vision transformer (ViT) was introduced in [15] to address image recognition tasks by taking 2D image patches with positional embeddings as inputs and undergoing pre-training on extensive datasets. Additionally, the data-efficient image transformer (DeiT), presented in [16], highlights the potential for training transformers on mid-sized datasets. Thus, considering the above limitations of current CNN models on pavement surface crack detection and the advantages of transformer architecture on modeling long-range pixel dependencies, we propose an improved Swin-Unet model for pavement surface crack detection. Namely, the skip attention module based on the channel attention mechanism is designed to efficiently focus on the crack region and assign larger weights to crack feature channels. Additionally, the residual Swin Transformer block is employed in the encoder architecture to model the crack features from a global perspective. In this manner, the effective crack feature information can be further exchanged.

This paper is structured as follows: in the Related Works Section, we present a concise overview of crack detection methods including the digital image processing-based method and the deep learning-based method. In the Proposed Network Section, we provide detailed explanations of the improved Swin-Unet model, the skip attention module, and the residual Swin Transformer block. In the Experiments and Results Section, we outline the performed experiments along with the corresponding results and analysis. In the last section, we offer a summary of the study and future work to be conducted.

2. Related Works

In this section, we provide a brief overview of the existing literature concerning image processing-based and deep learning-based applications for automatic pavement surface crack detection methods.

2.1. Image Processing-Based Methods

In general, image processing-based methods can be categorized into three main groups: intensity threshold-based algorithms [17,18], filter-based methods [19,20], and minimal path selection techniques [21,22]. Intensity threshold-based algorithms are straightforward to implement in the field and can yield satisfactory results in acquired images under specific

pavement conditions without much noise and interference. For instance, Cheng et al. [23] developed a method that involves reducing the sample space and employing interpolation to determine the threshold value. In [24], the thresholded pavement surface images were divided into non-overlapping blocks for entropy computation. Subsequently, a second dynamic thresholding process was applied to predict cracks. Besides, a pavement surface crack detection method [25] was proposed based on the neighboring difference histogram method, and the objective function for maximizing the difference background and crack pixels was constructed. However, these algorithms prove less effective when dealing with weak contrasts between pavement surface cracks and the background. Notably, lighting conditions significantly impact detection accuracy, especially in the case of unevenly distributed lighting conditions. In addition, accuracy is impacted in scenarios where the image background exhibits noise and intricate textures. Thus, intensity-threshold methods have difficulty detecting the whole crack pattern and fail to identify complicated pavement surface cracks.

Pavement crack detection has seen the widespread adoption of edge detectors, such as the Sobel filter and the Canny filter, owing to the inherent similarity between cracks and edges in morphology, as noted by researchers [26–28]. However, a notable limitation of these algorithms is their inability to accurately identify complete crack profiles, especially when cracks are set against complex textured backgrounds. The methodology described in [29] diverges from employing a single filter and instead utilizes multiple pre-designed filters adept at capturing cracks based on various attributes such as pixel intensity, shape, and orientation. The extraction of cracks from the background occurs through successive filtering processes using these pre-designed filters. This algorithm effectively transforms pavement images into a new space, preserving cracks while eliminating the background. Nevertheless, a drawback of this approach is the sensitivity of the matched filters to crack width, resulting in challenges in accurately identifying pixels at crack borders.

Minimal path selection techniques have also garnered attention in pavement surface crack detection [21,22,30]. From the perspective of computer science, these techniques involve identifying optimal paths between node pairs in a graph. In the context of pavement surface detection, they aid in establishing a threshold of minimal distances to distinguish continuous crack pixels. Specifically, Zou et al. developed CrackTree [31], addressing challenges such as shadows and discontinuity in pavement crack detection. However, these approaches presupposed that crack fragments were connected through a minimum path or minimum spanning tree, ignoring the influence of neighboring pixels. Amhaz et al. proposed a crack detection method [21] based solely on photometry and introduced an algorithm for minimal path searching with fast search speed. Test results have confirmed its capability to outperform the classic Dijkstra's algorithm in terms of search speed.

2.2. Deep Learning-Based Method

With the rapid development of artificial intelligence, more and more researchers have focused on deep learning-based methods, especially CNN models, to perform pavement surface crack detection. Generally, CNN-based methods have included block-wise and pixel-wise approaches, respectively. For the block-wise method, object detection models (e.g., YOLO [32], Faster R-CNN [33], RetinaNet [34], etc.) have been commonly adopted. Du et al. [35] adopted YOLO to predict pavement surface cracks with possible location and category on a large pavement survey dataset. The inference speed can achieve 0.0347 s/pic, which is nine times faster than Faster R-CNN, and the detection accuracy has reached 73.64%. Liu et al. [36] developed a two-step method to detect pavement surface cracks. In the first step, the modified YOLOv3 is adopted and can perform the crack detection in an accurate manner. Gou et al. [37] developed a pavement surface crack detection model based on the improved Faster R-CNN with the residual neural network and the anchor size modification strategy. Experimental results validated that the proposed model was less affected by the illumination and the image quality. Zhai et al. [38] proposed the improved Faster R-CNN method for pavement surface crack detection. The classification

and regression network was developed by incorporating the residual network, while the model incorporated a feature ensemble structure. A notable limitation of object detection-based models for pavement surface crack detection is the challenge in accurately evaluating distress regions. Consequently, an increasing number of studies have turned to semantic segmentation models to achieve pixel-level detection of pavement surface cracks for better representation and evaluation.

FCN has been one of the commonly adopted semantic segmentation approaches to conduct road crack detection and measurement [28,30]. For instance, Liu et al. [39] utilized U-Net, an encoder–decoder structure, for concrete crack detection. Specifically, it included the focal loss, showing better robustness and accuracy compared to previous DCNNs. Lau et al. [40] introduced a U-Net-based architecture with the ResNet-34 encoder, incorporating layer group freezing, variable learning rates, and incremental image size increments to improve the detection performance. Song et al. [41] presented the Multi-scale Feature Attention Network for automatic crack identification. With the fused crack features, the detection performance was largely improved. In [42], researchers developed the weighted cross-entropy (CE) loss function and the distribution equalization learning mechanism for pavement crack detection. With the weighted CE loss function, the training process was simple and efficient.

In [43], the authors developed a new strategy which applied a DeepLabv3+ model to train the original pavement crack image and predicted the CLAHE-enhanced pavement crack images. The sensitivity analysis was performed to indicate the impact of the data volume and the shooting angle of the camera. Wang et al. [44] developed a salient object detection-based method for pavement crack detection. Namely, the hierarchical feature fusion module was developed to integrate crack features, and the boundary refinement module was proposed to refine the crack boundary. Taking advantage of the rapid development of transformers, Liu et al. [45] designed CrackFormer by incorporating the novel self-attention modules and efficient positional embedding to enhance the long-range interactions between neighbor crack pixels. Through suppressing non-semantic features and sharpening semantic crack pixels, the detection performance on cracks was largely improved. Guo et al. [3] improved the pixel-level crack detection performance by unifying the Swin Transformer as the encoder and the SegFormer as the decoder. Based on experimental results on three popular crack datasets, it can be found that the thin crack and the crack impacted by environmental noises can also be accurately detected. Lu et al. [46] have advanced the field of crack classification by ingeniously incorporating pretrained Swin Transformer models into U-Net frameworks, a methodology termed Crack_PSTU. This innovative amalgamation significantly enriches the model's capacity to discern intricate details and contextual nuances within crack imagery, thereby markedly elevating the accuracy and reliability of crack detection. Even CNN-based models and transformer-based models have exhibited distinct characteristics in modeling crack pixels. However, there have been few studies exploring how to leverage both approaches in a single model for crack pixel detection. The method proposed in this study aimed to leverage the strengths of both approaches to refine the pavement surface crack detection performance.

3. Proposed Network

3.1. Overview of the Developed Architecture

The enhanced architecture, depicted in Figure 1, is an evolution of the original Swin-Unet [47–49]. It features an improved Swin Transformer arranged in a "U" shape, serving as both the encoder and the decoder. Drawing inspiration from ResNet, the basic unit of Swin-Unet incorporates a residual block flow. During training, input samples undergo an initial partitioning into non-overlapping patches with a size of 4×4. This process transforms the inputs into sequence embeddings, facilitating the efficient calculation of pavement crack features. Each feature patch has a dimension of $4 \times 4 \times 3$, resulting in 48 features. To project this feature dimension into an arbitrary dimension denoted as C (in our study, C = 96), a linear embedding layer is employed. The transformed patch tokens

traverse residual Swin Transformer blocks and patch merging layers, creating pavement crack feature representations with varying scales. The patch merging layer handles down-sampling and increases the feature dimension, while the residual Swin Transformer block focuses on feature representation learning.

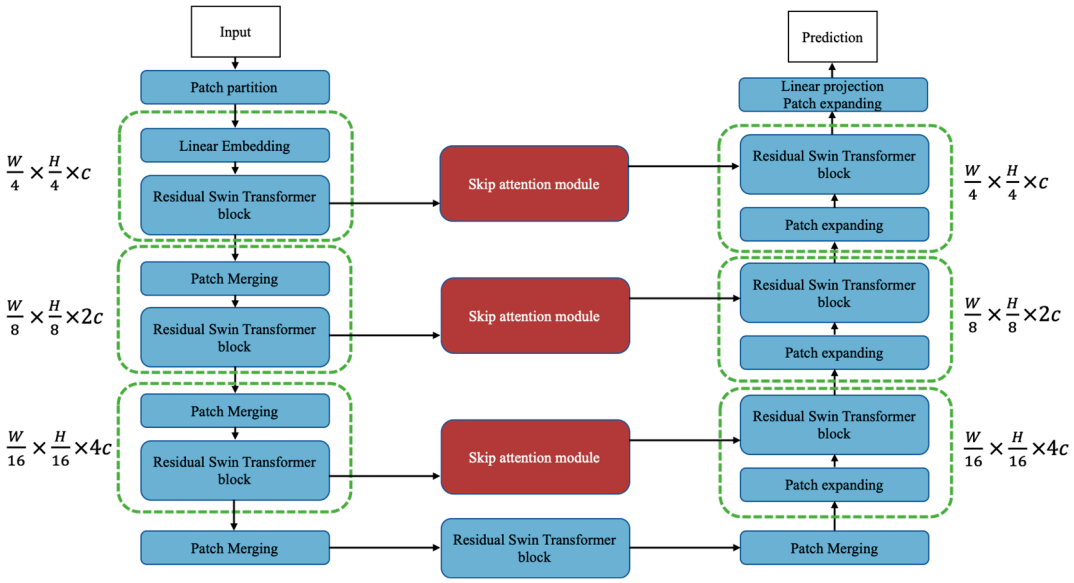

Figure 1. The architecture of the developed iSwin-Unet.

Capitalizing on the U-Net design, a symmetric decoder in the form of a "U" shape is meticulously crafted. The designed decoder incorporates residual Swin Transformer blocks and a patch expanding layer. To combat the loss of spatial information in pavement crack features, context features are extracted and fused with multiscale features from the encoder using skip attention layers. The skip attention layer serves a dual purpose—it suppresses redundant information from large background areas while enhancing the efficient utilization of crack features. In contrast to the patch merging layer, the patch expanding layer is dedicated to the up-sampling of pavement crack features. Specifically, this layer reshapes feature maps of adjacent dimensions into larger feature maps with a 2× up-sampling of resolution. The final layer involves the use of a last patch expanding layer for a 4× up-sampling, restoring the resolution of feature maps to match the original input resolution. In the last step, a linear projection layer is applied to the up-sampled pavement crack features, culminating in the generation of refined pavement crack detection results.

3.2. Residual Swin Transformer Block

Building upon the foundational framework of the multi-head self-attention (MSA) module inherent to the conventional transformer architecture, the original Swin Transformer block, serving as the basic unit of the novel model, is introduced. This block incorporates both the window-based multi-head self-attention (W-MSA) and the shifted window-based multi-head self-attention (SW-MSA) modules, thus augmenting the model's capacity for contextual understanding and feature extraction within localized regions of input sequences. As depicted in Figure 2, the basic components of the Swin Transformer block are presented. Specifically, it comprises a LayerNorm (LN) layer, a multi-head self-attention module, a residual connection, and a 2-layer MLP with GELU non-linearity. Specifically, the first block utilizes the W-MSA module while the second block adopts the SW-MSA module. On the basis of this window partitioning mechanism, a sequence of Swin Transformer blocks can be expressed in the following Equations (1)–(4).

$$\hat{z}^{l} = W - MSA\left(LN\left(z^{l-1}\right)\right) + z^{l-1} \quad (1)$$

$$z^{l} = MLP\left(LN\left(\hat{z}^{l}\right)\right) + \hat{z}^{l} \tag{2}$$

$$\hat{z}^{l+1} = SW - MSA\left(LN\left(z^{l}\right)\right) + z^{l} \tag{3}$$

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

where $\hat{z}^{l}$ and z^{l} are the outputs of the W-MSA and SW-MSA modules, respectively. Similar to the previous fashion of calculating self-attention, the similarity is computed based on the relative position encoding. The computation equation is shown in Equation (5).

$$Attention(Q, K, V) = SoftMax\left(\frac{QK^{T}}{\sqrt{d}} + B\right)V \tag{5}$$

where $Q, K, V \in \mathbb{R}^{M^2 \times d}$ indicate the query, key, and value, respectively. M^2 and d are the number of patches and the dimensions, respectively. Also, it needs to be mentioned that B comes from the bias matrix.

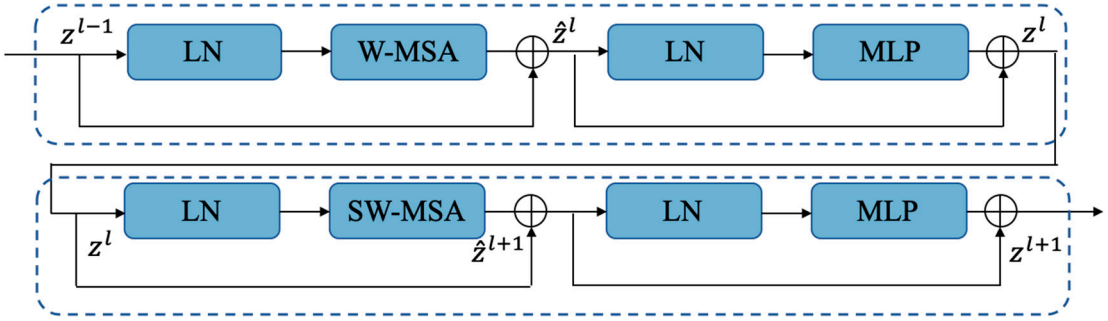

Figure 2. The architecture of the basic Swin Transformer block.

Inspired by the great success of ResNet and SwinIR, we develop the residual Swin Transformer block to enhance the crack feature aggregation. Figure 3 presents the architecture of the developed residual Swin Transformer block. This residual design offers two advantages. Firstly, the transformer structure can be conceptualized as a particular implementation of spatially varying convolution which can promote the feature extraction of the proposed model. Secondly, the inclusion of a residual connection establishes an identity-based link from diverse blocks to the decoder module, facilitating the amalgamation of pavement crack features across multiple levels. The following Equation (6) indicates the computation of the developed model.

$$\tilde{z}^{l} = Residual(z^{l}) + z^{l} \tag{6}$$

where $\tilde{z}^{l}$ is the output of the developed residual Swin Transformer block and z^{l} is the input.

3.3. Skip Attention Module

Semantic segmentation networks like U-Net, ResU-Net, Attention U-Net, Dense U-Net, etc. have dominated in network design, with their shared characteristic being the incorporation of skip layers to preserve spatial details and edge region information during decoding. While these skip connections effectively retain crucial crack features for prediction, they introduce a challenge by including redundant information in the extracted crack features. This redundancy disrupts crack feature extraction, which can result in irrelevant and false predictions during network training.

To conquer this challenge, we adopt the attention mechanism, a successful implementation in CNN-based networks. The Attention U-Net, introduced by Oktay et al. [50], integrates attention gates to suppress irrelevant regions during training, reducing redundant feature maps and inhibiting activations in irrelevant areas. While effective in mitigating the impact of redundant information, this approach involves a distinct drawback, as it decreases the proportion of features in the decoding process for the entire network. Recognizing the distinct design of attention modules in CNNs and transformers, we propose a novel skip attention module to connect the transformer-based encoder and decoder, aiming to enhance the overall performance of the network in semantic segmentation tasks with a pixel-level manner.

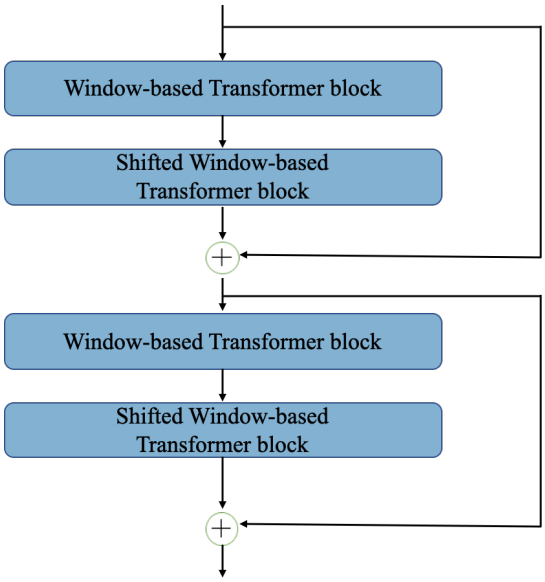

Figure 3. The proposed residual Swin Transformer block.

Figure 4 illustrates the architecture of the residual Swin Transformer. This design features a parallel structure that conducts spatial and channel attention operations concurrently. The process commences with spatial normalization, during which the attention weight (W_{att}) is computed for each encoder block to accentuate the informative tokens that carry salient crack features. Further, by calculating the attention weight ($W_{att} = softmax(\frac{Q_e K_e^T}{\sqrt{d}} + B)$) through the mapping from the encoding to the decoding, the area importance can be determined. As illustrated in the previous section, the attention value in the i-th scale can be calculated in the following Equation (7).

$$Att^i(Q_d, K_d, V_d) = (softmax\left(\frac{Q_e K_e^T}{\sqrt{d}} + B\right) + W_{att})V_d \tag{7}$$

where $Att^i(Q_d, K_d, V_d)$ represents the weight of the i-th decoder path. This weight calculation method is adopted in all three decoding paths to update the weight with the corresponding scale during the training. To enhance the interaction between two feature calculates, two series including the spatial branch and the channel branch are included to calculate the generated features. Firstly, the token average operation is conducted for the purpose of global representation. Then, the global representation is fused with the token computed from the original version for the reconsidering of the feature representation to perform the channel-wise attention calculation. The following Equations (8) and (9) reflect the reasoning process.

$$q = Z_{D_{globe}} W_j, k = Z\prime_D W_k, v = Z\prime_D W_l \tag{8}$$

$$A = softmax\left(\frac{qk^T}{\sqrt{\frac{C}{h}}}\right), CA\left(Z'_D\right) = Av \tag{9}$$

where W_j, W_k, and W_l mean the parameters that can be used in the training. C and h represent the dimension of the embedding space and the head, respectively. The CA represents the hybrid attention operation that we computed. With this method, not only can these two attention operations be achieved, but it also offers a distinctive method for formulating the manner of interaction in a non-linear fashion to connect the encoder and the decoder.

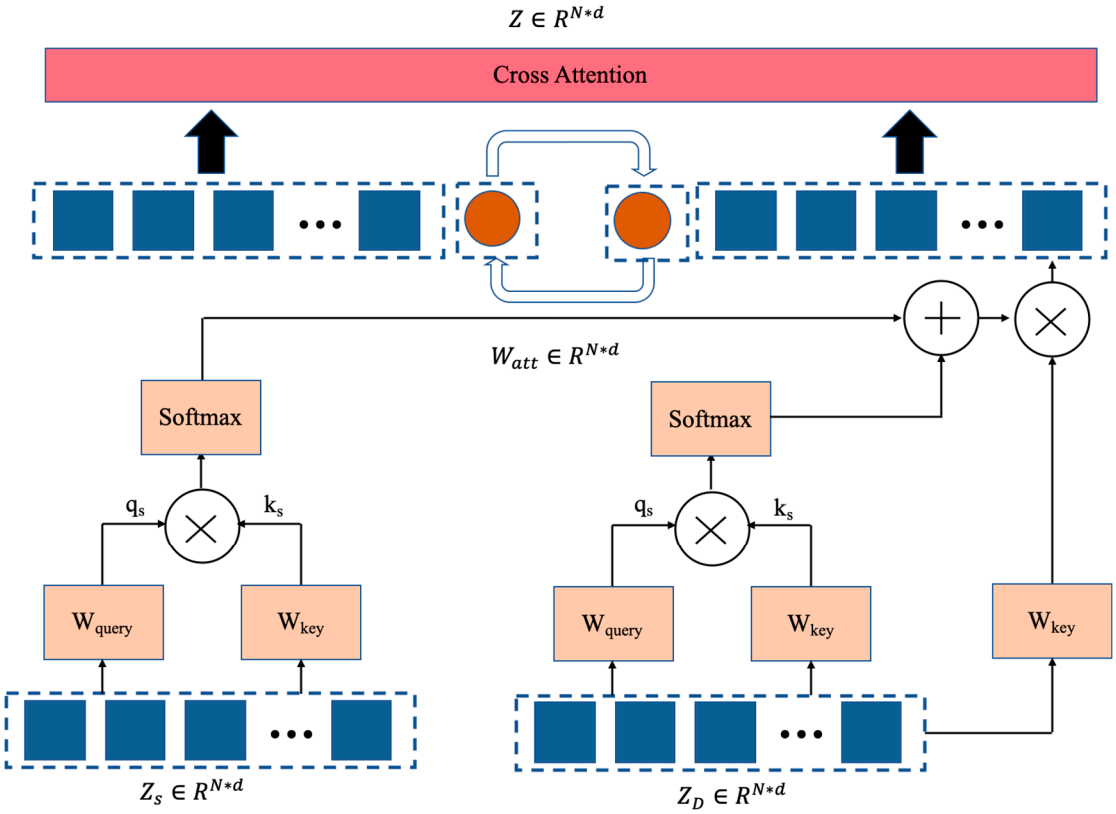

Figure 4. The structure of the skip attention module.

4. Experiments and Results

4.1. Experimental Setting

We conducted our experiments using a deep learning machine equipped with the NVIDIA 3080 Ti GPU, featuring 12 GB of memory. The operating system employed was Ubuntu (version 20.04). For model training, validation, and testing, we utilized the MMSegmentation library [51], built on the PyTorch framework. The study involved the comparison of two models, namely Unet and Swin Transformer, across visualization and inference tests. Default configurations and training parameters were adhered to for each model, with training iterations set at 12 epochs. The model training utilized three public pavement surface crack datasets.

4.2. Experimental Data

The CFD dataset [52,53] is a commonly adopted pavement crack dataset which consists of 118 RGB pavement images captured by an iPhone 5 from Beijing, China, and corresponding annotations created with binary pixels. The resolution of each image is 480 × 320 pixels. The phone camera has a focus of 4mm, an aperture of f/2.4, and an exposure time of 1/135s. The dataset provides annotated pixel-level labels, indicating the presence and location of cracks in the images. With annotated images, CFD serves as a valuable resource for

training and evaluating deep learning models for pavement crack detection. Image and label samples can be found in Figure 5a.

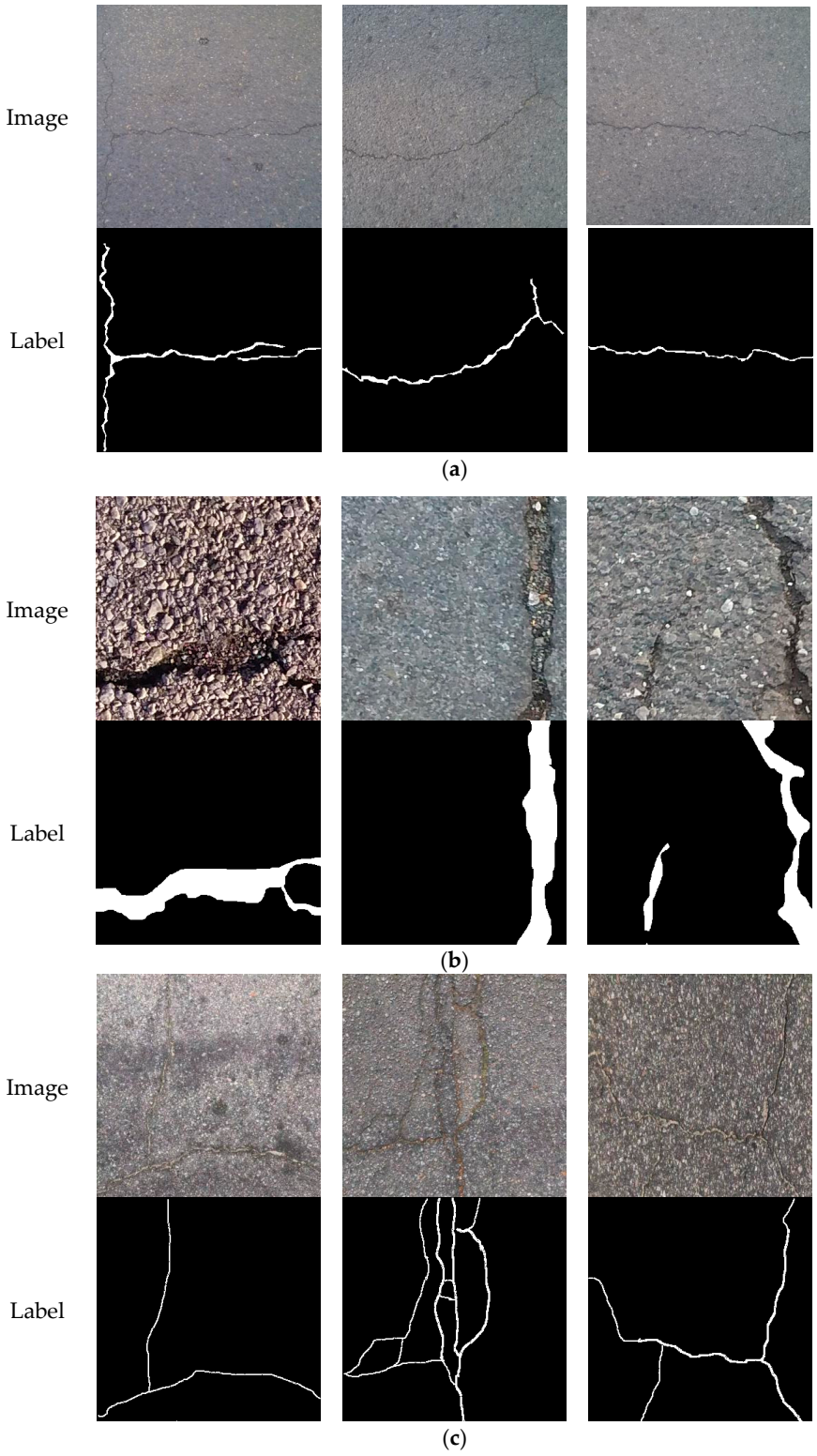

Figure 5. Image and label samples of three datasets. (**a**) CFD dataset; (**b**) Crack500 dataset; (**c**) CrackSC dataset.

The Crack500 dataset [54] includes 500 RGB images featuring pavement surface cracks, where each image has a resolution of 2000 × 1500 pixels. These images are captured using

cell phones on the main campus of Temple University, United States. Notably, the dataset contains pixel-level annotations that precisely delineate the location and boundaries of cracks in the images. With its meticulously curated content and detailed annotations, Crack500 stands as a valuable resource for training and evaluating deep learning models specifically designed for crack detection. Image and label samples can be found in Figure 5b.

The CrackSC dataset [3] consists of 197 RGB images depicting pavement surfaces with noticeable noise. Captured using an iPhone 8 near Enoree Ave, Columbia, SC, these images serve as a benchmark dataset for evaluating pavement surface crack segmentation algorithms. The primary goal is to encourage the development of accurate and robust methods for detecting pavement cracks, thereby advancing the field of computer vision in transportation infrastructure inspection and maintenance. Image and label samples can be found in Figure 5c.

Three widely employed evaluation metrics—precision, recall, and F1 score—are employed. The true positive (TP), true negative (TN), false positive (FP), and false negative (FN) indicators are utilized to calculate these three evaluation metrics. Precision and recall measure the accuracy of positive predictions and the completeness of positive predictions, respectively. Typically, a higher precision indicates a lower recall and vice versa. Thus, a balanced indicator is needed in the prediction performance evaluation. It is important to note that the F1 score is chosen for its ability to balance the trade-off between precision and recall, making it a superior indicator for reflecting model performance compared to the other two metrics. Given that crack shapes are typically long and thin, the conventional indicator of intersection over union (IoU) is deemed inappropriate for performance evaluation. Many studies introduce pixel tolerances (e.g., two-pixel or five-pixel) between real crack pixels and the background. However, such an approach often leads to the inaccurate evaluation of numerous pixels, failing to truly represent a model's proficiency in pixel classification. Therefore, we opt for a zero-pixel tolerance, and the experimental results indicate lower performance compared to previous studies. Furthermore, we compute the values of each class's indicators and use their mean values for comparison. The following equations outline the computation process for these indicators.

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{12}$$

$$mPrecision = \frac{\sum Precision}{N} \tag{13}$$

$$mRecall = \frac{\sum Recall}{N} \tag{14}$$

$$mF1 = \frac{\sum F1}{N} \tag{15}$$

where N is the number of classes. In this study, N equals 2, indicating the background and the pavement crack.

4.3. Training Performance

From Tables 1–3, we compare the performance of the adopted method with two other popular models on the CFD, Crack500, and CrackSC datasets. In Table 1, it is evident that our proposed model achieves the maximum values of mRecall, mF1, and mPrecision among all the models. Specifically, iSwin-Unet's mF1 value is 1.60%, 2.46%, and 2.52% higher than Swin-Unet, Swin Transformer, and Unet, respectively. Regarding mPrecision, iSwin-Unet's value is 1.26%, 2.52%, and 1.99% higher than Swin-Unet, Swin Transformer,

and Unet, respectively. In terms of mRecall, iSwin-Unet outperforms Swin-Unet, Swin Transformer, and Unet by 2.17%, 1.80%, and 0.37%, respectively. This indicates that our adopted method achieves a better recall score, which is crucial for detecting pavement surface cracks.

Table 1. Training performance on the CFD dataset.

Model	mF1 (%)	mPrecision (%)	mRecall (%)
iSwin-Unet	84.58	83.44	85.64
Swin-Unet	83.23	82.39	83.78
Swin Transformer	82.50	81.34	84.10
Unet	82.45	81.78	85.32

Table 2. Training performance on the Crack500 dataset.

Model	mF1 (%)	mPrecision (%)	mRecall (%)
iSwin-Unet	86.98	85.96	86.12
Swin-Unet	86.33	85.89	85.10
Swin Transformer	85.54	85.83	84.36
Unet	80.44	81.23	80.06

Table 3. Training performance on the CrackSC dataset.

Model	mF1 (%)	mPrecision (%)	mRecall (%)
iSwin-Unet	78.12	74.76	78.53
Swin-Unet	77.21	73.23	76.89
Swin Transformer	76.40	73.99	77.87
Unet	74.88	72.23	76.52

Table 2 illustrates the training performance of three models on the Crack500 dataset. With respect to the mF1 score, it is easy to find that our proposed model achieves 86.98%, which is 0.75%, 1.66%, and 7.52% higher than Swin-Unet, Swin Transformer, and Unet, respectively. In terms of the indicator of mPrecision, our proposed method is 0.08%, 0.15%, and 5.50% higher than Swin-Unet, Swin Transformer, and Unet, respectively. Additionally, the mRecall value of the developed method is 1.18%, 2.04%, and 7.04% higher than Swin-Unet, Swin Transformer, and Unet, respectively.

Table 3 presents the training performance of three models on the CrackSC dataset. It is noteworthy that our proposed model achieves an mF1 score of 78.12%, which is 1.16%, 2.20%, and 4.15% higher than Swin-Unet, Swin Transformer, and Unet, respectively. As for the metric of mPrecision, our adopted method is 2.05%, 1.03%, and 3.38% higher than Swin-Unet, Swin Transformer, and Unet, respectively. Further, it is worth highlighting that our adopted method achieves the highest value of mRecall, which stands at 78.53%.

4.4. Visualization Results

From Figures 6–8, the visualized results of the three aforementioned datasets are presented, featuring four randomly selected crack images from each dataset. In Figure 6, it is evident that the figures in the first and fourth rows exhibit similar shapes of cracks, while those in the second and third rows showcase more intricate crack patterns. In the first row, a clear observation reveals that the long and thin cracks with branches, generated by both our adopted method and Unet, closely resemble the ground truth labeling. However, the results segmented by Unet show a noticeable loss of crack information, producing an unsatisfactory pattern. Moving to the second row, where the input image contains a complex crack pattern, the prediction results of Unet and Swin Transformer exhibit a greater loss of details compared to the ground truth label. Notably, our improved Swin-Unet demonstrates better continuity on branch cracks when compared to Swin-Unet. In the

third row, the bottom section of disconnected cracks is apparent in the prediction results of Swin-Unet, Swin Transformer, and Unet, indicating weaker segmentation performances on this particular image. Regarding the fourth row featuring a curve-shaped crack, due to its relatively simple appearance, all models successfully predict the crack, with the exception of noise pixels predicted by Swin Transformer.

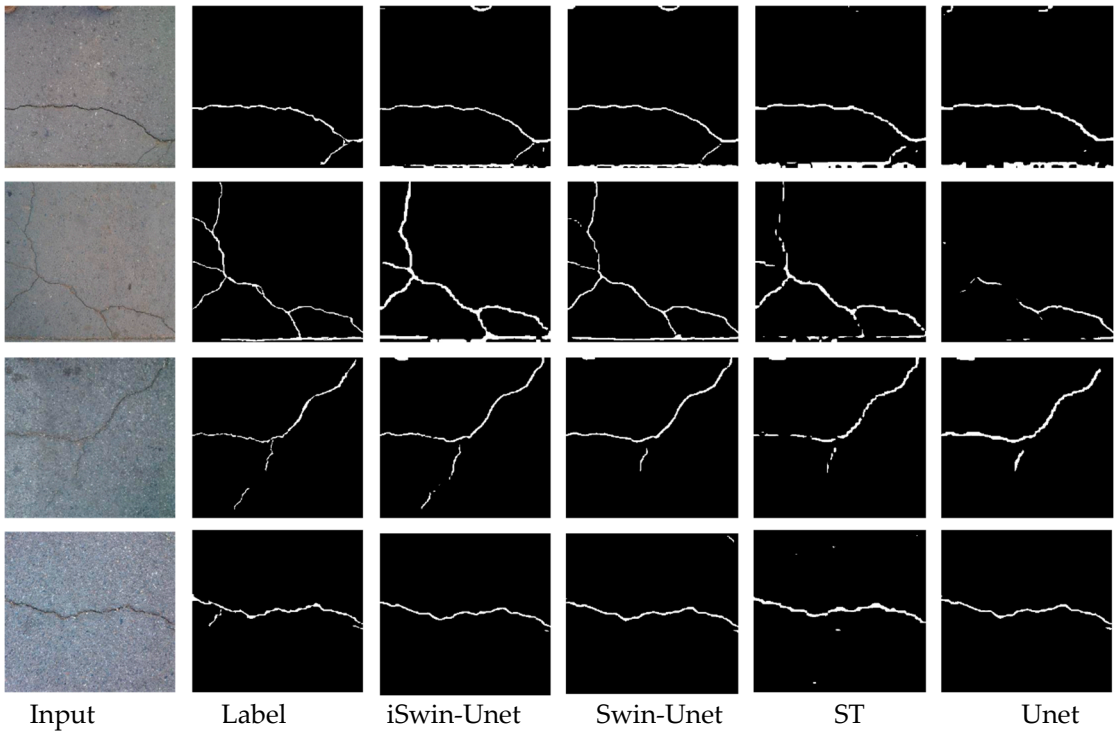

Figure 6. Prediction results of the CFD dataset.

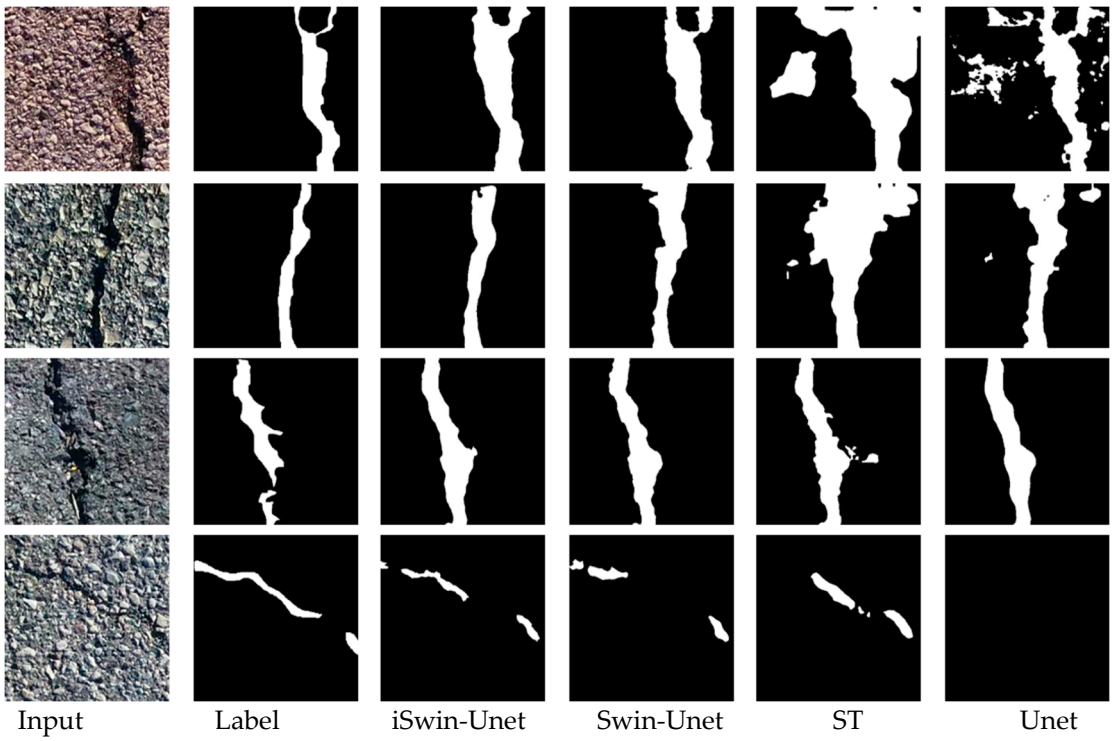

Figure 7. Prediction results of the Crack500 dataset.

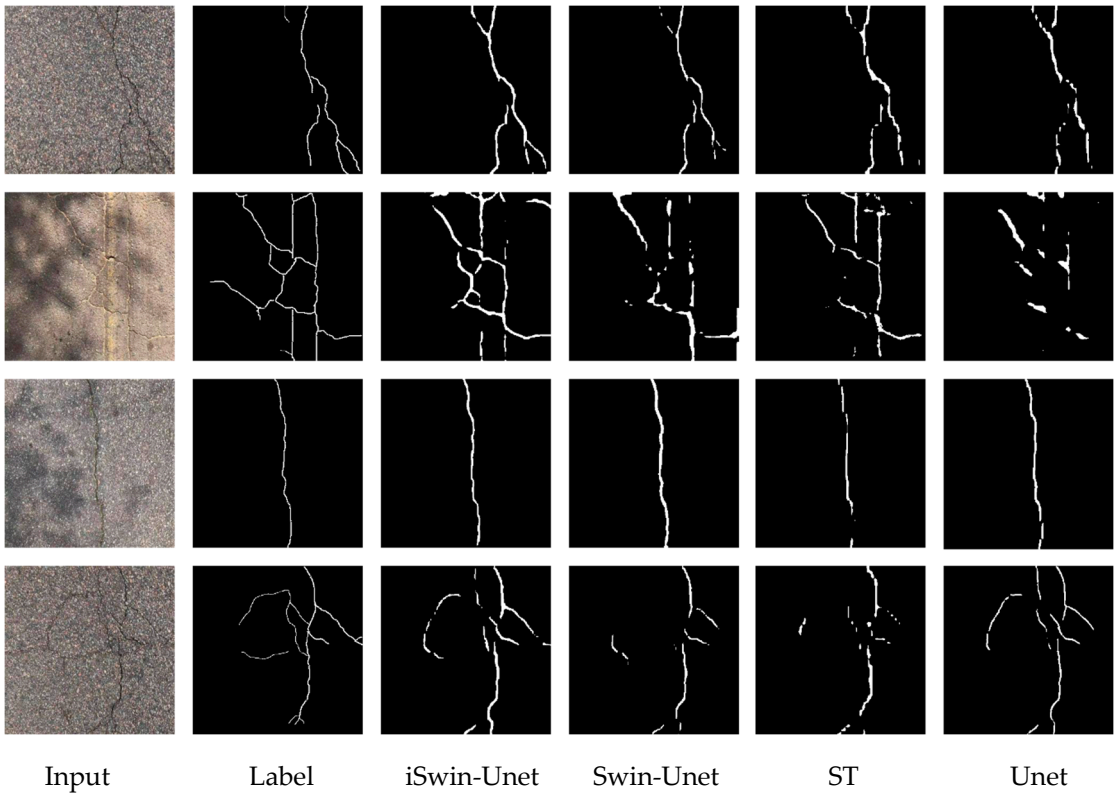

Figure 8. Prediction results of the CrackSC dataset.

In Figure 7, four randomly selected prediction results of Crack500 images are showcased, generated by our proposed iSwin-Unet, Swin-Unet, Swin Transformer, and Unet. In the first row, a thick crack appears on the right side of the image, and only our proposed iSwin-Unet and Swin-Unet achieve results similar to the ground truth labeling. In contrast, Swin Transformer produces the least accurate result, displaying a considerably different pattern compared to the ground truth labeling. Moving to the second row, the original input features a thick crack with a clear boundary between the crack and the background. Apart from our proposed model, all predictions fail to appropriately recover the full shape of the pavement crack. Notably, the results of Swin Transformer and Unet are much wider than the actual crack. In the third row, all models accurately predict a short and deep crack. However, Swin Transformer falsely predicts an apparent bulge, unlike other models. Turning to the fourth row, where a less obvious crack is present in the input, all models struggle due to the challenging condition in which the crack "hides" in the background texture. Remarkably, the prediction result of the proposed model is still the closest to the ground truth labeling, while Unet predicts no crack pixels in the visualization.

In Figure 8, four randomly selected prediction results of CrackSC images are displayed, generated by our proposed iSwin-Unet, Swin-Unet, Swin Transformer, and Unet. The CrackSC dataset presents a tougher challenge compared to the CFD and Crack500 datasets, as it includes images with leaves or shadows. In the first row, the original image features a branch-shaped crack, adding complexity to the prediction task. All models can predict the rough contour of the crack, and notably, our adopted model and Swin-Unet accurately predict the crack. Moving to the second row, a net-shaped crack with heavy shadows is presented. Our proposed model demonstrates better recoverability on the original shape with intersecting patterns compared to other models. Unet, in contrast, only predicts the thick crack segments. In the third row, a long and thin crack is observed on the right side of the image, with a tree shadow along the left side. With little interference from the shadow on the crack, the prediction results are relatively complete. Our improved model and the original Swin-Unet exhibit greater accuracy compared to Swin Transformer and Unet in

this scenario. In the fourth row, the prediction becomes more challenging due to the unclear boundary between the crack and the background, coupled with a complicated pattern. The segmentation results generated by our proposed model showcase superior performance on both the backbone and branches, indicating robustness across different scenarios.

4.5. Computational Efficiency

In this section, we conducted a comparative analysis of the computational efficiency of iSwin-Unet, Swin-Unet, ST, and Unet using indicators such as frames per second (FPS), parameters, and floating point operations (FLOPs), which are commonly used to evaluate inference performance in the test stage. Based on Table 4, it is evident that the developed iSwin-Unet model achieves the highest inference speed (50.65 FPS) with an image resolution of 320 × 480 on a single NVIDIA 3080Ti GPU. Specifically, it surpasses Swin-Unet, ST, and Unet by 4.6%, 0.6%, and 42.9%, respectively. In terms of model parameters, all models exhibit similar parameter sizes. The parameter count of the developed model is 6.1% smaller than Swin-Unet and 2.7% smaller than Unet, while it is 28.1% larger than the ST model. Regarding FLOPs, the developed model, Swin-Unet, and Unet show similar values, whereas the ST model has significantly fewer FLOPs. In summary, the proposed model achieves the highest inference speed for pavement crack detection, indicating its potential practical implementation.

Table 4. Indicators of the inference speed.

Model	FPS	Parameters (M)	FLOPs (G)
iSwin-Unet	50.65	28.12	115.67
Swin-Unet	48.33	29.83	118.34
ST	50.34	20.21	24.54
Unet	28.92	28.87	117.95

5. Conclusions

In this paper, a pixel-level semantic segmentation model is improved based on Swin-Unet to achieve better prediction performance for pavement crack detection. Specifically, to achieve this goal, the skip attention module based on the channel attention mechanism is developed to focus on the crack region and assign larger weights to pavement crack feature channels. The residual Swin Transformer block is developed in the encoder architecture to model the crack features from a global perspective. The proposed module enhances the ability to fuse features from multi-scales, resulting in more comprehensive and informative representations of pavement cracks.

To assess the performance of the adopted network, a total of four models were trained and tested on three public datasets (CFD, Crack500, and CrackSC). Evaluation metrics such as mPrecision, mRecall, and mF1 are used to validate the accurate detection performance of our proposed model, while visualized results are also employed for evaluation purposes. From both the training performance and visualization results, it can be found that our proposed model achieves better accuracy and efficiency in pavement crack detection by better capturing and prioritizing crack feature channels and enabling global modeling of pavement crack features, thereby marking a substantial advancement over existing models in the field.

In the future, our exploration will focus on two key aspects. Firstly, despite the model's effectiveness on workstations, there is a pressing need to optimize the computational efficiency of crack detection models. This involves not only reducing the model size but also enhancing its processing speed without compromising accuracy. Secondly, it is imperative to concentrate on the practical deployment of crack detection models, particularly in terms of scalability and integration into existing infrastructure management systems. Thus, future investigations will be dedicated to the formulation of standardized protocols for the deployment of models, along with the advancement of scalable, cloud-based infrastructures designed to process and analyze pavement condition data over broad road networks.

Author Contributions: Conceptualization, S.C., H.Y. and Z.F.; methodology, S.C., H.Y., Z.F. and G.X.; validation, Z.F., G.X., C.G., M.Z. and X.C.; investigation, G.X., X.C., M.Z. and C.G.; writing—original draft preparation, S.C. and Z.F.; supervision, H.Y.; funding acquisition, H.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Natural Science Foundation of China (grant number 52208437), Guangdong Basic and Applied Basic Research Foundation (grant numbers 2022A1515011537, 2022A1515011607, and 2020A1515110900), and the Fundamental Research Funds for the Central Universities (grant number 2022ZYGXZR056).

Data Availability Statement: The data presented in this study are available on request from the corresponding authors.

Acknowledgments: The authors would like to thank the technicians in the road laboratories of South China University of Technology for technical support and assistance in the experimental activities.

Conflicts of Interest: Authors Song Chen, Guangqing Xiao and Xilong Chen were employed by the company Guangzhou Guangjian Construction Engineering Testing Center Co., Ltd. Author Mingming Zhao was employed by the company Yantai Donghua Material Science Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Zhang, A.; Wang, K.C.; Li, B.; Yang, E.; Dai, X.; Peng, Y.; Fei, Y.; Liu, Y.; Li, J.Q.; Chen, C. Automated pixel-level pavement crack detection on 3D asphalt surfaces using a deep-learning network. *Comput.-Aided Civ. Infrastruct. Eng.* **2017**, *32*, 805–819. [CrossRef]
2. Guo, F.; Liu, J.; Lv, C.; Yu, H. A novel transformer-based network with attention mechanism for automatic pavement crack detection. *Constr. Build. Mater.* **2023**, *391*, 131852. [CrossRef]
3. Guo, F.; Qian, Y.; Liu, J.; Yu, H. Pavement crack detection based on transformer network. *Autom. Constr.* **2023**, *145*, 104646. [CrossRef]
4. Gong, H.; Liu, L.; Liang, H.; Zhou, Y.; Cong, L. A State-of-the-art survey of deep learning models for automated pavement crack segmentation. *Int. J. Transp. Sci. Technol.* **2023**, *13*, 44–57. [CrossRef]
5. Gong, H.; Sun, Y.; Hu, W.; Huang, B. Neural networks for fatigue cracking prediction using outputs from pavement mechanistic-empirical design. *Int. J. Pavement Eng.* **2021**, *22*, 162–172. [CrossRef]
6. Wang, W.; Wang, M.; Li, H.; Zhao, H.; Wang, K.; He, C.; Wang, J.; Zheng, S.; Chen, J. Pavement crack image acquisition methods and crack extraction algorithms: A review. *J. Traffic Transp. Eng. (Engl. Ed.)* **2019**, *6*, 535–556. [CrossRef]
7. Zakeri, H.; Nejad, F.M.; Fahimifar, A. Image based techniques for crack detection, classification and quantification in asphalt pavement: A review. *Arch. Comput. Methods Eng.* **2017**, *24*, 935–977. [CrossRef]
8. Cao, W.; Liu, Q.; He, Z. Review of pavement defect detection methods. *Ieee Access* **2020**, *8*, 14531–14544. [CrossRef]
9. Huyan, J.; Li, W.; Tighe, S.; Xu, Z.; Zhai, J. CrackU-net: A novel deep convolutional neural network for pixelwise pavement crack detection. *Struct. Control. Health Monit.* **2020**, *27*, e2551. [CrossRef]
10. Kheradmandi, N.; Mehranfar, V. A critical review and comparative study on image segmentation-based techniques for pavement crack detection. *Constr. Build. Mater.* **2022**, *321*, 126162. [CrossRef]
11. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
12. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015.
13. Chen, L.-C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.
14. Dung, C.V. Autonomous concrete crack detection using deep fully convolutional neural network. *Autom. Constr.* **2019**, *99*, 52–58. [CrossRef]
15. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2010**, arXiv:2010.11929.
16. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jégou, H. Training data-efficient image transformers & distillation through attention. In Proceedings of the 38th International Conference on Machine Learning, Virtual Event, 18–24 July 2021; Volume 139, pp. 10347–10357.
17. Peng, C.; Yang, M.; Zheng, Q.; Zhang, J.; Wang, D.; Yan, R.; Wang, J.; Li, B. A triple-thresholds pavement crack detection method leveraging random structured forest. *Constr. Build. Mater.* **2020**, *263*, 120080. [CrossRef]
18. Akagic, A.; Buza, E.; Omanovic, S.; Karabegovic, A. 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), Opatija, Croatia, 21–25 May 2018.

19. Hassan, N.; Mathavan, S.; Kamal, K. Road crack detection using the particle filter. In Proceedings of the 2017 23rd International Conference on Automation and Computing (ICAC), Huddersfield, UK, 7–8 September 2017.
20. Li, S.; Cao, Y.; Cai, H. Automatic pavement-crack detection and segmentation based on steerable matched filtering and an active contour model. *J. Comput. Civ. Eng.* **2017**, *31*, 04017045. [CrossRef]
21. Amhaz, R.; Chambon, S.; Idier, J.; Baltazart, V. Automatic crack detection on two-dimensional pavement images: An algorithm based on minimal path selection. *IEEE Trans. Intell. Transp. Syst.* **2016**, *17*, 2718–2729. [CrossRef]
22. Chen, Y.; Liang, J.; Gu, X.; Zhang, Q.; Deng, H.; Li, S. An improved minimal path selection approach with new strategies for pavement crack segmentation. *Measurement* **2021**, *184*, 109877. [CrossRef]
23. Cheng, H.; Shi, X.; Glazier, C. Real-time image thresholding based on sample space reduction and interpolation approach. *J. Comput. Civ. Eng.* **2003**, *17*, 264–272. [CrossRef]
24. Oliveira, H.; Correia, P.L. Automatic road crack segmentation using entropy and image dynamic thresholding. In Proceedings of the 2009 17th European Signal Processing Conference, Glasgow, UK, 24–28 August 2009.
25. Li, Q.; Liu, X. Novel approach to pavement image segmentation based on neighboring difference histogram method. In Proceedings of the 2008 Congress on Image and Signal Processing, Sanya, China, 27–30 May 2008.
26. Wang, W.; Li, H.; Wang, K.; He, C.; Bai, M. Pavement crack detection on geodesic shadow removal with local oriented filter on LOF and improved Level set. *Constr. Build. Mater.* **2020**, *237*, 117750. [CrossRef]
27. Zhang, A.; Li, Q.; Wang, K.C.; Qiu, S. Matched filtering algorithm for pavement cracking detection. *Transp. Res. Rec.* **2013**, *2367*, 30–42. [CrossRef]
28. Salman, M.; Mathavan, S.; Kamal, K.; Rahman, M. Pavement crack detection using the Gabor filter. In Proceedings of the 16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013), Hague, The Netherlands, 6–9 October 2013.
29. Cubero-Fernandez, A.; Rodriguez-Lozano, F.J.; Villatoro, R.; Olivares, J.; Palomares, J.M. Efficient pavement crack detection and classification. *EURASIP J. Image Video Process.* **2017**, *2017*, 39. [CrossRef]
30. Chatterjee, A.; Tsai, Y.-C. A fast and accurate automated pavement crack detection algorithm. In Proceedings of the 2018 26th European Signal Processing Conference (EUSIPCO), Rome, Italy, 3–7 September 2018.
31. Zou, Q.; Cao, Y.; Li, Q.; Mao, Q.; Wang, S. CrackTree: Automatic crack detection from pavement images. *Pattern Recognit. Lett.* **2012**, *33*, 227–238. [CrossRef]
32. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
33. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*. [CrossRef]
34. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; IEEE: New York, NY, USA, 2017; pp. 2980–2988.
35. Du, Y.; Pan, N.; Xu, Z.; Deng, F.; Shen, Y.; Kang, H. Pavement distress detection and classification based on YOLO network. *Int. J. Pavement Eng.* **2021**, *22*, 1659–1672. [CrossRef]
36. Liu, J.; Yang, X.; Lau, S.; Wang, X.; Luo, S.; Lee, V.C.S.; Ding, L. Automated pavement crack detection and segmentation based on two-step convolutional neural network. *Comput.-Aided Civ. Infrastruct. Eng.* **2020**, *35*, 1291–1305. [CrossRef]
37. Gou, C.; Peng, B.; Li, T.; Gao, Z. Pavement crack detection based on the improved faster-rcnn. In Proceedings of the 2019 IEEE 14th International Conference on Intelligent Systems and Knowledge Engineering (ISKE), Dalian, China, 14–16 November 2019.
38. Zhai, J.; Sun, Z.; Huyan, J.; Li, W.; Yang, H. Feature representation improved Faster R-CNN model for high-efficiency pavement crack detection. *Can. J. Civ. Eng.* **2022**, *50*, 114–125. [CrossRef]
39. Liu, Z.; Cao, Y.; Wang, Y.; Wang, W. Computer vision-based concrete crack detection using U-net fully convolutional networks. *Autom. Constr.* **2019**, *104*, 129–139. [CrossRef]
40. Lau, S.L.; Chong, E.K.; Yang, X.; Wang, X. Automated pavement crack segmentation using u-net-based convolutional neural network. *IEEE Access* **2020**, *8*, 114892–114899. [CrossRef]
41. Song, W.; Jia, G.; Jia, D.; Zhu, H. Automatic pavement crack detection and classification using multiscale feature attention network. *IEEE Access* **2019**, *7*, 171001–171012. [CrossRef]
42. Fang, J.; Qu, B.; Yuan, Y. Distribution equalization learning mechanism for road crack detection. *Neurocomputing* **2021**, *424*, 193–204. [CrossRef]
43. Wang, X.; Wang, T.; Li, J. Advanced crack detection and quantification strategy based on CLAHE enhanced DeepLabv3+. *Eng. Appl. Artif. Intell.* **2023**, *126*, 106880. [CrossRef]
44. Wang, Y.; Song, K.; Liu, J.; Dong, H.; Yan, Y.; Jiang, P. RENet: Rectangular convolution pyramid and edge enhancement network for salient object detection of pavement cracks. *Measurement* **2021**, *170*, 108698. [CrossRef]
45. Liu, H.; Miao, X.; Mertz, C.; Xu, C.; Kong, H. Crackformer: Transformer network for fine-grained crack detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; IEEE: New York, NY, USA, 2021; pp. 3783–3792.
46. Lu, W.; Qian, M.; Xia, Y.; Lu, Y.; Shen, J.; Fu, Q.; Lu, Y. Crack_PSTU: Crack detection based on the U-Net framework combined with Swin Transformer. In *Structures*; Elsevier: Amsterdam, The Netherlands, 2024; p. 106241.

47. Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; Wang, M. Swin-unet: Unet-like pure transformer for medical image segmentation. In Proceedings of the European conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer Nature: Cham, Switzerland, 2022; Volume 13803, pp. 205–218.
48. Liu, S.; Jin, J.; Yu, H.; Qian, G.; Zhang, B.; Shi, J.; Gao, Y. Promotional effect of shaped coal gangue composite phase change agents doping in asphalt on pavement properties. *Constr. Build. Mater.* **2024**, *411*, 134447. [CrossRef]
49. Liu, S.; Jin, J.; Yu, H.; Gao, Y.; Du, Y.; Sun, X.; Qian, G. Performance enhancement of modified asphalt via coal gangue with microstructure control. *Constr. Build. Mater.* **2023**, *367*, 130287. [CrossRef]
50. Oktay, O.; Schlemper, J.; Folgoc, L.L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Hammerla, N.Y.; Kainz, B.; et al. Attention u-net: Learning where to look for the pancreas. *arXiv* **2018**, arXiv:1804.03999.
51. MMSegmentation: Openmmlab Semantic Segmentation Toolbox and Benchmark. 2020. Available online: https://github.com/open-mmlab/mmsegmentation (accessed on 18 May 2022).
52. Shi, Y.; Cui, L.; Qi, Z.; Meng, F.; Chen, Z. Automatic road crack detection using random structured forests. *IEEE Trans. Intell. Transp. Syst.* **2016**, *17*, 3434–3445. [CrossRef]
53. Jin, J.; Gao, Y.; Wu, Y.; Liu, S.; Liu, R.; Wei, H.; Qian, G.; Zheng, J. Rheological and adhesion properties of nano-organic palygorskite and linear SBS on the composite modified asphalt. *Powder Technol.* **2021**, *377*, 212–221. [CrossRef]
54. Zhang, L.; Yang, F.; Zhang, Y.D.; Zhu, Y.J. Road crack detection using deep convolutional neural network. In Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA, 25–28 September 2016.