This is a repository copy of *DCUFormer: Enhancing pavement crack segmentation in complex environments with dual-cross/upsampling attention*.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/220258/

Version: Accepted Version

Article:
Shan, J., Huang, Y. orcid.org/0000-0002-1220-6896 and Jiang, W. (2025) DCUFormer: Enhancing pavement crack segmentation in complex environments with dual-cross/upsampling attention. Expert Systems with Applications, 264. 125891. ISSN 0957-4174

https://doi.org/10.1016/j.eswa.2024.125891

Reuse
This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here: https://creativecommons.org/licenses/

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

DCUFormer: Enhancing Pavement Crack Segmentation in Complex Environments with Dual-Cross/UpSampling Attention

Jinhuan Shan[a, b], Yue Huang[c], Wei Jiang[a, b, *]

a. Key Laboratory for Special Area Highway Engineering of Ministry of Education, Chang'an University, Xi'an 710064, China

b. School of Highway, Chang'an University, Xi'an 710064, China

c. Institute for Transport Studies, University of Leeds, Leeds LS2 9JT, UK

Email: jhshan@chd.edu.cn (J. Shan), y.huang1@leeds.ac.uk (Y. Huang), jiangwei@chd.edu.cn (W. Jiang).

*Corresponding author:
Wei Jiang, +86-13572239600 (Mobile), jiangwei@chd.edu.cn

Abstract: Efficient road inspection and maintenance are essential to extend pavement lifespan and enhance safety. However, automated crack detection remains challenging due to varied environmental conditions and differences in image collection equipment, making robust algorithm development a critical need. Vision Transformers, with their capacity to capture long-range dependencies, offer significant advantages for crack detection in complex scenarios by effectively extracting global features. Nevertheless, existing Transformer-based methods encounter difficulties in boundary delineation due to decoder design limitations, which lead to suboptimal fusion of low-level and high-level features. To address this issue, we propose a comprehensive approach that integrates semantic preservation, detail refinement, and detail delineation. These concepts are realized through our novel Dual-Cross Attention Module (DCA) and Upsampling Attention Module (UA). The DCA module progressively filters redundant details from low-level feature layers using high-level semantic information, while preserving boundary details to refine high-level feature boundaries. In addition, the UA module employs progressive local cross-attention in upsampling, facilitating more precise boundary definitions and surpassing conventional dynamic upsampling methods. Our approach, utilizing both lightweight (MiT-B0, LVT) and middle-weight (Swin-T) backbones, demonstrates state-of-the-art performance on three diverse datasets—Crack500, CrackSC, and UAV-Crack500—highlighting its robustness across varied conditions. This work contributes to advancing Transformer-based architectures for defect segmentation in complex engineering contexts, underscoring the critical role of improved feature fusion in crack detection. The code is available at: https://github.com/SHAN-JH/DCUFormer.

Keywords: Pavement crack, Vision Transformer, Semantic segmentation, Feature upsampling

1. Introduction

Roads play a critical role in transportation, directly affecting commuter safety and comfort. Regular inspection and timely maintenance are essential to prolonging road lifespan and ensuring safety, particularly in urban areas where asphalt pavements are susceptible to environmental degradation and traffic loads (Lei et al., 2024; J. Li et al., 2022; Munawar et al., 2021). Without proper intervention, cracks can expand, leading to severe structural instability due to moisture and air infiltration, underscoring the need for efficient, precise crack detection methods (Marcelino et al., 2018; Ragnoli et al., 2018).

With advances in artificial intelligence (AI), deep learning has greatly enhanced the efficiency and accuracy of pavement inspections (Dong et al., 2024; Y. Li et al., 2021; Roy & Bhaduri, 2023; Tong et al., 2023; Zhu et al., 2023). However, crack segmentation remains challenging, especially under complex conditions. Cracks often have irregular, thin shapes, indistinct edges, and low contrast with their surroundings, making them difficult to detect. Furthermore, various external factors such as lighting and stains add complexity of AI-based pavement crack detection (F. Guo et al., 2023; Z. Li et al., 2024).

To address these challenges, recent research has explored Transformer-based architectures, which excel in capturing global dependencies across images via self-attention. Unlike convolutional neural networks (CNNs), which incrementally build feature representations through limited receptive fields, Transformers can model long-range relationships within the entire image, making them advantageous for complex crack patterns (Duan et al., 2024; Islam et al., 2024; Younesi et al., 2024). Although promising, existing Transformer models like Swin Transformer (Z. Liu et al., 2021) and MiT (Xie et al., 2021) lack effective decoders for fusing low-level and high-level features. The integration of detailed local information with global semantic context is essential for accurate crack segmentation in complex scenarios. Effectively combining these complementary aspects could significantly enhance segmentation performance by leveraging the strengths of both types of information.

To tackle this issue, we propose a novel Dual-Cross Attention Module (DCA) and an Upsampling Attention Module (UA) to enhance feature fusion and detail preservation (Fig. 1). Our DCA module uniquely combines high-level and low-level features, differing from prior models like FeedFormer (Shim et al., 2023) and U-MixFormer (Yeom & von Klitzing, 2023) by using a two-step cross-attention approach. First, it injects high-level semantic information into the low-level feature space to retain contextual information (semantic preservation). Then, it transmits low-level structural details to high-level feature maps, refining edges and eliminating redundant background information (detail refinement). This method addresses the need for accurate crack segmentation by preserving semantic context while amplifying essential edge details.

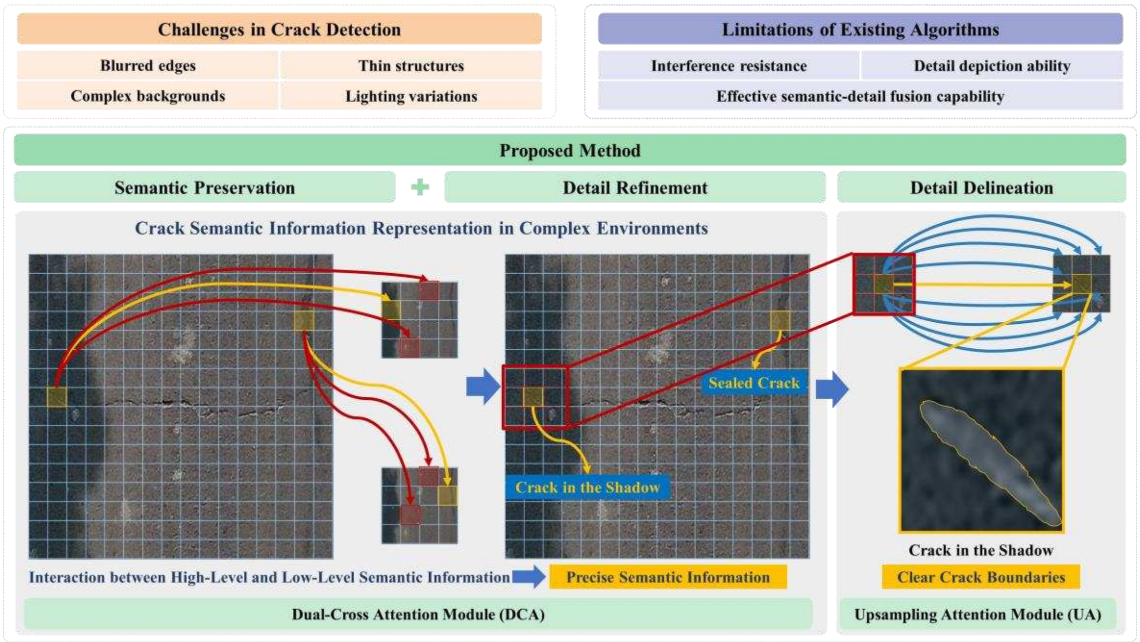

Challenges in Crack Detection
Blurred edges
Thin structures
Complex backgrounds
Lighting variations
Limitations of Existing Algorithms
Interference resistance
Detail depiction ability
Effective semantic-detail fusion capability
Proposed Method
Semantic Preservation
Detail Refinement
Detail Delineation
Crack Semantic Information Representation in Complex Environments
Sealed Crack
Crack in the Shadow
Crack in the Shadow
Interaction between High-Level and Low-Level Semantic Information
Precise Semantic Information
Clear Crack Boundaries
Dual-Cross Attention Module (DCA)
Upsampling Attention Module (UA)

Fig. 1. Overview of challenges in crack detection, limitations of existing algorithms, and the advantages of our proposed method.

Furthermore, our UA module improves upon traditional upsampling methods by applying local cross-attention within same resolution feature maps. Unlike methods that rely on high-resolution features for upsampling, which can result in suboptimal attention mapping, our approach leverages detail preservation and similarity requirements within the cross-attention framework to better delineate fine textures and boundaries in complex crack images. These advancements are compared against popular decoder and upsampling modules, demonstrating state-of-the-art (SOTA) performance. The primary contributions of this work are as follows:

(1) We propose the Dual-Cross Attention Module (DCA), designed to enhance the integration of low-level detail with high-level semantic information. The DCA improves the understanding of high-level semantic information in low-level feature maps, eliminates redundant information in lower-level features, and reconstructs or amplifies important details that may be lost or blurred due to the increasing depth of neural networks.

(2) We introduce the Upsampling Attention Module (UA), a novel upsampling module based on attention mechanisms. This module leverages progressive local cross-attention for precise and effective upsampling, enabling improved learning and prediction of edges and texture details.

(3) The model's performance was evaluated on three datasets with significant variations in crack morphology and environmental interference: Crack500, CrackSC, and our UAV-Crack500. Utilizing MiT-B0, LVT, and Swin-T as backbones, our model outperformed existing high-performance decoder models, offering new perspectives for Transformer-based feature refinement and upsampling design.

The structure of the paper is as follows: Section 2 reviews current decoder designs based on CNNs and Transformers as well as upsampling methods; Section 3 introduces our model architecture; Section 4 presents test and visualization results on three datasets, along with ablation experiments; Section 5 concludes the content of the paper and discusses future research directions.

2. Related Works

In semantic segmentation tasks, the encoder-decoder architecture is fundamental. The encoder extracts features, capturing edges, textures, shapes, and semantic information, often through progressive downsampling to reduce computational demand and capture global contextual information. However, direct use of downsampled feature maps can blur boundary information. To address this, decoders are designed to reconstruct the image, gradually restoring resolution and recovering lost spatial details for high-accuracy segmentation with fine boundaries.

This section reviews CNN-based and Transformer-based decoders, and upsampling methods, highlighting their efficiency and accuracy in recovering spatial details and boundary information, while also pointing out their limitations.

2.1 CNN-based Decoder Heads

CNN architectures utilize downsampling to enhance computational efficiency, feature representation, and model generalization. Various methods have been proposed to restore downsampled feature maps to their original resolution. The Fully Convolutional Network (FCN) (Long et al., 2015) directly upsamples feature maps downsampled by factors of 32 or 16, resulting in coarse restorations and blurred boundaries. U-Net (Ronneberger et al., 2015) employs stepwise upsampling and lateral connections to gradually restore spatial details and structural information, showing excellent performance across various segmentation domains. The Pyramid Scene Parsing Network (PSPNet) (Zhao et al., 2017) utilizes a Pyramid Pooling Module (PPM) to integrate context information at different scales, significantly improving segmentation accuracy in complex backgrounds and multi-scale object scenarios. DeepLabv3 (Chen et al., 2017) incorporates atrous convolution to capture multi-scale context information through the Atrous Spatial Pyramid Pooling (ASPP) module, while DeepLabv3+ (Chen et al., 2018) combines low-level and high-level features to enhance detail resolution capability. Panoptic FPN (Kirillov et al., 2019), with an FPN backbone (Lin et al., 2017), uses a top-down and skip-connection architecture similar to UNet, but with an asymmetrical, lightweight design, adjusting different-level feature maps to have the same number of channels, thus reducing computational load and parameter count. These methods improve the decoder's capability to restore fine image details through effective multi-scale information fusion.

Despite the introduction of techniques such as atrous convolutions (Chen et al., 2017) and deformable convolutions (Dai et al., 2017) in CNN decoder structures to expand the receptive field, their global perception capability remains insufficient. This limitation often results in false negative predictions when segmenting thin and elongated cracks in complex environments. In classical CNN architectures, to restore the resolution of high-level feature maps, bilinear interpolation is typically employed for upsampling. Although low-level features are integrated through concatenation or addition, this approach can still lead to issues with unclear boundaries.

2.2 Transformer-based Decoder Heads

While deeper CNNs capture broader contextual information, they still primarily focus on local features and may lack global awareness in complex scenes. Transformer-based models address this limitation through self-attention mechanisms, enabling superior performance in capturing global dependencies. These models typically employ Transformer/CNN backbones for initial feature extraction, followed by advanced decoder structures that leverage Transformer mechanisms to further enhance the extraction of detailed information and semantic enrichment. SenFormer (Bousselham et al., 2022) builds on the FPN structure, incorporating a Transformer-based learner to extract features from different decoder levels. Mask2Former (Cheng et al., 2022) introduces a pixel decoder module that gradually upsamples features, feeding them into a Transformer decoder to enhance small object recognition. FeedFormer (Shim et al., 2023) uses high-level encoder features as queries and lowest-level encoder features as keys and values in its Transformer decoder, enhancing structure by integrating fine spatial details from low-level features with high-level semantic information. This approach effectively restores important details in the segmentation process. U-MixFormer (Yeom & von Klitzing, 2023) integrates the U-Net structure with

Transformer operations, replacing lateral connections with Transformer decoders and mixing features from both encoder and previous decoder stages. These models demonstrate the evolution towards more sophisticated architectures that effectively balance global context capture and local feature preservation, pushing the boundaries of performance in visual semantic segmentation tasks.

Transformer-based decoder heads enhance global information decoding through attention mechanisms, strengthening the semantic information in high-level feature maps while preserving important details. However, previous research has typically focused either on deepening the semantics of feature maps or on characterizing fine details, without effectively combining these two aspects. This dichotomy in approach suggests a potential gap in the field, where a more integrated method could potentially yield improved results by simultaneously addressing both semantic enrichment and detail preservation.

2.3 Upsampling Methods

In the decoder stage, upsampling methods are typically employed to recover image detail information. Traditional upsampling methods include bilinear interpolation and nearest neighbor interpolation, which are non-learnable and use predefined kernels for upsampling operations. Other methods such as deconvolution (Noh et al., 2015), pixel shuffle (Shi et al., 2016), and unpooling (Badrinarayanan et al., 2017) are also widely used. Although the convolutions in deconvolution and pixel shuffle are learnable, their kernels operate on the entire feature map and cannot be dynamically generated. Unpooling can perform upsampling based on indices saved during downsampling and can adjust dynamically according to input, but its zero-filling operation compromises semantic information.

Recently, researchers have proposed several new dynamic upsampling methods, such as CARAFE (Wang et al., 2019), FADE (Lu, Liu, Fu, et al., 2022), SAPA (Lu, Liu, Ye, et al., 2022), DySample (W. Liu et al., 2023), and ReSFU (Zhou et al., 2024). CARAFE dynamically generates upsampling operators based on encoder feature maps;

FADE further combines encoder and decoder feature maps to guide the upsampling process; SAPA utilizes a similarity-aware point affiliation mechanism to design an upsampling operator, achieving both semantic smoothness and boundary sharpness; DySample dynamically generates sampling point positions from a point sampling perspective to guide upsampling; ReSFU achieves more fine-grained upsampling through query-key feature alignment and a fine-grained neighbor selection strategy. These methods show certain advancements compared to fixed upsampling methods, primarily generating query-key pairs to guide upsampling using encoder or decoder feature maps.

However, these dynamic upsampling methods still have some limitations. As pointed out by ReSFU, query-key pairs from different feature maps are not fully aligned in detail and semantic spaces, leading to suboptimal upsampling results. Although ReSFU attempts to perform query-key feature alignment, discrepancies in semantic and detail spaces still exist. This is because the detail space contains more high-frequency information such as structure and color, while the semantic space is smooth. To perform query-key attention calculations more effectively, cross-processing of information is needed beforehand. Subsequently, local cross-attention can further restore crack edge details.

3. Proposed Architecture

3.1 Overall Architecture

Based on the aforementioned approach, we propose our model – DCUFormer (Fig. 2). DCUFormer is designed to address the challenges in dense prediction tasks, particularly focusing on the effective fusion of low-level and high-level feature maps. The architecture incorporates mechanisms for semantic preservation, detail refinement, and detail delineation, aiming to achieve a balance between preserving high-level semantic information and enhancing fine-grained details. The model leverages a hierarchical structure to extract multi-scale features while employing novel techniques to overcome the limitations of traditional upsampling and feature fusion methods. By implementing a progressive fusion strategy and utilizing cross-attention mechanisms, DCUFormer strives to maintain the integrity of semantic information from high-level features while accurately delineating detailed structures guided by low-level features.

The model accepts feature maps with four levels, which align with the outputs from popular backbone networks such as Swin Transformer, MiT (Mix Transformer), and LVT (Light Vision Transformer). This design choice ensures compatibility with diverse state-of-the-art backbones, allowing for flexible integration into various deep learning pipelines. Assuming the input image size is $H \times W \times C$, the different levels of output feature maps are $\frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}} \times C_i$, denoted as E_i.

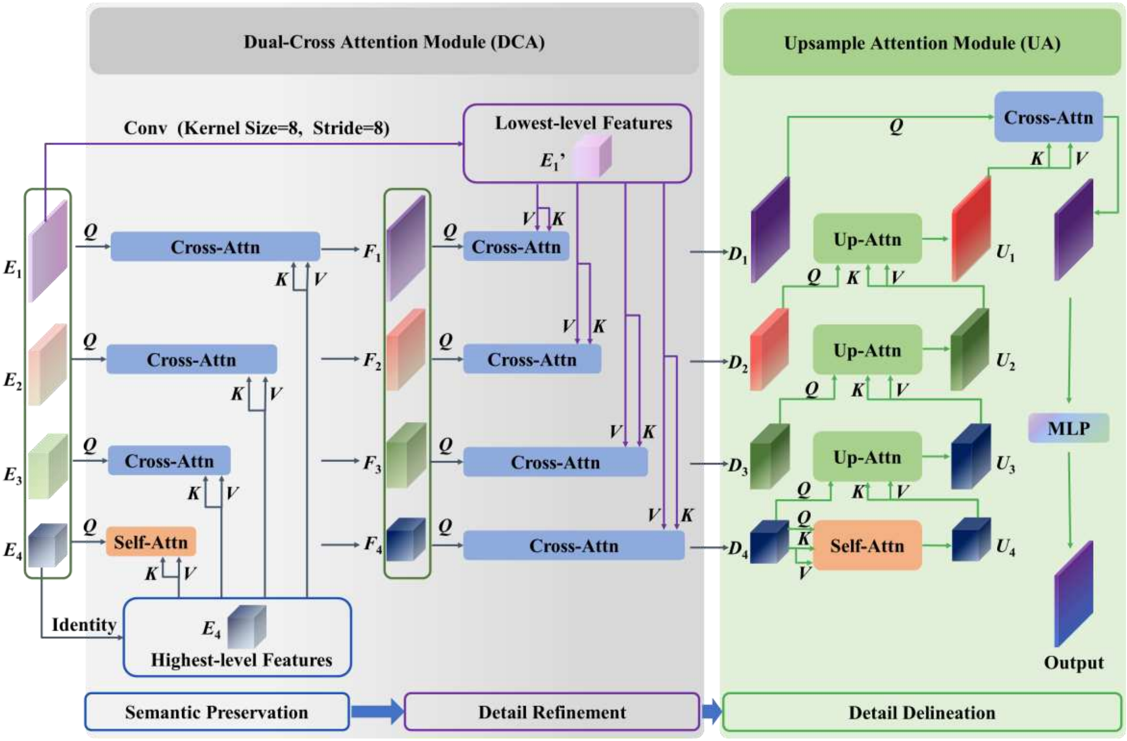


Fig. 2. DCUFormer architecture.

Our model architecture leverages the Dual-Cross Attention Module (DCA) and the Upsampling Attention Module (UA) to effectively integrate and refine features extracted by the encoder, enhancing the semantic segmentation performance.

Initially, feature maps from different hierarchical levels of the encoder are fed into the DCA, enhancing the low-level feature maps' understanding of high-level semantic information while eliminating redundant information.

Following the DCA, the refined feature maps from different levels are processed by the Upsampling Attention Module (UA). Within a U-shaped architecture, high-level feature maps are connected laterally and undergo upsampling attention mechanisms. This process results in upsampled lower-level feature maps, where the learnable upsampling attention mechanism ensures the gradual restoration of detailed information.

3.2 Dual-Cross Attention Module (DCA)

Considering that high-level feature maps obtained from the encoder are rich in semantic information while low-level feature maps contain detailed structural and boundary information, the Dual-Cross Attention Module (DCA) fully utilizes both highest-level feature maps E_4 and lowest-level feature maps E_1.

Initially, the feature maps E_i serve as the query, with the highest-level feature map E_4 acting as both key and value for cross-attention computation. Subsequently, the resulting feature maps F_i from different levels serve as the query, and the lowest-level feature map E_1, after undergoing convolution operations with a kernel size and stride of 8 and having its channels expanded to match E_4, acts as both key and value for a second round of cross-attention computation. This integration ensures a more comprehensive representation by combining both high-level semantic information and low-level detailed information.

3.3 Upsampling Attention Module (UA)

Currently, for upsampling operations, most models adopt the simple and explicit method of bilinear interpolation; however, this method is non-learnable and tends to smooth out boundary information to some extent. To fully utilize the feature maps obtained from the previous layer's upsampling as well as their lateral connections for upsampling operations, we propose the Upsampling Attention Module (UA) (Fig. 3).

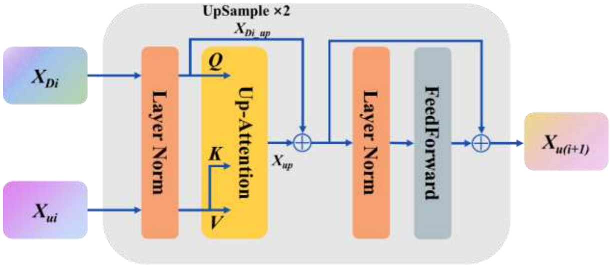


Fig. 3. Upsampling Attention Block.

In this module, the laterally connected feature maps from the previous layer (D_i) serve as the query, and the upsampled feature maps from the previous layer (U_i) serve as both key and value. They first undergo layer normalization before proceeding to the Upsampling Attention Operation. To accommodate the residual connection after upsampling, the D_i map is upsampled by a factor of 2 and then added to the map processed by the attention mechanism. This is followed by computation in a feed-forward neural network to achieve nonlinear fitting. Unlike traditional non-learnable methods, the Upsampling Attention Module (UA) leverages higher-layer contextual information for precise and effective upsampling through the attention mechanism, enabling better learning and prediction of edge and texture detail information.

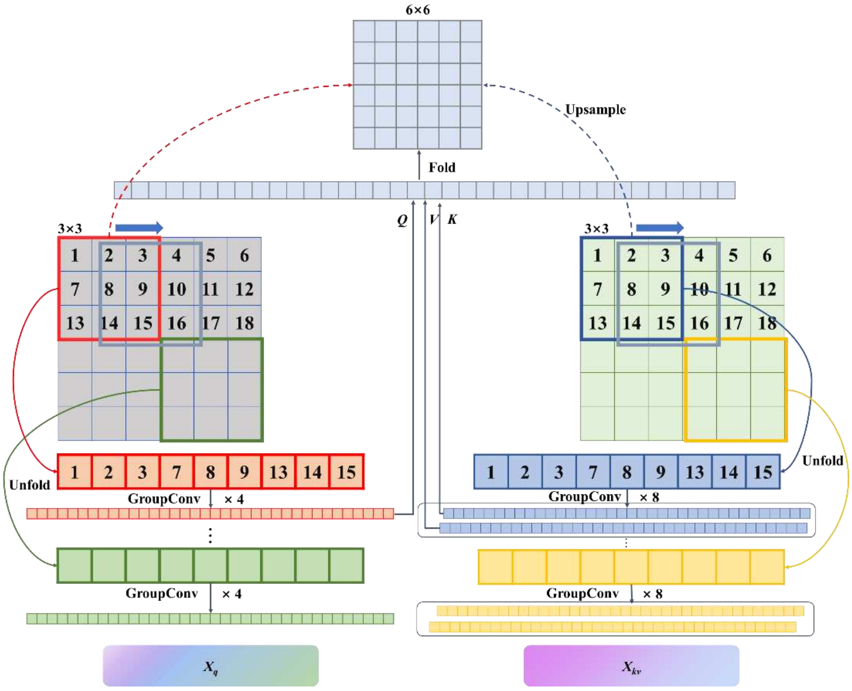


Fig. 4. Upsampling Attention Operation.

The steps for the Upsampling Attention Operation (Fig. 4) are as follows: first, the input feature maps of the *i*-th layer D_i and U_i undergo layer normalization followed by an unfold operation with a kernel size of 3×3, stride of 1, and padding of 1. Assume the sizes of input maps D_i and U_i are both $X \in \mathbb{R}^{C_i \times H_i \times W_i}$, where C_i represents the number of channels, and H_i and W_i represent height and width, respectively. The unfolding operation can be regarded as transforming each local $k \times k$ window in the feature map X into a column vector, thereby generating a new matrix $X_{\text{unfold}} \in \mathbb{R}^{(k^2 C_i) \times N}$ (Formula (1)-(2)), where N is the number of columns after unfolding. This process does not change the spatial dimensions due to the use of stride 1 and padding 1 in the unfold operation. Consequently, $N = H_i \times W_i$.

$$X_{D_i\_un} = \text{Unfold}\left(X_{D_i}\right) \tag{1}$$

$$X_{U_i\_un} = \text{Unfold}\left(X_{U_i}\right) \tag{2}$$

Subsequent to the unfold operation, grouped convolution (Formula (3)-(4)) is employed to facilitate feature learning for upsampling, with each group consisting of the unfolded 9-pixel blocks. The output channel dimension for D_i, when it functions as the query, is established at 36, reflecting a doubling in the upsampling rate, explicitly calculated as $(3 \times 2) \times (3 \times 2)$. In the case of U_i, designated as both key and value, the output channels are accordingly doubled to 36 channels for the key and 36 channels for the value, to accommodate the upsampled feature representation.

$$X_q = W_{Up\_D_i} \cdot X_{D_i\_un} \tag{3}$$

$$X_{kv} = W_{Up\_U_i} \cdot X_{U_i\_un} \tag{4}$$

The weights for D_i and U_i in the upsampling operation are denoted as $W_{Up\_D_i} \in \mathbb{R}^{(k^2 C_i \cdot \alpha^2) \times (k^2 C_i)}$ and $W_{Up\_U_i} \in \mathbb{R}^{(2k^2 C_i \cdot \alpha^2) \times (k^2 C_i)}$; α represents the scaling factor.

Subsequently, we reshape and permute the dimensions of the unfolded feature maps to fit the dimensions required for the subsequent operations. The reshaped q_i, k_i, v_i $(q_i, k_i, v_i \in \mathbb{R}^{B\times num\_heads\times(H_i\times W_i)\times(\alpha^2\times k^2)\times\frac{C_{out\_i}}{num\_heads}})$ have dimensions suited for computing the attention mechanism (Formula (5)), where 36 denotes the number of channels for each of the unfolded pixel groups, facilitating the attention operation across 3×3 pixel areas. This allows for a detailed feature learning process, effectively capturing both spatial and semantic information within these regions. This attention mechanism helps to selectively emphasize the most relevant features within the upsampled feature space, incorporating a richer contextual understanding that goes beyond local pixel information.

$$\mathrm{MultiHead\_Attention}(q_i, k_i, v_i) = \mathrm{softmax}(\tfrac{q_i k_i^T}{\sqrt{d_k}})v_i \tag{5}$$

Where d_k represents the dimensionality of the key vectors, ensuring that the attention scores are appropriately normalized, avoiding disproportionately large values that could dominate the softmax output, thereby maintaining a balanced attention distribution across the features.

After the attention computation, the processed feature maps are subject to two subsequent folding operations aimed at restoring the attended feature maps to their original spatial configuration. The first folding operation employs a kernel size of 2×2, with a stride of 2 and no padding, effectively producing an upsampled feature map with dimensions doubled in both height and width ($\mathbb{R}^{B\times(C_{out\_i}\times k^2)\times(\alpha H_i\times\alpha W_i)}$).

The second folding operation then re-integrates the 9-pixel neighborhood back into the feature map using a kernel size of 3×3, with a stride of 1 and padding of 1. Unlike the first fold, this operation does not alter the size of the feature map; instead, it focuses on rearranging the pixels to their precise locations based on the attention-driven importance. By doing so, it ensures that the detailed spatial relationships and contextual information, accentuated through the attention mechanism, are accurately represented within the upsampled feature map. This dual-stage folding process is crucial for achieving a refined reconstruction of the feature map that retains both the enhanced details and the original spatial integrity.

After obtaining the upsampled feature map X_{up}, it is added to the bilinearly upsampled feature map $X_{D_i\_up}$. This operation enriches the pathways through which the upsampled feature map is generated, incorporating both a learnable upsampling method and a direct bilinear upsampling shortcut branch. This approach effectively enhances upsampling capability and mitigates gradient vanishing issues during deep network training. This is followed by a standard layer normalization and feedforward operation, ultimately producing the upsampled $X_{u(i+1)}$. The overall process is illustrated in the given Formula (6)-(7).

$$X_i = \mathrm{Up\_Atten}(\mathrm{LN}(X_{D_i}, X_{U_i})) + \mathrm{Up}(X_{D_i}) \tag{6}$$

$$X_{U(i+1)} = \mathrm{FFL}(\mathrm{LN}(X_i)) + X_i \tag{7}$$

4. Experimental results and analysis

4.1 Datasets

Imaging equipment variability and altitude significantly impact image quality (Fig. 5). Aerial photography yields broader area coverage but often results in the loss of minor features and details, with greater susceptibility to weather and lighting conditions, leading to reduced image contrast and color saturation. Conversely, low-altitude imagery captures more detailed information but is limited in scope and contains redundant data, potentially compromising model efficiency. To evaluate the model's segmentation performance on complex scene cracks, we utilized three distinct imaging devices and pavement crack datasets from various scenarios, including Crack500, CrackSC, and our UAV-Crack500.

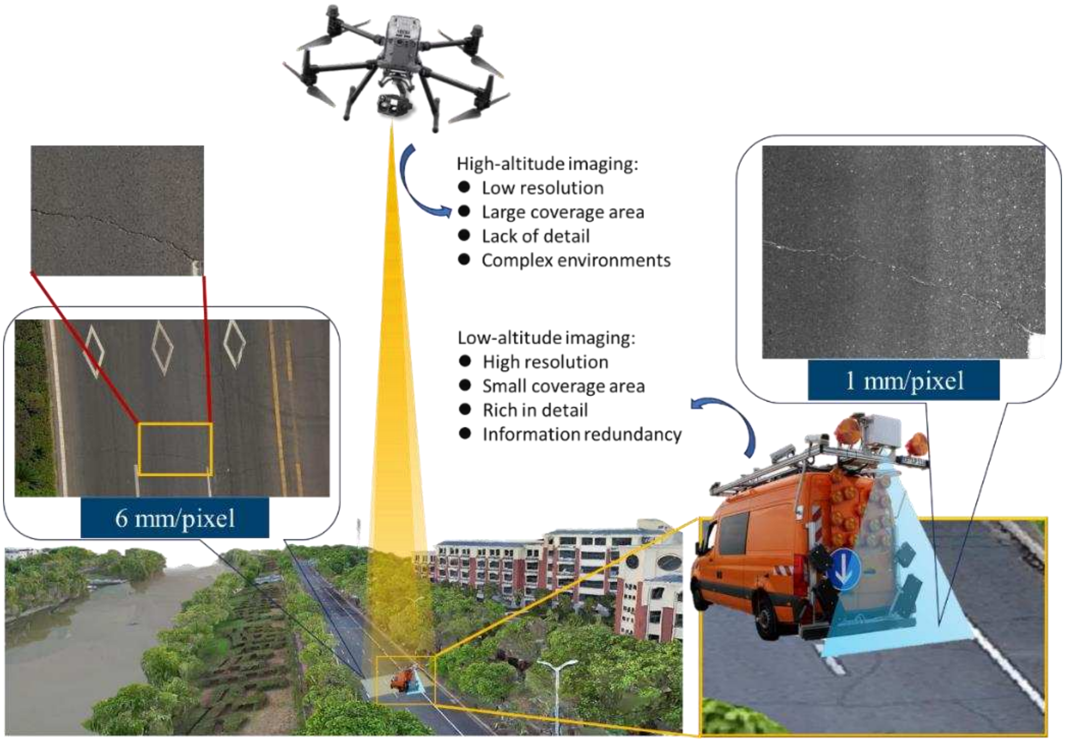


Fig. 5. Comparison of pavement crack images at different imaging altitudes.

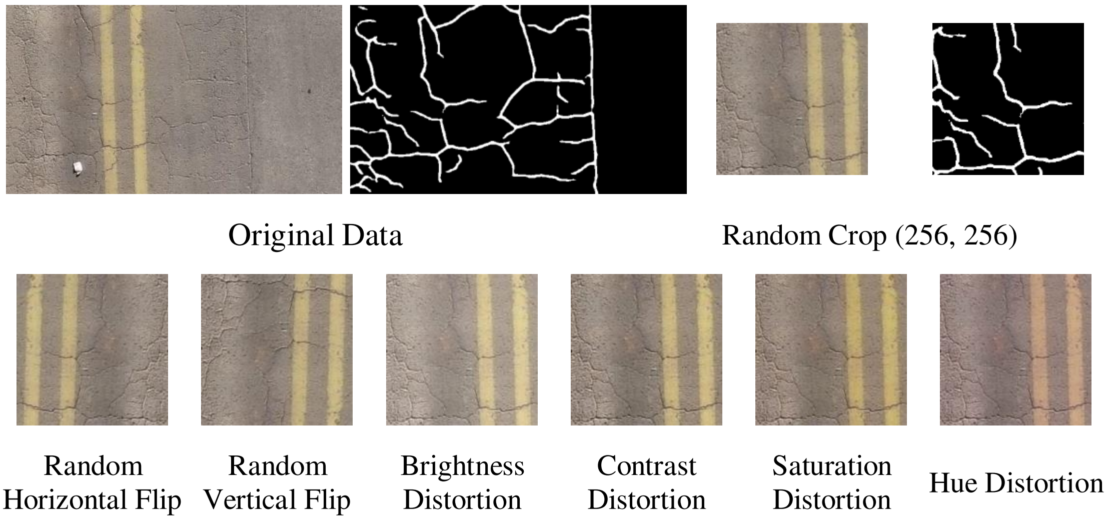
Crack500 dataset (Yang et al., 2020) is composed of 500 high-resolution photographs of road damages, each with an original resolution of 2000 × 1500 pixels, captured using cell phones on the main campus of Temple University. To economize on training expenses while enhancing the crack pixel ratio, the original images were segmented into 16 non-overlapping regions, with only those containing over 1000 crack pixels retained. In total, 1896 images were selected for the training set, 348 for the validation set, and 1124 for the test set.

CrackSC dataset (F. Guo et al., 2023) consists of 197 road damage images (320 × 480 pixels) captured by an iPhone 8 around Enoree Ave, Columbia, SC. This dataset emphasizes complex pavement distress scenes with interference factors like shadows, leaves, and moss, which pose significant challenges to crack detection. Without a predefined dataset division by the authors, we divided it into 99 training images, 19 validation images, and 79 testing images, adhering to a 5:1:4 distribution ratio.

UAV-Crack500 dataset (Shan et al., 2024), collected and annotated by us using EISeg (Hao et al., 2022), is focused on pavement distress imagery obtained from drones.

Captured at an altitude of 50 m, the original image resolution is 2688 × 1512 pixels, covering approximately 16 m × 9 m. The aerial perspective results in a lower ratio of crack pixels, with the images being blurred and more susceptible to external environmental noise, adding to the segmentation challenge. The images were divided into 16 non-overlapping regions, from which 500 images displaying significant distress features and disturbances were selected, comprising 250 images for the training set, 50 for the validation set, and 200 for the testing set.

In real-world scenarios, data is inherently diverse; however, the datasets collected often have inherent limitations and do not cover a wide range of scenes. Through data augmentation, models can be trained to grasp deeper semantic information beyond simple low-level features (such as color and contours). Moreover, limited dataset sizes can lead to overfitting, particularly in large models like Transformers. Data augmentation creates new, unseen examples, thereby enhancing the model's generalization and robustness and preventing overfitting. This paper employs three data augmentation techniques: Random Crop, Random Flip (Horizontal and Vertical), and Photometric Distortion (adjusting Brightness, Contrast, Saturation, and Hue), to achieve sample diversity. The specific alterations to images and masks are detailed in Fig.6.



Original Data

Random Crop (256, 256)

Random Horizontal Flip

Random Vertical Flip

Brightness Distortion

Contrast Distortion

Saturation Distortion

Hue Distortion

Fig. 6. Examples of data augmentation.

4.2 Training and Evaluation Settings

For fairness in our experiments, all our procedures were conducted within the public codebase—MMSegmentation v1.2.0 framework[1], using an NVIDIA Tesla T4 GPU (16G) for model construction, training, and testing. The following details the rationale behind our parameter choices and optimization strategies:

In our approach, the image crop size of 256 × 256 during both training and testing was selected to balance computational efficiency with capturing sufficient contextual details from the input data, while the batch size of 16 optimized GPU memory usage and maintained stable gradient estimates. For testing, we used the “slide” prediction mode with a crop size of 256 × 256 and a stride of 128, which not only ensured consistency with the training process but also enhanced accuracy by averaging overlapping predictions, reducing edge artifacts.

The AdamW optimizer was selected for its effective handling of sparse gradients and adaptive learning rates. A learning rate of 6e-5 was determined through preliminary experimentation, ensuring stable convergence. The exponential decay averages for gradients were set at 0.9 and 0.999, with a weight decay of 0.01 added to regularize the model and mitigate overfitting risks. Training spanned 30,000 iterations, with the first 1,500 iterations featuring a linear learning rate warm-up to facilitate a smooth adaptation to the optimization process. Afterward, a polynomial learning rate decay (power = 1) was applied for progressive fine-tuning.

To improve segmentation accuracy, especially in imbalanced datasets, we employed a combination of binary cross-entropy (BCE) and dice loss. BCE handles pixel-wise classification, while dice loss addresses overlap-based loss, offering a balanced approach that enhances model performance on challenging segmentation tasks.

[1] https://github.com/open-mmlab/mmsegmentation

Reproducibility was ensured by using consistent random seeds across all experiments, and results were averaged over three trials to minimize the effects of random variations. This comprehensive setup ensured reliable and robust model evaluation.

4.3 Implementation Details

The selection of backbone networks for this study was guided by three critical factors: dataset characteristics, hierarchical architecture, and global information extraction capabilities. Given the relatively small scale of pavement crack semantic segmentation datasets, light to medium-weight backbones were prioritized to mitigate the risk of overfitting, which is particularly pertinent when dealing with limited data. Backbones with hierarchical network architectures were selected due to their demonstrated efficacy in processing multi-scale information, allowing for more nuanced feature extraction across different levels of abstraction. This approach boosts both model efficiency and accuracy. Recent advancements in Vision Transformer-based architectures have shown significant advantages in global information extraction. Balancing these considerations, MiT-B0 and LVT were employed as light-weight options, and Swin-T as a medium-weight alternative for the experiments. This selection allows for performance evaluation across different computational complexities while leveraging the strengths of Vision Transformer-based architectures. The hierarchical structures of these chosen backbones further contribute to mitigating overfitting and enhancing processing efficiency.

For comparison, we selected the SegFormer Head and U-MixFormer Head to contrast with our decoder model. SegFormer utilizes a straightforward MLP for channel rearrangement followed by concatenation and another MLP to arrive at the final prediction. Meanwhile, U-MixFormer, which has shown impressive performance in the visual domain, employs an upsampling and lateral connection structure similar to UNet, thereby excelling in detail and boundary recovery. To establish the superiority of the proposed method, comparisons were conducted with current high-performing segmentation models, including SegNeXt (M.-H. Guo et al., 2022), Mask2Former (Cheng et al., 2022), and VWFormer (Yan et al., 2024). This comprehensive approach ensures a robust evaluation of the method against state-of-the-art alternatives while addressing the specific challenges posed by pavement crack datasets. Through this systematic experimental design and comprehensive comparative analysis, the study aims to provide valuable insights and innovative approaches to the field of semantic segmentation, particularly in the challenging application scenario of pavement crack detection.

For performance evaluation, we utilize Average Accuracy (aAcc), Mean Intersection over Union (mIoU), Mean Accuracy (mAcc), Mean Precision (mPr), Mean Recall (mRe), and Mean F_1 (mF_1) Score as our metrics. The best models typically showcase paired superior performance in both mIoU and Mean F_1 Score, and we select the model that performs optimally on these two metrics as best model. Furthermore, considering that cracks do not have clear and distinct pixel boundaries and that the dataset annotation process is subject to human error, leading to possible pixel deviations, we follow the practice of other studies (Weng et al., 2019; Panella et al., 2022; Zhang et al., 2022) by applying a 2-pixel tolerance in our model evaluation. This means that if the model's predictions are within two pixels of the ground truth, they are considered true positives.

4.4 Comparison with State-of-the-art Segmentation Approaches

The model was tested on three distinct datasets sourced from Crack500, CrackSC, and UAV-Crack500, with the results displayed in Tables 1 to 3. Based on the mIOU and mF_1 scores, it is evident that our model, DCUFormer, surpassed existing models across different backbones, achieving state-of-the-art (SOTA) results.

For the Crack500 dataset, models with a Swin-T backbone exhibit similar performance, linked to the dataset's characteristics of larger and more prevalent cracks. Swin-T's effective feature extraction via sliding windows permits simpler feature interpretation in the decoder phase, resulting in comparable outcomes among the models. However, for datasets with complex scenes and blurred boundaries, such as CrackSC and UAV-Crack500, Swin-T's strong feature extraction capacity requires a decoder that excels in feature interpretation and fusion, leading our model with Swin-T as the backbone to achieve superior results on the CrackSC and UAV-Crack500 datasets.

Figures 7 to 9 visualize the performance of our top models (measured by mIoU and mF_1) in both light-weight and middle-weight categories, compared to other state-of-the-art models.

In the Crack500 dataset, although cracks are larger and more prominent, the similarity between crack pixels and background road surface pixels leads to a tendency for models to produce false negatives, resulting in discontinuous cracks. However, our model achieved better prediction results, identifying cracks more accurately and with better connectivity.

In the CrackSC dataset, cracks are finer and accompanied by shadows and stains. Under shadows, road and crack pixels are almost indistinguishable, often leading models to false negatives by misclassifying them as pavement pixels. Under influences like stains and leaves, due to their color and shape similarities to cracks, models are prone to false positives. As shown in Fig. 8, our model can effectively refine crack information from shadows based on global context and distinguish between disturbances such as stains and leaves.

The UAV-Crack500 dataset, captured from high altitudes, suffers from atmospheric lighting interference, diminishing clarity and color saturation of distant objects. This effect diminishes the contrast between crack and background pixels, with cracks occupying a smaller proportion and having blurred boundaries. These conditions complicate the segmentation task. However, as demonstrated in Fig. 9, our model maintains commendable performance, effectively distinguishing cracks from shadows and accurately separating disruptive elements such as shadows along markings and transition zones around manhole covers.

Table 1. Performance comparison with the state-of-the art methods on Crack500.

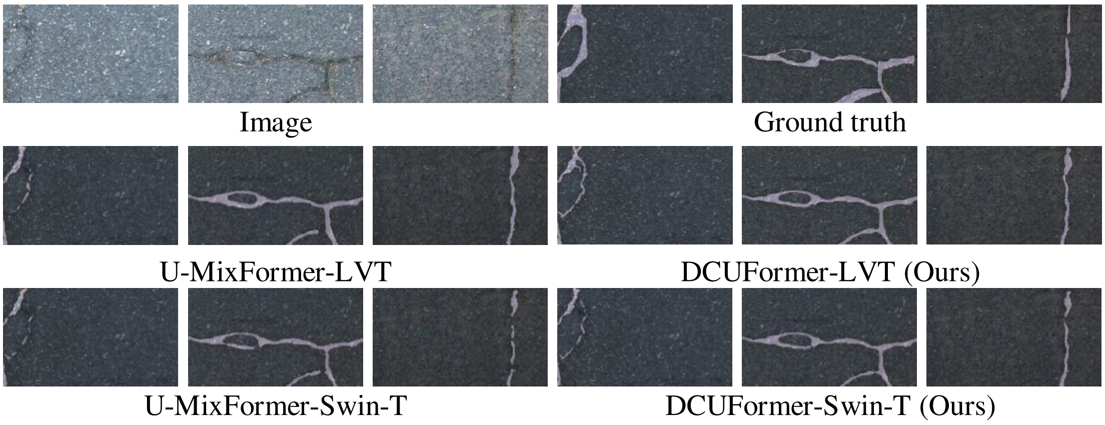
	Method	Encoder	aAcc	mIoU	mAcc	mPr	mRe	mF$_1$
Light-weight	Segformer	MiT-B0	97.50	81.41	87.95	89.89	87.95	88.89
	U-MixFormer	MiT-B0	97.56	81.92	88.63	89.90	88.63	89.26
	Segformer	LVT	97.57	81.69	87.92	**90.35**	87.92	89.09
	U-MixFormer	LVT	**97.58**	82.03	88.69	90.00	88.69	89.33
	SegNeXt	MSCAN-T	97.43	81.35	88.65	89.07	88.65	88.86
	DCUFormer (Ours)	MiT-B0	**97.58**	82.11	88.76	90.04	88.76	89.39
	DCUFormer (Ours)	LVT	**97.58**	**82.15**	**89.10**	89.74	**89.10**	**89.42**
Middle-weight	Segformer	Swin-T	97.57	82.06	88.81	89.92	88.81	89.35
	Mask2Former	Swin-T	97.07	79.49	87.95	87.08	87.95	87.51
	U-MixFormer	Swin-T	**97.62**	82.06	88.21	**90.57**	88.21	89.35
	VWFormer	Swin-T	97.47	81.99	**89.81**	88.84	**89.81**	89.32
	DCUFormer (Ours)	Swin-T	97.61	**82.07**	88.39	90.39	88.39	**89.36**

Table 2. Performance comparison with the state-of-the art methods on CrackSC.

	Method	**Encoder**	**aAcc**	**mIoU**	**mAcc**	**mPr**	**mRe**	**mF$_1$**
Light-weight	SegFormer	MiT-B0	98.81	78.07	80.85	93.82	80.85	86.15
	U-MixFormer	MiT-B0	98.84	78.97	81.91	93.76	81.91	86.86
	SegFormer	LVT	98.81	78.01	80.66	94.05	80.66	86.09
	U-MixFormer	LVT	**98.86**	79.24	81.95	**94.30**	81.95	87.07
	SegNeXt	MSCAN-T	98.63	73.35	75.56	93.66	75.56	82.12
	DCUFormer (Ours)	MiT-B0	98.83	78.97	82.01	93.59	82.01	86.86
	DCUFormer (Ours)	LVT	**98.86**	**79.85**	**83.06**	93.55	**83.06**	**87.54**
Middle-weight	SegFormer	Swin-T	98.76	75.71	77.86	**94.54**	77.86	84.19
	Mask2Former	Swin-T	98.83	80.15	83.82	92.85	83.82	87.77
	U-MixFormer	Swin-T	98.85	78.85	81.67	93.96	81.67	86.76
	VWFormer	Swin-T	98.77	76.87	79.53	93.68	79.53	85.16
	DCUFormer (Ours)	Swin-T	**98.89**	**80.84**	**84.14**	93.69	**84.14**	**88.29**

Table 3. Performance comparison with the state-of-the art methods on UAV-Crack500.

	Method	Encoder	aAcc	mIoU	mAcc	mPr	mRe	mF1
Light-weight	SegFormer	MiT-B0	99.21	84.18	87.19	94.95	87.19	90.68
	U-MixFormer	MiT-B0	99.22	84.49	87.58	94.89	87.58	90.90
	SegFormer	LVT	99.20	83.84	86.52	95.37	86.52	90.44
	U-MixFormer	LVT	99.21	83.87	86.47	**95.52**	86.47	90.47
	SegNeXt	MSCAN-T	99.21	84.63	87.46	95.31	87.46	91.00
	DCUFormer (Ours)	MiT-B0	**99.24**	**85.19**	**88.74**	94.42	**88.74**	**91.38**
	DCUFormer (Ours)	LVT	**99.24**	84.92	88.00	95.01	88.00	91.19
Middle-weight	SegFormer	Swin-T	99.19	83.68	86.48	95.16	86.48	90.34
	Mask2Former	Swin-T	98.91	79.49	84.46	90.59	84.46	87.26
	U-MixFormer	Swin-T	99.22	84.23	87.39	94.75	87.39	90.72
	VWFormer	Swin-T	99.20	84.37	87.11	95.41	87.11	90.82
	DCUFormer (Ours)	Swin-T	**99.27**	**85.45**	**88.40**	**95.29**	**88.40**	**91.55**



Image

Ground truth

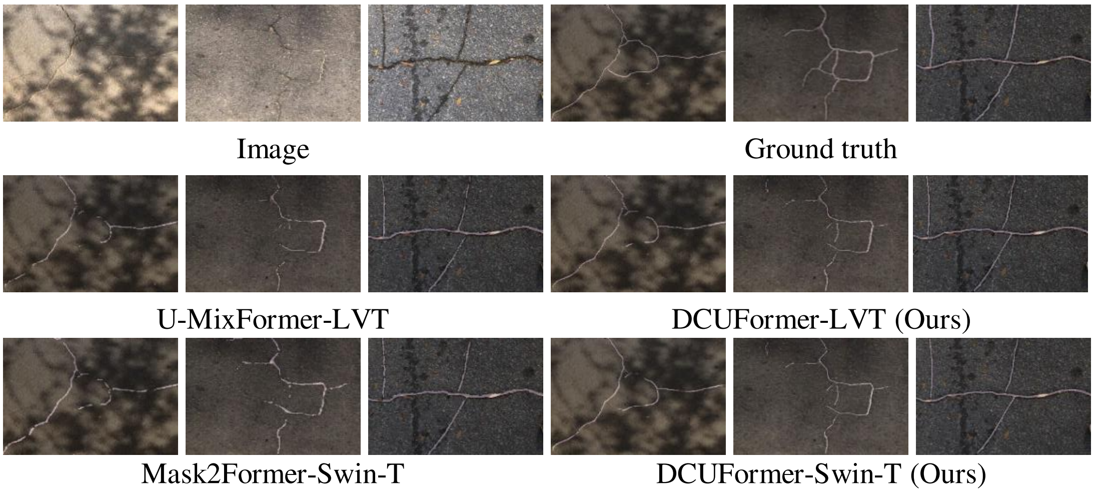
U-MixFormer-LVT

DCUFormer-LVT (Ours)

U-MixFormer-Swin-T

DCUFormer-Swin-T (Ours)

Fig. 7. Qualitative results on Crack500.



Image

Ground truth

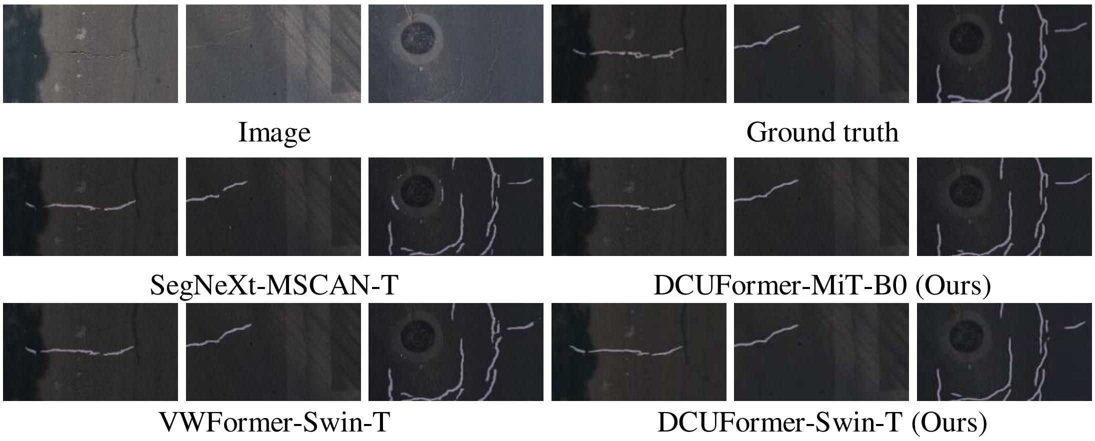
U-MixFormer-LVT

DCUFormer-LVT (Ours)

Mask2Former-Swin-T

DCUFormer-Swin-T (Ours)

Fig. 8. Qualitative results on CrackSC.



Image

Ground truth

SegNeXt-MSCAN-T

DCUFormer-MiT-B0 (Ours)

VWFormer-Swin-T

DCUFormer-Swin-T (Ours)

Fig. 9. Qualitative results on UAV-Crack500.

4.5 Comparison with State-of-the-art Upsampling Approaches

To validate the superiority of the Upsampling Attention Module (UA), this study conducted comparative experiments on upsampling modules using SegFormer-B0. The comparison included novel and effective dynamic upsampling modules in semantic segmentation, such as SAPA, DySample, ReSFU, as well as the conventional but efficient bilinear interpolation operation (Table 4). Although SAPA, DySample, and ReSFU achieved state-of-the-art results on large-scale datasets (such as ADE20K and Cityscapes), they did not perform as well on smaller crack datasets, failing to surpass the effectiveness of direct bilinear interpolation.

SAPA and ReSFU compute queries from the previous encoder layer and perform semantic alignment with keys from the current layer. However, for thin cracks with pixels similar to the background, guiding upsampling through query-key pairs from different sources does not achieve effective alignment. This approach may even compromise the semantic information of previously extracted boundaries, erroneously classifying them as background. DySample utilizes local information from input features to dynamically adjust sampling strategies. Despite its simplicity and dynamic nature, the lack of comprehensive pixel interaction hinders its ability to differentiate semantic information of pixels near crack boundaries.

This limitation is evident in the varying segmentation results across different crack datasets. SAPA, DySample, and ReSFU perform comparably to bilinear interpolation on the Crack500 dataset, where cracks are relatively large with clear boundaries. However, their performance significantly degrades compared to direct bilinear interpolation on datasets like CrackSC, featuring thin cracks in complex environments, and UAV-Crack500, which contains low-resolution and blurry crack images.

The proposed Upsampling Attention Module (UA) innovatively combines cross-attention upsampling of same-level semantic feature maps with bilinear interpolation residual connections, effectively addressing key issues in dynamic upsampling. The UA module achieves semantic-level query-key alignment, enhancing the model's comprehension of high-level features.

Furthermore, UA introduces bilinear interpolation residual connections, which not only enhance gradient flow but also prove particularly effective in distinguishing semantically similar foreground and background elements. This approach utilizes bilinear interpolation information to rectify semantic errors that may arise from the cross-attention mechanism. While bilinear interpolation can produce smoothing effects, it also preserves certain boundary information. By leveraging the advantages of both methods, UA achieves clear boundary semantics.

Compared to other dynamic upsampling models, UA maintains computational efficiency while balancing semantic consistency, detail preservation, and model robustness by integrating original features with attention mechanism outputs. This approach not only enhances model performance in complex visual tasks but also provides new insights into addressing challenging problems such as fine boundary recognition and semantic segmentation.

Table 4. Performance comparison with different upsampling modules.

SegFormer-B0	**Params**	**FLOPs**	**Crack500**		**CrackSC**		**UAV-Crack500**	
			mIoU	**mF$_1$**	**mIoU**	**mF$_1$**	**mIoU**	**mF$_1$**
Bilinear-MLP	3.7M	7.9G	81.41	88.89	78.07	86.15	84.18	90.68
SAPA-MLP	3.8M	8.5G	81.34	88.86	73.90	82.62	83.52	90.23
DySample-MLP	3.8M	8.0G	81.33	88.84	72.64	81.47	81.16	88.51
ReSFU-MLP	3.9M	10.0G	81.04	88.63	72.08	80.95	81.93	89.09
UA (Ours)	7.0M	6.3G	**82.05**	**89.35**	**78.48**	**86.48**	**84.61**	**90.98**

4.6 Ablation Studies

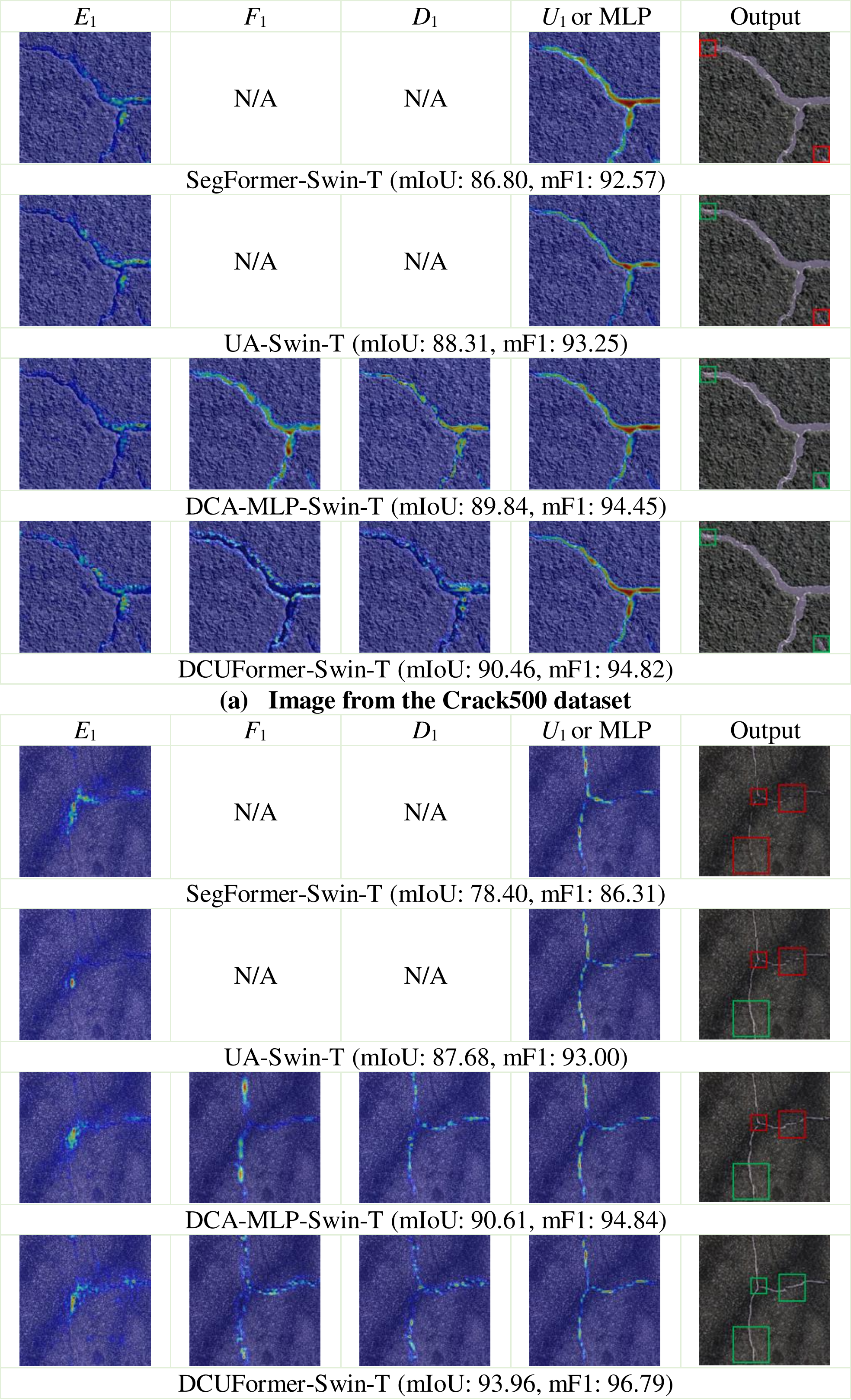
We conducted ablation studies on the different modules of our approach. Using SegFormer-B0 as the baseline, we integrated the Upsampling Attention Module (UA) directly onto the encoder, allowing for direct prediction using the UA module on the four different levels of feature maps. Furthermore, we experimented with directly concatenating the four-level feature maps obtained from our Dual-Cross Attention Module (DCA) and then predicting outcomes through an MLP. The final model incorporates both the Dual-Cross Attention Module (DCA) and the Upsampling Attention Module (UA) as parts of the decoder module, which constitutes our proposed method, DCUFormer. This integration aims to harness the strengths of both modules to enhance the model's ability to accurately segment and delineate intricate features such as cracks, especially in challenging environments, thereby significantly improving the segmentation accuracy and detail capture compared to conventional methods.

According to the results in Table 5, our model significantly improves segmentation precision across different backbones (encoders). On the Crack500 dataset, our model can enhance performance up to 0.70% mIOU and 0.50% mF$_1$; on the CrackSC dataset, improvements can reach up to 5.13% mIOU and 4.1% mF$_1$; and on the UAV-Crack500 dataset, we observe a maximum increase of 1.01% mIOU and 0.70% mF$_1$. Notably, our model exhibits the most substantial improvement with the Swin-T encoder for the CrackSC and UAV-Crack500 datasets, but the least for the Crack500 dataset. This could be due to the larger proportion of cracks and clearer crack boundaries in the Crack500 dataset, where even other lightweight encoders can perform well in feature extraction. The CrackSC and UAV-Crack500 datasets, characterized by finer and more blurred crack boundaries, gain advantages from the hierarchical Transformer structure of Swin's Windows Multi-Head Self-Attention and Shifted Windows Multi-Head Self-Attention. This architecture improves the identification of crack boundaries and leverages contextual information to mitigate interference from diverse environmental factors.

Table 5. Ablation results.

Method	Encoder	Params	FLOPs	Crack500		CrackSC		UAV-Crack500	
				mIoU	**mF$_1$**	**mIoU**	**mF$_1$**	**mIoU**	**mF$_1$**
SegFormer	MiT-B0	3.7M	7.9G	81.41	88.89	78.07	86.15	84.18	90.68
UA	MiT-B0	7.0M	6.3G	82.05	89.35	78.48	86.48	84.61	90.98
DCA-MLP	MiT-B0	10.9M	7.3G	81.99	89.30	78.50	86.69	84.50	90.91
DCUFormer	MiT-B0	10.8M	9.2G	**82.11**	**89.39**	**78.97**	**86.86**	**85.19**	**91.38**
SegFormer	LVT	3.6M	6.7G	81.69	89.09	78.01	86.09	83.84	90.44
UA	LVT	7.3M	11.2G	81.89	89.23	79.53	87.30	84.44	90.87
DCA-MLP	LVT	11.7M	11.2G	81.65	89.07	78.01	86.09	84.52	90.97
DCUFormer	LVT	11.7M	15.9G	**82.15**	**89.42**	**79.85**	**87.54**	**84.92**	**91.19**
SegFormer	Swin-T	28.2M	30.8G	82.06	89.35	75.71	84.19	83.68	90.34
UA	Swin-T	54.2M	44.3G	81.97	89.22	78.26	86.29	85.19	91.38
DCA-MLP	Swin-T	86.0M	52.0G	81.74	89.13	79.51	87.28	84.92	91.20
DCUFormer	Swin-T	85.5M	62.1G	**82.07**	**89.36**	**80.84**	**88.29**	**85.45**	**91.55**

LayerCAM, which assigns element-wise weights for generating class activation maps, was applied to our model for interpretability. Class activation maps of the highest-resolution feature maps (E_1, F_1, D_1, U_1) in Swin-T-based models were visualized using images from the Crack500 and CrackSC datasets, as shown in Fig. 10.


E_1
F_1
D_1
U_1 or MLP
Output
N/A
N/A
SegFormer-Swin-T (mIoU: 86.80, mF1: 92.57)
N/A
N/A
UA-Swin-T (mIoU: 88.31, mF1: 93.25)
DCA-MLP-Swin-T (mIoU: 89.84, mF1: 94.45)
DCUFormer-Swin-T (mIoU: 90.46, mF1: 94.82)
E_1
F_1
D_1
U_1 or MLP
Output
N/A
N/A
SegFormer-Swin-T (mIoU: 78.40, mF1: 86.31)
N/A
N/A
UA-Swin-T (mIoU: 87.68, mF1: 93.00)
DCA-MLP-Swin-T (mIoU: 90.61, mF1: 94.84)
DCUFormer-Swin-T (mIoU: 93.96, mF1: 96.79)

(a) Image from the Crack500 dataset

(b) Images from the CrackSC dataset

Fig. 10. LayerCAM visualizations of feature maps E_1, F_1, D_1, U_1 from the model based on Swin-T backbone.

In the original SegFormer-Swin-T model, the highest-layer feature maps exhibit poor delineation of details. The direct resizing followed by MLP-based segmentation leads to suboptimal performance in regions influenced by shadows and complex backgrounds (highlighted in red), causing segmentation discontinuities. The incorporation of the UA module progressively injects regional semantic information through localized cross-attention mechanisms, restoring high-resolution detail layer by layer. This process results in significant detail recovery, improving the overall delineation of cracks. However, when segmenting cracks affected by shadows or similar to pavement textures (highlighted in red), the local cross-attention mechanism shows some limitations, with certain discontinuities persisting despite improvements over the original model.

To further enhance performance, the DCA module was introduced. After the first cross-attention mechanism (with E_4 as the key and value, and E_1 as the query), the resulting feature map F_1 shows enhanced crack perception, with activations more concentrated around the cracks, thereby eliminating redundant information in lower-level features and preserving semantic information. However, this step alone is insufficient for precise crack localization due to the lower resolution of the high-level feature maps. Through the second cross-attention mechanism, with F_1 as the query and E_1 as the key and value, the resulting D_1 feature map further focuses on the center and edges of the cracks. This improvement occurs because F_1, rich in semantic information, computes the similarity with E_1, which contains detailed spatial information, allowing E_1 to guide F_1 in reconstructing or amplifying important details, thereby achieving finer detail refinement.

It is worth noting that with the combined DCA and UA modules, the model's first cross-attention operation in DCA focuses more on the crack boundaries rather than the center. After the second cross-attention operation, the activations gradually shift towards the crack center, achieving greater precision. By integrating the UA module, the model accurately identifies both the crack region and refines the crack edges, resulting in a comprehensive process from semantic preservation to detail refinement and, ultimately, to the delineation of fine details. This improvement allows the model to overcome environmental interferences such as shadows, ensuring precise crack segmentation and significantly enhancing performance.

4.7 Computational Efficiency

We utilized the fvcore library (https://github.com/facebookresearch/fvcore) developed by the Facebook AI Research (FAIR) team to compare the parameters (Params) and floating-point operations (FLOPs) of our model with those of other state-of-the-art models (Tables 5 and 6) with input size of (3, 512, 512).

As shown in Table 5, although our UA module has the highest number of parameters, it exhibits the lowest FLOPs. This is due to our use of regional grouped convolution for regional feature extraction and the implementation of a regional cross-attention mechanism for upsampling. Compared to multi-layer perceptron (MLP) and other global dynamic upsampling methods, our approach results in lower FLOPs, thus providing a computational advantage. Furthermore, our UA module outperforms advanced upsampling operations and traditional bilinear interpolation in terms of performance.

Table 6 illustrates that, compared to the baseline model, our model shows a significant increase in both parameters and FLOPs. This is primarily because the DCA dual cross-attention module substantially increases the FLOPs, while the UA upsampling attention module significantly adds to the parameter count. However, our models based on lightweight encoders (such as MiT-B0 and LVT) perform better than those using the Swin-T middle-weight encoder (e.g., SegFormer, Mask2Former, U-MixFormer). This indicates that our proposed model can effectively leverage features extracted by lightweight networks to enhance performance without relying on excessively heavy encoders.

Nevertheless, there is still room for improvement in our model compared to lightweight models. Future research will focus on simplifying the DCA and UA modules by utilizing sparse attention and lightweight convolution, aiming to achieve true lightweight performance.

Table 6. Efficiency comparison.

	Method	**Encoder**	**Params**	**FLOPs**	**mIoU**		
					Crack500	CrackSC	UAV-Crack500
Light-weight	SegFormer	MiT-B0	3.7M	7.9G	81.41	78.07	84.18
	U-MixFormer	MiT-B0	6.4M	5.2G	81.92	78.97	84.49
	SegFormer	LVT	3.6M	6.7G	81.69	78.01	83.84
	U-MixFormer	LVT	6.8M	7.8G	82.03	79.24	83.87
	SegNeXt-T	MSCAN-T	4.2M	6.3G	81.35	73.35	84.63
	DCUFormer (Ours)	MiT-B0	10.8M	9.2G	82.11	78.97	**85.19**
	DCUFormer (Ours)	LVT	11.7M	15.9G	**82.15**	**79.85**	84.92
Middle-weight	SegFormer	Swin-T	28.2M	30.8G	82.06	75.71	83.68
	Mask2Former	Swin-T	47.0M	74.0G	79.49	80.15	79.49
	U-MixFormer	Swin-T	52.3M	40.2G	82.06	78.85	84.23
	VWFormer	Swin-T	35.1M	57.8G	81.99	76.87	84.37
	DCUFormer (Ours)	Swin-T	85.5M	62.1G	**82.07**	**80.84**	**85.45**

5. Conclusion and Future Research

Crack detection is an essential method for maintaining the normal operation and safety of civil engineering structures. However, current automated detection methods are significantly influenced by environmental conditions and equipment performance, and the robustness of these algorithms needs to be enhanced to meet higher standards. To efficiently utilize encoder feature maps, preserve semantic information, and enhance image details, we propose a three-step approach: semantic preservation, detail refinement, and detail delineation. This methodology aims to further improve the effective identification of cracks and accurate segmentation of boundaries in complex backgrounds. Consequently, we introduce two novel modules: a Dual-Cross Attention Module (DCA) and an Upsampling Attention Module (UA). The DCA incorporates semantic preservation and detail refinement capabilities, functioning as a feature extraction cross-attention network. It effectively infuses high-level semantic information into lower-level feature maps, enhancing their semantic understanding, and integrates lower-level structural and detail information back into the high-level semantic information, thereby reconstructing or reinforcing the details that might be lost or blurred due to increased depths of the neural networks. The UA focuses on detail delineation, employing a cross-attention mechanism among neighboring pixels for precise upsampling. This allows the model to learn the information of the upsampled image through the attention mechanism, making boundary semantics clearer compared to bilinear interpolation and other dynamic feature upsampling operators.

We evaluated our approach using both lightweight backbones (MIT-B0 and LVT) and a middle-weight backbone (Swin-T) across three diverse crack datasets: Crack500, CrackSC, and UAV-Crack500. These datasets encompass various crack formations and environmental conditions. By comparing our method with the current state-of-the-art feature extraction and dynamic upsampling algorithms, the results indicate that our approach achieves state-of-the-art (SOTA) performance.

While this study primarily focuses on pavement crack segmentation, the proposed method demonstrates broad application potential across various engineering domains. In manufacturing and construction industries, a wide array of defects—such as surface scratches and stains in manufactured products, welding cracks and line breaks in electronic components, and material cracks in steel structures, walls, and road/bridge surfaces—share common characteristics that pose significant challenges to detection and segmentation processes. These shared challenges primarily stem from three factors: (1) the high similarity between defect pixels and background pixels, (2) the variability introduced by imaging equipment parameters and environmental conditions, and (3) the diverse and often elongated morphology of defects. Collectively, these factors have historically impeded the efficacy of existing models in accurately distinguishing foreground (defect) from background pixels. Our DCA and UA Module could enhance the model's capacity for information extraction, and facilitate the gradual restoration of fine crack pixels, respectively. The synergistic operation of these modules significantly improves segmentation accuracy, thereby advancing the state-of-the-art in defect detection across multiple engineering applications.

Building upon insights gained from experimentation, future research in crack detection should address two key challenges: enhancing model generalization and optimizing lightweight efficiency. The significant variations observed in crack morphology and light-shadow conditions across different datasets, stemming from diverse data collection and processing techniques, underscore the need for more adaptable algorithms. Developing models capable of handling the even greater variability of environmental conditions and crack formations in natural settings will be essential. Simultaneously, despite our current light-weight models outperforming medium-weight counterparts, further optimization is necessary. Utilizing sparse attention mechanisms and lightweight convolution operations could achieve true lightweight efficiency. Such advancements could lead to significant breakthroughs in crack detection technology, balancing performance and efficiency.

Author contributions

Jinhuan Shan: Methodology, Software, Data Curation, Writing - Original Draft**.** **Yue Huang**: Conceptualization, Validation, Writing - Review & Editing**. Wei Jiang**: Conceptualization, Resources, Supervision, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this study.

Acknowledgments

This project was jointly supported by the National Natural Science Foundation of China (Grant No. 52122809, Grant No. 52038001, Grant No. 52311530081).

References

Badrinarayanan, V., Kendall, A., & Cipolla, R. (2017). SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(12), 2481–2495. https://doi.org/10.1109/TPAMI.2016.2644615

Bousselham, W., Thibault, G., Pagano, L., Machireddy, A., Gray, J., Chang, Y. H., & Song, X. (2022). *Efficient Self-Ensemble for Semantic Segmentation* (No. arXiv:2111.13280). arXiv. http://arxiv.org/abs/2111.13280

Chen, L.-C., Papandreou, G., Schroff, F., & Adam, H. (2017). *Rethinking Atrous Convolution for Semantic Image Segmentation* (No. arXiv:1706.05587). arXiv. http://arxiv.org/abs/1706.05587

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. *In Proceedings of the European Conference on Computer Vision (ECCV)*, 833–851. https://doi.org/10.1007/978-3-030-01234-2_49

Cheng, B., Misra, I., Schwing, A. G., Kirillov, A., & Girdhar, R. (2022). Masked-attention Mask Transformer for Universal Image Segmentation. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1280–1289. https://doi.org/10.1109/CVPR52688.2022.00135

Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., & Wei, Y. (2017). Deformable Convolutional Networks. *2017 IEEE International Conference on Computer Vision (ICCV)*, 764–773. https://doi.org/10.1109/ICCV.2017.89

Dong, J., Wang, N., Fang, H., Lu, H., Ma, D., & Hu, H. (2024). Automatic augmentation and segmentation system for three-dimensional point cloud of pavement potholes by fusion convolution and transformer. *Advanced Engineering Informatics*, *60*, 102378. https://doi.org/10.1016/j.aei.2024.102378

Duan, Z., Luo, X., & Zhang, T. (2024). Combining transformers with CNN for multi-focus image fusion. *Expert Systems with Applications*, *235*, 121156. https://doi.org/10.1016/j.eswa.2023.121156

Guo, F., Qian, Y., Liu, J., & Yu, H. (2023). Pavement crack detection based on transformer network. *Automation in Construction*, *145*, 104646. https://doi.org/10.1016/j.autcon.2022.104646

Guo, M.-H., Lu, C.-Z., Hou, Q., Liu, Z.-N., Cheng, M.-M., & Hu, S.-M. (2022). SegNeXt: Rethinking convolutional attention design for semantic segmentation. *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 1140–1156.

Hao, Y., Liu, Y., Chen, Y., Han, L., Peng, J., Tang, S., Chen, G., Wu, Z., Chen, Z., & Lai, B. (2022). *EISeg: An Efficient Interactive Segmentation Tool based on PaddlePaddle* (No. arXiv:2210.08788). arXiv. http://arxiv.org/abs/2210.08788

Islam, S., Elmekki, H., Elsebai, A., Bentahar, J., Drawel, N., Rjoub, G., & Pedrycz, W. (2024). A comprehensive survey on applications of transformers for deep learning tasks. *Expert Systems with Applications*, *241*, 122666. https://doi.org/10.1016/j.eswa.2023.122666

Kirillov, A., Girshick, R., He, K., & Dollár, P. (2019). *Panoptic Feature Pyramid Networks*. 6392–6401. https://doi.org/10.1109/CVPR.2019.00656

Lei, Q., Zhong, J., Wang, C., & Li, X. (2024). Integrating Crack Causal Augmentation Framework and Dynamic Binary Threshold for imbalanced crack instance segmentation. *Expert Systems with Applications*, *240*, 122552. https://doi.org/10.1016/j.eswa.2023.122552

Li, J., Zhang, Z., Wang, X., & Yan, W. (2022). Intelligent decision-making model in preventive maintenance of asphalt pavement based on PSO-GRU neural network. *Advanced Engineering Informatics*, *51*, 101525. https://doi.org/10.1016/j.aei.2022.101525

Li, Y., Che, P., Liu, C., Wu, D., & Du, Y. (2021). Cross-scene pavement distress detection by a novel transfer learning framework. *Computer-Aided Civil and Infrastructure Engineering*, *36*(11), 1398–1415. https://doi.org/10.1111/mice.12674

Li, Z., Lan, Y., & Lin, W. (2024). Footbridge damage detection using smartphone-recorded responses of micromobility and convolutional neural networks. *Automation in Construction*, *166*, 105587. https://doi.org/10.1016/j.autcon.2024.105587

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature Pyramid Networks for Object Detection. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 936–944. https://doi.org/10.1109/CVPR.2017.106

Liu, W., Lu, H., Fu, H., & Cao, Z. (2023). Learning to Upsample by Learning to Sample. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 6004–6014. https://doi.org/10.1109/ICCV51070.2023.00554

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 9992–10002. https://doi.org/10.1109/ICCV48922.2021.00986

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully Convolutional Networks for Semantic Segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3431–3440. https://openaccess.thecvf.com/content_cvpr_2015/html/Long_Fully_Convolutional_Networks_2015_CVPR_paper.html

Lu, H., Liu, W., Fu, H., & Cao, Z. (2022). FADE: Fusing the Assets of Decoder and Encoder for Task-Agnostic Upsampling. *Computer Vision – ECCV 2022*, 231–247. https://doi.org/10.1007/978-3-031-19812-0_14

Lu, H., Liu, W., Ye, Z., Fu, H., Liu, Y., & Cao, Z. (2022). SAPA: Similarity-Aware Point Affiliation for Feature Upsampling. *Advances in Neural Information Processing Systems*. https://openreview.net/forum?id=hFni381edL

Marcelino, P., Lurdes Antunes, M. de, & Fortunato, E. (2018). Comprehensive performance indicators for road pavement condition assessment. *Structure and Infrastructure Engineering*, *14*(11), 1433–1445. https://doi.org/10.1080/15732479.2018.1446179

Munawar, H. S., Hammad, A. W. A., Haddad, A., Soares, C. A. P., & Waller, S. T. (2021). Image-Based Crack Detection Methods: A Review. *Infrastructures*, *6*(8), Article 8. https://doi.org/10.3390/infrastructures6080115

Noh, H., Hong, S., & Han, B. (2015). Learning Deconvolution Network for Semantic Segmentation. *2015 IEEE International Conference on Computer Vision (ICCV)*, 1520–1528. https://doi.org/10.1109/ICCV.2015.178

Panella, F., Lipani, A., & Boehm, J. (2022). Semantic segmentation of cracks: Data challenges and architecture. *Automation in Construction*, *135*, 104110. https://doi.org/10.1016/j.autcon.2021.104110

Ragnoli, A., De Blasiis, M. R., & Di Benedetto, A. (2018). Pavement Distress Detection Methods: A Review. *Infrastructures*, *3*(4), Article 4. https://doi.org/10.3390/infrastructures3040058

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In N. Navab, J. Hornegger, W. M. Wells, & A. F. Frangi (Eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (pp. 234–241). Springer International Publishing. https://doi.org/10.1007/978-3-319-24574-4_28

Roy, A. M., & Bhaduri, J. (2023). DenseSPH-YOLOv5: An automated damage detection model based on DenseNet and Swin-Transformer prediction head-enabled YOLOv5 with attention mechanism. *Advanced Engineering Informatics*, *56*, 102007. https://doi.org/10.1016/j.aei.2023.102007

Shan, J., Jiang, W., Huang, Y., Yuan, D., & Liu, Y. (2024). Unmanned Aerial Vehicle (UAV)-Based Pavement Image Stitching Without Occlusion, Crack Semantic Segmentation, and Quantification. *IEEE Transactions on Intelligent Transportation Systems*, *25*(11), 17038–17053. https://doi.org/10.1109/TITS.2024.3424525

Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A. P., Bishop, R., Rueckert, D., & Wang, Z. (2016). Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1874–1883. https://doi.org/10.1109/CVPR.2016.207

Shim, J., Yu, H., Kong, K., & Kang, S.-J. (2023). FeedFormer: Revisiting Transformer Decoder for

Efficient Semantic Segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, *37*(2), 2263–2271. https://doi.org/10.1609/aaai.v37i2.25321

Tong, Z., Ma, T., Zhang, W., & Huyan, J. (2023). Evidential transformer for pavement distress segmentation. *Computer-Aided Civil and Infrastructure Engineering*, *38*(16), 2317–2338. https://doi.org/10.1111/mice.13018

Wang, J., Chen, K., Xu, R., Liu, Z., Loy, C. C., & Lin, D. (2019). CARAFE: Content-Aware ReAssembly of FEatures. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 3007–3016. https://doi.org/10.1109/ICCV.2019.00310

Weng, X., Huang, Y., & Wang, W. (2019). Segment-based pavement crack quantification. *Automation in Construction*, *105*, 102819. https://doi.org/10.1016/j.autcon.2019.04.014

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., & Luo, P. (2021). SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. *Advances in Neural Information Processing Systems*, *34*, 12077–12090. https://proceedings.neurips.cc/paper_files/paper/2021/hash/64f1f27bf1b4ec22924fd0acb550c235-Abstract.html

Yan, H., Wu, M., & Zhang, C. (2024, April 26). Multi-Scale Representations by Varying Window Attention for Semantic Segmentation. *The Twelfth International Conference on Learning Representations*. The Twelfth International Conference on Learning Representations. https://openreview.net/forum?id=lAhWGOkpSR

Yang, F., Zhang, L., Yu, S., Prokhorov, D., Mei, X., & Ling, H. (2020). Feature Pyramid and Hierarchical Boosting Network for Pavement Crack Detection. *IEEE Transactions on*

Intelligent Transportation Systems, *21*(4), 1525–1535. https://doi.org/10.1109/TITS.2019.2910595

Yeom, S.-K., & von Klitzing, J. (2023). *U-MixFormer: UNet-like Transformer with Mix-Attention for Efficient Semantic Segmentation* (No. arXiv:2312.06272). arXiv. http://arxiv.org/abs/2312.06272

Younesi, A., Ansari, M., Fazli, M., Ejlali, A., Shafique, M., & Henkel, J. (2024). A Comprehensive Survey of Convolutions in Deep Learning: Applications, Challenges, and Future Trends. *IEEE Access*, *12*, 41180–41218. https://doi.org/10.1109/ACCESS.2024.3376441

Zhang, Y., Wu, J., Li, Q., Zhao, X., & Tan, M. (2022). Beyond crack: Fine-grained pavement defect segmentation using three-stream neural networks. *IEEE Transactions on Intelligent Transportation Systems*, *23*(9), 14820–14832. IEEE Transactions on Intelligent Transportation Systems. https://doi.org/10.1109/TITS.2021.3134374

Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017). Pyramid Scene Parsing Network. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2881–2890. https://openaccess.thecvf.com/content_cvpr_2017/html/Zhao_Pyramid_Scene_Parsing_CVPR_2017_paper.html

Zhou, M., Wang, H., Zheng, Y., & Meng, D. (2024). *A Refreshed Similarity-based Upsampler for Direct High-Ratio Feature Upsampling* (No. arXiv:2407.02283). arXiv. http://arxiv.org/abs/2407.02283

Zhu, G., Liu, J., Fan, Z., Yuan, D., Ma, P., Wang, M., Sheng, W., & Wang, K. C. P. (2023). A lightweight encoder–decoder network for automatic pavement crack detection. *Computer-*

Aided Civil and Infrastructure Engineering, *39*(12), 1743–1765. https://doi.org/10.1111/mice.13103