

# Improvement in Road Crack Detection Based on Multiple Attention Mechanisms

Junqing Wang<sup>1</sup>, Qi Li<sup>2</sup> and Lin Meng<sup>2,\*</sup>

<sup>1</sup>Graduate School of Science and Engineering, Ritsumeikan University, 1-1-1 Noji-higashi, Kusatsu, Shiga, 525-8577, Japan

<sup>2</sup>College of Science and Engineering, Ritsumeikan University, 1-1-1 Noji-higashi, Kusatsu, Shiga, 525-8577, Japan

## Abstract

With the rapid expansion of global road networks, pavement maintenance is increasingly challenged by fine and irregular cracks caused by traffic loads and environmental conditions. Traditional inspection methods, including manual patrols and classical image processing, are often inefficient and lack sensitivity to subtle crack patterns. To address these limitations, we propose a novel road crack recognition framework that integrates object detection with semantic segmentation. The detection module enhances YOLOv11 by incorporating a Diverse Branch Block and a Triplet Attention Module to improve multi-scale feature extraction with low computational cost. The segmentation module extends TransUNet by replacing the standard Transformer Encoder with BiFormer Block and embedding a parallel Swin Transformer Block, enabling effective global-local context fusion. Experimental results demonstrate that the improved detection model achieves 83.9% mAP@50 and 65.8% mAP@[50:95] on the Ultralytics CrackSeg dataset. Meanwhile, the enhanced segmentation model attains 78.46% mean Intersection-over-Union on the CRACK500 dataset. These findings confirm the effectiveness of the proposed multi-attention architecture for accurate and scalable road crack analysis.

## Keywords

Road Crack Detection, YOLOv11, TransUNet, Deep learning

## 1. Introduction

The rapid expansion and aging of road infrastructure have intensified the demand for accurate and scalable pavement crack detection methods to ensure timely maintenance and structural safety. However, cracks are often fine, irregular, and low-contrast, making them difficult to detect with traditional manual or rule-based methods.

Deep learning-based object detection models [1, 2, 3], such as YOLOv8 [4] and hybrid CNN-Transformer frameworks [5], have been applied to road crack detection with encouraging results. However, these models frequently exhibit limited precision in detecting small, fragmented, or low-contrast cracks, particularly under complex surface textures or non-uniform lighting conditions. On the other hand, semantic segmentation models like DeepLabv3+ with attention [6] and domain-adaptive approaches [7] can achieve pixel-level delineation but still

*The 7th International Symposium on Advanced Technologies and Applications in the Internet of Things (ATAIT 2025), August 10-11, 2025, Kusatsu, Shiga, Japan*

\*Corresponding author.

✉ gr0662pf@ed.ritsumei.ac.jp (J. Wang); liqi24@fc.ritsumei.ac.jp (Q. Li); menglin@fc.ritsumei.ac.jp (L. Meng)

ORCID 0000-0002-1963-5263 (Q. Li); 0000-0003-4351-6923 (L. Meng)

© 2025 Copyright for this paper by its authors.

Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

face challenges in segmenting thin cracks accurately and consistently, especially in real-world conditions.

To address these precision limitations from both tasks, we propose two improved models for road crack recognition: an enhanced object detector based on YOLOv11 incorporating a Diverse Branch Block and Triplet Attention Module, and a segmentation network that upgrades TransUNet by replacing its Transformer Encoder with BiFormer Block and adding a parallel Swin Transformer Block.

Overall, our main contributions can be summarized below:

- We embed the Diverse Branch Block into the YOLOv11 backbone and insert the Triplet Attention Module in the neck to enrich multi-path features and joint spatial-channel attention.
- We replace the Transformer layer of TransUNet Encoder with BiFormer Block and add a parallel Swin Transformer Block path to create a dual-encoder that captures complementary global-local context.

The remainder of this article is organized as follows. Section 2 reviews the previous road crack detection and the related YOLO models and semantic segmentation models. Details of our proposed methods are introduced in Section 3. Section 4 presents the experiments and the datasets. Finally, Section 5 concludes the paper.

## 2. Related Work

Object detection and semantic segmentation are two mainstream vision approaches used in automated pavement crack analysis. Object detection locates crack regions with bounding boxes, offering real-time inference and localization capabilities. Semantic segmentation, in contrast, provides pixel-level classification, enabling precise extraction of crack shapes and boundaries. Due to their respective strengths—efficiency in detection and accuracy in structural delineation—both methods have been widely applied in pavement inspection scenarios. They help overcome challenges posed by the fine, irregular, and low-contrast nature of cracks, especially under varying illumination and surface textures.

YOLO models are widely applied in crack detection due to their real-time speed and localization capabilities[8, 9]. Xia et al. [10] enhanced YOLOv8 with attention mechanisms to improve detection of multi-scale bridge cracks, though performance on narrow cracks remained limited. Yu et al. [11] proposed a coordinate-attention YOLOv8 variant for concrete cracks, achieving better localization but showing reduced robustness under noisy textures. Ren and Zhong [12] integrated feature fusion and attention into YOLO for building crack detection, improving recall at the expense of increased computation. These approaches demonstrate progress, yet detecting fine or fragmented cracks remains challenging.

Semantic segmentation enables pixel-level crack extraction and is effective for detecting fine or irregular patterns. Yoon et al. [13] applied an attention-augmented UNet++ to port pavement cracks, achieving good accuracy but limited generalization. Tan et al. [14] introduced ETAFHrNet, a transformer-based model for asymmetric cracks, showing strong performance but high computational demand. Zhang et al. [15] proposed DLANet for sealed crack segmentation, which achieved fine boundary delineation but required extensive annotations. Despite

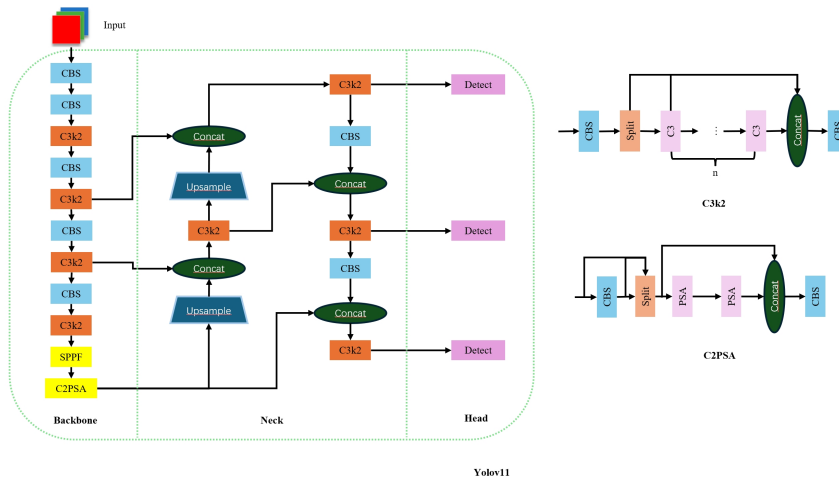
their precision, segmentation models often face trade-offs between accuracy, efficiency, and data dependence. To address the distinct limitations of existing detection and segmentation models—namely, insufficient sensitivity to fine-grained features in detection and poor efficiency or generalizability in segmentation—we propose two tailored network improvements targeting each task respectively.

### 3. Methodology

In this study, we propose two improved models for road crack recognition: an object detection network based on YOLOv11 and a segmentation network built upon TransUNet. In the detection branch, the original convolutional backbone is replaced with a Diverse Branch Block to enhance multi-scale feature extraction, while a Triplet Attention Module is integrated into the neck to strengthen spatial and channel-wise attention. In the segmentation branch, the Transformer encoder in TransUNet is replaced with BiFormer Block for improved global context modeling, and a parallel Swin Transformer Block is introduced to capture hierarchical local features.

#### 3.1. YOLOv11 Object Detection Model

YOLOv11 is a recent advancement in the YOLO series that aims to improve detection performance through architectural refinement and enhanced feature representation. It maintains the fast, single-stage design characteristic of previous YOLO models while introducing better optimization for small objects, high-density scenes, and complex backgrounds. These improvements position YOLOv11 as an effective baseline for real-time object detection tasks such as road crack detection [16]. Its overall architecture is shown in Figure 1:



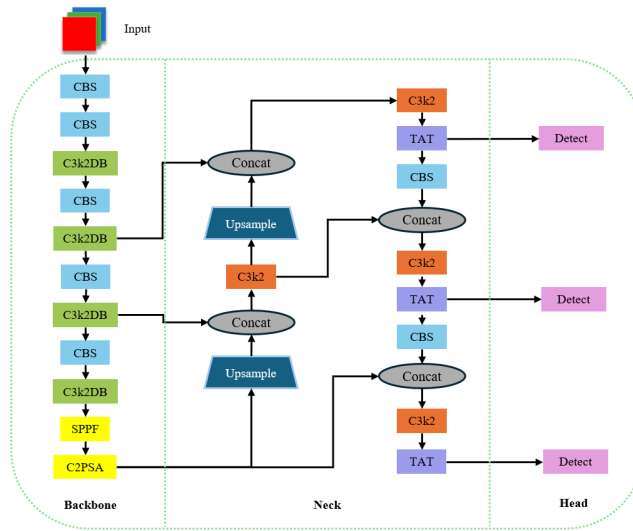
**Figure 1:** Overall architecture of the YOLOv11

YOLOv11 introduces architectural improvements to enhance detection accuracy and maintain real-time performance. Its backbone employs the SPPF and C2PSA modules to expand receptive

fields and strengthen spatial attention. The neck adopts a feature pyramid structure with C3k2 and upsampling layers for effective multi-scale feature fusion, while the head generates predictions at different scales using lightweight CBS Blocks. These enhancements improve robustness to small and dense objects, making YOLOv11 well-suited for fine-grained crack detection tasks.

### 3.2. Improved YOLOv11 Model

As shown in Figure 2, we propose an improved YOLOv11 model based on Diverse Branch Block and Triplet Attention.



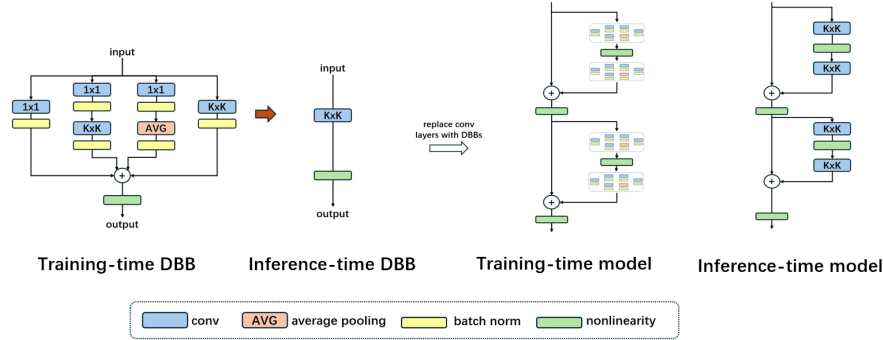
**Figure 2:** Overall architecture of the improved YOLOv11.

YOLOv11 demonstrates strong real-time performance and competitive detection accuracy, benefiting from its efficient single-stage design and optimized feature processing pipeline. However, it still encounters challenges in detecting fine-grained or low-contrast targets, particularly in cluttered or textured scenes, where the standard backbone may lack sufficient feature diversity and spatial focus. Inspired by Fan et al.[17], we adopt the Diverse Branch Block to enhance multi-scale feature representation through diverse receptive fields. Additionally, following the approach of Li et al.[18], we integrate the Triplet Attention Module to reinforce joint spatial and channel attention. These improvements are embedded into the YOLOv11 backbone and neck, respectively, leading to better localization of subtle targets while maintaining the lightweight structure suitable for real-time detection.

#### 3.2.1. Diverse Branch Block

Diverse Branch Block is a reparameterizable convolution module designed to enrich the representational capacity of convolutional layers by introducing multi-branch structures during

training. As illustrated in Figure 3, Diverse Branch Block consists of four parallel branches: a standard  $k \times k$  convolution, a  $1 \times 1$  convolution followed by  $k \times k$ , a  $1 \times 1$  convolution with average pooling, and a standalone average pooling path. All branches are followed by batch normalization and aggregated before a nonlinear activation.



**Figure 3:** Training and inference structure of the Diverse Branch Block.

During inference, these branches are mathematically merged into a single equivalent convolution kernel, allowing Diverse Branch Block to maintain inference efficiency. This design enables richer gradient flow and feature diversity during training without incurring runtime cost. Inspired by Li et al.[19], we replace the C3k2 modules in the YOLOv11 backbone with Diverse Branch Block to improve multi-scale feature encoding and enhance detection sensitivity for fine cracks and irregular textures.

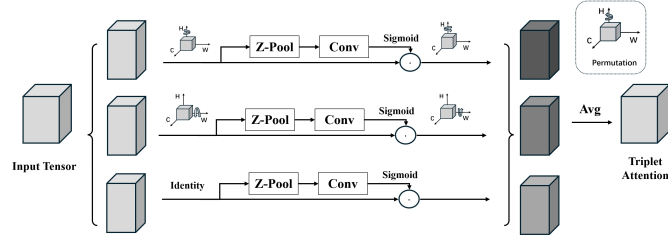
### 3.2.2. Triplet Attention Module

Triplet Attention Module (TAT) is an attention mechanism designed to capture spatial and channel-wise dependencies more effectively by employing a triplet-branch structure. As illustrated in Figure 4, it comprises three parallel branches, each computing attention along different axis pairs: (height  $\times$  channel), (width  $\times$  channel), and (height  $\times$  width). These branches apply convolutional transformations followed by sigmoid activation and inter-branch aggregation to capture richer feature interactions. According to Misra et al.[20], this structure improves the network’s ability to focus on salient regions while preserving spatial information across all directions.

In our model, we insert TAT into the neck of YOLOv11 to enhance spatial-channel attention before final prediction, improving its ability to detect fine and context-sensitive cracks.

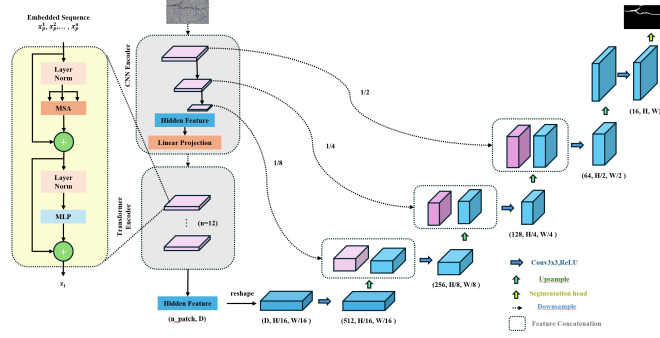
### 3.3. TransUNet Segmentation Model

TransUNet is a hybrid segmentation network that integrates CNN-based encoders with Transformer-based global attention, effectively combining local detail extraction and long-range context modeling. It outperforms traditional U-Net models, especially in scenarios with irregular shapes or low-contrast boundaries. These capabilities make TransUNet a strong baseline for pixel-wise segmentation tasks such as road crack detection, where both fine-grained



**Figure 4:** Illustration of the triplet attention.

localization and structural context are essential [21]. The overall architecture of the TransUNet framework is shown in Figure 5:



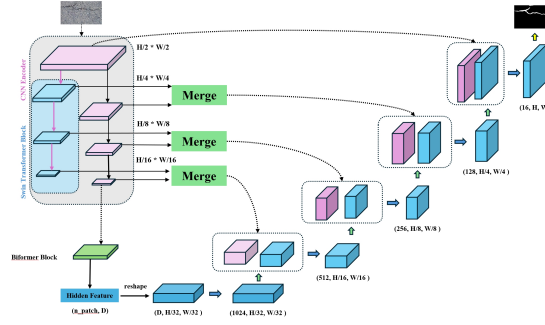
**Figure 5:** Overall architecture of the TransUNet framework.

The architecture adopts an encoder-decoder design, where a CNN backbone is used for hierarchical feature extraction and a Transformer bottleneck enhances global feature representation. The decoder integrates skip connections to refine spatial detail and recover semantic resolution. This synergy between convolutional and self-attention mechanisms improves boundary delineation and semantic coherence, making TransUNet particularly effective for dense prediction tasks like road crack segmentation.

### 3.4. Improved TransUNet Model

As shown in Figure 6, we propose an improved TransUNet model based on BiFormer Block and Swin Transformer Block.

While TransUNet effectively combines CNN encoders and Transformer-based bottlenecks, it still faces challenges in preserving fine-grained spatial features, particularly in road crack segmentation where boundary precision and structural continuity are vital. To address these issues, we propose an enhanced TransUNet framework that incorporates Swin Transformer Blocks[22] to capture local context through hierarchical window-based self-attention, replaces the ViT bottleneck with a lightweight BiFormer Block[23] to model long-range dependencies more efficiently, and introduces a Merge module combining Coordinate Attention and SENetV2

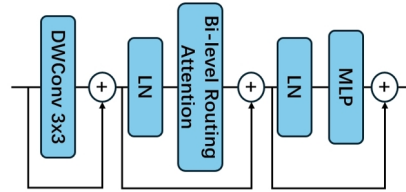


**Figure 6:** Overall architecture of the improved TransUNet.

to improve feature fusion by enhancing both spatial focus and channel-wise recalibration.

### 3.4.1. BiFormer Block

BiFormer is a lightweight vision transformer that introduces a Bi-level Routing Attention (BRA) mechanism to balance global representation and computational efficiency. As shown in Figure 7, the BiFormer Block incorporates depthwise convolution (DWConv), layer normalization (LN), and a multilayer perceptron (MLP) alongside the BRA module. The BRA adaptively selects query-key pairs to reduce redundant attention computation, enabling both global context modeling and efficient feature routing. Residual connections are used throughout to preserve gradient flow and feature continuity. Inspired by Wang et al.[23], we embed the BiFormer Block at the bottleneck of the TransUNet architecture to capture long-range dependencies with reduced complexity. This enhances contextual awareness and segmentation accuracy, especially in challenging cases like elongated or disconnected road cracks.

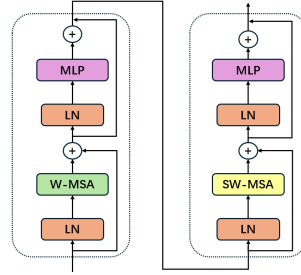


**Figure 7:** Details of a BiFormer Block.

### 3.4.2. Swin Transformer Block

The Swin Transformer Block represents a hierarchical vision transformer architecture that efficiently models both local and global dependencies through a window-based multi-head self-attention mechanism. As illustrated in Figure 8, the block sequentially employs regular window-based multi-head self-attention (W-MSA) and shifted window-based multi-head self-attention (SW-MSA), enabling cross-window contextual interaction while preserving linear

computational complexity with respect to image size. Each attention layer is followed by a multi-layer perceptron (MLP), with both modules encapsulated within residual connections and layer normalization to enhance optimization stability. By incorporating shifted windows, the architecture introduces inductive biases such as locality and translational equivariance, which are absent in standard Transformer designs. Within our framework, the integration of Swin Transformer Blocks significantly strengthens the encoder’s representational capacity, particularly in modeling fine-grained structures and preserving boundary continuity critical to tasks like road crack segmentation.



**Figure 8:** Two Successive Swin Transformer Blocks.

### 3.4.3. Merge Method

The proposed Merge Method is a hierarchical attention-based fusion mechanism designed to integrate multi-scale features from parallel encoder branches. It comprises two sequential attention modules: a coordinate attention block [24] and a spatial-and-excitation (SaE) module [25].

Initially, feature maps from different encoding stages are concatenated along the channel dimension to preserve spatial alignment. The coordinate attention block encodes directional information via separate height- and width-wise pooling, followed by channel transformation, thereby enhancing position-aware channel attention for improved localization of crack boundaries.

Subsequently, the SaE module models inter-channel dependencies through multi-path convolutions, capturing contextual diversity across varying receptive fields. This dual-attention fusion boosts feature selectivity and spatial consistency, ultimately enhancing the effectiveness of hierarchical feature aggregation for downstream segmentation tasks.

## 4. Experiment

In this section, we evaluate the effectiveness of the proposed improved YOLO and TransUNet models using two benchmark datasets: the official Ultralytics Crack Segmentation Dataset and a reduced version of the Crack500 Dataset[26]. Experimental results demonstrate that the enhanced architectures offer superior performance in crack segmentation tasks, validating the efficacy of the proposed modifications.



#### 4.1. Dataset

The improved YOLO model is evaluated on the official Ultralytics Crack Segmentation Dataset, which contains a single crack category and is divided into 3,717 training images, 200 validation images, and 112 testing images. For the improved TransUNet model, we use the reduced version of the Crack500 dataset due to the high computational cost and memory requirements of the full dataset, which includes two semantic classes—crack and background—with 2,413 images for training and 603 for testing.

#### 4.2. Implementation details

All experiments were conducted on a computing platform equipped with an NVIDIA GeForce RTX 4070 GPU, utilizing PyTorch 2.0.1 and CUDA 11.8 as the deep learning framework. The improved YOLO model was implemented based on the Ultralytics 8.3.0 framework, with training performed over 100 epochs using a batch size of 16, while maintaining default hyperparameter settings. For the improved TransUNet model, the same hardware and software environment was employed. Training followed an iteration-based strategy with 20,000 iterations, using a batch size of 4. The Adam optimizer was adopted with an initial learning rate set to 0.0001.

#### 4.3. Evaluation metrics

Typical metrics used for object detection tasks have been used to evaluate models in this study, including Precision, Recall, mAP50, and mAP50-95. The definitions of these metrics are as follows:

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives} \quad (1)$$

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives} \quad (2)$$

$$AP = \sum_{n=1}^N (R_{n+1} - R_n) Precision_{max}(R_{n+1}) \quad (3)$$

$$mAP = \frac{1}{C} \sum_j AP_j \quad (4)$$

To further assess the crack detection capability of the improved TransUNet model, the mean Intersection over Union (mIoU) was incorporated as an additional evaluation metric alongside Precision and Recall. Its definition is shown below:

$$mIoU = \frac{1}{k} \sum_{i=0}^k \frac{TP}{FN + FP + TP} \quad (5)$$

The Average Precision (AP) of all classes is the area of the region below the precision-recall curve.  $R_n$  represents the recall of the  $n$ th value, and  $Precision_{max}(R_{n+1})$  represents the highest precision value in the range  $R_n$  to  $R_{n+1}$ . The mAP is calculated by averaging the AP of each class in the dataset. mAP50 is obtained by averaging the AP (IoU = 0.5) of all classes,

and mAP50-95 is obtained by averaging the mAPs at different IoUs between 0.5 and 0.95. mIoU measures the average overlap between the predicted and ground truth regions across all classes.

#### 4.4. Improved YOLOV11 Performance Comparison

To validate the effectiveness of the proposed model in crack detection, YOLOv5 and YOLOv8 are introduced as baseline comparison methods. The performance of each model on the official Ultralytics Crack Segmentation Dataset is summarized in Table 1.

**Table 1**

Performance comparison of different YOLO-based models

Model	Para	P	R	mAP50	mAP50-95	GFLOPs
YOLOv5	2.18	0.799	0.785	0.804	0.596	5.8
YOLOv8	2.68	0.860	0.731	0.811	0.621	6.8
YOLOv11	2.58	0.877	0.751	0.818	0.621	6.3
<b>Ours</b>	<b>2.81</b>	<b>0.871</b>	<b>0.786</b>	<b>0.839</b>	<b>0.658</b>	<b>6.9</b>

Compared to YOLOv5, our model improves mAP50 by 4.35% and mAP50-95 by 6.2%, indicating enhanced detection accuracy. It also outperforms YOLOv8 and YOLOv11 in mAP50-95, demonstrating better localization. These results confirm the effectiveness of our architectural improvements for crack detection.

To validate the effectiveness of the proposed enhancements, ablation study results for each individual improvement are presented in Table 2.

**Table 2**

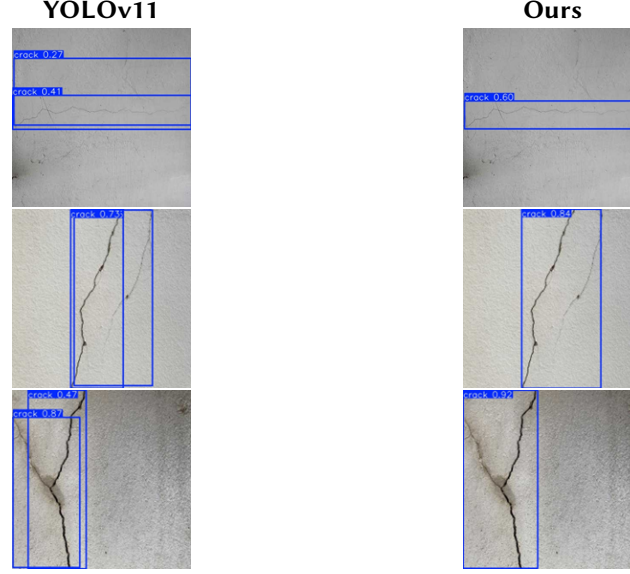
Ablation Experiments of Yolov11

Model	Para	P	R	mAP50	mAP50-95	GFLOPs
YOLOv11	2.58	0.877	0.751	0.818	0.621	6.3
YOLOv11 + Diverse Branch Block	2.71	0.873	0.773	0.824	0.622	6.7
YOLOv11 + Triplet Attention	2.58	0.888	0.755	0.826	0.654	6.4
<b>Ours</b>	<b>2.81</b>	<b>0.871</b>	<b>0.786</b>	<b>0.839</b>	<b>0.658</b>	<b>6.9</b>

The introduction of the Diverse Branch Block leads to a 0.6% increase in mAP50 and a 0.1% gain in mAP50-95, highlighting its contribution to multi-scale feature enhancement. Incorporating the TAT module further improves mAP50 by 0.8% and mAP50-95 by 3.3%, demonstrating its effectiveness in spatial and channel attention. Their combination yields the highest performance, validating the synergy of both modules.

To intuitively demonstrate the effectiveness of the proposed improvements, a visual comparison is conducted between the baseline YOLOv11 and the enhanced model. Representative results are presented in the figure 9.

We observe that the proposed method achieves more accurate crack localization and higher confidence scores compared to YOLOv11. It produces fewer false detections and better captures complete crack structures, particularly in fine or low-contrast regions, demonstrating improved detection robustness and spatial precision.



**Figure 9:** Visual comparison between YOLOv11 and the proposed method.

#### 4.5. Improved TransUNet Performance Comparison

All comparative models in this study were reproduced using the MMSegmentation 0.29.1 framework. To validate the effectiveness of the proposed model in crack segmentation, UNet and UNet++ are introduced as baseline comparison methods. The performance of each model on the reduced Crack500 Dataset is summarized in Table 3. All comparative models in this study were reproduced using the MMSegmentation 0.29.1 framework.

**Table 3**

Comparison of different segmentation models

Method	Precision	Recall	mIoU
TransUnet	87.25	69.12	69.12
Unet	86.25	83.09	75.77
Unet++	86.90	84.04	76.79
<b>Ours</b>	<b>86.94</b>	<b>86.48</b>	<b>78.46</b>

Compared to TransUnet, our model improves mRecall by 17.36% and mIoU by 9.34%, indicating enhanced segmentation completeness and region overlap. It also outperforms Unet and Unet++ in all metrics, demonstrating superior precision–recall balance. These results confirm the effectiveness of our architectural improvements for semantic segmentation.

To validate the effectiveness of the proposed enhancements, ablation study results for each individual improvement are presented in Table 4 .

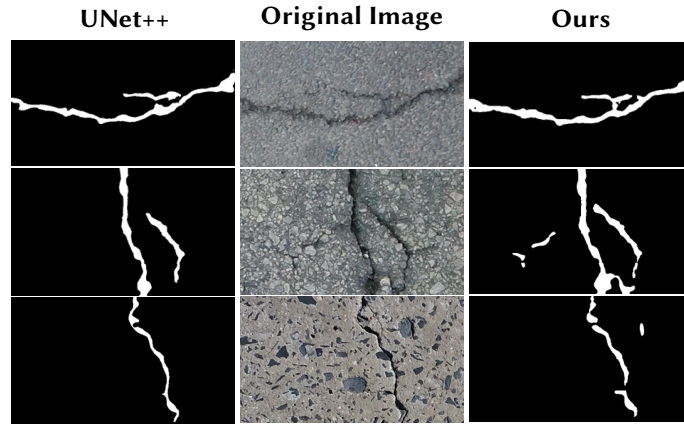
The integration of the Biformer module leads to an increase of 15.74% in mRecall and 8.11% in mIoU compared to the baseline TransUnet, highlighting its effectiveness in improving contextual understanding. Incorporating the Swin Transformer further enhances mIoU by 0.45%, demonstrating its strength in capturing long-range dependencies. The combination of our

**Table 4**  
Ablation Experiments of TransUNet

Method	Precision	Recall	mIoU
TransUNet	87.25	69.12	69.12
TransUNet+BiFormer	86.70	84.86	77.23
TransUNet+Swin Transformer	87.88	84.49	77.68
<b>Ours</b>	<b>86.94</b>	<b>86.48</b>	<b>78.46</b>

architectural refinements achieves the highest mRecall and mIoU, validating the complementary benefits of both modules for segmentation accuracy.

To intuitively demonstrate the effectiveness of the proposed improvements, a visual comparison is conducted between the UNet++ and the enhanced model. Representative results are presented in the figure10 .



**Figure 10:** Visual comparison of segmentation between UNet++ and the proposed method.

We observe that the proposed method delivers more precise crack segmentation compared to UNet++. It better preserves the continuity and topology of fine cracks, especially in noisy or complex backgrounds. The results exhibit fewer broken or fragmented regions and reduced false positives, indicating enhanced segmentation accuracy and structural consistency.

## 5. Conclusion

In this study, we proposed improved network architectures for both road crack detection and segmentation tasks. For object detection, we enhanced the backbone of YOLOv11 by introducing the Diverse Branch Block and integrated the Triplet Attention Module into the neck to improve spatial and channel attention. For semantic segmentation, we modified the original TransUNet by incorporating Swin Transformer Blocks and a BiFormer Block, and designed a hierarchical Merge mechanism based on coordinate and excitation attention to strengthen multi-scale feature fusion.

Experiments conducted on the Ultralytics Crack Segmentation Dataset and the reduced

Crack500 dataset demonstrate that our improved YOLOv11 model achieves superior mAP50 and mAP50–95 performance compared to YOLOv5, YOLOv8, and the original YOLOv11. Similarly, the improved TransUNet outperforms baseline segmentation networks in terms of mIoU, particularly in capturing fine-grained crack boundaries.

In summary, our enhanced detection model significantly boosts sensitivity and robustness for detecting small or subtle cracks, while the improved segmentation model offers superior spatial accuracy and contextual understanding, proving the effectiveness of our design in real-world road crack analysis tasks.

## Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

## References

- [1] Z. Li, Y. Ge, L. Meng, A multi-scale information fusion framework with interaction-aware global attention for industrial vision anomaly detection and localization, *Information Fusion* 124 (2025) 103356.
- [2] Z. Li, Y. Yan, X. Wang, Y. Ge, L. Meng, A survey of deep learning for industrial visual anomaly detection, *Artificial Intelligence Review* 58 (2025) 279.
- [3] X. Yue, L. Meng, Yolo-msa: A multiscale stereoscopic attention network for empty-dish recycling robots, *IEEE Transactions on Instrumentation and Measurement* 72 (2023) 1–14.
- [4] T. L. Mien, N. D. Tu, N. Van Lam, Deploying yolov8 for real-time road crack detection on smart road length measurement devices, *Journal of Future Artificial Intelligence and Technologies* 2 (2025) 135–144.
- [5] Y. Deng, H. Yu, P. Niu, F. Guo, Enhancing pavement crack detection using a hybrid convolutional neural network-transformer architecture, *Transportation Research Record* (2025) 03611981251329046.
- [6] O. Isreal, Integrating spatial and channel attention in deeplabv3 for fine-grained road crack and lane marking segmentation (2025).
- [7] H. Zhang, Y. Hu, J. Hu, J. Jin, P. Liu, Crackadaptnet: End-to-end domain adaptation for crack detection and quantification, *Measurement* (2025) 117716.
- [8] Y. Ge, Z. Li, X. Yue, H. Li, L. Meng, Dataset purification-driven lightweight deep learning model construction for empty-dish recycling robot, *IEEE Transactions on Emerging Topics in Computational Intelligence* (2025) 1–16.
- [9] X. Yue, L. Meng, Yolo-sm: A lightweight single-class multi-deformation object detection network, *IEEE Transactions on Emerging Topics in Computational Intelligence* 8 (2024) 2467–2480.
- [10] H. Xia, Q. Li, X. Qin, W. Zhuang, H. Ming, X. Yang, Bridge crack detection algorithm designed based on yolov8, *Applied Soft Computing* 149 (2025) 110118. doi:10.1016/j.asoc.2024.110118.
- [11] G. Yu, X. Yan, C. Ma, D. Yang, Z. Li, An improved yolov8-based method for concrete

- surface crack detection, *Nondestructive Testing and Evaluation* 40 (2025) 211–225. doi:10.1080/10589759.2025.2499032.
- [12] W. Ren, Z. Zhong, Building construction crack detection with bccd-yolo enhanced feature fusion and attention mechanisms, *Scientific Reports* 15 (2025) 5665. doi:10.1038/s41598-025-05665-y.
  - [13] H. Yoon, H. K. Kim, S. Kim, Ppdd: Egocentric crack segmentation in the port pavement with deep learning-based methods, *Applied Sciences* 15 (2025) 5446. doi:10.3390/app15105446.
  - [14] C. Tan, J. Liu, Z. Zhao, R. Liu, H. Zhang, Etafhrnet: A transformer-based multi-scale network for asymmetric pavement crack segmentation, *Applied Sciences* 15 (2025) 6183. doi:10.3390/app15116183.
  - [15] A. A. Zhang, Y. Wei, D. Wang, Y. Peng, Pixel-level efficient detection of pavement seal cracks using dlanet, *Journal of Infrastructure Systems* (2025). doi:10.1061/JITSE4.ISENG-2537.
  - [16] R. Khanam, M. Hussain, Yolov11: An overview of the key architectural enhancements, *arXiv preprint arXiv:2410.17725* (2024).
  - [17] Y. Fan, K. Zhi, H. An, R. Gu, X. Ding, J. Tang, Disease monitoring and characterization of feeder road network based on improved yolov11, *Electronics* 14 (2025) 1818.
  - [18] Q. Li, T. Wu, T. Xu, J. Lei, J. Liu, A novel yolo algorithm integrating attention mechanisms and fuzzy information for pavement crack detection, *International Journal of Computational Intelligence Systems* 18 (2025) 1–25.
  - [19] X. Ding, X. Zhang, J. Han, G. Ding, Diverse branch block: Building a convolution as an inception-like unit, in: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 10886–10895.
  - [20] D. Misra, T. Nalamada, A. U. Arasanipalai, Q. Hou, Rotate to attend: Convolutional triplet attention module, in: *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2021, pp. 3139–3148.
  - [21] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, Y. Zhou, Transunet: Transformers make strong encoders for medical image segmentation, *arXiv preprint arXiv:2102.04306* (2021).
  - [22] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10012–10022.
  - [23] L. Zhu, X. Wang, Z. Ke, W. Zhang, R. W. Lau, Biformer: Vision transformer with bi-level routing attention, in: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 10323–10333.
  - [24] Q. Hou, D. Zhou, J. Feng, Coordinate attention for efficient mobile network design, in: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 13713–13722.
  - [25] M. Narayanan, Senetv2: Aggregated dense layer for channelwise and global representations, *arXiv preprint arXiv:2311.10807* (2023).
  - [26] F. Yang, L. Zhang, S. Yu, D. Prokhorov, X. Mei, H. Ling, Feature pyramid and hierarchical boosting network for pavement crack detection, *IEEE Transactions on Intelligent Transportation Systems* 21 (2019) 1525–1535.