CrackDINO: A DINOv3-based Hybrid Framework for Fine-Grained Crack Segmentation

Jiajie Li[1], Bin Hu[1(✉)], Jinlai Zhang[2(✉)], Shifa Tang[1], Kejia Wang[1]

[1] School of Computer Science and Technology, Changsha University of Science and Technology, Changsha, 410114, Hunan, China
[2] College of Mechanical and Vehicle Engineering, Changsha University of Science and Technology, Changsha, 410114, Hunan, China

Abstract. Precise pixel-wise segmentation of pavement cracks serves as a fundamental requirement for smart road inspection and structural infrastructure health monitoring. However, due to the low contrast, complex background interference, and elongated structural characteristics of cracks, existing segmentation methods often suffer from discontinuous predictions and missed detection of tiny crack regions. Specifically, CNN architectures cannot effectively model global long-distance feature correlations; in contrast, Transformer models easily discard subtle local spatial textures when splitting input images into discrete patch tokens. To mitigate all the above drawbacks, this work constructs a dedicated crack segmentation framework named CrackDINO built upon the pre-trained DINOv3 vision model. Specifically, we design an RGB-guided Multi-scale Feature Pyramid (RGMFP) module to enhance hierarchical semantic interaction across different feature resolutions. In addition, a Crack-aware Stable Attention (CAS) module is introduced to strengthen weak crack responses and improve discriminative representation for thin and low-contrast crack regions. Furthermore, a Cascaded Hierarchical Multi-scale Decoder (CHMD) is proposed to progressively recover spatial details and preserve crack continuity during feature reconstruction. We conduct comprehensive comparative experiments on two public crack segmentation benchmarks, Crack500 and CrackForest. The experimental results prove that our CrackDINO outperforms a set of classic mainstream segmentation networks: U-Net, DeepLabV3+, SegFormer, TransUNet and Swin-Unet. On the Crack500 Dataset, CrackDINO gains an mIoU of 62.61% and an F1-score of 74.16%. On the CrackForest Dataset, the proposed method obtains an mIoU of 60.08% and an F1-score of 74.79%, demonstrating favorable robustness and generalization performance for fine-grained crack segmentation.

Keywords: Pavement crack segmentation, DINOv3, Transformer, Multi-scale feature fusion, Attention mechanism, Deep learning.

1 Introduction

Pavement cracks are among the most common types of distress in road infrastructure, and their timely detection is essential for ensuring traffic safety, ride comfort, and structural durability [12, 24]. Traditional manual inspection methods are labor-intensive and highly dependent on human experience, while conventional image processing approaches often suffer from poor robustness under complex backgrounds and varying illumination conditions [1]. Consequently, automated crack analysis has attracted increasing attention in recent years.

Current pavement crack analysis technologies are usually divided into three mainstream branches: crack category classification, target detection and pixel-level segmentation. Crack detection methods, such as Faster R-CNN [27] and YOLO [26], mainly focus on locating crack regions using bounding boxes. In contrast, crack segmentation provides pixel-level crack localization and enables more accurate measurement of crack morphology, including crack width, length, and continuity. For this reason, pixel-level crack segmentation has evolved into a core research field for automated road surface inspection equipment.

Deep learning crack segmentation schemes fall into two primary branches: convolutional neural network pipelines and Transformer-driven architectures. CNN-based methods exhibit strong capability in local texture extraction and edge representation, while Transformer-based architectures are more effective in modeling long-range contextual dependencies and global semantic relationships [17, 10]. However, existing methods still face challenges in preserving crack continuity, enhancing weak crack responses, and effectively integrating global semantic information with fine-grained local details, especially under low-contrast and complex background conditions.

To tackle the above limitations, this paper develops CrackDINO, a hybrid segmentation architecture built upon pre-trained DINOv3 [31] dedicated to subtle crack segmentation tasks. The proposed framework enhances multi-scale semantic interaction, strengthens weak crack responses, and progressively recovers spatial details to preserve crack continuity and improve segmentation accuracy. Quantitative tests on open-source pavement crack datasets demonstrate that CrackDINO outperforms classic CNN and Transformer segmentation networks, especially when handling narrow, low-visibility crack targets.

The main contributions of this work are summarized as follows:

- We propose CrackDINO, a DINOv3-based hybrid crack segmentation framework that effectively combines global semantic modeling with local structural preservation for fine-grained crack extraction.
- We design an RGB-guided Multi-scale Feature Pyramid (RGMFP) and a Crack-aware Stable Attention (CAS) mechanism to enhance multi-scale semantic interaction and strengthen weak crack feature representation.
- We introduce a Cascaded Hierarchical Multi-scale Decoder (CHMD) to progressively recover spatial details and preserve crack continuity across scales, achieving favorable segmentation performance on public crack benchmarks.

2 Related Work

Numerous research breakthroughs have been achieved for pavement crack segmentation in recent years. Current mainstream algorithms can be split into three categories: CNN-oriented segmentation pipelines, Transformer-driven crack detection architectures, as well as multi-scale fusion and attention-augmented networks. We analyze typical representative works and summarize their inherent drawbacks in the subsequent subsections.

2.1 CNN-based Crack Segmentation

Convolutional Neural Networks (CNNs) are extensively deployed for pavement crack segmentation, as they possess remarkable performance in extracting local texture cues and outlining object edge features. Early semantic segmentation frameworks, such as Fully Convolutional Networks (FCN) [21] and U-Net [28], established the foundation for dense crack prediction by introducing encoder-decoder architectures and skip connections. Subsequently, numerous CNN-based crack segmentation models were proposed, including CrackU-Net [38], DeepCrack [39], and FPHBN [37], which further improved crack localization and boundary recovery.

To enhance contextual modeling capability, the DeepLab series [6] introduced atrous convolution and spatial pyramid pooling to enlarge receptive fields while preserving spatial resolution. Attention-based CNN variants further improved feature representation by emphasizing crack-sensitive regions. For example, DMA-Net [32] incorporated multi-scale attention mechanisms into DeepLab frameworks to improve localization robustness. Pyramid attention fusion strategies were also introduced to improve crack boundary clarity and suppress background interference [33]. In addition, Tiny-Crack-Net [9] utilized multi-scale feature fusion and attention mechanisms to enhance segmentation performance for fine and irregular cracks.

Despite satisfying segmentation results obtained by CNN-based schemes, the local receptive range of convolution kernels restricts the modeling of long-distance crack structural correlations. Consequently, these methods often suffer from discontinuous predictions, blurred crack boundaries, and insufficient representation of thin crack structures, especially under complex backgrounds and low-contrast conditions.

2.2 Transformer-based Crack Segmentation

Recently, Transformer architectures have shown remarkable success in computer vision tasks due to their powerful global dependency modeling capability. Vision Transformer (ViT) first introduced pure Transformer structures into visual representation learning, while subsequent variants such as Swin Transformer [20], SegFormer [36], Mask2Former [8], and MAE [13] further improved hierarchical representation learning, self-supervised representation capability, and dense prediction performance.

Transformer-based architectures have also been explored in crack segmentation tasks. CrackFormer [19] introduced Transformer blocks into crack segmentation frame-

works to better capture long-range crack continuity and contextual semantics. Moreover, large-scale pre-trained visual foundation models, including the DINO series [4, 23] and the Segment Anything family [16, 25], have demonstrated promising generalization ability in dense prediction tasks through large-scale self-supervised pretraining.

Compared with CNN-based methods, Transformers are more effective in modeling global semantic dependencies and large-scale contextual relationships. However, their patch-wise tokenization strategy inevitably weakens local texture perception and fine-grained boundary modeling. As a result, Transformer-based crack segmentation methods may still suffer from missed detections of tiny cracks [40-42], discontinuous crack structures, and insufficient sensitivity to weak crack responses in low-contrast regions.

2.3 Multi-scale Fusion and Attention Mechanisms

To improve feature representation ability, multi-scale fusion and attention mechanisms have been extensively investigated in dense prediction tasks. Feature Pyramid Network (FPN) [18], UPerNet [35], and related hierarchical fusion architectures enhance segmentation performance by aggregating semantic features from different spatial scales, thereby improving small-object perception and boundary refinement capability.

Meanwhile, attention mechanisms have been widely introduced to strengthen discriminative feature learning. Representative methods include Squeeze-and-Excitation (SE) Networks [15], Convolutional Block Attention Module (CBAM) [34], and Coordinate Attention [14], which enhance feature representation through channel-wise and spatial-wise feature recalibration. These mechanisms have confirmed effectiveness in improving semantic perception and suppressing background interference.

Despite their success, existing attention mechanisms generally adopt simultaneous enhancement and suppression strategies, which may inadvertently weaken low-response crack structures. Since thin cracks often exhibit weak texture responses and low contrast, excessive feature suppression can easily lead to crack discontinuity and missed detections. In addition, many existing multi-scale fusion strategies still lack sufficient interaction between local texture features and high-level semantic representations.

3 Methodology

This part begins with an overview of the complete structural layout of our CrackDINO framework, which is specifically designed for fine-grained road crack segmentation. Subsequently, the RGB-guided Multi-scale Feature Pyramid (RGMFP), Crack-aware Stable Attention (CAS), and Cascaded Hierarchical Multi-scale Decoder (CHMD) modules are presented in detail.

3.1 Overview of CrackDINO

The overall architecture of the proposed CrackDINO framework is illustrated in Fig. 1. Although DINOv3 provides strong global semantic representations

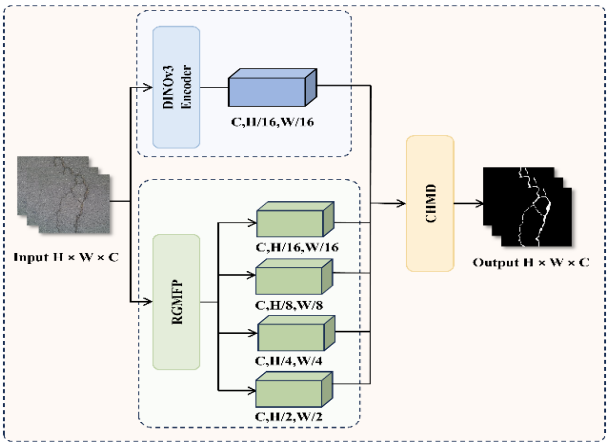

Fig. 1. Overview of the proposed CrackDINO framework.

through self-supervised vision Transformers, its patch-token representations are inherently coarse and often fail to preserve fine-grained crack continuity and boundary information. To overcome such drawbacks, we propose CrackDINO, a DINOv3-based crack segmentation framework that integrates an RGB-guided Multi-scale Feature Pyramid (RGMFP), Crack-aware Stable Attention (CAS), and a Cascaded Hierarchical Multi-scale Decoder (CHMD) for robust multi-scale crack representation learning.

Concretely, we feed input road images into the DINOv3 backbone to extract multi-layer semantic token embeddings. Meanwhile, the proposed RGMFP module constructs hierarchical RGB spatial priors from shallow convolutional representations to compensate for the local structural information lost during Transformer tokenization. To further enhance subtle crack responses while avoiding excessive suppression of weak crack regions, the CAS module is embedded into the RGB pyramid branches to adaptively emphasize crack-sensitive spatial and channel features. Subsequently, the extracted DINOv3 semantic features and RGB-guided multi-scale features are progressively fused by the proposed CHMD module through hierarchical cross-modal decoding and multi-stage feature aggregation. Finally, the fused representations are decoded into dense crack prediction maps for accurate crack segmentation.

3.2 RGB-guided Multi-scale Feature Pyramid

Although Transformer-based encoders provide strong global semantic representations, their patch-wise tokenization weakens local texture perception and boundary sensitivity for thin crack structures. To mitigate this defect, we design a compact RGB-guided Multi-scale Feature Pyramid (RGMFP), as visualized in Figure 2. The proposed module progressively extracts hierarchical RGB structural representations and enhances multi-scale spatial feature interaction for fine-grained crack modeling. Specifically, the proposed RGMFP constructs a four-stage convolutional pyramid to progressively generate four hierarchical feature representations, namely R_2, R_4, R_8, and R_{16}, whose spatial

resolutions correspond to 1/2, 1/4, 1/8, and 1/16 of the input image, respectively. The proposed Crack-aware Stable Attention (CAS) module is embedded after the R_4 and R_8 stages to further enhance weak crack responses and preserve crack continuity. For the initial pyramid branch, the input image is processed by stacked 3×3 convolutional layers, where the first convolution adopts stride 2 for spatial downsampling while preserving shallow edge structures. The subsequent convolutions further refine local texture and boundary representations. Group Normalization (GN) and GELU activation are employed after each convolution to improve feature stability and nonlinear representation capability. Subsequently, the following pyramid stages further enlarge the receptive field and progressively capture richer contextual semantics while preserving crack-sensitive structural information.

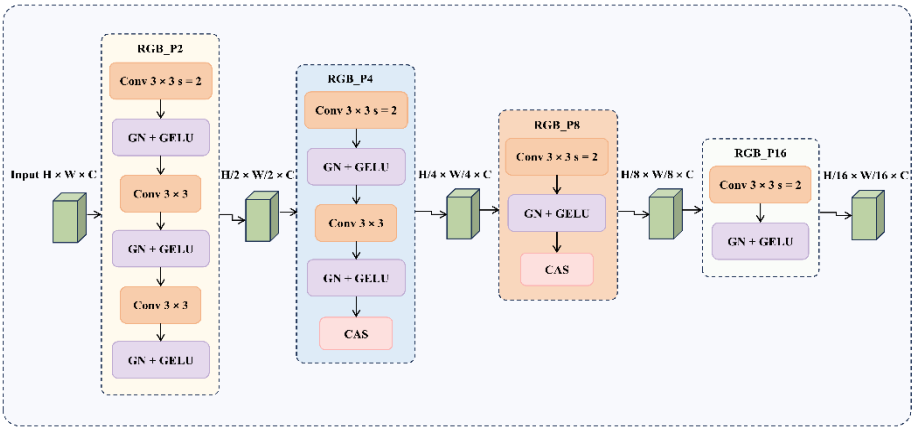

Fig. 2. Architecture of the proposed RGMFP module.

As illustrated in Fig. 3, the proposed CAS module first applies global average pooling and global max pooling to generate complementary channel descriptors. These pooled feature maps are concatenated together and fed into lightweight convolution blocks to compute channel attention coefficients. Meanwhile, average pooling and max pooling are utilized to derive spatial attention cues, which are then refined via spatial convolution operations. Different from conventional multiplicative attention mechanisms, CAS adopts a boost-only modulation strategy to avoid weakening thin crack structures. The refined feature representation is formulated as

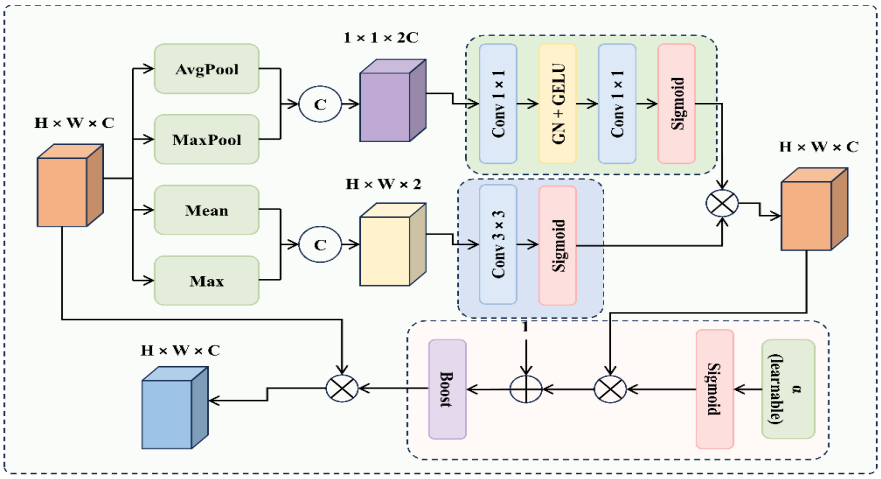

Fig. 3. Architecture of the proposed CAS module. C denotes channel concatenation, ⊗ denotes element-wise multiplication, ⊕ denotes element-wise addition.

$$F_{out} = F \odot (1 + \alpha A) \tag{1}$$

where *Fout* denotes the refined feature representation, A denotes the joint attention map generated by CAS, α is a learnable modulation coefficient, and ⊙ represents element-wise multiplication. To avoid excessive response amplification, the attention gain is constrained within a predefined upper bound, thereby enabling stable enhancement of crack-sensitive regions while preserving the original structural information. Notably, the deepest feature representation R_{16} is spatially aligned with the DINOv3 patch token representation at 1/16 resolution, facilitating subsequent cross-modal hierarchical fusion in the decoder. By integrating shallow RGB structural features with deep Transformer semantic representations, the proposed RGMFP effectively improves crack continuity, boundary completeness, and fine-grained crack representation capability.

3.3 Cascaded Hierarchical Multi-scale Decoder

Although Transformer-based encoders provide powerful global semantic representations, their low-resolution token features often fail to preserve fine-grained crack boundaries and structural continuity. In addition, conventional lightweight decoders that rely on direct upsampling are insufficient for recovering hierarchical spatial details, especially for thin and low-contrast crack regions. To address these limitations, we propose a Cascaded Hierarchical Multi-scale Decoder (CHMD), as illustrated in Fig. 4. The proposed decoder progressively fuses DINOv3 semantic representations with RGB-guided multi-scale features extracted by RGMFP, enabling hierarchical reconstruction of crack structures from coarse semantics to fine-grained boundaries.

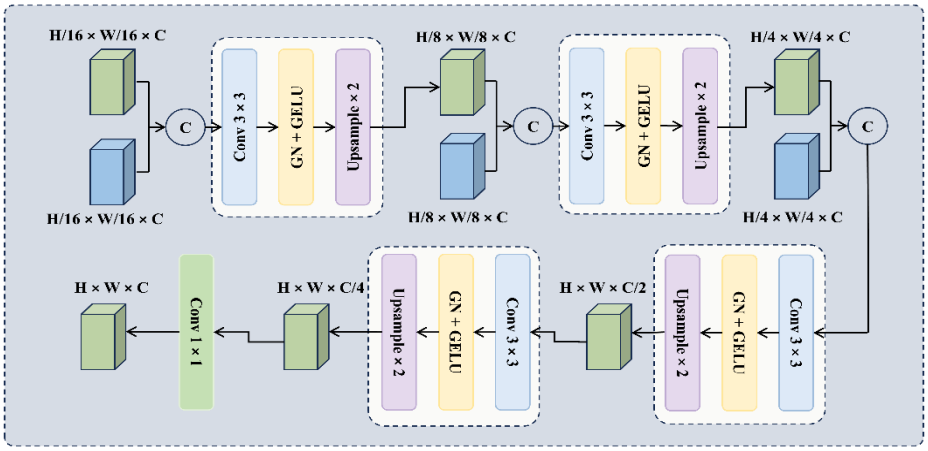

Fig. 4. Architecture of the proposed CHMD module. C denotes channel concatenation.

Specifically, the DINOv3 feature is first fused with the deepest RGB-guided feature through channel concatenation and convolutional refinement to generate an initial semantic representation. Subsequently, the decoder performs multi-scale feature refinement through bilinear upsampling and feature fusion with R_8 and R_4. Each decoding stage contains convolution, Group Normalization (GN), and GELU activation to enhance feature stability and structural consistency. Compared with conventional lightweight decoders that directly perform aggressive semantic upsampling, the proposed CHMD adopts a progressive hierarchical reconstruction strategy, which gradually injects RGB structural priors during the decoding process. Such a design effectively compensates for the insufficient local detail perception of Transformer features while preserving global semantic consistency.

Finally, the refined feature representation is progressively upsampled to recover high-resolution spatial details, and the final segmentation prediction map is generated through a 1×1 convolution layer. By integrating deep semantic representations and shallow structural priors across multiple resolutions, the proposed CHMD effectively improves crack boundary integrity, thin crack sensitivity, and topology continuity under complex background conditions.

3.4 Loss Function

Road crack segmentation usually suffers from severe foreground-background imbalance, low contrast, and fragmented thin crack structures. Following previous Transformer-based segmentation frameworks such as TransUNet [5], we employ a hybrid loss function composed of BCE Loss, Dice Loss, and Tversky Loss to improve crack continuity and enhance sensitivity to weak crack regions.

The overall segmentation loss is formulated as:

$$L_{seg} = \lambda_1 L_{BCE} + \lambda_2 L_{Dice} + \lambda_3 L_{Tversky}, \tag{2}$$

where L_{BCE} provides pixel-level supervision, L_{Dice} improves regional overlap consistency, and $L_{Tversky}$ further suppresses false-negative predictions for thin crack structures.

Specifically, the BCE Loss is defined as:

$$L_{BCE} = -\left(\frac{1}{N}\right) * \sum_{i=1}^{N}\left[w_p * y_i * \log(p_i) + (1 - y_i) * \log(1 - p_i)\right], \quad (3)$$

where p_i and y_i denote the predicted probability and ground-truth label of the i-th pixel, respectively, N represents the total number of pixels, and w_p is the dynamically adjusted positive sample weight.

The Dice Loss is formulated as:

$$L_{Dice} = 1 - \frac{2\sum_i p_i y_i + \varepsilon}{\sum_i p_i + \sum_i y_i + \varepsilon}, \quad (4)$$

which enhances structural consistency and crack continuity.

The Tversky Loss [29] is defined as:

$$L_{Tversky} = 1 - \frac{TP + \varepsilon}{TP + \alpha FP + \beta FN + \varepsilon}, \quad (5)$$

where α and β are adopted to adjust the trade-off between prediction precision and recall rate.

In our implementation, the hybrid loss weights are empirically set as $\lambda_1 = 0.4$, $\lambda_2 = 0.3$, and $\lambda_3 = 0.3$ to balance pixel-level supervision and structural consistency. For the Tversky Loss, we set $\alpha = 0.3$ and $\beta = 0.7$ to place greater emphasis on reducing false-negative predictions for thin crack structures. Meanwhile, the positive sample weight w_p is dynamically adjusted according to the foreground-background distribution during training, which further improves the segmentation performance for sparse and low-contrast crack regions.

4 Experiments

4.1 Datasets

In this study, we evaluate our method on two widely used crack segmentation datasets: CrackForest Dataset (CFD) [30] and Crack500 [37], to comprehensively assess its effectiveness.

CrackForest Dataset. The CrackForest Dataset (CFD) is a classical small-scale crack detection dataset, consisting of 118 images with a resolution of 480×320 and pixel-level annotations. It mainly contains thin and elongated cracks with irregular widths, often accompanied by background disturbances such as shadows, stains, and texture noise. Such complex scene features bring great difficulties for the model to extract subtle crack structural details accurately. We partition all samples of this dataset into training, validation and test subsets following the 6:2:2 proportional split rule.

Crack500 Dataset. Crack500 is a larger-scale dataset that originally contains 500 high-resolution images collected from diverse road scenes. Compared to CFD,

Crack500 exhibits more complex crack patterns, including network-like cracks, branching structures, and cracks with varying widths. In addition, the background is more challenging due to illumination variations, occlusions, and noise interference. To increase the number of training samples and improve model generalization, each image is divided into 16 smaller patches, resulting in patches with a resolution of 640 × 320. To reduce annotation ambiguity and stabilize optimization, patches containing extremely sparse or missing crack pixels are filtered out, yielding a total of 3303 valid samples. The filtered sample pool is then split into training, validation and test subsets following the 6:2:2 proportional partition rule.

The two benchmarks possess obvious disparities in data volume, crack geometric shapes and background noise levels, which supports all-round validation of our network on subtle crack extraction tasks and complicated actual road scenes.

4.2 Experimental Details

This subsection elaborates full experimental configurations, including experimental hardware environment, data preprocessing pipeline, training hyperparameters, evaluation criteria and comparison baseline networks.

Platform. All experiments were conducted on a server equipped with NVIDIA Tesla V100 GPUs (16GB memory). The models were implemented using Python 3.10, with PyTorch 2.5.0 and CUDA 13.0 support. We utilized popular deep learning libraries including torchvision, timm, and transformers for model implementation and evaluation.

Data Preprocessing. Following standard image classification practices, we normalize input images using the ImageNet mean and standard deviation ($\mu = [0.485, 0.456, 0.406], \sigma = [0.229, 0.224, 0.225]$). To preserve the thin crack geometry, we avoid aggressive resizing and instead pad the images to the nearest multiple of the DINO patch size (16×16) using reflection padding. For data augmentation, we apply horizontal/vertical flipping, random rotation, and color jittering. Before feeding the images into the transformer backbone, we scale the pixel values to the [0, 1] range and perform the aforementioned normalization.

Hyperparameters. We train the model for 40 epochs with an initial learning rate of 4×10^{-5} using the AdamW optimizer [22]. The weight decay is set to 1×10^{-4}, with AMSGrad enabled. Due to the high resolution of the input images, a batch size of 1 per GPU is used. The learning rate scheduler follows a cosine annealing strategy, with a 3-epoch linear warmup, decaying to a minimum value of 4×10^{-7}. Gradient clipping with a norm of 2.0 is applied to stabilize the training process. The segmentation head has a hidden dimension of 448, and dropout is applied with a rate of 0.22.

Evaluation Metrics. We adopt mean Intersection over Union (mIoU), Precision, Recall, F1-score, and Pixel Accuracy as the primary segmentation metrics. The formulas are defined as follows:

$$mIoU = \frac{TP}{TP+FP+FN} \tag{6}$$

$$Precision = \frac{TP}{TP+FP} \tag{7}$$

$$Recall = \frac{TP}{TP+FN} \tag{8}$$

$$F1 - score = \frac{2 \cdot Precision \cdot Recall}{Precision+Recall} \tag{9}$$

$$Pixel\ Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{10}$$

In this formula, TP, FP, FN and TN separately stand for true positive samples, false positive samples, false negative samples and true negative samples.

4.3 Main Results

Table 1 and Table 2 present the quantitative comparison results on the CFD and Crack500 Dataset, respectively.

To comprehensively evaluate the proposed framework, we compare CrackDINO with several representative CNN-based and Transformer-based segmentation methods, including U-Net [28], DeepLabV3+ [7], SegFormer [36], TransUNet [5], and Swin-Unet [2]. Since these methods employ different backbone architectures and implementation settings, the comparison mainly reflects the overall segmentation capability of different frameworks under standard benchmark protocols.

In addition, to further validate the effectiveness of the proposed decoder design, we construct a DINOv3+UNet baseline using the same DINOv3 backbone and training configuration as our method. This controlled comparison allows us to fairly evaluate the contribution of the proposed RGB-guided Multi-scale Feature Pyramid (RGMFP), Crack-aware Stable Attention (CAS), and Cascaded Hierarchical Multi-scale Decoder (CHMD).

CrackForest Dataset. As shown in Table 1, CrackDINO achieves the best overall segmentation performance on the CrackForest Dataset with an mIoU of 0.6008 and an F1-score of 0.7479. Compared with the DINOv3+UNet baseline, the proposed framework reveals clear improvements in crack continuity preservation and fine-grained crack representation capability. In particular, CrackDINO consistently outperforms conventional CNN-based methods such as U-Net and DeepLabV3+, while also surpassing Transformer-based approaches including SegFormer, TransUNet, and Swin-Unet. These results indicate that the proposed multi-scale RGB feature enhancement and hierarchical decoding strategy are effective for capturing thin and fragmented crack structures.

Crack500 Dataset. Table 2 further demonstrates the superior capacity of our designed network on more challenging large-scale crack scenarios. CrackDINO achieves the best performance with an mIoU of 0.6261 and an F1-score of 0.7564. Compared with the DINOv3+UNet baseline (mIoU 0.5783), the proposed framework achieves an 8.3% relative improvement in mIoU, demonstrating the effectiveness of the proposed RGMFP, CAS, and CHMD modules. Moreover, CrackDINO consistently achieves superior segmentation accuracy and crack continuity compared with existing CNN-based and Transformer-based methods, especially in complex backgrounds and low-contrast crack regions.

Table 1. Comparison results on the CrackForest Dataset. Results of TransUNet and Swin-Unet are adopted from [11].

Method	mIoU	Prec	Recall	Acc	F1
U-Net [28]	0.3958	0.4221	0.8224	0.9775	0.5639
DeepLabV3+ [7]	0.4108	0.4469	0.8336	0.9796	0.5797
SegFormer [36]	0.4957	0.5662	0.8016	0.9862	0.6603
TransUNet [5]	0.5243	0.6639	0.7122	0.9484	0.6791
Swin-Unet [2]	0.4690	0.6344	0.6574	0.9416	0.6297
DINOv3+UNet	0.4180	0.5036	0.7426	0.9833	0.5896
CrackDINO (Ours)	**0.6008**	**0.7233**	**0.8027**	**0.9912**	**0.7479**

Table 2. Comparison results on the Crack500 Dataset. Results of TransUNet and Swin-Unet are adopted from [11].

Method	mIoU	Prec	Recall	Acc	F1
U-Net [28]	0.5091	0.5899	0.7852	0.9631	0.6543
DeepLabV3+ [7]	0.4885	0.5634	0.7891	0.9598	0.6388
SegFormer [36]	0.5440	0.6308	0.7962	0.9636	0.6866
TransUNet [5]	0.5035	0.7014	0.6587	0.9577	0.6520
Swin-Unet [2]	0.4971	0.6915	0.6519	0.9529	0.6426
DINOv3+UNet	0.5783	0.6991	0.7663	0.9720	0.7106
CrackDINO (Ours)	**0.6261**	**0.7464**	**0.7982**	**0.9749**	**0.7564**

4.4 Results Analysis

In order to further explore the effectiveness of our proposed CrackDINO in handling diverse crack scenarios, we conduct an in-depth qualitative analysis of the segmentation results on two benchmark datasets, focusing on visual comparisons across different crack types including linear cracks, reticular branching cracks, and low-contrast fine cracks.

1) CrackForest Dataset Visualization:

To verify the superiority of our enhanced decoder in preserving fine crack structures, we visualize the segmentation results on the CrackForest Dataset (CFD). As shown in Fig. 5, for reticular branching cracks (top row), CrackDINO successfully preserves the complete reticular topology with sharp and continuous predictions, while comparison methods exhibit fractures or blurring.

For low-contrast fine cracks (middle row), the proposed method demonstrates clearer advantages. The blue box in GT marks a subtle secondary branch that is extremely difficult to detect due to low contrast against the complex background. Only CrackDINO is able to preserve this subtle branch structure, while the baseline DINOv3+UNet, UNet, DeepLabV3+, and SegFormer all fail to preserve the subtle branch structure, resulting in broken or incomplete predictions.

For linear cracks with weak boundaries (bottom row), most methods exhibit discontinuity. UNet produces fragmented results with gaps, while DeepLabV3+ tends to unnaturally thicken the crack line. CrackDINO produces more continuous and structurally consistent boundaries.

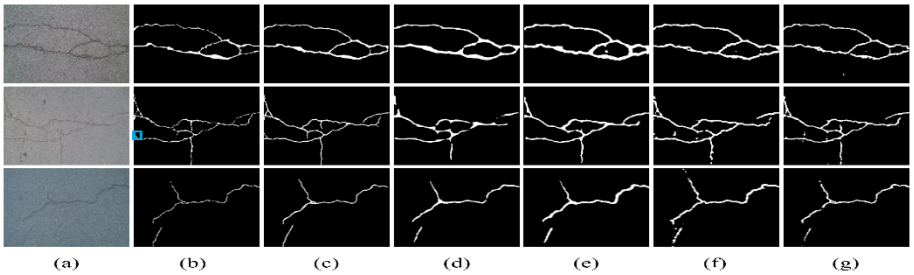

Fig. 5. Qualitative comparison on CrackForest Dataset: (a) original image, (b) GT, (c) CrackDINO (ours), (d) DINOv3+UNet (baseline), (e) UNet, (f) DeepLabV3+, (g) SegFormer. The blue box indicates a subtle secondary branch that is missed by all methods except CrackDINO.

2) Crack500 Dataset Visualization:

Fig. 6 presents the qualitative results on the Crack500 Dataset with higher-resolution images and more complex asphalt textures. For thick main cracks (top row), UNet and DeepLabV3+ exhibit fractures, while SegFormer is disturbed by noise. CrackDINO generates the most precise and continuous crack profiles.

For multi-scale mixed cracks with complex junctions (middle row), UNet and DeepLabV3+ produce overly thick predictions with blurred boundaries at intersections. The baseline DINOv3+UNet captures main structures but lacks precision at bifurcation points. CrackDINO accurately captures both thick main cracks and thin secondary branches with sharp junctions.

For extremely thin cracks (bottom row), the superiority of CrackDINO becomes more noticeable. DeepLabV3+ misclassifies shadows as cracks (red box) and produces overly thick predictions, while UNet produces fragmented results. Only CrackDINO successfully detects the complete crack line with consistent thickness and no false positives, validating the effectiveness of our method in maintaining high recall for fine structures while controlling false positives.

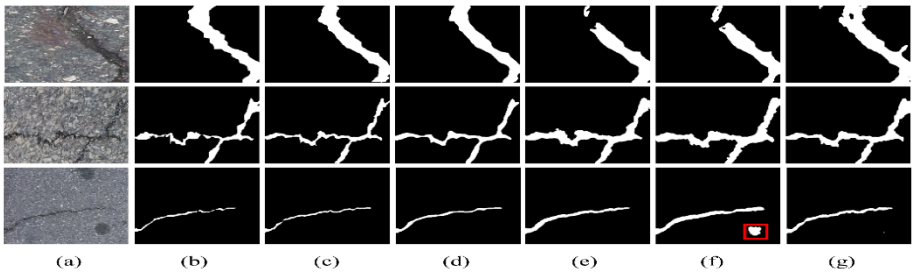

Fig. 6. Qualitative comparison on Crack500 Dataset: (a) original image, (b) GT, (c) CrackDINO (ours), (d) DINOv3+UNet (baseline), (e) UNet, (f) DeepLabV3+, (g) SegFormer. The red box indicates false positives from shadow misclassification by DeepLabV3+.

3) Comparison with SAM3:

We further compare CrackDINO with SAM3 [3], as shown in Fig. 7. Although SAM3 demonstrates strong general segmentation capability, it still struggles to preserve fine crack structures and accurate crack topology in complex pavement scenarios.

Compared with the baseline DINOv3+UNet, SAM3 misses more subtle crack branches and produces noticeable false positives near crack boundaries. The baseline model successfully detects one subtle branch but still fails to preserve the complete crack topology. In contrast, CrackDINO achieves more complete crack reconstruction and maintains finer structural continuity, demonstrating superior capability in handling thin and low-contrast crack regions.

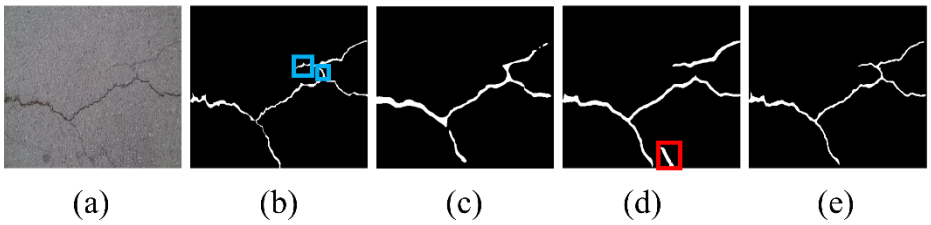

Fig. 7. Qualitative comparison between the baseline DINOv3+UNet, SAM3, and the proposed CrackDINO: (a) original image, (b) GT, (c) DINOv3+UNet (baseline), (d) SAM3, (e) CrackDINO (ours). The blue boxes indicate subtle crack branches missed by different methods, while the red box indicates a false positive produced by SAM3.

4.5 Ablation Experiments

As shown in Table 3 and Fig. 8, we conduct ablation experiments on the CrackForest Dataset to evaluate the effectiveness of the proposed CHMD, RGMFP, and CAS modules. The baseline model is constructed using the DINOv3 backbone with a conventional U-Net decoder head. Based on the baseline framework, the proposed modules are progressively introduced for evaluation.

1) Effect of CHMD: We first introduce the proposed CHMD decoder into the baseline model to enhance hierarchical feature recovery and spatial detail reconstruction. Compared with the baseline, the F1-score improves from 0.5896 to 0.6822, while the mIoU increases from 0.4180 to 0.5211. These improvements show that the cascaded multi-scale decoding strategy effectively preserves crack continuity and enhances the segmentation capability for thin crack structures.

2) Effect of RGMFP: Based on the CHMD decoder, we further incorporate the RGMFP module to strengthen multi-scale semantic interaction across different feature levels. As shown in Table 3, the F1-score further improves from 0.6822 to 0.7316, and the mIoU increases from 0.5211 to 0.5881. The results suggest that the proposed feature pyramid structure effectively improves fine-grained crack representation and semantic fusion ability.

3) Effect of CAS: Finally, the CAS module is introduced to enhance weak crack responses and suppress background interference. Compared with the model without

CAS, the final model achieves further improvements across all evaluation metrics, reaching 0.6008 mIoU and 0.7479 F1-score. In particular, the precision increases from 0.7088 to 0.7233, verifying that the CAS module effectively improves the discriminative capability for tiny and low-contrast crack regions.

Overall, the experimental results verify that each proposed component contributes positively to crack segmentation performance. As illustrated in Fig. 8, the mIoU consistently increases with the progressive introduction of CHMD, RGMFP, and CAS, demonstrating the effectiveness and complementarity of the proposed modules.

Table 3 Ablation study on the CrackForest Dataset.

Models	mIoU	Prec	Recall	Acc	F1
Baseline	0.4180	0.5036	0.7426	0.9833	0.5896
+ CHMD	0.5211	0.6371	0.7566	0.9844	0.6822
+ CHMD + RGMFP	0.5881	0.7088	0.8001	0.9908	0.7316
+ Ours	**0.6008**	**0.7233**	**0.8027**	**0.9912**	**0.7479**

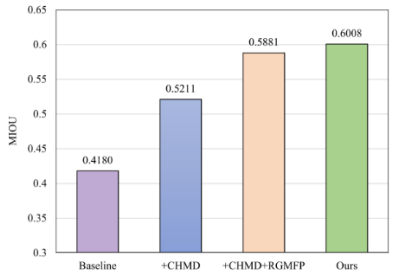

Fig. 8 Ablation study results on the CrackForest Dataset.

5 Conclusion

This work designs CrackDINO, a hybrid segmentation architecture built upon DINOv3 to realize precise pixel-level extraction of subtle pavement cracks. The proposed method is designed to address the challenges of detecting thin, low-contrast, and topologically complex crack structures. By combining the strong global semantic modeling capability of DINOv3 with enhanced multi-scale feature fusion and detail-aware decoding strategies, the proposed framework effectively improves crack continuity and segmentation accuracy. Specifically, we introduce the RGB-guided Multi-scale Feature Pyramid (RGMFP) to enhance semantic interaction across different feature levels, Crack-aware Stable Attention (CAS) module to strengthen weak crack responses and suppress background interference, and the Cascaded Hierarchical Multi-scale Decoder (CHMD) to progressively recover spatial details and preserve crack structures during decoding. Extensive experiments on public crack datasets demonstrate that the proposed method achieves favorable performance compared with existing segmentation approaches. In particular, the proposed framework shows strong robustness in detecting tiny and low-contrast crack regions. Although the proposed method achieves promising

results, there are still several limitations. The current framework mainly focuses on segmentation accuracy and does not fully consider computational efficiency and real-time deployment. For follow-up research, we intend to investigate compact lightweight network structures, optimized cross-scale feature fusion modules, and stronger cross-scene generalization to fit real intelligent road inspection systems.

References

1. Li, T., Wang, G., Yu, B., Liu, Y., & Sun, W. Don't let the information slip away. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 8504-8513).2026.
2. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M. Swin-unet: Unet-like pure transformer for medical image segmentation. In: Proceedings of the European Conference on Computer Vision Workshops (ECCVW) (2022)
3. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K. V., Khedr, H., Huang, A., Lei, J., Ma, T., Guo, B., Kalla, A., Marks, M., Greer, J., Wang, M., Sun, P., Rädle, R., Afouras, T., Mavroudi, E., Xu, K., Wu, T.H., Zhou, Y., Momeni, L., Hazra, R., Ding, S., Vaze, S., Porcher, F., Li, F., Li, S., Kamath, A., Cheng, H.K., Dollár, P., Ravi, N., Saenko, K., Zhang, P., Feichtenhofer, C. Sam 3: Segment anything with concepts. arXiv preprint arXiv:2511.16719 (2025), https://arxiv.org/abs/2511.16719
4. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A. Emerging properties in self-supervised vision transformers. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 9630–9640 (2021). https://doi.org/10.1109/ICCV48922.2021.00951
5. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y. Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
6. Chen, L.C., Papandreou, G., Schroff, F., Adam, H. Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017)
7. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 801–818 (2018)
8. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R. Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
9. Chu, H., Wang, W., Deng, L. Tiny-crack-net: A multiscale feature fusion network with attention mechanisms for segmentation of tiny cracks. Computer-Aided Civil and Infrastructure Engineering, 37(14): 1914–1931 (2022). https://doi.org/10.1111/mice.12881
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16×16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2020), https://openreview.net/forum?id=YicbFdNTTy
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, 24(22): 7134 (2024). https://doi.org/10.3390/s24227134
12. Gong, H., Liu, L., Liang, H., Zhou, Y., Cong, L. A state-of-the-art survey of deep learning models for automated pavement crack segmentation. International Journal of Transportation Science and Technology, 13(c): 44–57 (2024)

13. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R. Masked autoencoders are scalable vision learners. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 15979–15988 (2022). https://doi.org/10.1109/CVPR52688.2022.01553
14. Hou, Q., Zhou, D., Feng, J. Coordinate attention for efficient mobile network design. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 13708–13717 (2021). https://doi.org/10.1109/CVPR46437.2021.01350
15. Hu, J., Shen, L., Sun, G. Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7132–7141 (2018). https://doi.org/10.1109/CVPR.2018.00745
16. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R. Segment anything. arXiv preprint arXiv:2304.02643 (2023)
17. Lau, S.L.H., Chong, E.K.P., Yang, X., Wang, X. Automated pavement crack segmentation using u-net-based convolutional neural network. IEEE Access, 8: 114892–114899 (2020). https://doi.org/10.1109/ACCESS.2020.3003638
18. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S. Feature pyramid networks for object detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 936–944 (2017). https://doi.org/10.1109/CVPR.2017.106
19. Liu, H., Miao, X., Mertz, C., Xu, C., Kong, H. Crackformer: Transformer network for fine-grained crack detection. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 3763–3772 (2021). https://doi.org/10.1109/ICCV48922.2021.00376
20. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 9992–10002 (2021). https://doi.org/10.1109/ICCV48922.2021.00984
21. 21. Long, J., Shelhamer, E., Darrell, T. Fully convolutional networks for semantic segmentation. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3431–3440 (2015). https://doi.org/10.1109/CVPR.2015.7298965
22. Loshchilov, I., Hutter, F. Fixing weight decay regularization in adam. arXiv preprint arXiv:1711.05101 (2017)
23. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Howes, R., Huang, P.Y., Xu, H., Sharma, V., Li, S.W., Galuba, W., Rabbat, M., Assran, M., Ballas, N., Synnaeve, G., Misra, I., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P. Dinov2: Learning robust visual features without supervision (2023)
24. Rathee, M., Bačić, B., Doborjeh, M. Automated road defect and anomaly detection for traffic safety: A systematic review. Sensors, 23(12): 5656 (2023). https://doi.org/10.3390/s23125656
25. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K.V., Carion, N., Wu, C.Y., Girshick, R., Dollár, P., Feichtenhofer, C. Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024), https://arxiv.org/abs/2408.00714
26. Redmon, J., Divvala, S., Girshick, R., Farhadi, A. You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779–788 (2016). https://doi.org/10.1109/CVPR.2016.91
27. Ren, S., He, K., Girshick, R., Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(6): 1137–1149 (2017). https://doi.org/10.1109/TPAMI.2016.2577031

28. Ronneberger, O., Fischer, P., Brox, T. U-net: Convolutional networks for biomedical imag e segmentation. In: Medical Image Computing and Computer-Assisted Intervention – MIC CAI 2015, pp. 234–241 (2015). https://doi.org/10.1007/978-3-319-24574-4_28
29. Salehi, S.S.M., Erdogmus, D., Gholipour, A. Tversky loss function for image segmentation using 3d fully convolutional deep networks. In: MLMI Workshop, MICCAI, pp. 379–387 (2017)
30. Shi, Y., Cui, L., Qi, Z., Meng, F., Chen, Z. Automatic road crack detection using random s tructured forests. IEEE Transactions on Intelligent Transportation Systems, 17(12): 3434–3445 (2016)
31. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Sza franiec, M., Yi, S., Ramamonjisoa, M., Massa, F., Haziza, D., Wehrstedt, L., Wang, J., Dar cet, T., Moutakanni, T., Sentana, L., Roberts, C., Vedaldi, A., Tolan, J., Brandt, J., Coupri e, C., Mairal, J., Jégou, H., Labatut, P., Bojanowski, P. Dinov3 (2025), https://arxiv.org/ab s/2508.10104
32. Sun, X., Xie, Y., Jiang, L., Cao, Y., Liu, B. Dma-net: Deeplab with multi-scale attention fo r pavement crack segmentation. IEEE Transactions on Intelligent Transportation Systems, 23(10): 18392–18403 (2022). https://doi.org/10.1109/TITS.2022.3158670
33. Wang, W., Su, C. Convolutional neural network-based pavement crack segmentation using pyramid attention network. IEEE Access, 8: 206548–206558 (2020). https://doi.org/10.11 09/ACCESS.2020.3037667
34. Woo, S., Park, J., Lee, J.Y., Kweon, I.S. Cbam: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19 (2018)
35. Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J. Unified perceptual parsing for scene understa nding. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 418–434 (2018)
36. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P. Segformer: Simple an d efficient design for semantic segmentation with transformers. Advances in Neural Infor mation Processing Systems (NeurIPS), 34: 12077–12090 (2021)
37. Yang, F., Zhang, L., Yu, S., Prokhorov, D., Mei, X., Ling, H. Feature pyramid and hierarc hical boosting network for pavement crack detection. IEEE Transactions on Intelligent Tra nsportation Systems (2019)
38. Zhang, L., Yang, F., Zhang, Y.D., Zhu, Y.J. Road crack detection using deep convolutiona l neural network. In: 2016 IEEE International Conference on Image Processing (ICIP), pp. 3708–3712 (2016). https://doi.org/10.1109/ICIP.2016.7533052
39. Zou, Q., Zhang, Z., Li, Q., Qi, X., Wang, Q., Wang, S. Deepcrack: Learning hierarchical c onvolutional features for crack detection. IEEE Transactions on Image Processing, 28(3): 1 498–1512 (2019). https://doi.org/10.1109/TIP.2018.2878966
40. Zhao, F., Xu, D., Ren, Z., Shao, X., Wu, Q., Liu, Y., ... & Mizuno, K. Mamba-based super -resolution and semi-supervised YOLOv10 for freshwater mussel detection using acoustic video camera: A case study at Lake Izunuma, Japan. Ecological Informatics, 103324. 202 5.
41. Zhao, F., Chen, Y., Xi, D., Liu, Y., Wang, J., Tabeta, S., & Mizuno, K. (2025). Enhanced hermit crabs detection using super-resolution reconstruction and improved YOLOv8 on U AV-captured imagery. Marine Environmental Research, 210, 107313. 2025.
42. Zhao, F., Ren, Z., Wang, J., Wu, Q., Xi, D., Shao, X., ... & Mizuno, K. Smart UAV-assiste d rose growth monitoring with improved YOLOv10 and Mamba restoration techniques. S mart Agricultural Technology, 10, 100730. 2025.