applied sciences

MDPI

Article

Crack Segmentation Model for Low-Quality Crack Images Based on Feature Integration and Triple Attention

Yonghua Xie and Yuyang Wang *

School of Computer Science, Nanjing University of Information Science and Technology, Nanjing 210044, China; yonghua@nuist.edu.cn

* Correspondence: 202312490632@nuist.edu.cn; Tel.: +86-18114022292

Featured Application

This study can provide reliable technical support for road crack detection in intelligent transportation systems. The proposed method is suitable for automatic inspection vehicles equipped with high-definition cameras. It can accurately mark the location of road cracks even in low-contrast pavement backgrounds. This method provides a good basis for preventing potential pavement distress in later service stages and provides key data for the predictive maintenance of roads in intelligent urban infrastructure management.

Abstract

To address the problem of road crack detection in low-quality pavement images, existing semantic segmentation methods still have shortcomings such as missed crack detection and inaccurate localization due to weak crack boundaries, low contrast, and complex pavement texture. To address these limitations, this study proposes a crack segmentation model based on feature integration and a triple attention mechanism. The model uses DeepLabv3+ as the backbone network and introduces the proposed three-dimensional interactive attention module after feature extraction. The attention module enhances the extraction of key features related to the spatial location and morphological details of cracks, thereby improving the ability of crack location. A hierarchical feature integration branch is introduced in the cross-layer connection, and a dimension-aware selective fusion module is used to enhance the saliency of small cracks in complex backgrounds. In addition, the proposed multi-group dilation feature fusion module is introduced to improve the multi-scale modeling of small and slender cracks and reduce background interference. The experimental results on Crack500 and GAPS384 datasets show that the proposed model achieves better overall segmentation performance than the comparison model, especially in reducing the missed detection of weak, small, and discontinuous cracks in low-quality pavement images. Complexity analysis further shows that the proposed model maintains practical inference efficiency rather than relying on too large a model size. These results show that the proposed method provides an effective solution for low-quality road crack segmentation, but it still needs to be further verified in actual detection scenarios.

Keywords: crack segmentation; semantic segmentation; encoder–decoder; feature fusion; attention mechanism; strip convolution

Check for updates

Academic Editor: Christos Bouras

Received: 3 April 2026
Revised: 16 May 2026
Accepted: 19 May 2026
Published: 22 May 2026

Copyright: © 2026 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.

1. Introduction

Road surface cracks are one of the most common types of damage to road structures. Pavement cracks can directly reduce pavement smoothness and affect the pavement friction

coefficient. According to the analysis of 2022 Iowa traffic data, when the road surface roughness deteriorates from good to poor, the accident rate rises from 134 to 466 per billion miles, an increase of 248%. When the friction coefficient decreases from advantageous to poor, the accident rate increases from 118 times to 571 times, an increase of about 384%. These quantitative results indicate that reductions in pavement smoothness and friction coefficient caused by cracks and other pavement distress can significantly increase the risk of traffic accidents [1]. Therefore, the development of an efficient and accurate method to detect pavement cracks is highly practical.

Road crack detection methods are mainly divided into image processing and deep learning methods. Kheradmandi N et al. [2] pointed out that although the image processing method is simple and computationally efficient, it achieves better segmentation results under relatively good conditions such as clear cracks and background contrast. However, under complex conditions involving rough pavement textures and low-contrast small cracks, image-processing methods often exhibit reduced robustness. To overcome these limitations, in recent years, with the rapid advancement of artificial intelligence, many researchers have begun to focus on deep learning. Among them, convolutional neural networks (CNNs) can effectively address challenges in crack detection tasks such as morphological diversity, uneven illumination, and complex background interference due to their powerful feature learning and end-to-end modeling capabilities. Many classic semantic segmentation models, such as UNet [3], which was originally applied to medical image segmentation, and DeepLabv3+ [4], which can capture multiscale context, have been proposed. After specific improvements, these models have been successfully applied to tasks such as road crack detection. Yuan H et al. [5] incorporated an efficient channel attention (ECA) mechanism into a classical UNet network. By assigning weights to individual channels, the model can identify cracks more accurately and enhance its ability to discern complex noise in crack images. Yang L et al. [6] proposed the EGA-UNet network, which introduces the GSConv, A-RepViT, and SPPF downsampling mechanisms. This not only strengthens the multiscale crack feature extraction capability but also achieves a lightweight design and real-time high-precision segmentation. In addition, feature map fusion was employed to alleviate edge blurring in the segmentation results. Jun F et al. [7] designed the ACAU-Net network, which integrates a densely dilated convolutional module and a crack attention module to obtain more crack and contextual information, thereby increasing the accuracy of crack localization. Zhang Y et al. [8] designed the CrackSeg network, which introduces the ELF module into a high-resolution network, enhancing the crack edges and texture features while filtering out shadow information. Li L et al. [9] proposed DFP-Net, which integrates a three-layer feature pyramid and cascaded feature fusion to achieve high-precision crack segmentation while alleviating the challenges of crack diversity and class imbalance. BGCrack [10] enhances crack-edge feature modeling through a high-frequency information enhancement module and global information perception, thereby improving segmentation accuracy in complex backgrounds. CrackNet [11] integrates CNN and transformer architectures, introduces strip pooling to enhance the features of slender cracks, and employs a dynamic loss function to alleviate class imbalance, leading to a significant improvement in crack segmentation accuracy. Research has shown that the selective integration of high-level and low-level features [12], along with the application of attention mechanisms [13], enables a more comprehensive feature representation and effectively reduces the rate of missed crack detection.

In summary, the existing methods of attention mechanisms, multi-scale feature extraction, and feature fusion have improved the performance of pavement crack segmentation. However, three key limitations remain insufficiently addressed when these methods are applied to low-quality crack image segmentation. First, many attention modules model

channel information and spatial information separately, making it difficult to capture the cross-dimensional dependencies necessary for the directional continuity of slender cracks. Second, the traditional multi-scale feature extraction method usually uses a square convolution kernel. While expanding the receptive field, it may introduce a large number of surrounding pavement textures and weaken the weak crack response. Thirdly, simply fusing high-level semantic features and shallow detail features, ignoring the intermediate layer features, can easily lead to the insufficient characterization of small-scale cracks in complex backgrounds.

To address these limitations, this study proposes a hypothesis: by combining cross-dimensional attention, anisotropic multi-scale feature extraction, and hierarchical feature integration, it can reduce missed crack detection and improve crack localization accuracy, especially in low-contrast and complex pavement-texture scenes. In order to verify this hypothesis, we chose DeepLabv3+ as the backbone network. First, triple-dimensional interactive attention (TDIA) is introduced at the end of the encoder to explicitly model the C × H, C × W, and H × W interactions required for directional crack localization. Subsequently, a multi-groups dilation feature fusion (MDFF) module is introduced after the attention mechanism to extract anisotropic multi-scale features of slender cracks and reduce surrounding pavement-texture interference. Furthermore, a hierarchical feature integration branch was added to the decoder, which included a dimension aware selective integration (DASI) module, to adaptively integrate shallow, intermediate, and deep features and make weak crack responses more significant. These three designs correspond to the main difficulties of low-quality crack images, namely weak localization, slender multi-scale morphology, and the loss of fine details during decoding. The effectiveness of this method was verified by comparative experiments, ablation studies, and the visual analysis of two public crack datasets.

2. Method

2.1. Overall Architecture

To address these challenges, we propose a crack segmentation model based on feature integration and triple attention. The model uses DeepLabv3+, renowned for its multi-scale context capture capability, as the backbone network and introduces three core modules on this foundation. The overall network architecture is shown in Figure 1.

The network first employs a ResNet-50 backbone, augmented with dilated convolutions, to extract feature maps rich in semantic information. The residual structure in ResNet contributes to the optimization of the network and can extract hierarchical visual representations more effectively. In addition, dilated convolution is introduced into the ResNet network to expand the receptive field while maintaining the resolution of the feature map. This design has been widely used in semantic segmentation [14], which can not only alleviate the problem of deep network gradient disappearance but also retain the characteristics of small crack structures in the feature extraction process.

Following the output of the ResNet-50 feature extraction network, a triple-dimensional interactive attention mechanism (TDIA) is introduced. This module models the interdependencies among the channel, height, and width dimensions, enhancing the capacity of the network to spatially localize crucial crack features. Subsequent to the TDIA module, a multi-groups dilation feature fusion module (MDFF) was employed. This module enriches the diversity of multi-scale crack feature representations, thereby improving the model's ability to capture complex crack characteristics. During the decoding stage, a feature integration branch was added to the original decoding path. This branch first integrates feature maps from different stages (Stage 1, Stage 2, and Stage 3) of ResNet-50 and introduces a dimension aware selective integration module (DASI) to enable the network to adaptively

select and integrate features from different layers, thereby enhancing the saliency of small cracks in complex backgrounds.

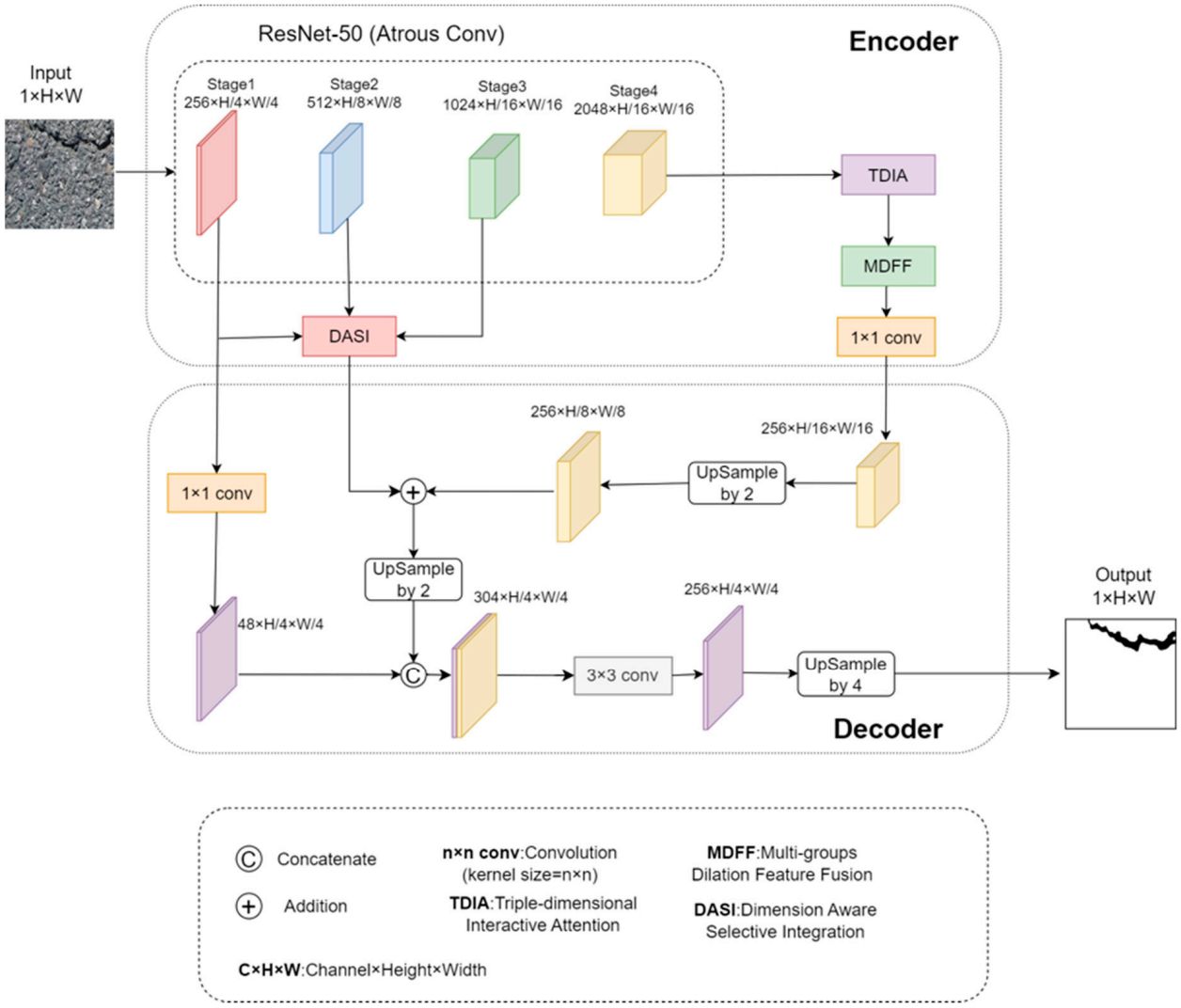


Figure 1. Crack segmentation model based on feature integration and triple attention mechanism.

Before presenting the detailed mathematical formulations of the three modules, their intuitive roles in the proposed framework are summarized in Table 1. This summary is intended to help readers understand the functional motivation of each module and its relationship to the challenges of low-quality pavement crack segmentation. Specifically, TDIA focuses on directional crack localization, MDFF focuses on anisotropic multi-scale slender crack modeling, and DASI focuses on the adaptive recovery of weak crack details through cross-layer feature integration.

Table 1. Intuitive roles of the proposed modules in the segmentation framework.

Module	Main Challenge Addressed	Intuitive Function	Expected Contribution
TDIA	Weak localization of slender and low-contrast cracks	Models cross-dimensional interactions among channel, height, and width to enhance directional crack continuity	Improves crack localization and reduces missed detections

Table 1. *Cont.*

Module	Main Challenge Addressed	Intuitive Function	Expected Contribution
MDFF	Background interference introduced by square receptive fields in multi-scale extraction	Splits the input feature map into channel groups; each group is independently processed by an improved ASPP-style multi-branch block with strip dilated convolutions	Captures multi-scale slender crack features while reducing redundant full-channel computation
DASI	Insufficient recovery of weak and small crack details during decoding	Adaptively fuses shallow, middle, and deep features under dimension-aware selection	Enhances weak crack responses and preserves fine crack boundaries

2.2. Triple-Dimensional Interactive Attention (TDIA)

Attention mechanisms are inspired by the human cognitive ability to selectively focus on important information. In deep neural networks, they model interdependencies among feature channels and generate spatial attention weights, allowing the network to enhance target-relevant features while suppressing irrelevant background interference. By integrating attention mechanisms, deep convolutional neural networks can better capture critical features and achieve improved performance in large-scale visual tasks.

Although existing attention mechanisms (such as convolutional block attention module (CBAM) [15] and coordinate attention (CA) [16]) have made significant progress in computer vision tasks, they have limitations when used in road cracks. The crack shape usually belongs to a slender linear topology and has a low contrast with the background. The previous attention mechanism focused on the independent weighting of channels and space, relying on isotropic square receptive fields. When dealing with the directional characteristics of cracks, surrounding pavement texture information can easily be introduced, thereby weakening the continuity and directional spatial localization ability of the crack features. To objectively investigate this issue, in Section 4.6 of this study, TDIA was further quantitatively compared with CBAM [15] and CA [16] under the same backbone network and training settings.

Therefore, this study proposes a triple-dimensional interactive attention mechanism subsequent to the feature extraction part of the ResNet-50 backbone network. This module aims to establish the interactive relationships among the three dimensions of the channel, height, and width. We can model the channel and spatial attention more efficiently without reducing dimensionality, thereby improving our ability to locate key targets.

The difference between TDIA and CBAM and coordinate attention is its explicit cross-dimensional interaction strategy. CBAM models channel attention and spatial attention sequentially, while coordinate attention embeds position information into channel attention through two one-dimensional direction encodings. In contrast, TDIA constructs three pairs of interactive branches, namely C × H, C × W, and H × W branches, to jointly model channel semantics, directional spatial extension, and spatial continuity. The C × H and C × W branches are designed to capture the interaction between the channel dimension C and height H or width W, respectively, whereas the last H × W branch focuses on modeling the internal relationship between the two spatial dimensions H and W. Finally, the attention weights of the interaction dimensions calculated on the three branches are multiplied. This three-way interactive design is of crucial importance for road crack segmentation. Since pavement cracks usually exhibit strong directional extensibility, such as the longitudinal and transverse distribution of long strips. The C × H and C × W branches can capture the interdependencies between channel characteristics and vertical or horizontal directions, respectively, thus retaining the morphological continuity of slender cracks in the feature

map. At the same time, the convolution kernel used in the three branches can effectively filter the interference information generated by the complex road surface so as to better maintain the continuous shape of the crack and improve the crack localization accuracy under weak contrast. The attention structure diagram of the TDIA is shown in Figure 2.

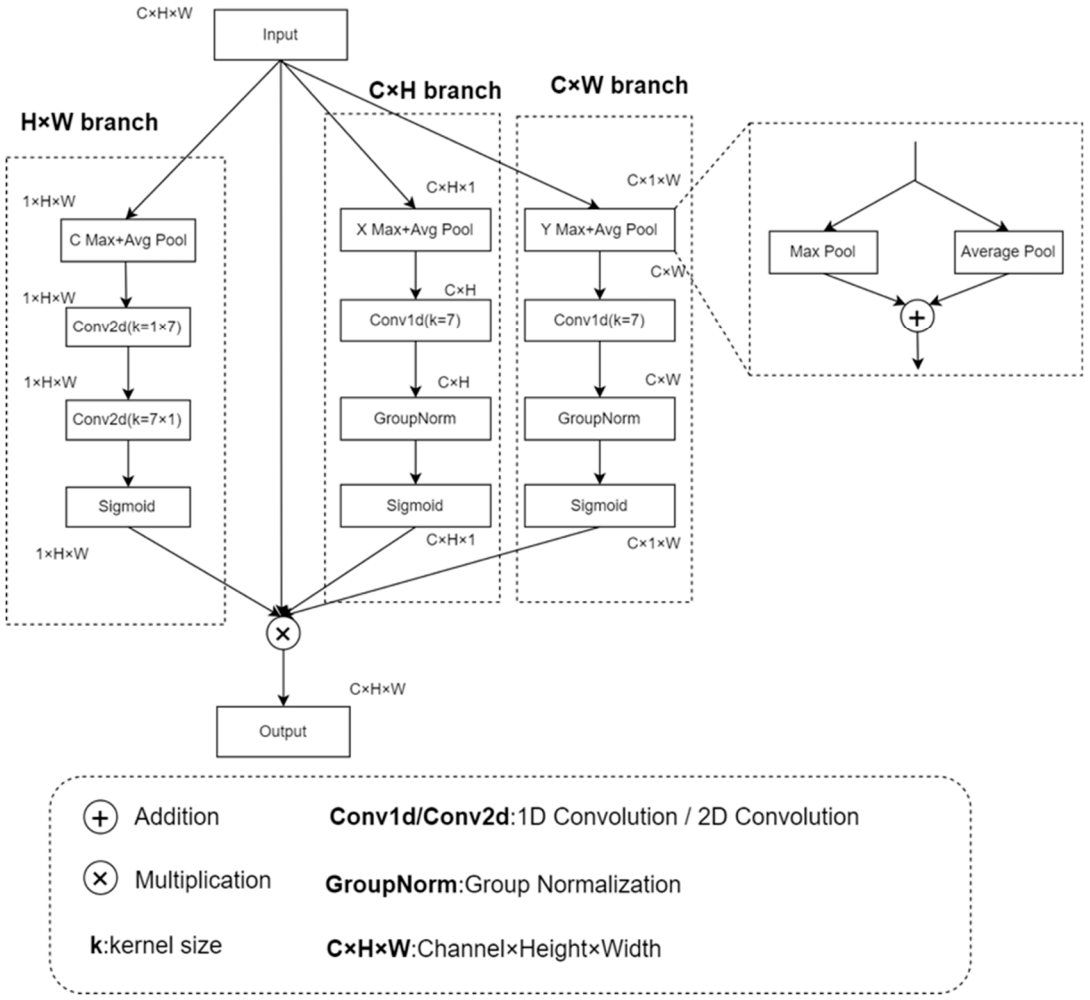


Figure 2. Structure of TDIA module. Note: The module contains three interactive branches, C × H, C × W, and H × W, which jointly model channel semantics, directional crack continuity, and spatial crack structure.

The TDIA module can be understood as a three-branch attention block designed for the elongated geometry of pavement cracks. Instead of computing channel attention and spatial attention separately, TDIA builds three pairwise interactions. The C × H branch emphasizes crack continuity along the vertical direction, the C × W branch emphasizes crack continuity along the horizontal direction, and the H × W branch enhances the spatial structure of crack regions. The outputs of these three branches are then multiplied to generate a crack-aware attention map. In this way, TDIA aims to strengthen weak and slender crack responses while suppressing irrelevant pavement textures.

For clarity, $X \in R^{C \times H \times W}$ denotes the input feature map, where C, H, and W represent the channel number, height, and width, respectively. The three attention maps generated by the C × H, C × W, and H × W branches are denoted as y^H, y^W, and y^C, respectively.

The first branch, the C × H branch, was used to establish the interaction between the channel and the height dimensions. First, strip average pooling and strip max pooling were

applied to the input feature X along the width dimension, and the results were summed to obtain a C × H × 1 feature map encoding the localization information, as in Equation (1).

$$g_H(X) = \frac{1}{H} \sum_{0 \le i \le H} X_{h,i} + \max_H(X) \quad (1)$$

In Equation (1), H represents the height, and the feature map $g_H(X) \in R^{C \times 1 \times W}$ is obtained by summing the results of average pooling ($\frac{1}{H} \sum_{0 \le i \le H} X_{h,i}$) and max pooling ($\max_H(X)$), with the height as the unit.

On the basis of the location information embedded feature map obtained from Equation (1), we generated the attention feature map via the following encoding method. Firstly, a one-dimensional convolution (Conv1D) with a convolution kernel size of 7 was used to enhance the position feature map. This design refers to the common design of representative attention modules. For example, CBAM [15] uses a 7 × 7 convolution to generate a spatial attention map in the spatial attention branch; the triplet attention module [17] also uses a convolution kernel size of 7. Therefore, this design can achieve a balance between a larger receptive field and limited computational overhead. Group normalization (GN) [18] was subsequently used to improve the generalizability of training with mini-batches. Finally, the sigmoid activation function was used to generate attention weights across the height dimension. The encoding method is shown in Equation (2).

$$y^H = \sigma(GN(Conv1D_7(g_H))) \quad (2)$$

In Equation (2), $\sigma(x)$ is the sigmoid function.

Similar to the C × H branch, the C × W branch first applies strip average pooling and strip max pooling to the input feature map $X \in R^{C \times H \times W}$, as defined in Equation (3). It then employs a similar encoding method to generate an attention feature map for the width dimension, as shown in Equation (4).

$$g_W(X) = \frac{1}{W} \sum_{0 \le j \le W} X_{j,w} + \max_W(X) \quad (3)$$

$$y^W = \sigma(GN(Conv1D_7(g_W))) \quad (4)$$

In Equation (3), W represents the width, and the feature map $g_W(X) \in R^{C \times H \times 1}$ is obtained by summing the results of average pooling ($\frac{1}{W} \sum_{0 \le j \le W} X_{j,w}$) and max pooling ($\max_W(X)$) with the width as the basic unit. In Equation (4), $\sigma(x)$ is a sigmoid function.

The final branch computes the attention feature maps between two spatial dimensions. First, the input tensor $X \in R^{C \times H \times W}$ is compressed along the channel dimension via both average pooling and maximum pooling over the channels to obtain the spatial information feature map, as shown in Equation (5). Subsequently, the spatial information feature map was processed by two strip convolutions with kernel sizes of 1 × 7 and 7 × 1. The combination of horizontal and vertical strip convolutions enables the network to capture directional features along both spatial axes, enhancing the representation of slender and narrow crack structures in the spatial dimension [19]. Simultaneously, it avoids irrelevant background interference caused by the square receptive field of traditional convolution and improves the directionality of feature extraction. Finally, the final spatial attention feature map is output through the sigmoid activation function, as shown in Equation (6).

$$g_C(X) = AvgPool(X) + MaxPool(X) \quad (5)$$

$$y^C = \sigma(Conv_{7\times1}(Conv_{1\times7}(g_C))) \quad (6)$$

In Equation (5), by considering the channels as the basic unit, the results of average pooling and max pooling are applied and then summed to obtain the feature map $g_C(X) \in R^{1\times H\times W}$. In Equation (6), $\sigma(x)$ is a sigmoid function.

Finally, we multiply the three attention feature maps, y^H, y^W, and y^C obtained from Equations (2), (4), and (6), respectively, to obtain the comprehensive attention feature map. This comprehensive weight map is then multiplied by the input feature map X to obtain the output T of the TDIA module, as shown in Equation (7).

$$T = X \times y^C \times y^H \times y^W \quad (7)$$

The final output of the TDIA module is $T \in R^{C\times H\times W}$. This module captures cross-dimensional interactions across channel, height, and width. It enhances key crack features and improves localization while preserving original semantic information.

In summary, TDIA is designed to enhance crack-related features through three complementary interactions. The C × H and C × W branches preserve the directional continuity of slender cracks along vertical and horizontal directions, while the H × W branch strengthens the spatial crack structure. Therefore, TDIA provides a more task-specific attention mechanism for low-quality crack images than conventional channel–spatial attention modules.

2.3. Multi-Groups Dilation Feature Fusion (MDFF)

After feature extraction, the traditional DeepLabv3+ model feeds the feature map into the atrous spatial pyramid pooling (ASPP) module before sending the output feature map to the decoder. The ASPP module consists of 5 branches, each consisting of a 1 × 1 convolution, three 3 × 3 dilated convolutions with dilation rates of 6, 12, and 18, and global average pooling. The three parallel dilated convolutions can extract features from different receptive fields and obtain feature representations of crack targets at different scales. Global average pooling was used to obtain global contextual information, and the scale of the feature map was adjusted by upsampling to be consistent with the other four branches.

However, since the traditional ASPP module [4] uses a fixed square receptive field, it encounters significant limitations in complex pavement environments. Cracks typically manifest as slender linear structures. Consequently, square convolutions inevitably introduce substantial background noise into the feature maps when expanding the receptive field, leading to the dilution of fine crack features. Furthermore, the ASPP module applies the same convolution operation across all channels, thereby failing to adequately model morphological diversity. Therefore, we improved the ASPP module and proposed a multi-groups dilation feature fusion module (MDFF) by combining the grouping strategy idea [20]. The structure diagram is shown in Figure 3.

To make the feature extraction process clearer, the MDFF module can be understood as a grouped improved ASPP-like structure. First, the input feature map is divided into four sub-feature maps along the channel dimension, and each sub-feature map has a size of C/4 × H × W. Then, each sub-feature map is independently processed by the same multi-branch feature extraction block. For each group, one branch uses a 1 × 1 convolution to preserve local channel information, while the other three branches use strip dilated convolutions with dilation rates of 4, 8, and 12 to capture crack features at different receptive-field scales. The outputs of the four branches within each group are fused to obtain one corresponding output sub-feature map. Finally, the four output sub-feature maps are concatenated along the channel dimension to recover the final output feature map with size C × H × W.

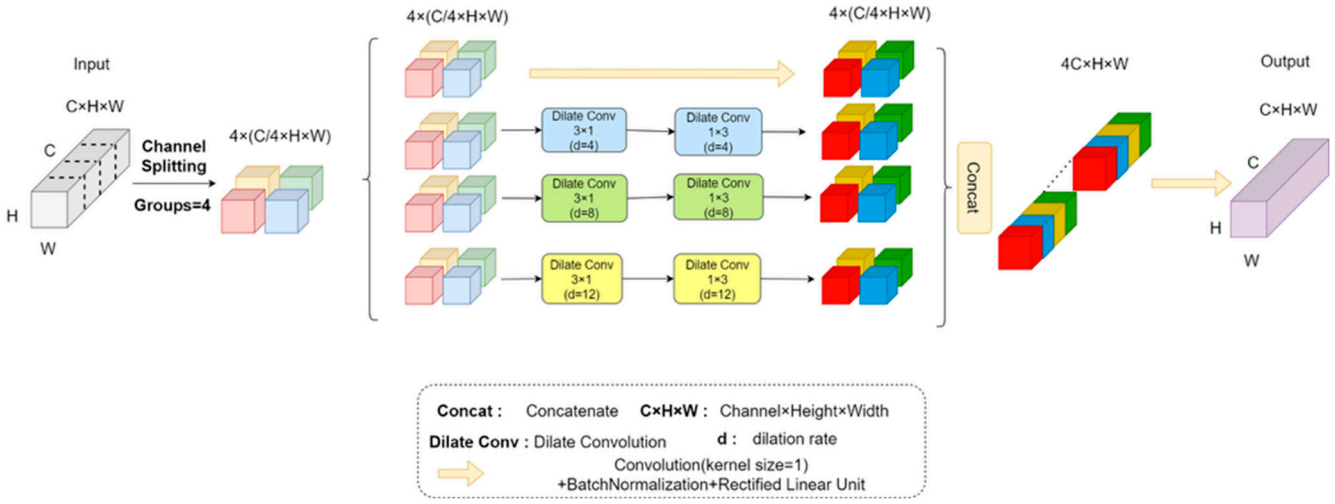


Figure 3. Structure of MDFF module. Note: The input feature map is first divided into four sub-feature maps along the channel dimension. Each sub-feature map is independently processed by an improved ASPP-style multi-branch block consisting of one 1 × 1 convolution branch and three strip dilated convolution branches with dilation rates of 4, 8, and 12. The group-wise outputs are then concatenated to generate the final output feature map.

Specifically, the output $T \in R^{C\times H\times W}$ of the TDIA attention module is first used as the input feature map of the module and is divided into N groups (N = 4) along the channel dimension to generate four sub-feature maps $x_i \in R^{C/4\times H\times W}(i = 1,2,3,4)$. The C/N ratio (N = 4) is the number of channels in each group. Each of the four channel groups is independently processed by the same four-branch improved ASPP-style block. In this block, the 1 × 1 convolution branch and the three strip convolution branches with dilation rates of 4, 8, and 12 jointly generate the corresponding group-wise output feature map. One branch adjusts the number of channels by using a 1 × 1 convolution to obtain the output. The other three branches were processed by strip convolution modules (with dilation rates d of 4, 8, and 12), and $x'_{i(D=1)} \in R^{C/4\times H\times W}(i = 1,2,3,4)$ was finally calculated. The calculation process for the strip convolution module is shown in Equation (8).

$$x'_{i(D=d)} = DConv_{1\times 3}^{D=d}(DConv_{3\times 1}^{D=d}(x_i))(d = 1,4,8,12) \quad (8)$$

In Equation (8), x_i denotes the i-th sub-feature map obtained after channel splitting. The strip dilated convolution operation is applied to each channel group under a given dilation rate d, where d takes values of 4, 8, and 12 in the strip dilated convolution branches. The resulting branch-level feature representation is denoted as $x'_i \in R^{C/4\times H\times W}$ (where i = 1, 2, 3, 4). This operation enables each channel group to extract elongated crack features at different receptive-field scales.

After the strip convolution module, the outputs of each branch $x'_{i(D=d)} \in R^{C/4\times H\times W}$ are concatenated along the channel dimension to obtain $\hat{x}_i \in R^{C\times H\times W}$, as shown in Equation (9). Multiple subfeature maps were recombined. Subsequently, the feature maps corresponding to the four heads $\hat{x}_1$, $\hat{x}_2$, $\hat{x}_3$ and $\hat{x}_4$ are concatenated. At this stage, the size of the feature map becomes $4C \times H \times W$. Then, through 1 × 1 convolution, batch normalization (BN), and nonlinear activation (ReLU), the sub-feature maps interact across groups and are recombined into the output feature map $X_O \in R^{C\times H\times W}$. The recombination operation is shown in Equation (10).

$$\hat{x}_i = Concat(x'_{i(D=d)})(d = 1,4,8,12) \quad (9)$$

$$X_O = \mathrm{Re}LU(BN(Conv_{1\times1}(Concat(\hat{x}_1, \hat{x}_2, \hat{x}_3, \hat{x}_4)))) \quad (10)$$

In Equations (9) and (10), $\hat{x}_1$, $\hat{x}_2$, $\hat{x}_3$, and $\hat{x}_4$ denote the feature maps generated from the four channel groups. *Concat* represents channel-wise concatenation, and $X_O \in R^{C\times H\times W}$ denotes the final output feature map after 1×1 convolution, batch normalization, and ReLU activation.

In summary, MDFF first divides the input feature map into several channel groups and then applies an improved ASPP-style multi-branch operation to each group independently. Within each group, strip dilated convolutions with different dilation rates are used to extract multi-scale slender crack features, while the 1×1 convolution branch preserves local channel information. The group-wise outputs are finally concatenated to form the complete output feature map. This design improves multi-scale crack representation while reducing redundant full-channel computation and limiting the interference introduced by surrounding pavement textures.

2.4. Hierarchical Feature Integration Branches

The traditional DeepLabv3+ model fuses only high-level features rich in semantic information located at the tail of the encoder and detailed features from low-level features (Stage 1) during the decoding stage. However, this two-layer feature map fusion strategy is limited. First, the deepest feature map underwent multiple downsampling operations, resulting in the loss of fine crack details. Conversely, the shallowest feature map, which is rich in detail, lacks sufficient semantic context. The crack feature information contained in the equally important middle layer features (Stages 2 and 3) was ignored. If only high-level and low-level feature maps are fused, it is difficult to capture the complex and varied features of cracks, and the faint features of small cracks are easily obscured in a complex background.

To address this issue, we introduced a feature integration branch into the backbone network. This network integrates three key feature maps from Stages 1, 2, and 3 of the ResNet-50 feature extraction part, with 256, 512, and 1024 channels, respectively. First, we used 1×1 convolutions to adjust the number of channels in the three feature maps to 128, 256, and 512. Subsequently, we incorporated the dimension aware selective integration (DASI) module, designed by Shibiao Xu et al. [21], to conduct adaptive feature fusion and produce the final output of this branch. The hierarchical feature integration branch is shown in Figure 4.

Dimension aware selective integration (DASI) module is a channel-partitioning selection mechanism. It can adaptively define the weight of the feature map of each layer by referring to the scale and shape of cracks of different sizes, and selectively fuse these feature maps, as shown in Figure 5.

Intuitively, the DASI module is used to decide how much information should be taken from shallow detail features and deep semantic features under the guidance of intermediate-layer features. For crack segmentation, shallow features contain edge and texture details, deep features contain semantic discrimination, and intermediate features provide scale-related structural information. Instead of directly concatenating these features, DASI adaptively assigns their contributions, so that small cracks can retain local details while still benefiting from high-level semantic context.

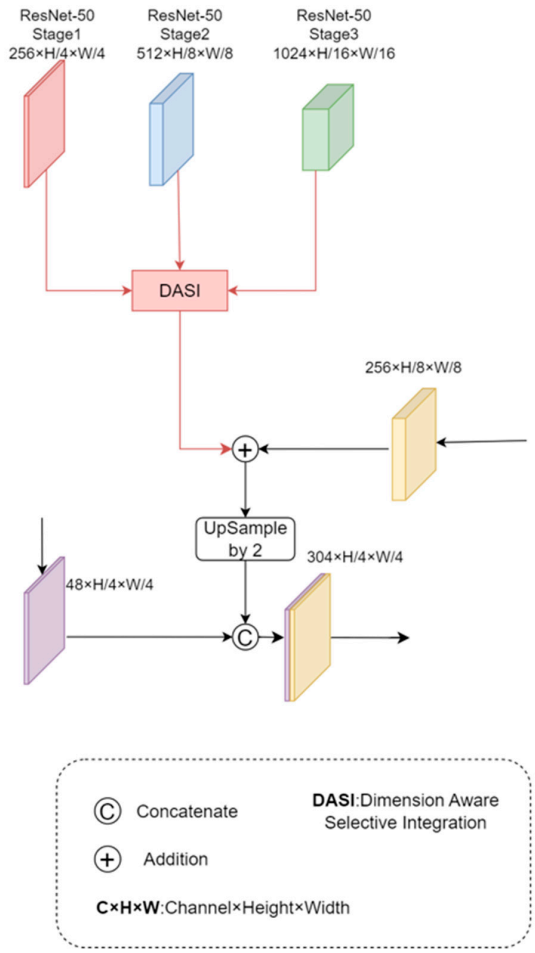


Figure 4. Structure of hierarchical feature integration branches.

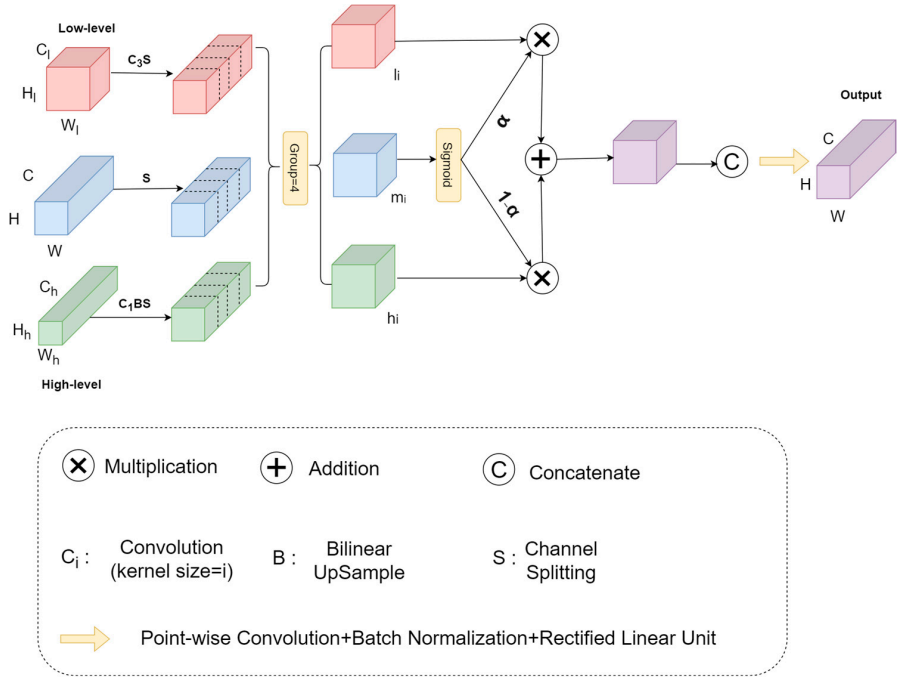


Figure 5. Structure of the DASI module. Note: The shallow, intermediate, and deep feature maps are first aligned and split along the channel dimension. The intermediate feature is used to generate adaptive weights, which guide the selective fusion of local detail information and high-level semantic information.

The dimension aware selective integration module comprises two main steps. First, convolution is applied to the high-dimensional feature map $I_h \in R^{C_h \times H_h \times W_h}$, and bilinear interpolation upsampling is performed on the low-dimensional feature map $I_l \in R^{C_l \times H_l \times W_l}$ such that their features can be aligned with the intermediate-layer feature map $I_m \in R^{C_m \times H_m \times W_m}$. Subsequently, these three feature maps were divided into four equal parts according to the channel dimension, resulting in $h_i \in R^{C/4 \times H \times W} (i = 1, 2, 3, 4)$ and $m_i \in R^{C/4 \times H \times W} (i = 1, 2, 3, 4)$, where i denotes the i-th segmentation feature. First, the weights $\alpha \in R^{C/4 \times H \times W}$ of the middle-layer feature m_i were calculated via the sigmoid activation function. Then, the weights of the high-dimensional and low-dimensional features are dynamically allocated via α and added together to obtain the feature integration result $m_i \in R^{C/4 \times H \times W}$ for each segment. The calculation process is shown in Equation (11).

$$\alpha = \sigma(m_i), m_i' = \alpha l_i + (1 - \alpha) h_i \tag{11}$$

If $\alpha > 0.5$ is true, the integrated features are more biased toward local details. If $\alpha < 0.5$ is true, the integrated features are more biased toward global contextual features.

Then, all $m_i \in R^{C/4 \times H \times W}$ are concatenated into $O_m' \in R^{C \times H \times W}$, and the calculation process is shown in Equation (12).

$$O_m' = Concat(m_i'), \hat{O}_m = \sigma(BN(I_m + Conv(O_m'))) \tag{12}$$

The output feature map $\hat{O}_m \in R^{C \times H \times W}$ integrates features from the shallow, middle, and deep layers of the encoder. It not only preserves detailed information, but also possesses strong semantic expressive power. Through the adaptive weight allocation mechanism of the DASI module, the network can dynamically adjust the contribution of each layer's feature map according to the scale and morphology of the different cracks, thereby effectively mitigating the missed detection phenomenon caused by improper feature fusion.

In summary, DASI acts as an adaptive cross-layer feature selector. It avoids the limitation of directly fusing only shallow and deep features in the original DeepLabv3+ decoder. By introducing intermediate-layer guidance, DASI helps the network preserve fine crack boundaries and enhance weak crack responses in complex pavement textures.

3. Datasets and Experimental Design

3.1. Datasets

This experiment focused on road crack detection. Therefore, we used the Crack500 and GAPS384 crack datasets as our evaluation datasets. The Crack500 dataset contains 3368 images of road surface cracks collected under different backgrounds, angles, and lighting conditions [22]. Crack500 contains pavement crack images collected under different backgrounds and angles. The shape and scale of cracks vary greatly, and most of the images have complex pavement textures. Therefore, it is suitable for evaluating the segmentation performance of the model on cracks with different morphologies and the stability of crack detection under complex pavement textures. The GAPS384 dataset contains 509 images characterized by faint crack features and poor illumination [23]. The pictures in the GAPS384 dataset are obtained from the road surface of the German federal highway captured by the vehicle camera. It contains many low-quality crack images with weak crack characteristics, low contrast, and stains. This dataset provides a more challenging scenario for testing the robustness of the model under low contrast and noisy road conditions.

However, the representativeness of these datasets still has limitations. Crack datasets are usually constructed from different collection sources and can usually only represent a limited subset of crack materials, surface types, imaging equipment, and environmental

conditions. Models trained based on public crack datasets may still have domain bias when applied to unseen road types or different imaging conditions. In this study, although these two datasets provide useful diversity in terms of crack morphology, low contrast, and complex pavement textures, they do not cover all possible real inspection scenarios, such as nighttime imaging, rainy or slippery roads, motion blur caused by high-speed vehicle acquisition, severe occlusion, different pavement materials, or cross-regional road conditions.

Data augmentation methods include random combinations of rotation, flipping, cropping, brightness adjustment, and contrast adjustment. After data augmentation, the Crack500 dataset contained 7160 images with a resolution of 512 × 512, including 5688 images in the training set, 348 images in the validation set, and 1124 images in the test set. After augmenting the GAPS384 dataset and removing images that did not contain cracks, 3916 images with a resolution of 512 × 512 pixels were obtained, including 3440 images in the training set, 320 images in the validation set, and 156 images in the test set.

3.2. Experimental Environment and Evaluation Indicators

This experiment was conducted on a server equipped with a Tesla V100S PCIE 32GB GPU (NVIDIA Corporation, Santa Clara, CA, USA). Inference and testing were performed on a separate computer with an Intel(R) Core (TM) i9-14900HX processor (Intel Corporation, Santa Clara, CA, USA) and 32GB RAM. The model was implemented via PyTorch 2.0 (PyTorch Foundation, San Francisco, CA, USA), and a neural network was trained. During training, we trained for 50 epochs, set the batch size to 4, set the learning rate to 0.0001, and used the Adam optimizer algorithm. Each model was trained three times using three random seeds: 0, 42, and 2026, to ensure the robustness of model training.

This study employed four metrics to evaluate the model performance: $mIoU$ (mean intersection over union), $Recall$, $Precision$, and $F1-score$. These metrics can be used to evaluate the accuracy of the model in segmenting cracks comprehensively. The equations are as follows:

$$mIoU = \frac{1}{k+1}\sum_{i=0}^{k}\frac{P_{ii}}{\sum_{i=0}^{k}P_{ij}+\sum_{i=0}^{k}P_{ji}-P_{ii}} \tag{13}$$

$$Recall = \frac{TP}{TP+FN} \tag{14}$$

$$Precision = \frac{TP}{TP+FP} \tag{15}$$

$$F1 = \frac{2\times Precision\times Recall}{Precision+Recall} \tag{16}$$

In Equation (13), k is the number of segmentation classes, and P_{ij} denotes the prediction of class j pixels.

In Equations (14)–(16), TP represents the number of pixels correctly predicted as cracks, TN represents the number of pixels correctly predicted as background, FP represents the number of pixels incorrectly predicted as cracks, and FN represents the number of pixels incorrectly predicted as background pixels.

3.3. Loss Function

To address the class imbalance between cracks and the background, we employed the Tversky loss [24] as the loss function. This loss function balances the weights between false

positives (FPs) and false negatives (FNs) by adjusting two hyperparameters. The formula for the loss function is as follows:

$$TverskyLoss = 1 - \frac{\sum_{i=1}^{N} p_{0i}g_{0i}}{\sum_{i=1}^{N} p_{0i}g_{0i} + \alpha\sum_{i=1}^{N} p_{0i}g_{1i} + \beta\sum_{i=1}^{N} p_{1i}g_{0i}} \quad (17)$$

where N represents the total number of pixels in the sample and where p_{0i} represents the probability that the i-th element is a positive sample of the crack.g_{0i} indicates that the i-th element is 1 for a positive sample of the crack and 0 for a negative sample of the background, and g_{1i} indicates the opposite.

For the Crack500 dataset, which has a relatively balanced class distribution, $\alpha = \beta = 0.5$ was employed. For the GAPS384 dataset, the crack pixels were significantly outnumbered by the background pixels due to class imbalance. Together with the low crack contrast, this resulted in predictions biased toward high precision at the expense of low recall. Therefore, $\alpha = 0.3, \beta = 0.7$ was employed to increase the penalty weight for false negatives, thereby improving the generalizability of the highly imbalanced dataset.

4. Experiment Results

4.1. Comparative Experiment of Different Models

To verify whether the model proposed in this study can effectively solve the problem of missed crack detection and inaccurate localization, we compared it with different models (e.g., DeepLabv3+ [4], DeepCrackAT [25], CrackScopeNet [26], CrackFormer [27], CT-CrackSeg [28], and UCTransNet [29]). Here, the performance measures reported by our method are expressed as the mean ± standard deviation. The results of the comparative experiments on the Crack500 dataset are shown in Table 2.

Table 2. Comparison of evaluation metrics for various network models on the Crack500 dataset.

Model	mIoU/%	Recall/%	Precision/%	F1-Score/%
DeepLabv3+	74.77 ± 0.08	69.8 ± 0.32	69.03 ± 0.26	69.41 ± 0.29
DeepCrackAT	75.11 ± 0.05	72.44 ± 0.31	67.75 ± 0.29	70.02 ± 0.08
CrackScopeNet	75.52 ± 0.08	70.2 ± 1.04	70.72 ± 1.02	70.46 ± 0.15
CrackFormer	75.78 ± 0.06	73.5 ± 0.61	68.64 ± 0.23	70.99 ± 0.13
CT-CrackSeg	75.96 ± 0.12	74.97 ± 1.04	68.03 ± 0.43	71.33 ± 0.17
UCTransNet	75.87 ± 0.1	73.84 ± 0.36	68.64 ± 0.27	71.14 ± 0.18
Ours	76.39 ± 0.06	75.51 ± 0.56	68.81 ± 0.44	72 ± 0.09

For the Crack500 dataset, our model benefits from the triple-dimensional interactive attention and dimension aware selective integration mechanisms, enabling it to effectively handle cracks with varying scales and morphologies, resulting in improvements in crack recall. These findings demonstrate that the model presented in this study can effectively alleviate the problem of missed crack detection. It performed the best among all network models. The mIOU reached 76.39% (±0.06%), an average improvement of 0.89% over those of the other networks, and improvements of 0.87% and 0.61% over those of CrackScopeNet and CrackFormer, respectively. The F1-score reached 72.00% (±0.09%), an average improvement of 1.44% over those of the other networks. The recall rate reached 75.51% (±0.56%). This shows that the network can ensure that the average recall rate of cracks is the highest without a significant decrease in accuracy, and it is not easy to miss detection.

In contrast, the proposed model still has some limitations in the performance of precision, and its value was 68.81% (±0.44%). This limitation is mainly due to the high semantic similarity between the complex background interference and the crack texture in the Crack500 dataset. Especially when the three-dimensional interactive attention and feature-selective integration mechanisms are used to enhance the extraction of small cracks,

the model may be sensitive to road shadows, oil stains, or repair marks, thus introducing some false detection. However, the model shows strong stability in improving crack continuity and reducing the missed detection rate. The experimental results confirm that the proposed model alleviates the detection problem of morphologically variable cracks at the expense of a small amount of accuracy and achieves a better overall performance balance while ensuring a leading recall rate.

The comparative experimental results for the GAPS384 dataset are shown in Table 3. For the GAPS384 dataset, our network achieved the best results in terms of these three metrics, with mIOU reaching 71.63% (±0.17%), which is an average improvement of 2.14% compared with the comparison networks, and an improvement of 2.78% and 1.53% compared with CrackScopeNet and CrackFormer, respectively. The F1 score reached 61.5% (±0.34%), which was 3.99% higher than other networks on average, and 5.38% and 2.82% higher than CrackScopeNet and CrackFormer, respectively. The comparative results indicate that the proposed model can effectively deal with the problem of missed crack detection even in low-contrast and low-quality crack images. The images in this dataset generally suffer from weak crack features and poor lighting, whereas some samples contain stained shadows, which increases the complexity of the segmentation task. This finding demonstrates that the triple-dimensional interactive attention mechanism introduced in this study's model can accurately focus on the key location information of cracks, even when faced with uneven lighting and strain interference. The dimension aware selective integration module improved the visibility of cracks in complex backgrounds by adaptively fusing multilayer features.

Table 3. Comparison of evaluation metrics for various network models on the GAPS384 dataset.

Model	mIoU/%	Recall/%	Precision/%	F1-Score/%
DeepLabv3+	69.15 ± 0.17	49.38 ± 0.76	66.74 ± 0.47	56.76 ± 0.46
DeepCrackAT	69.41 ± 0.14	53.76 ± 1.57	61.22 ± 1.31	57.25 ± 0.31
CrackScopeNet	68.85 ± 0.18	49.94 ± 1.81	64.06 ± 1.46	56.12 ± 0.77
CrackFormer	70.1 ± 0.14	53.63 ± 0.54	64.77 ± 0.24	58.68 ± 0.18
CT-CrackSeg	69.95 ± 0.22	53.02 ± 1.77	65.18 ± 1.01	58.47 ± 0.36
UCTransNet	69.57 ± 0.23	51.40 ± 2.25	65.81 ± 2.94	57.72 ± 0.28
Ours	71.63 ± 0.17	56.5 ± 1.62	67.47 ± 0.91	61.5 ± 0.34

Figure 6 shows the P–R plots corresponding to the test set results for the Crack500 and GAPS384 datasets. As shown in the figure, the curve of the network in this study was closer to the upper-right corner for both datasets. This demonstrates that our model maintained the highest recall and precision rates across different thresholds. This indicates that our model can maintain high detection accuracy across various thresholds.

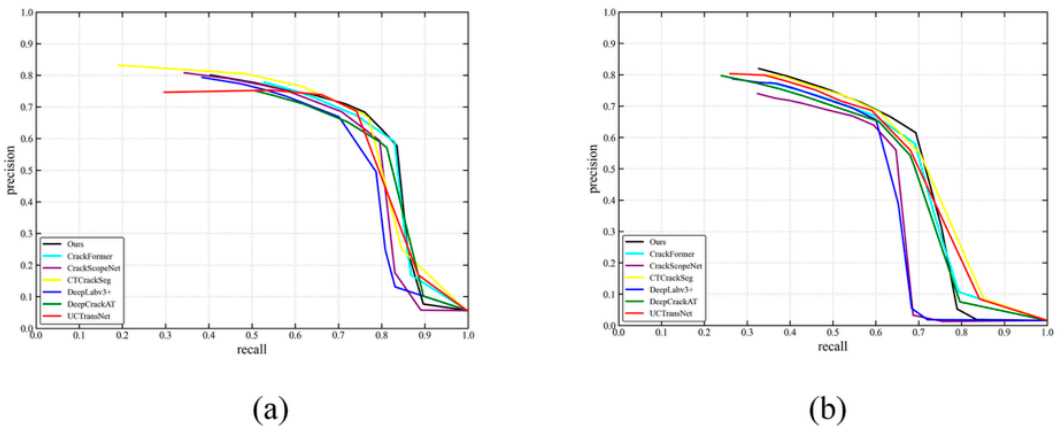


Figure 6. P–R plots of various models on two datasets: (**a**) Crack500; (**b**) GAPS384.

Although the numerical improvement over some recent methods was moderate, this result should be interpreted together with the characteristics of the datasets and the evaluation metrics. On Crack500, many comparison models already achieve strong performance, and the remaining errors are mainly concentrated in small, weak, and texture-confused crack regions. Therefore, a moderate improvement in mIoU and F1-score still indicates that the model can better handle difficult samples rather than only improve easy crack regions. More importantly, the proposed model achieved the highest recall on both datasets, suggesting that it is more effective in reducing missed detections, which is particularly important for pavement inspection tasks.

4.2. Ablation Experiment

To evaluate the impact of each proposed module, our method was based on the DeepLabv3+ backbone network, in which we integrated various performance-enhancing modules. Our method includes the multi-groups dilation feature fusion (MDFF) module, the triple-dimension interactive attention (TDIA) module, and the dimension aware selective integration (DASI) module. In addition, improved models with only one or two modules will be tested to better evaluate the contribution of each improvement to model performance. The effectiveness of each module was evaluated on the Crack500 and GAPS384 datasets, and the results are shown in Tables 4 and 5.

Table 4. Ablation experiment results from the Crack500 dataset.

MDFF	TDIA	DASI	mIoU/%	Recall/%	Precision/%	F1-Score/%
−	−	−	74.77 ± 0.08	69.8 ± 0.32	69.03 ± 0.26	69.41 ± 0.29
√	−	−	75.11 ± 0.07	71.65 ± 0.99	68.26 ± 0.62	69.92 ± 0.18
−	√	−	75.62 ± 0.11	72.39 ± 0.47	69.14 ± 0.42	70.72 ± 0.12
−	−	√	75.42 ± 0.11	71.83 ± 1.91	69.14 ± 1.49	70.46 ± 0.1
√	√	−	75.98 ± 0.05	75.28 ± 0.91	67.87 ± 0.36	71.38 ± 0.06
√	−	√	75.92 ± 0.15	73.62 ± 1.47	69.05 ± 0.58	71.26 ± 0.34
−	√	√	76.24 ± 0.07	73.95 ± 0.28	69.64 ± 0.48	71.73 ± 0.09
√	√	√	76.39 ± 0.06	75.51 ± 0.56	68.81 ± 0.44	72 ± 0.09

Table 5. Ablation experiment results from the GAPS384 dataset.

MDFF	TDIA	DASI	mIoU/%	Recall/%	Precision/%	F1-Score/%
−	−	−	69.15 ± 0.17	49.38 ± 0.76	66.74 ± 0.47	56.76 ± 0.46
√	−	−	69.53 ± 0.10	50.74 ± 0.61	65.84 ± 1.37	57.31 ± 0.10
−	√	−	70.49 ± 0.21	53.95 ± 0.97	66.61 ± 1.72	59.62 ± 0.08
−	−	√	70.51 ± 0.26	53.05 ± 1.72	68.10 ± 1.34	59.64 ± 0.17
√	√	−	71.09 ± 0.12	55.28 ± 2.25	66.87 ± 2.30	60.53 ± 0.21
√	−	√	71.05 ± 0.20	53.98 ± 0.67	68.33 ± 0.60	60.31 ± 0.13
−	√	√	71.2 ± 0.15	53.89 ± 0.32	69.35 ± 0.86	60.64 ± 0.27
√	√	√	71.63 ± 0.17	56.5 ± 1.62	67.47 ± 0.91	61.5 ± 0.34

Note: In Tables 4 and 5, √ and − indicate the inclusion and exclusion of the corresponding module, respectively.

As shown in Tables 4 and 5, each improvement enhanced the segmentation accuracy to varying degrees. With the gradual integration of different modules, the segmentation accuracy progressively increased. In the Crack500 dataset, compared with the backbone, our model improved the mIoU, recall, and F1 by 1.62%, 5.71%, and 2.59%, respectively. In the GAPS384 dataset, compared with the backbone, the mIoU, recall, and F1 improved by 2.48%, 7.12%, and 4.74%, respectively. These results indicate that all three modules contribute to improving the model performance, especially the recall rate, which is moderately improved, demonstrating that these modules can alleviate the problem of missed crack detection to

some extent. Among them, the triple-dimensional interactive attention (TDIA) module and the dimension aware selective integration (DASI) module play crucial roles in improving model performance. If the TDIA module is ablated, the encoded semantic information will be unable to provide accurate localization information. If the DASI module is ablated, the feature maps of multiple layers cannot be integrated together in an appropriate proportion, making it difficult to distinguish small cracks from the background and reducing the detection performance.

The performance gains can be explained by the complementary roles of the three modules. TDIA mainly improves crack localization by strengthening the interaction between channel semantics and spatial directions, which helps preserve the continuity of slender cracks and reduces fragmented predictions. MDFF improves multi-scale representation by using strip dilated convolutions within channel groups, making the receptive field more consistent with the elongated morphology of cracks and reducing the influence of surrounding pavement textures. DASI further improves weak-detail recovery by adaptively integrating shallow, intermediate, and deep features, so that small cracks can retain boundary details while benefiting from semantic context. Therefore, the improvement in recall is consistent with the design objective of reducing missed detections in low-quality crack images.

4.3. Model Complexity Analysis

In addition to the segmentation accuracy, the complexity of the model is also an important consideration in actual pavement crack detection. In order to evaluate the cost-effectiveness of the proposed modules, this study further compared the proposed model with the representative benchmark network in terms of the number of parameters (Params), the number of floating-point operations (FLOPs), the inference time, and the inference speed (FPS).

All models were evaluated in the same input resolution and hardware environment. The number of parameters reflects the storage requirements of the model, the number of floating-point operations represents the theoretical calculation cost, and the inference time directly reflects the deployment efficiency in the crack image processing process. The detailed complexity comparison is shown in Table 6.

Table 6. Model complexity comparison.

Model	Params/M	FLOPs/G	FPS
DeepLabv3+	40.35	69.14	56.55
DeepCrackAT	5.39	84.81	33.94
CrackScopeNet	3.16	11.07	74.66
CrackFormer	4.96	81.55	8.27
CT-CrackSeg	17.76	95.47	9.06
UCTransNet	67.22	171.71	8.70
Backbone + MDFF	25.44	55.58	68.24
Backbone + TDIA	40.37	69.18	55.63
Backbone + DASI	41.61	72.96	51.82
Ours	26.72	59.36	60.38

Table 6 shows the comparison results of the computational complexity between this model and other comparison models. It can be seen that although the model introduces three modules of TDIA, MDFF, and DASI, it does not bring excessive computational burden. The parameter quantity of the proposed model was 26.72 M, the FLOPs was 59.36 G, and the inference speed reached 60.38 FPS. Compared with the DeepLabv3 + benchmark model, the parameter number of the proposed model was reduced from 40.35 M to 26.72 M, the FLOPs were reduced from 69.14 G to 59.36 G, and the inference speed was increased from

56.55 FPS to 60.38 FPS. These results indicate that the module introduced in this study does not simply increase the model size in exchange for performance improvement, but improves the effectiveness of feature interaction and feature fusion while maintaining low computational overhead.

Furthermore, it can be seen from the complexity results of the ablation model that different modules have different effects on the computational overhead. Firstly, the parameters of Backbone + MDFF were 25.44 M, FLOPs ere 55.58 G, and FPS was 68.24, which were better than the original DeepLabv3 +. These results indicate that the MDFF module not only enhances the modeling ability of multi-scale features of slender cracks, but also reduces the number of parameters and calculations after replacing the traditional ASPP structure. The main reason is that MDFF uses a channel grouping strategy and strip dilated convolution to model different sub-features in parallel, which avoids the problem of the repeated calculation of all channels by multiple complete 3×3 dilated convolution branches in traditional ASPP. Therefore, MDFF is not an additional stacked complex structure but achieves more efficient computing allocation while improving the multi-scale feature extraction method.

In contrast, the parameters of Backbone + TDIA and Backbone + DASI were 40.37 M and 41.61 M, respectively; the FLOPs were 69.18 G and 72.96 G, respectively; and the FPS were 55.63 and 51.82, respectively. Compared with the original DeepLabv3+, TDIA had almost no significant increase in the number of parameters and calculations, indicating that it achieves low-cost attention enhancement through strip pooling, one-dimensional convolution, and cross-dimensional interaction modeling. The DASI module brings some additional computational overhead due to the need to fuse the feature maps of different stages and perform channel division, upsampling, and selective fusion. However, combined with the ablation experiment results, it can be found that DASI plays an important role in the detection of small cracks and weak feature cracks, especially in enhancing the collaborative expression between shallow detail information and deep semantic information, thereby reducing the risk of missed detection in complex backgrounds. Therefore, the limited computational overhead added by DASI is reasonable.

When the three modules work together, the parameters and FLOPs of the complete model do not increase significantly with the increase in the number of modules, but are lower than the original DeepLabv3+. This is mainly due to the efficient replacement of the traditional ASPP structure by MDFF, and its reduced computational complexity offsets the additional overhead introduced by TDIA and DASI to a certain extent. Therefore, the complete model can still obtain faster inference speed while maintaining a medium parameter scale. This also shows that TDIA, MDFF, and DASI are not simple structural superposition relationships, but form a more complementary functional division: TDIA mainly enhances the spatial localization ability of cracks, MDFF models multi-scale slender crack characteristics with lower complexity, and DASI strengthens cross-layer feature selective fusion. The three together improve the model's ability to identify low-contrast cracks and complex background cracks.

From the perspective of cost–benefit, lightweight models such as CrackScopeNet have fewer parameters and higher inference speed, but such models have certain deficiencies in crack recall ability, which may increase the risk of missed detection in actual inspection. For road maintenance tasks, the problem of missed crack detection is often more worthy of attention than a slight increase in computational overhead, because unidentified cracks may continue to expand during subsequent service and develop into more serious pavement distresses. In contrast, although models such as CrackFormer, CT-CrackSeg and UCTransNet have strong feature modeling capabilities, their FLOPs are higher and their inference speed is significantly lower. For example, the FLOPs of CrackFormer was 81.50 G, and the FPS was only 8.27; the parameters and FLOPs of UCTransNet reached

67.22 M and 171.71 G, respectively, and the FPS was only 8.70. Compared with these models, the proposed model achieved a more balanced performance between parameter quantity, computational complexity, and inference speed.

Therefore, the improved model proposed in this study has a reasonable cost–benefit ratio. Its performance improvement is not at the expense of significantly increasing the complexity of the model but improves the crack segmentation effect at a medium parameter scale and a fast inference speed and helps to reduce the missed detection rate of small cracks and weak contrast cracks. These results indicate that the additional structural design brought by the TDIA, MDFF, and DASI modules is necessary and reasonable. However, if further deployed on embedded devices with limited computing resources, model pruning, knowledge distillation, or lightweight structure design still needs to be carried out in subsequent research to further improve the actual deployment efficiency.

4.4. Hyperparameter Sensitivity Analysis

In order to further verify the rationality of the training configuration, this study analyzed the sensitivity of two key training hyperparameters, namely the learning rate and batch size. The learning rate directly affects the convergence speed and optimization stability of the network, and the batch size affects the gradient estimation, memory usage, and training stability. In the crack segmentation task, inappropriate hyperparameter settings may lead to unstable model convergence or make it difficult for the model to fully learn weak crack features. Therefore, this study conducted comparative experiments on different learning rates and batch sizes under the condition of maintaining the network structure, data division, input size, optimizer, and training rounds. The experimental results are shown in Table 7.

Table 7. Hyperparameter sensitivity analysis table.

Hyper Parameter		Crack500		GAPS384	
Learning Rate	**Batch Size**	**mIoU/%**	**F1-Score/%**	**mIoU/%**	**F1-Score/%**
1×10^{-3}	4	75.57	70.69	70.21	58.82
1×10^{-4}	**4**	**76.40**	**72.01**	**71.80**	**61.84**
1×10^{-5}	4	75.3	70.26	69.79	57.87
1×10^{-4}	2	75.13	69.98	69.7	57.76
1×10^{-4}	**4**	**76.40**	**72.01**	**71.80**	**61.84**
1×10^{-4}	8	76.15	71.66	70.83	59.94

Note: Bold values indicate the best performance among the compared settings.

Table 7 shows the sensitivity analysis results of the two key training hyperparameters of learning rate and batch size. It can be seen that when the batch size was fixed at 4 and the learning rate was 1×10^{-4}, the best results were obtained on both datasets. On the Crack500 dataset, the mIoU and F1 scores of this setting were 76.40% and 72.01%, respectively, which were better than the learning rates of 1×10^{-3} and 1×10^{-5}. The same trend was also shown on the GAPS384 dataset. Under the configuration of learning rate 1×10^{-4}, the mIoU and F1 scores were 71.80% and 61.84%, respectively, which were the highest. These results indicate that the larger learning rate may lead to the instability of the optimization process, while the too small learning rate may make the model fall into the local optimal solution, and it is difficult to fully learn the weak crack characteristics within a fixed training round.

When the learning rate was fixed at 1×10^{-4}, the batch size also affects the model performance. Compared with batch sizes of 2 and 8, batch size of 4 achieved the best results on both datasets. Especially on the GAPS384 dataset, the performance difference was more obvious, indicating that the low-contrast crack image is more sensitive to the

training configuration. The smaller batch size may lead to unstable gradient estimation, while the larger batch size may weaken the model's ability to learn subtle crack changes. Therefore, the final experiment used a learning rate of 1×10^{-4} and a batch size of 4, which achieved a good balance between optimization stability and segmentation accuracy.

4.5. MDFF Module Internal Parameters Comparison Experiment

In order to further verify the rationality of the internal parameter setting of the MDFF module, this study conducted a comparative experiment on its key parameters. The MDFF module mainly involves two design factors: the number of channel groupings and the dilation rate combination of strip dilated convolution. The number of packet channels will affect the diversity of feature splitting and the information interaction between different groups. The dilation rate determines the range of the receptive field when the model captures different scale crack structures. An inappropriate grouping number or dilation rate combination may weaken the ability of multi-scale feature expression and may also introduce too much background interference. Therefore, this study made an experimental comparison of different parameter combinations under the condition of maintaining the same network structure, data division, training strategy, and input size. The experimental results are shown in Table 8.

Table 8. Comparison of the MDFF module internal parameters.

Hyper Parameter		Crack500		GAPS384	
Dilated Rate (d)	**Groups (N)**	**mIoU/%**	**F1-Score/%**	**mIoU/%**	**F1-Score/%**
{2,4,6}	4	75.66	70.90	69.82	56.05
{4,8,12}	**4**	**76.40**	**72.01**	**71.80**	**61.84**
{6,12,18}	4	76.03	71.31	70.78	59.26
{4,8,12}	2	76.07	71.45	70.74	59.22
{4,8,12}	**4**	**76.40**	**72.01**	**71.80**	**61.84**
{4,8,12}	8	75.92	71.11	70.56	58.89

Note: Bold values indicate the best performance among the compared settings.

Table 8 shows that the internal parameters of the MDFF module had a significant effect on the segmentation results. When the number of groups was fixed at N = 4, the dilation rate combination {4,8,12} achieved the best results on both datasets. On Crack500, the combination achieved 76.40% mIoU and 72.01% F1-score; on GAPS384, a 71.80% mIoU and 61.84% F1-score were obtained. Compared with the smaller dilation rate combination {2,4,6}, {4,8,12} can provide a larger receptive field range, which is beneficial to capture the crack characteristics of different scales. Compared with the larger dilation rate combination {6,12,18}, it can avoid introducing too much background interference, which is particularly important for low-contrast cracks in GAPS384.

When the dilation rate was fixed at {4,8,12}, the number of groups also affected the performance of the model. N = 4 was better than N = 2 and N = 8 on both datasets. A small number of groups may limit the diversity of multi-branch feature extraction, while an excessive number of groups will reduce the number of channels in each group and weaken the semantic expression ability of crack features. Therefore, the dilation rate combination {4,8,12} and the number of groups N = 4 achieved a better balance between multi-scale receptive field modeling, feature diversity, and background noise suppression. This group of parameters was finally used as the configuration of the MDFF module.

4.6. Quantitative Comparison Experiment and Analysis of Attention

To further verify the effectiveness of the proposed TDIA module, we compared it with two representative attention mechanisms, namely CBAM [15] and CA [16], under

the same DeepLabv3+ backbone, training strategy, and dataset settings. This experiment aims to evaluate whether the cross-dimensional interaction strategy in TDIA can provide more effective crack feature localization than conventional channel–spatial attention or coordinate attention mechanisms. The quantitative comparison results on the Crack500 and GAPS384 datasets are shown in Table 9.

Table 9. Comparison of evaluation metrics for various attention.

Model	Crack500		GAPS384	
	mIoU/%	F1-Score/%	mIoU/%	F1-Score/%
DeepLabv3+	74.79	69.42	69.11	56.54
+CBAM	75.36	70.36	69.57	57.55
+CA	75.42	70.42	69.9	58.12
+TDIA	75.66	70.81	70.70	59.70

As shown in Table 9, all attention mechanisms improved the segmentation performance compared with the original DeepLabv3+ backbone, indicating that attention-based feature enhancement is beneficial for crack segmentation. On the Crack500 dataset, the baseline DeepLabv3+ achieved an mIoU of 74.79% and an F1-score of 69.42%. After introducing CBAM and CA, the mIoU increased to 75.36% and 75.42%, respectively, while the F1-score increased to 70.36% and 70.42%. In comparison, the proposed TDIA module achieved the best performance, with an mIoU of 75.66% and an F1-score of 70.81%. Compared with the baseline, TDIA improved the mIoU and F1-score by 0.87% and 1.39%, respectively. Compared with CBAM and CA, TDIA also achieved a higher mIoU and F1-score, demonstrating its stronger ability to enhance crack-related features.

The advantage of TDIA was more evident on the GAPS384 dataset, which contained more low-contrast and low-quality crack images. The baseline DeepLabv3+ obtained an mIoU of 69.11% and an F1-score of 56.54%. CBAM improved these values to 69.57% and 57.55%, while CA further improved them to 69.90% and 58.12%. In contrast, TDIA achieved an mIoU of 70.70% and an F1-score of 59.70%, outperforming the baseline by 1.59% and 3.16%, respectively. Compared with CBAM, TDIA improved the mIoU and F1-score by 1.13% and 2.15%; compared with CA, the improvements were 0.80% and 1.58%, respectively.

These results indicate that the proposed TDIA module provides more effective feature enhancement than conventional attention mechanisms. CBAM and CA mainly emphasize channel–spatial weighting or coordinate-aware feature encoding, whereas TDIA explicitly models the interactions among channel, height, and width dimensions. This cross-dimensional interaction helps preserve the directional continuity of slender cracks and strengthens the spatial localization of weak crack features. Therefore, TDIA is particularly effective for low-quality crack images with weak contrast, complex pavement texture, and discontinuous crack structures.

In order to verify the ability of triple-dimensional interactive attention to extract and integrate key features of cracks in the encoder, in this study, the TDIA was compared with the backbone network, CBAM [15] spatial channel attention, and CA [16] in the heat map. The visualization of the heat map is shown in Figure 7.

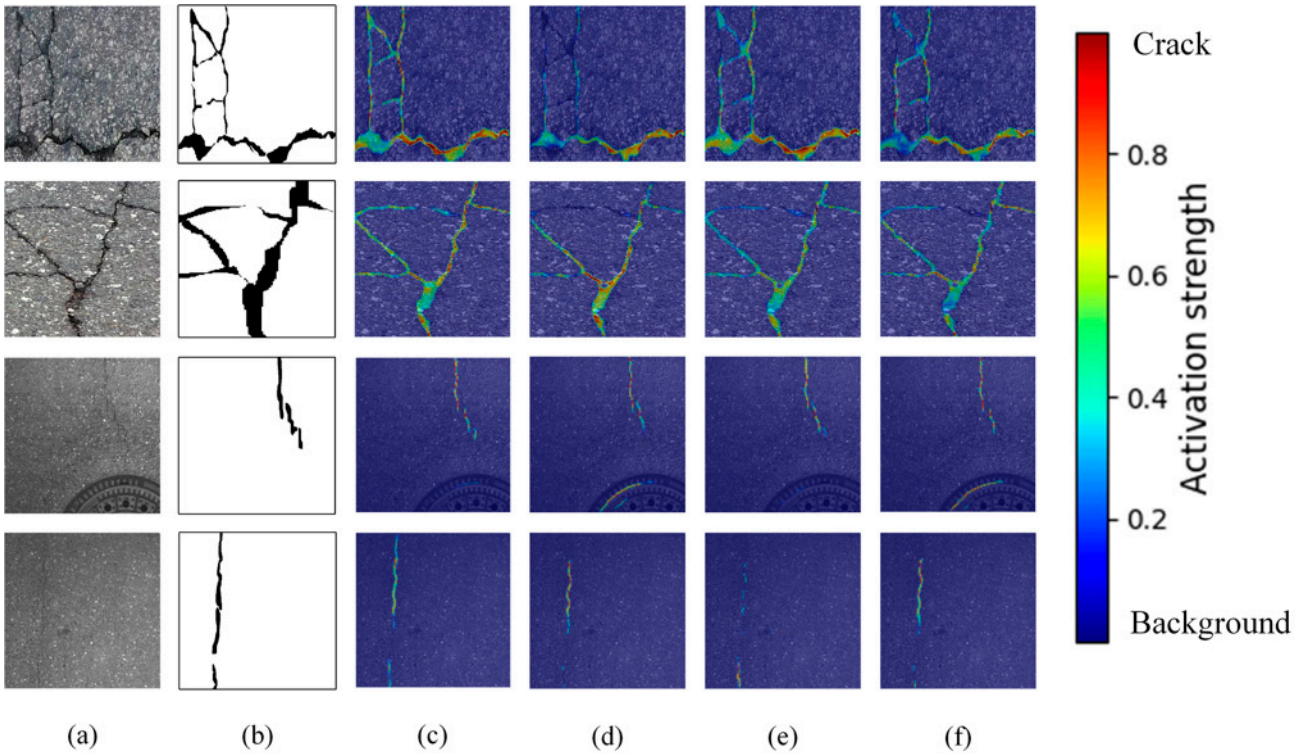


Figure 7. The heat map comparison diagram of each attention mechanism: (**a**) image; (**b**) label; (**c**) ours; (**d**) backbone; (**e**) CBAM; (**f**) CA.

It can be seen from Figure 7 that the triple-dimensional interactive attention mechanism designed in this study has obvious advantages in aggregating key features of cracks and suppressing background noise. Compared to the original backbone network, although the CBAM and CA attention mechanisms exhibited a certain level of perception for key crack features, the orange-red part of the heat map was concentrated toward the crack. However, in the face of more complex pavement backgrounds or low-contrast cracks, the square receptive field is prone to incorporating noise pixels around the cracks, which will not only introduce a small amount of background noise interference, but also be prone to the phenomenon of low-contrast crack feature dispersion or even loss.

In contrast, the TDIA attention of the proposed model performed better in the localization of cracks and the key feature extraction of low-contrast cracks. It can be seen from the figure that the high confidence region generated by TDIA attention can be accurately fitted in the crack area, and is almost not disturbed by background noise. This is mainly due to the unique design of TDIA attention. Through the parallel calculation of the three-dimensional relationship of height, width, and channel, and the combination of strip pooling and strip convolution operation, it effectively breaks through the limitations of the traditional attention square receptive field and can accurately collect the directional characteristics of cracks, so as to achieve high-precision crack feature locking under low contrast conditions.

5. Discussion

5.1. Qualitative Performance Analysis and Failure Cases

The visualization results for the Crack500 dataset are shown in Figure 8. Our model can reduce missed crack detections while maintaining a low rate of false positives. Compared with the segmentation maps of DeepLabv3+, the crack segmentation results of our model were more accurate. Especially for the first, third, and fourth lines of crack images, even in the case of complex pavement texture, the proposed model did not lose the important

features of cracks, and could better retain the main structure and edge direction of cracks. These results indicate that the extraction of multi-scale feature information can improve the segmentation accuracy of cracks of different sizes. Combined with the improved feature integration method and the synergy of the three-dimensional interactive attention mechanism, the network can correctly identify important crack feature information and pavement texture interference.

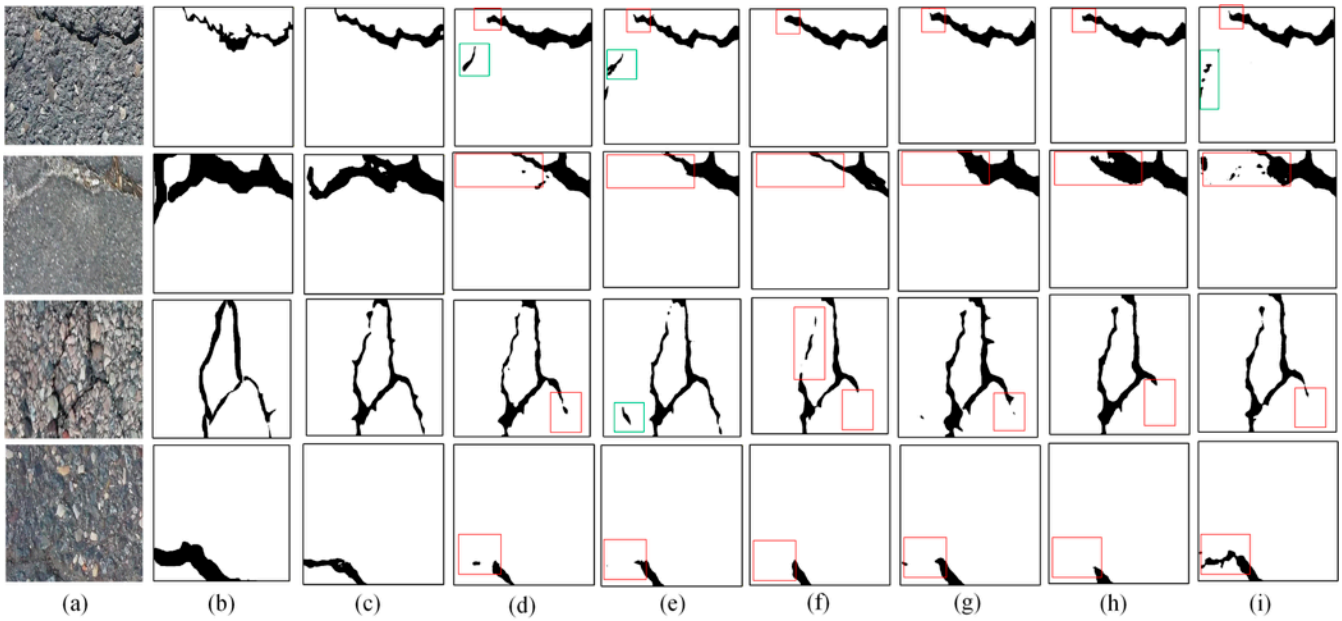


Figure 8. Segmentation results of various models on representative samples from the Crack500 dataset: (**a**) image; (**b**) label; (**c**) ours; (**d**) DeepLabv3+; (**e**) DeepCrackAT; (**f**) CrackScopeNet; (**g**) CrackFormer; (**h**) CT-CrackSeg (**i**) UCTransNet. Note: In the figure, red boxes mark missed crack detections, and green boxes mark false crack detections.

However, Figure 8 also reflects that the model still has some limitations. In the second row of samples, some cracks located at the edge of the image still appeared to be missing detection. This may be due to the lack of complete context information in the edge region, which makes it difficult for the model to fully judge whether the region belongs to the continuous crack structure. At the same time, the image cropping and size normalization process may also weaken the local characteristics of the cracks at the boundary, making it more likely to be misjudged as the background. In the fourth row sample, when the crack had a high similarity with the pavement texture and the crack contrast was low, although the proposed model could identify the approximate location of the crack, it failed to fully recover the width and detail boundary of the crack. These results indicate that the model's ability to locate low-contrast cracks had been improved, but in the case of blurred crack boundaries, strong texture noise, or large changes in crack width, there may still be problems of boundary shrinkage or incomplete local segmentation.

The visualization of the GAPS384 dataset is shown in Figure 9. It can be seen that the proposed model could still be accurately detected in the low-contrast crack image, and the crack detection accuracy was better than other networks. This suggests that allowing the model to adaptively and selectively integrate feature maps from different stages improves the saliency of cracks in complex backgrounds. As shown in the first row of images in Figure 9, compared with the other models, our model clearly delineated the contours of the cracks and moderately reduced the discontinuities in the segmentation results. This finding demonstrates that the synergistic effect of the triple-dimensional interactive attention

and multi-scale modules enhances the model's ability to accurately locate the key spatial information of cracks under harsh conditions.

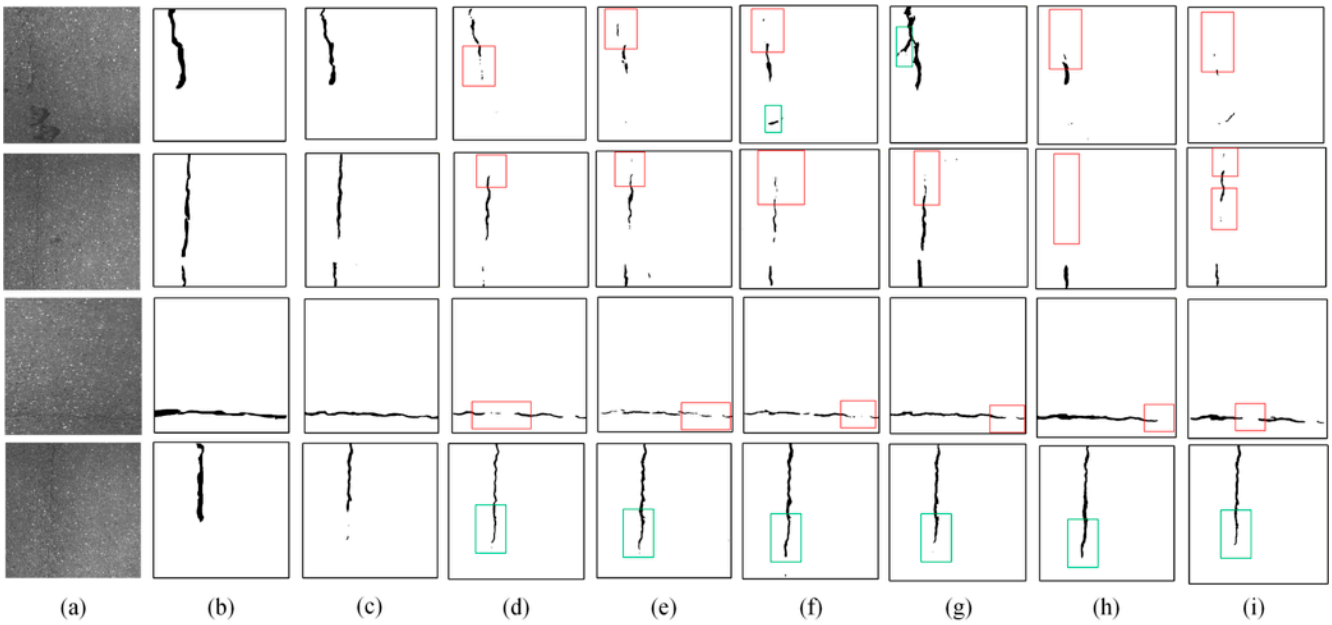


Figure 9. Segmentation results of various models on representative samples from the GAPS384 dataset: (**a**) image; (**b**) label; (**c**) ours; (**d**) DeepLabv3+; (**e**) DeepCrackAT; (**f**) CrackScopeNet; (**g**) CrackFormer; (**h**) CT-CrackSeg (**i**) UCTransNet. Note: In the figure, red boxes mark missed crack detections, and green boxes mark false crack detections.

However, in the third-row sample, the improvement of the proposed model was relatively limited compared with DeepLabv3+ and Transformer-based models such as CrackFormer, CT-CrackSeg, and UCTransNet. This can be partly attributed to the strong contextual modeling capabilities of these comparison methods. Specifically, DeepLabv3+ contains an ASPP module, which extracts multi-scale contextual information through atrous convolutions with different dilation rates. Meanwhile, Transformer-based models are effective in modeling long-range dependencies and can capture the overall orientation and global structural information of cracks over a large spatial range. Therefore, these methods already have strong segmentation capability for cracks with relatively clear morphology, large scale, or obvious continuity.

In addition, the crack in the third-row sample exhibited a long and continuous structure, and its main contour was relatively easy to identify. In this case, the performance differences among models were mainly reflected in the edge details and local continuity, rather than in the ability to detect the main crack body. Therefore, the advantages of the proposed model were more pronounced in challenging samples with low contrast, small cracks, local discontinuities, and strong interference from complex backgrounds. For large-scale cracks with clear continuity and well-defined global structures, ASPP-based and Transformer-based models can already provide strong contextual modeling capabilities, resulting in relatively small qualitative differences among different methods. This also indicates that the proposed method is not intended to replace the strengths of existing models in all scenarios; rather, it focuses on reducing missed detections and discontinuous segmentation in low-quality crack images.

Compared with ASPP-based and Transformer-based models, the advantage of the proposed method was more evident in samples where the crack response was weak, locally discontinuous, or strongly confused with pavement texture. In such cases, global context alone may identify the main crack direction, but the proposed combination of directional

attention, strip-based multi-scale extraction, and adaptive cross-layer fusion helps preserve local crack continuity and weak boundary details.

5.2. Generalization Analysis and Dataset Bias in Road Crack Segmentation

Although the model proposed in this study achieved competitive results on both Crack500 and GAPS384 datasets, its generalization ability still needs to be carefully discussed. This study mainly focused on the task of road crack segmentation, especially on low-quality road crack images with complex pavement texture, blurred crack boundary, low contrast, and background interference. Many images in the Crack500 dataset contain complex pavement textures, diverse crack patterns, and weak crack edges. These factors increase the difficulty of crack segmentation. In contrast, GAPS384 places more emphasis on crack segmentation under low-quality imaging conditions, such as weak crack features, low contrast, stains, and background noise. Therefore, the proposed model maintains relatively stable performance on both datasets, indicating that it has certain adaptability to different road crack image conditions.

The performance consistency on the two datasets is related to the way the proposed components adapt related designs to low-quality crack segmentation. Compared with conventional channel–spatial attention modules, TDIA explicitly models pairwise interactions among the channel, height, and width dimensions, which helps preserve directional continuity in slender cracks. Compared with the conventional ASPP structure, MDFF uses grouped strip dilated convolutions to extract elongated crack features at multiple receptive-field scales while reducing the background responses introduced by square receptive fields. Compared with decoder strategies that mainly combine shallow and deep features, the hierarchical integration branch with DASI introduces intermediate-layer guidance and adaptive selection, which helps retain local crack details while maintaining semantic discrimination. These designs target common difficulties in low-quality road crack images, including weak crack responses, complex pavement textures, and discontinuous crack structures.

However, the current results are more suitable to be understood as evidence of the performance consistency of the model on two road crack datasets, rather than proof of the generalization ability of the complete cross-dataset in a strict sense. In this study, the models were trained and tested on Crack500 and GAPS384, respectively. Therefore, these results show that the network structure proposed in this study is effective under two different road crack datasets, but it cannot fully prove that the model trained on one dataset can be directly migrated to another unseen dataset without retraining. In actual road inspection, the data distribution may change significantly due to differences in pavement materials, camera parameters, shooting height, vehicle speed, lighting conditions, weather conditions, and regional road characteristics. These factors may introduce domain offset problems that are not fully covered by these two public datasets.

In addition, the application scope of the proposed model also needs to be clearly defined. The proposed model is designed for road pavement crack segmentation tasks, especially for low-quality road crack images. Its effectiveness on other types of crack datasets, such as bridge cracks, concrete surface cracks, tunnel lining cracks, wall cracks, or industrial material cracks, has not been fully verified. These crack images may have different surface materials, shooting distances, crack widths, background textures, and labeling standards. Therefore, although the proposed model performed well on the road crack dataset, there may still be some limitations when it is directly applied to other crack segmentation scenarios.

Future work should further improve the generalization of the crack segmentation model to adapt to more diverse external crack datasets. In addition, data enhancement for severe weather and complex lighting conditions and semi-supervised learning using

unlabeled real road images also help to reduce dataset bias and improve the robustness of crack segmentation models in practical road inspection applications.

5.3. Practical Deployment Considerations

The proposed model is designed for road crack segmentation, especially for low-quality crack images with complex pavement textures, weak crack boundaries, and low contrast. From the perspective of practical deployment, this method has potential application value in vehicle-mounted road inspection systems equipped with high-definition cameras. In such systems, road surface images can be collected continuously by inspection vehicles and then processed by the segmentation model to locate crack regions automatically. The predicted crack masks can provide basic information for subsequent pavement condition assessment, crack length and area estimation, and maintenance decision-making.

The computational complexity results also indicate that the proposed model has a certain degree of deployment feasibility. Although TDIA, MDFF, and DASI are introduced into the network, the complete model does not lead to excessive computational burden. The model achieved 26.72 M parameters, 59.36 G FLOPs, and an inference speed of 60.38 FPS in the experimental environment. Compared with Transformer-based models such as CrackFormer, CT-CrackSeg, and UCTransNet, the proposed model achieved a higher inference speed while maintaining competitive segmentation performance.

However, practical deployment involves more than model accuracy and inference speed. In real road inspection scenarios, the segmentation results need to be further combined with camera calibration, image stitching, localization information, and scale conversion. For example, pixel-level crack masks should be converted into physical measurements such as crack length, width, and area before they can be directly used for pavement condition evaluation. When vehicle-mounted cameras are used, GPS or inertial measurement information is also required to associate detected cracks with specific road locations. Therefore, the proposed model should be regarded as a core perception component of an automatic road inspection system, rather than a complete road maintenance solution by itself.

In addition, the current experimental environment is different from real deployment conditions. The FPS reported in this study was measured under specific hardware and image resolution settings, and the actual inference speed may vary when the model is deployed on embedded devices, edge computing platforms, or vehicle-mounted processors. Moreover, real road images may be affected by motion blur, strong shadows, nighttime illumination, rainy or wet surfaces, road markings, repaired pavement areas, and occlusions caused by vehicles or debris. These factors may reduce the reliability of crack segmentation and increase the risk of false positives or missed detections. Therefore, further validation using real vehicle-mounted inspection data is still necessary.

For future deployment, model compression and system-level optimization should be further investigated. Techniques such as pruning, quantization, knowledge distillation, and lightweight backbone replacement may help reduce computational costs and improve inference efficiency on resource-constrained devices. In addition, integrating temporal information from continuous road video frames may improve segmentation stability and reduce frame-level noise. Future work should also combine crack segmentation with damage quantification, geographic localization, and pavement maintenance decision systems to promote the practical application of deep learning-based crack detection in intelligent transportation infrastructure.

6. Conclusions and Future Work

In this study, a crack segmentation model for low-quality pavement crack images was proposed. In the framework based on DeepLabv3+, feature interaction, anisotropic multi-scale feature extraction, and hierarchical feature fusion are integrated. The model mainly includes three components: a three-dimensional interactive attention module (TDIA), a multi-group dilation feature fusion module (MDFF), and a dimension-aware selective fusion module (DASI). The experimental results on the Crack500 and GAPS384 datasets show that the proposed method can improve the overall segmentation performance, especially in recall and F1-score, indicating that the model helps to reduce the missed detection of weak cracks, small cracks, and discontinuous cracks. The ablation experiment, attention comparison experiment, and model complexity analysis further show that the method achieves a performance improvement at a reasonable computational cost, rather than relying on excessive model size.

From the perspective of architecture design, the proposed framework can be understood as a task-oriented response to the main difficulties in low-quality crack segmentation. TDIA explicitly models the interaction between the three dimensions of channel, height, and width, which helps to maintain the directional continuity and spatial positioning ability of slender cracks. MDFF uses grouped strip dilated convolution to replace the traditional full-channel square dilated convolution to reduce the introduction of surrounding road texture noise while achieving anisotropic multi-scale feature extraction. DASI further enhances the hierarchical feature selection ability. By adaptively fusing shallow, middle, and deep features, the model maintains strong semantic discrimination ability while retaining local fracture details. Therefore, these three modules are not simply stacked but have complementary effects, corresponding to three aspects: direction positioning, multi-scale slender structure modeling, and cross-layer feature selection.

Nevertheless, there are still some limitations in this study. The current verification was mainly based on two public road crack datasets, which have not been fully verified in a wide range of real inspection scenarios. Future work will further validate the proposed methods in more diverse and challenging scenarios, including cross-domain road images, different road materials, nighttime lighting, rainy or slippery roads, motion blur, shadows, and occlusion. In addition, deployment verification based on a UAV or vehicle platform should be carried out to evaluate the real-time performance and robustness of the model in the actual inspection system. Future research can further combine crack segmentation with physical damage quantification, geolocation, and pavement maintenance decision-making systems so that pixel-level crack masks can be transformed into actual decision-making information that can be used for intelligent transportation infrastructure management.

Author Contributions: Conceptualization, Y.X. and Y.W.; methodology, Y.X. and Y.W.; software, Y.W.; validation, Y.W.; formal analysis, Y.W.; investigation, Y.W.; resources, Y.X.; data curation, Y.W.; writing—original draft preparation, Y.W.; writing—review and editing, Y.X. and Y.W.; visualization, Y.W.; supervision, Y.X.; project administration, Y.X.; funding acquisition, Y.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant number 62076123.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Datasets used in this study were made publicly available for comparison purposes and can be found at the following link: https://github.com/s47y8/road-crack-datasets (accessed on 3 April 2026).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Lebaku, P.K.R.; Gao, L.; Sun, J.; Wang, X.; Kang, X. Assessing the Influence of Pavement Performance on Road Safety Through Crash Frequency and Severity Analysis. *Int. J. Pavement Res. Technol.* **2025**, 1–22. [CrossRef]
2. 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]
3. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2015), Munich, Germany, 5–9 October 2015*; Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F., Eds.; Springer: Cham, Switzerland, 2015; pp. 234–241. [CrossRef]
4. Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In *Computer Vision—ECCV 2018*; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2018; Volume 11211, pp. 833–851. [CrossRef]
5. Yuan, H.; Jin, T.; Ye, X. Modification and Evaluation of Attention-Based Deep Neural Network for Structural Crack Detection. *Sensors* **2023**, *23*, 6295. [CrossRef] [PubMed]
6. Yang, L.; Deng, J.; Duan, H.; Yang, C. An Efficient Semantic Segmentation Method for Road Crack Based on EGA-UNet. *Sci. Rep.* **2025**, *15*, 33818. [CrossRef] [PubMed]
7. Jun, F.; Jiakuan, L.; Yichen, S.; Ying, Z.; Chenyang, Z. ACAU-Net: Atrous Convolution and Attention U-Net Model for Pavement Crack Segmentation. In *Proceedings of the 2022 International Conference on Computer Engineering and Artificial Intelligence (ICCEAI)*; IEEE: Shijiazhuang, China, 2022; pp. 561–565. [CrossRef]
8. Zhang, Y.; Liu, C. Generative Adversarial Network Based on Domain Adaptation for Crack Segmentation in Shadow Environments. *Comput.-Aided Civ. Infrastruct. Eng.* **2025**, *40*, 3997–4013. [CrossRef]
9. Li, L.; Liu, R.; Ali, R.; Chen, B.; Lin, H.; Li, Y.; Zhang, H. DFP-Net: A Crack Segmentation Method Based on a Feature Pyramid Network. *Appl. Sci.* **2024**, *14*, 651. [CrossRef]
10. He, Z.; Chen, W.; Zhang, J.; Wang, Y.-H. Crack Segmentation on Steel Structures Using Boundary Guidance Model. *Autom. Constr.* **2024**, *162*, 105354. [CrossRef]
11. Fan, Y.; Hu, Z.; Li, Q.; Sun, Y.; Chen, J.; Zhou, Q. CrackNet: A Hybrid Model for Crack Segmentation with Dynamic Loss Function. *Sensors* **2024**, *24*, 7134. [CrossRef] [PubMed]
12. Yu, Z.; Yu, L.; Zheng, W.; Wang, S. EIU-Net: Enhanced Feature Extraction and Improved Skip Connections in U-Net for Skin Lesion Segmentation. *Comput. Biol. Med.* **2023**, *162*, 107081. [CrossRef] [PubMed]
13. Niu, Y.; Fan, S.; Cheng, X.; Yao, X.; Wang, Z.; Zhou, J. Road Crack Detection by Combining Dynamic Snake Convolution and Attention Mechanism. *Appl. Sci.* **2024**, *14*, 8100. [CrossRef]
14. Chen, L.-C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking Atrous Convolution for Semantic Image Segmentation. *arXiv* **2017**. [CrossRef]
15. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In *Computer Vision–ECCV 2018*; Lecture Notes in Computer Science; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Springer International Publishing: Cham, Switzerland, 2018; Volume 11211, pp. 3–19. [CrossRef]
16. Hou, Q.; Zhou, D.; Feng, J. Coordinate Attention for Efficient Mobile Network Design. In *Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Nashville, TN, USA, 2021; pp. 13708–13717. [CrossRef]
17. Misra, D.; Nalamada, T.; Arasanipalai, A.U.; Hou, Q. Rotate to Attend: Convolutional Triplet Attention Module. In *Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*; IEEE: Waikoloa, HI, USA, 2021; pp. 3138–3147. [CrossRef]
18. Wu, Y.; He, K. Group Normalization. In *Proceedings of the European Conference on Computer Vision (ECCV 2018), Munich, Germany, 8–14 September 2018*; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Springer: Cham, Switzerland, 2018; pp. 3–19. [CrossRef]
19. Yuan, X.; Zheng, Z.; Li, Y.; Liu, X.; Liu, L.; Li, X.; Hou, Q.; Cheng, M.-M. Strip R-CNN: Large Strip Convolution for Remote Sensing Object Detection. In *Proceedings of the 40th AAAI Conference on Artificial Intelligence (AAAI 2026), Philadelphia, PA, USA, 22 February–1 March 2026*; AAAI Press: Washington, DC, USA, 2026; Volume 40, pp. 12259–12267. [CrossRef]
20. Ma, J.; Jiang, W.; Tang, X.; Zhang, X.; Liu, F.; Jiao, L. Multiscale Sparse Cross-Attention Network for Remote Sensing Scene Classification. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5605416. [CrossRef]
21. Xu, S.; Zheng, S.; Xu, W.; Xu, R.; Wang, C.; Zhang, J.; Teng, X.; Li, A.; Guo, L. HCF-Net: Hierarchical Context Fusion Network for Infrared Small Object Detection. In *Proceedings of the 2024 IEEE International Conference on Multimedia and Expo (ICME)*; IEEE: Niagara Falls, ON, Canada, 2024; pp. 1–6. [CrossRef]

22. Yang, F.; Zhang, L.; Yu, S.; Prokhorov, D.; Mei, X.; Ling, H. Feature Pyramid and Hierarchical Boosting Network for Pavement Crack Detection. *IEEE Trans. Intell. Transp. Syst.* **2020**, *21*, 1525–1535. [CrossRef]
23. Eisenbach, M.; Stricker, R.; Seichter, D.; Amende, K.; Debes, K.; Sesselmann, M.; Ebersbach, D.; Stoeckert, U.; Gross, H.-M. How to Get Pavement Distress Detection Ready for Deep Learning? A Systematic Approach. In *Proceedings of the 2017 International Joint Conference on Neural Networks (IJCNN)*; IEEE: Anchorage, AK, USA, 2017; pp. 2039–2047. [CrossRef]
24. Salehi, S.S.M.; Erdogmus, D.; Gholipour, A. Tversky Loss Function for Image Segmentation Using 3D Fully Convolutional Deep Networks. In *Machine Learning in Medical Imaging*; Springer: Cham, Switzerland, 2017. [CrossRef]
25. Lin, Q.; Li, W.; Zheng, X.; Fan, H.; Li, Z. DeepCrackAT: An Effective Crack Segmentation Framework Based on Learning Multi-Scale Crack Features. *Eng. Appl. Artif. Intell.* **2023**, *126*, 106876. [CrossRef]
26. Zhang, T.; Qin, L.; Zou, Q.; Zhang, L.; Wang, R.; Zhang, H. CrackScopeNet: A Lightweight Neural Network for Rapid Crack Detection on Resource-Constrained Drone Platforms. *Drones* **2024**, *8*, 417. [CrossRef]
27. Liu, H.; Yang, J.; Miao, X.; Mertz, C.; Kong, H. CrackFormer Network for Pavement Crack Segmentation. *IEEE Trans. Intell. Transp. Syst.* **2023**, *24*, 9240–9252. [CrossRef]
28. Tao, H.; Liu, B.; Cui, J.; Zhang, H. A Convolutional-Transformer Network for Crack Segmentation with Boundary Awareness. In *Proceedings of the 2023 IEEE International Conference on Image Processing (ICIP)*; IEEE: Kuala Lumpur, Malaysia, 2023; pp. 86–90. [CrossRef]
29. Wang, H.; Cao, P.; Wang, J.; Zaiane, O.R. UCTransNet: Rethinking the Skip Connections in U-Net from a Channel-Wise Perspective with Transformer. *Proc. AAAI Conf. Artif. Intell.* **2022**, *36*, 2441–2449. [CrossRef]

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.