



# Journal of Hunan University (Natural Sciences)

Vol. 53 No. 4  
April 2026

Available online at  
<https://jonuns.com>



ELSEVIER  
Scopus



Clarivate  
WEB OF SCIENCE

Open Access Article

 <https://doi.org/10.55463/issn.1674-2974.53.4.2>

## Integrating YOLOv11 and DualUNet for Precise Bridge Crack Detection and Segmentation

Zhou Wang<sup>1</sup>, Yuan Li<sup>1</sup>, Huailiang Cheng<sup>1</sup>, Jie Zhou<sup>1</sup>, Jun Song<sup>2\*</sup>

<sup>1</sup>Hunan Provincial Administration of Quality and Safety Supervision for Transportation Construction,  
Changsha, Hunan 410000, China,

<sup>2</sup>Hunan CCCC Jingwei Information Technology Co., Ltd., Changsha, Hunan 410000, China,

\* Corresponding author: [495011917@qq.com](mailto:495011917@qq.com)

### Article History:

Received: February 15, 2026

Revised: March 29, 2026

Accepted: April 10, 2026

Published: April 25, 2026

**Abstract:** Concrete bridge crack detection plays a critical role in intelligent infrastructure inspection and structural safety assessment. However, existing automated approaches still face significant challenges, including the high miss rate of fine cracks, strong interference from complex backgrounds, and insufficient boundary delineation accuracy.

To address these limitations, this study proposes a novel two-stage collaborative detection–segmentation framework that integrates an enhanced YOLOv11 detector with a dual-branch DualUNet architecture. The framework follows a “coarse localization–guided fine segmentation” strategy. In the detection stage, YOLOv11 is improved by



Copyright: © 2026 by the authors. Licensee JHU

This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution License  
<http://creativecommons.org/licenses/by/4.0/>

incorporating a P2 high-resolution feature layer and a C2PSA attention module, enabling accurate localization of fine cracks while effectively suppressing background noise and generating reliable region-of-interest (RoI) priors.

In the segmentation stage, a dual-branch DualUNet with a shared ResNet34 encoder is developed. The global branch captures contextual semantic information from the full image, whereas the RoI branch refines local crack features by integrating detection priors with a Multi-Scale Squeeze-and-Excitation (MSSE) module. This design enhances the representation of fine structural details and mitigates the limitations of single-stage segmentation methods.

To address the severe class imbalance caused by sparse crack pixels, a hybrid loss function combining weighted binary cross-entropy and Dice loss is adopted. Additionally, a boundary supervision mechanism is introduced to improve contour accuracy.

Experimental results on the Crack500 dataset demonstrate that the proposed framework achieves superior performance, with recall and precision reaching 0.79 and 0.75, respectively. Compared with baseline models, the proposed method improves Boundary IoU by 26%, indicating significantly enhanced edge delineation. Visual results further confirm that the method effectively suppresses background interference while preserving crack continuity, making it suitable for practical bridge inspection applications.

**Keywords:** Concrete crack detection; Bridge inspection; YOLOv11; Dual U-Net; Attention mechanism; Boundary segmentation.

---

## 融合 YOLOv11 与 DualUNet 的桥梁裂缝精准检测与分割框架

**摘要:** 混凝土桥梁裂缝检测在智能基础设施巡检和结构安全评估中具有重要意义。然而, 现有自动化方法仍面临诸多挑战, 包括细微裂缝漏检率较高、复杂背景干扰严重以及边界刻画精度不足等问题。

针对上述问题, 本文提出了一种新型两阶段协同检测与分割框架, 将改进的 YOLOv11 检测网络与双分支 DualUNet 架构相结合。该框架采用“粗定位引导精细分割”的策略。在检测阶段, 通过引入 P2 高分辨率特征层和 C2PSA 注意力模块, 对 YOLOv11 进行改进, 从而在有效抑制背景噪声的同时, 实现对细微裂缝的精确定位, 并为后续分割提供可靠的感兴趣区域 (RoI) 先验信息。

在分割阶段, 构建了一个具有共享 ResNet34 编码器的双分支 DualUNet 结构。全局分支用于提取整幅图像的语义上下文信息, 而 RoI 分支则结合检测框先验信息与多尺度 Squeeze-and-Excitation (MSSE) 模块, 对局部裂缝细节特征进行增强。该设计有效提升了细节特征表达能力, 弥补了单阶段分割模型在局部细节恢复方面的不足。

为缓解裂缝像素极度稀疏所导致的类别不平衡问题, 本文采用了加权二元交叉熵 (Binary Cross-Entropy, BCE) 与 Dice Loss 相结合的混合损失函数。此外, 引入边界监督机制以进一步提升裂缝轮廓的分割精度。

在 Crack500 数据集上的实验结果表明, 所提出的方法具有优越的性能, 召回率和精确率分别达到 0.79 和 0.75。与基线模型相比, 该方法在边界交并比 (Boundary IoU) 指标上提升了 26%, 显著增强了边界刻画能力。定性可视化结果进一步验证了该方法在抑制背景干扰的同时, 能够有效保持裂缝的连续性, 具有良好的工程应用前景。

**关键词:** 混凝土裂缝检测; 桥梁检测; YOLOv11; 双 U-Net; 注意力机制; 边界分割

---

## 1. Introduction

As a vital component of transportation infrastructure, bridges are subjected to multiple long-term effects, including vehicular loads, temperature fluctuations, and environmental erosion, leading to a progressive degradation in their structural performance over their service life. Cracks are the most prevalent form of surface damage in concrete bridges, diminishing the effective load-bearing area of the cross-section and providing infiltration channels for external moisture and corrosive media. This phenomenon accelerates the corrosion of internal steel reinforcement and causes the spalling of the concrete cover, thereby posing a severe threat to the overall safety and durability of the structure. Consequently, the accurate and timely detection and identification of bridge cracks serve as a crucial technical prerequisite for ensuring structural integrity and prolonging service life.

Traditional bridge crack detection primarily relies on close-range manual visual inspections, supplemented by instruments such as crack width gauges for observation and documentation [1][2]. However, this methodology exhibits significant limitations. Firstly, spatial constraints restrict access to critical inspection zones, such as bridge pylons, high piers, and girder soffits, making it exceedingly difficult for inspection personnel to reach them. This often leads to inspection blind spots, where personnel may easily overlook underlying hazards. Secondly, manual inspection is inherently inefficient, rendering the comprehensive examination of large-span and multi-span bridges both time-consuming and labor-intensive. Furthermore, the quality and reliability of the inspection outcomes depend heavily on the subjective expertise of the personnel, resulting in poor reproducibility. As a result, traditional methods are increasingly inadequate for meeting the high-frequency, standardized maintenance requirements of modern, large-scale infrastructure networks, which necessitate more efficient and objective inspection techniques to ensure safety and reliability [3].

With the rapid development of computer vision and deep learning technologies, convolutional neural networks (CNNs) have been widely introduced into automatic concrete crack detection and pixel-level segmentation tasks [4-7]. Within the technical trajectory

of object detection, single-stage detection algorithms, prominently represented by the YOLO series, have garnered considerable attention due to their excellent balance between speed and accuracy [8]. For example, YOLOv8 uses an anchor-free detection head and a task-aligned sample assignment strategy to make it easier to find targets that are not perfectly shaped, like long cracks [9]. Building upon this foundation, numerous studies have applied improved YOLOv8 models to crack detection in bridges and concrete structures [10-12]. Examples include integrating a Bidirectional Feature Pyramid Network (BiFPN) into the neck to enhance multi-scale crack feature fusion capabilities [11, 12] or incorporating coordinate attention mechanisms into the backbone network to improve the representation of minute defects [13, 14]. These works have demonstrated the critical value of multi-scale feature fusion and coordinate attention mechanisms for small-object perception in tasks involving bridge cracks, concrete surface defects, and other engineering small-object detection applications. Furthermore, while continuing the anchor-free architecture, YOLOv11 introduces modules such as C2PSA [15]. The YOLOv11-KW-TA-FP model proposed by Song et al. [16] effectively improves detection precision and recall through the introduction of dynamic convolutions and a triplet attention mechanism. Similarly, the MEP-YOLOv11 model constructed by Zhang et al. [17] significantly enhances the localization capability for slender cracks and the accuracy of edge perception by introducing the C3K2-MSEIE multi-scale edge information enhancement module and a P2 small object detection layer. Gao et al. [18] substantially improved localization accuracy while maintaining high inference speeds by integrating the C3K2-SG module into the YOLOv11 backbone, replacing SPPF with FPSConv in the neck, and designing an Inner\_MPDIoU bounding box regression loss.

In the trajectory of semantic segmentation, the encoder-decoder architecture, represented by U-Net, possesses natural advantages in fine boundary reconstruction and has emerged as the mainstream baseline architecture for pixel-level crack segmentation tasks [19]. To further elevate performance, the Feature Pyramid Hierarchical Boosting Network (FPHBN) proposed by Yang et al. [20] emphasizes hard-example

crack regions through a hierarchical sample re-weighting mechanism, thereby improving pixel-level crack detection performance. The literature [21] also documents the introduction of atrous (dilated) convolution modules with multiple dilation rates into the U-Net encoder; this significantly enlarges the receptive field while maintaining feature map resolution, thereby strengthening the connectivity modeling capability for slender cracks. In recent years, the self-attention mechanism of Transformers has also been introduced into the field of crack segmentation to model long-range dependencies and global contexts within high-resolution images, thereby improving the continuity representation at the terminal ends of minute cracks [22, 23]. For example, CAFANet [24] utilizes a cross-attention feature alignment module to establish pixel-level correlations, effectively enhancing the boundary integrity of pavement cracks.

In segmentation models utilizing full-image inputs, the uniform scaling process further degrades minute crack regions, making it difficult for the decoder to accurately reconstruct crack boundaries. Furthermore, in typical crack datasets, the proportion of crack pixels is frequently below 1:100 [25]. If an unweighted cross-entropy loss is applied, the network is highly susceptible to converging on a local optimum where all pixels are predicted as background. To address this class imbalance, existing studies [26, 27] have proposed various improvement strategies. For instance, some works explicitly increase the weight of hard-to-classify samples using focal loss to suppress the training bias dominated by negative samples. Numerous crack segmentation studies [28, 29] have demonstrated that a linear combination of binary cross-entropy (BCE) and Dice loss functions effectively alleviates the pixel sparsity issue on datasets such as Crack500, balancing the overall intersection over union (IoU) with boundary quality. Additionally, architectures like the IoU Adaptive Deformable R-CNN [30] dynamically adjust sample weights based on the IoU between predicted and ground-truth bounding boxes, significantly enhancing detection accuracy against complex backgrounds. This adaptive weighting philosophy is equally applicable to pixel-level crack prediction to mitigate the interference of low-quality predictions. Overall, while existing research predominantly improves detection capabilities

through multi-scale fusion and attention mechanisms, the inadequate delineation of crack boundaries remains a persistent challenge, particularly in distinguishing between actual cracks and noise in the data.

The aforementioned challenges indicate that a singular detection or segmentation paradigm cannot simultaneously fulfill the dual engineering requirements of precise small-object localization and fine pixel-level segmentation for bridge crack detection. Consequently, two-stage collaborative frameworks integrating "detection and segmentation" have emerged as a focal point of research. Mask R-CNN, proposed by He et al. [31], first realized the collaborative output of bounding box detection and pixel-level mask prediction; its RoIAlign mechanism established the foundational logic of "first locating the Region of Interest (RoI), then executing fine segmentation." Other investigations [32] have utilized the bounding box outputs from single-stage detectors (e.g., the YOLO series) as spatial priors for subsequent segmentation networks (e.g., U-Net or DeepLab). This approach achieves a synergy between rapid coarse localization and local fine segmentation through a two-stage "crop-segment-restore" pipeline. However, most existing methodologies train these two models independently, which suffers from a lack of end-to-end information interaction. Furthermore, the segmentation branch frequently lacks targeted feature enhancement designs for the RoI, failing to fully exploit the complementary advantages of full-image semantic constraints and local detailed features.

To address the aforementioned limitations, this paper proposes a two-stage framework for the fine-grained recognition of bridge cracks, driven by the collaborative integration of an improved YOLOv11 and a dual-branch DualUNet. The primary contributions of this work are summarized as follows:

1. To address the challenge of missed detections of elongated cracks from a full-image perspective, we construct a customized YOLOv11 localization network tailored for extremely fine cracks. Building upon the utilization of the P2 high-resolution detection layer to preserve shallow local features, a C2PSA attention module is further integrated into both the backbone network and the feature fusion layer. This architectural design effectively mitigates interference from complex bridge deck

backgrounds, significantly enhances the localization stability for minute cracks, and provides a highly accurate spatial Region of Interest (RoI) before subsequent fine-grained segmentation.

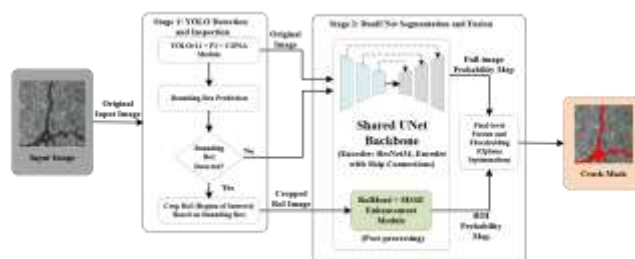
2. A dual-branch DualUNet collaborative segmentation architecture is designed to construct a global-image branch and a local-detail RoI branch utilizing a shared ResNet34 encoder. By fully exploiting the Region of Interest (RoI) generated by the detector as a spatial prior, this framework enables the segmentation network to perform high-resolution, fine-grained segmentation within local regions while simultaneously preserving global contextual constraints.
3. A Multi-Scale Squeeze-and-Excitation (MSSE) module is proposed to enhance directional features. Addressing the physical characteristics of cracks, specifically their narrow, elongated morphology and highly variable topologies, this module extracts multi-scale texture information via parallel  $3 \times 3$ ,  $1 \times 7$ , and  $7 \times 1$  directional convolutions. By coupling the MSSE module with a Squeeze-and-Excitation (SE) channel attention mechanism for adaptive feature recalibration, the network's sensitivity to the edges and width variations of slender cracks is significantly augmented.

This study systematically validates the proposed method on the public Crack500 dataset, employing a comprehensive evaluation suite comprising five metrics: precision, recall, Dice coefficient, intersection over union (IoU), and boundary IoU. Notably, the Boundary IoU metric is specifically utilized to quantify the pixel-level precision of the crack boundaries. Experimental results demonstrate that the two-stage collaborative framework achieves superior comprehensive performance across all metrics compared to standalone detection or segmentation methods. This substantiates the efficacy of the paradigm "customized detection and localization guiding fine-grained segmentation," thereby providing novel technical support for the intelligent perception and evaluation of surface defects in bridge structures.

## 2. Methods

### 2.1. Overall Framework

The core design rationale of the overall framework is to organically integrate the rapid, coarse localization capabilities of object detection with the fine, pixel-level descriptive power of semantic segmentation via a structured information transfer mechanism. This integration establishes a collaborative inference pipeline characterized by "macroscopic localization guiding local fine-grained segmentation." In the first stage, a customized YOLOv11 crack detector rapidly localizes crack regions at the full-image scale, outputting the bounding box coordinates for the Regions of Interest (RoIs). In the second stage, the dual-branch DualUNet segmentation network utilizes both the full image and the extracted RoIs as dual inputs. By employing a shared encoder and specialized enhancement modules, it achieves a synergy between global semantic perception and local detailed segmentation. Finally, an adaptive probability fusion module conducts a weighted fusion of the probability maps generated by the two branches, yielding the ultimate pixel-level crack mask following post-processing. The overall framework is illustrated in Figure 1.

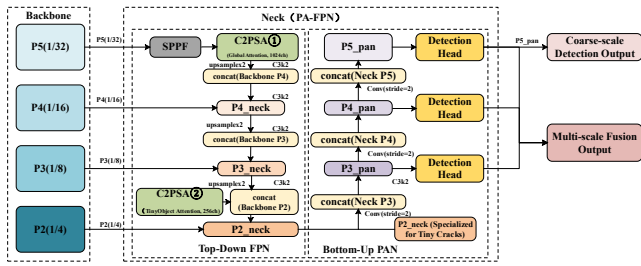


**Figure 1. The overall framework of the proposed method**

### 2.2. Improved YOLOv11 Crack Detector

Building upon the anchor-free architecture of YOLOv8, YOLOv11 incorporates the C2PSA (Cross-Stage Partial with Position-Sensitive Attention) module [14] and adopts Task-Aligned Learning (TAL) to replace the traditional Intersection over Union (IoU)-based positive and negative sample assignment mechanism. This adaptation allows the bounding box regression to better accommodate scenarios involving small, irregularly shaped crack targets. The standard YOLOv11 Feature Pyramid Network (FPN) outputs at the P3, P4, and P5 levels (corresponding to down-sampling rates of 1/8, 1/16, and 1/32 of the original image, respectively). However, the receptive field of P3, the highest-resolution feature map, remains relatively

large, creating a pronounced perceptive blind spot for small targets such as fine, thread-like cracks that are merely a few pixels wide. To explicitly address the characteristics of bridge cracks, which are elongated, exceptionally narrow, and heavily intermingled with concrete textures, targeted structural modifications are introduced to YOLOv11 through both detection-layer expansion and attention enhancement. The specific architecture is illustrated in Figure 2.



**Figure 2. YOLOv11 P2-Enhanced Crack Detection Network Architecture**

### 2.2.1. The P2 Small Object Detection Layer

To enhance the model's perceptual capacity for ultra-fine cracks, a P2 detection layer is integrated into the standard YOLOv11 Feature Pyramid Network (FPN) architecture[14], expanding the detection scales from the original three-scale configuration (P3/P4/P5) to a four-scale configuration (P2/P3/P4/P5). The P2 layer corresponds to a 1/4 down-sampling rate of the original image, yielding a feature map dimension of  $C_{P2} \times \frac{H}{4} \times$

$\frac{W}{4}$ . Compared to P3, it preserves richer shallow geometric details and high-frequency texture information, thereby equipping the network with the foundational resolution necessary to effectively respond to crack targets as narrow as approximately 4 pixels.

In the top-down feature fusion pathway, the P2 features are generated by concatenating and fusing the unsampled features from the higher P3 layer with the shallow, high-resolution output from the second stage of the backbone network. This process can be expressed as follows:

$$F_{P2} = \text{Conv}_{3 \times 3}(\text{Up}(F_{P3}) \oplus F_{\text{stage2}}) \quad (1)$$

$\text{Up}(\cdot)$  denotes the upsampling operator,  $\oplus$  represents the channel concatenation operation, and  $F_{\text{stage2}}$  signifies the high-resolution features extracted from the second stage of the backbone network. By

adopting the design philosophy of the PANet bidirectional feature pyramid, this architecture systematically covers the perceptive blind spots for small targets inherent in the standard three-scale FPN without significantly increasing inference latency. Consequently, it provides denser feature support for fine-grained crack regions.

### 2.2.2. Attention Mechanism Enhancement: C2PSA Channel Attention Module

Fundamentally, C2PSA is a channel attention mechanism integrated with positional encoding. Tailored for scenarios characterized by complex concrete surface textures and substantial noise, this mechanism effectively suppresses background responses irrelevant to cracks and amplifies the activation intensity of channels corresponding to the fine-line structures of the cracks. Consequently, it elevates detection accuracy while effectively mitigating the false positive rate. To operate in synergy with the P2 small-target detection layer, the C2PSA module is deployed simultaneously within the deep feature extraction stage of the backbone network and at the feature fusion node of the P2 detection layer, thereby bolstering the network's discriminative response to critical crack regions.

For an input feature map  $F \in \mathbb{R}^{C \times H \times W}$ , the C2PSA module achieves feature recalibration by incorporating a positional encoding, denoted as PE, alongside a channel squeeze-and-excitation structure. This process can be formalized as follows:

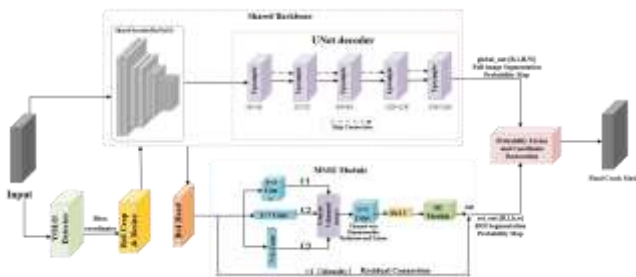
$$F_{\text{att}} = \text{Attention}(F + \text{PE}) \odot \sigma(W_{\text{se}} \text{GAP}(F)) \quad (2)$$

$\text{GAP}(\cdot)$  represents Global Average Pooling,  $\sigma(\cdot)$  denotes the Sigmoid activation function,  $W_{\text{se}}$  indicates the linear transformation weights for the channel Squeeze-and-Excitation, and  $\odot$  signifies element-wise multiplication. The first term models inter-channel correlations and spatial positional context via position-sensitive multi-head attention. In contrast, the second term achieves adaptive recalibration of channel importance through the SE (Squeeze-and-Excitation) mechanism. Consequently, the improved YOLOv11 detector provides a spatial prior with higher confidence and superior localization precision for the subsequent ROI branch.

## 2.3. Dual-Branch DualUNet Segmentation Network

As the regions further degrade, it becomes increasingly difficult for the decoder to precisely reconstruct boundaries from abstract features. Concurrently, there exists an extreme quantitative imbalance between crack pixels and background pixels, with a ratio of approximately 1:100. To process such severely imbalanced data while simultaneously accommodating full-image macroscopic semantics and RoI-level local high-resolution features, a DualUNet segmentation network comprising a shared UNet encoder and an RoI post-processing head is constructed[31].

As illustrated in Figure 3, the core architecture of DualUNet is implemented based on the UNet from the `segmentation_models_pytorch` library. The encoder utilizes a pre-trained ResNet34 initialized with large-scale pre-trained weights from ImageNet. This strategy transfers rich, general visual priors to the crack segmentation task, effectively mitigating the risk of overfitting on medium-scale datasets such as Crack500. The global-image branch takes the original bridge image, scaled to  $256 \times 256$ , as input. Through the shared UNet backbone, it directly outputs a full-image scale crack logit map, which is subsequently processed via a sigmoid function to generate the crack probability map.



**Figure 3. DualUNet Segmentation Structure**

### 2.3.1. RoI Branch and Coordinate Mapping

The RoI branch utilizes the bounding boxes generated by the improved first-stage YOLOv11 detector as spatial priors, extracting the corresponding RoI patches from the original image and uniformly resizing them to  $256 \times 256$  via bilinear interpolation prior to network ingestion. These RoI images are subsequently processed through the identical shared UNet backbone utilized by the global-image branch, yielding RoI-scale crack logit maps. These maps undergo further post-processing and enhancement

within the RoI Head module, ultimately producing the final RoI-scale crack probability maps:

$$P_{\text{RoI}} \in [0,1]^{H_{\text{RoI}} \times W_{\text{RoI}}} \quad (3)$$

Utilizing the bounding box coordinates and the scaling factor,  $P_{\text{RoI}}$  is projected back into the full-image space via inverse coordinate mapping, yielding an RoI probability map with dimensions identical to the original image:

$$P_{\text{RoI}}^{\text{full}} \in [0,1]^{H \times W} \quad (4)$$

Regions not covered by any Region of Interest (RoI) are assigned a probability of zero, while the pixel-wise maximum value is retained for areas where multiple RoIs overlap. This strategy not only ensures the high-resolution characterization of local regions by the RoI branch but also achieves consistent alignment with the probability map of the global-image branch at the full-image scale.

### 2.3.2. RoI Head Module

To mitigate the discrepancies in statistical distribution and semantic granularity between the features generated by the shared encoder and the local structures of the Region of Interest (RoI), the RoI Head performs an initial local semantic refinement on the encoded RoI features. The RoI Head is constructed by cascading two convolutional blocks, where each layer comprises a  $3 \times 3$  convolution and a ReLU. The single-layer update process can be formulated as:

$$F'_{\text{RoI}} = F_{\text{RoI}} + \text{Conv}_{3 \times 3} \left( \text{BN}(\text{ReLU}(F_{\text{RoI}})) \right) \quad (5)$$

Two such consecutive residual convolutional blocks endow the RoI features with a more discriminative local semantic representation before their input into the MSSE module. This refinement is instrumental in ensuring that the subsequent multi-scale directional convolutions and channel recalibration mechanisms achieve optimal efficacy.

### 2.3.3. Multi-Scale Squeeze-and-Excitation (MSSE) Module

Designed specifically for the enhancement of local crack features within the RoI branch, the Multi-Scale Squeeze-Excitation (MSSE) module[32] is rooted in the geometric priors of crack targets: cracks propagate along specific trajectories, exhibit varying texture widths

across different observation scales, and possess pronounced asymmetrical structural characteristics in both horizontal and vertical directions. Traditional  $3 \times 3$  convolutions exhibit relatively inadequate directional selectivity, rendering them ill-equipped to simultaneously capture the manifestations of slender cracks across diverse orientations. Consequently, the MSSE module introduces three parallel branches of multi-scale directional convolutions to the RoI feature,  $F''_{RoI}$  generated by the shared encoder. These branches consist of a standard  $3 \times 3$  convolution, a  $1 \times 7$  horizontal strip convolution, and a  $7 \times 1$  vertical strip convolution:

$$F_{3 \times 3} = \text{Conv}_{3 \times 3}(F''_{RoI}) \quad (6)$$

$$F_{1 \times 7} = \text{Conv}_{1 \times 7}(F''_{RoI}) \quad (7)$$

$$F_{7 \times 1} = \text{Conv}_{7 \times 1}(F''_{RoI}) \quad (8)$$

The features extracted from these three parallel branches are concatenated along the channel dimension, fused via a  $1 \times 1$  convolution and a ReLU activation function, and subsequently integrated with the input features through a residual connection employing a fixed gain coefficient of  $\alpha=0.5$ :

$$F_{\text{cat}} = \text{Concat}[F_{3 \times 3}, F_{1 \times 7}, F_{7 \times 1}] \quad (9)$$

$$F_{\text{out}} = \alpha \cdot \text{SE}(\text{ReLU}(\text{Conv}_{1 \times 1}(F_{\text{cat}}))) + F''_{RoI} \quad (10)$$

As illustrated in Figure 4, by employing  $1 \times 7$  and  $7 \times 1$  strip convolutions to respectively enhance texture sensitivity in the horizontal and vertical directions, in conjunction with the isotropic receptive field of the  $3 \times 3$  convolution, the MSSE module achieves the synchronous extraction of multi-directional and multi-scale crack features. Furthermore, the SE operation[33] adaptively suppresses redundant features and amplifies discriminative, crack-related channels along the channel dimension. This renders the RoI branch highly sensitive to the edges and width variations of minute cracks, thereby establishing a high-quality local feature foundation for the subsequent fusion of its probability map with that of the global-image branch.

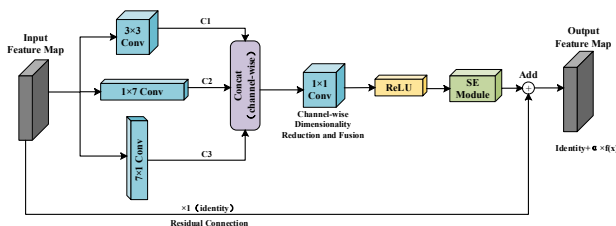


Figure 4. Structure of the MSSE module

## 2.4. Loss Function Design

To address the extreme class imbalance between crack and background pixels in crack segmentation tasks, a ratio of approximately 1:100, a joint loss function termed crack\_loss[29][30] was designed by combining weighted Binary Cross-Entropy (BCE) and Dice loss. The weighted BCE loss applies additional weights to the positive samples (crack pixels) to compensate for this class imbalance:

$$w(y) = y \cdot (\alpha - 1) + 1 \quad (11)$$

$y \in \{0,1\}$  denotes the pixel label, and  $\alpha = 5.0$  represents the weight coefficient for positive samples. Specifically, a weight of  $\alpha = 5.0$  is assigned when the pixel is a crack, whereas a weight of 1.0 is assigned to background pixels, thereby imposing a greater penalty on the model for the misclassification of crack pixels during the training process. The weighted BCE loss is defined as follows:

$$L_{\text{BCE}} = -\frac{1}{N} \sum_i w(y_i) [y_i \log \hat{y}_i + (1 - y_i) \log (1 - \hat{y}_i)] \quad (12)$$

The Dice loss directly optimizes the overlap between the predicted and ground-truth masks, demonstrating greater robustness against small targets and extreme class imbalance:

$$L_{\text{Dice}} = 1 - \frac{2 \sum_i \hat{y}_i y_i + \epsilon}{\sum_i \hat{y}_i + \sum_i y_i + \epsilon} \quad (13)$$

The final joint loss function is defined as the equally weighted sum of these two components:

$$L = 0.5 \cdot L_{\text{BCE}} + 0.5 \cdot L_{\text{Dice}} \quad (14)$$

This joint loss is applied simultaneously to both the global branch output  $\hat{y}_{\text{global}}$  and the RoI branch output  $\hat{y}_{\text{RoI}}$ ; thus, the total training loss is given by:

$$L_{\text{total}} = L(\hat{y}_{\text{global}}, y) + \beta \cdot L(\hat{y}_{\text{RoI}}, y_{\text{RoI}}) \quad (15)$$

$\beta = 0.5$  denotes the loss weight for the RoI branch, serving to balance the gradient contributions from both branches.

## 2.5. Inference Fusion Strategy

Following the completion of training, the inference phase employs the following three-step fusion pipeline:

Step 1: RoI localization via YOLO detection. An improved YOLOv11 detection algorithm is executed on the input image to acquire a list of candidate crack bounding boxes with confidence scores exceeding a

predefined threshold. In the absence of positive detections, the process degrades to utilizing the global U-Net output solely as the final prediction.

Step 2: Probability map fusion. For each YOLO-detected bounding box, the corresponding RoI is cropped from the original image and fed into the DualUNet. Upon acquiring the RoI probability map, it is restored to the global image space via coordinate mapping; for overlapping regions among multiple bounding boxes, the pixel-wise maximum is computed to generate the full-image RoI probability map,  $P_{RoI}^{full}$ . Concurrently, a single DualUNet inference is performed on the entire image to obtain  $P_{global}$ . The final probability map is subsequently derived through the following pixel-wise maximum fusion:

$$P_{final} = \max(w_g \cdot P_{global}, P_{RoI}^{full}) \quad (16)$$

Where  $w_g \in [0,1]$  represents the scaling weight applied to the global probability, which serves to suppress false-positive predictions from the global branch. This enables the high-confidence local predictions from the RoI branch to effectively override and correct the global results.

Step 3: Binarization and small connected-component filtering. The fused probability map is binarized using a specified threshold, and an 8-connected component analysis is subsequently employed to filter out isolated noise regions with an area smaller than a predefined number of pixels:

$$M_{final} = \text{Filter}_{S_{min}}(\mathbb{I}(P_{final} > \tau)) \quad (17)$$

### 3. Experiments and Result Analysis

#### 3.1. Experimental Environment

All model optimization and training experiments in this study were conducted within a unified experimental environment. The operating system was Windows 11. The hardware configuration included a 12 vCPU Intel Xeon Platinum 8352V processor (base frequency 2.10 GHz), a single vGPU-32GB graphics card (32 GB VRAM), and 90 GB of system memory. Python 3.9 was utilized as the programming language, alongside PyTorch 2.1.0 as the deep learning framework for training and inference, supported by CUDA version 11.7. The construction of the segmentation network encoder and the loading of ImageNet pre-trained weights were implemented using the

segmentation\_models\_pytorch (smp) library. The object detection component was based on YOLOv11, released by Ultralytics. Automated hyperparameter searching was conducted using the Optuna framework[34] (based on the Tree-structured Parzen Estimator, or TPE, algorithm), while real-time metric monitoring and visual recording during the training process were facilitated by the Weights & Biases (WandB) platform. Dataset loading and data format conversion were accomplished using libraries such as torchvision, opencv-python, and numpy, and the experimental results were visually presented utilizing the matplotlib library.

The publicly available Crack500 dataset was employed as the experimental benchmark. This dataset comprises 500 images of real-world pavement and bridge cracks, each accompanied by a pixel-level, manually annotated binary segmentation mask. It encompasses various crack morphologies with diverse widths and orientations. Given that crack pixels account for an average of approximately 1% of the total image area, this extreme imbalance between positive and negative samples makes it an ideal benchmark for validating class-imbalance processing strategies. The dataset was partitioned into a training set (350 images), a validation set (50 images), and a test set (100 images) at a ratio of 7:1:2. To accommodate the data flow requirements of the two-stage training paradigm, bounding boxes encompassing the crack regions were automatically generated via connected-component analysis, building upon the original pixel-level annotations of the Crack500 dataset. This process established YOLO-format object detection annotations, which were subsequently utilized for the supervised training of the first-stage detector. The Region of Interest (RoI) labels required for the second stage were obtained by performing comprehensive inference on the training set using the trained, improved YOLOv11. This methodology achieves a seamless integration of the data flow across the two training stages, thereby eliminating any additional reliance on manual RoI annotations for the second-stage training.

#### 3.2. Evaluation Metrics

The experiment employs the following five quantitative metrics to comprehensively evaluate the model's performance:

**Precision:** This metric measures the proportion of pixels predicted as cracks that are genuinely ground-truth cracks, thereby reflecting the model's capability to suppress false-positive predictions:

$$\text{Precision} = \frac{TP + \varepsilon}{TP + FP + \varepsilon} \quad (18)$$

**Recall:** This metric measures the proportion of genuine ground-truth crack pixels that are correctly predicted, reflecting the model's capability to effectively cover the crack regions:

$$\text{Recall} = \frac{TP + \varepsilon}{TP + FN + \varepsilon} \quad (19)$$

**Dice Coefficient:** Equivalent to the F1 score, this metric calculates the harmonic mean of precision and recall. It serves as the most commonly utilized comprehensive performance indicator in segmentation tasks:

$$\text{Dice} = \frac{2 \cdot TP + \varepsilon}{2 \cdot TP + FP + FN + \varepsilon} \quad (20)$$

**Intersection over Union (IoU):** This metric directly measures the degree of regional overlap between the predicted mask and the ground truth mask, exhibiting sensitivity to both over-segmentation and under-segmentation:

$$\text{IoU} = \frac{TP + \varepsilon}{TP + FP + FN + \varepsilon} \quad (21)$$

**Boundary Intersection over Union (Boundary IoU):** This metric calculates the Intersection over Union between the boundary pixel set of the predicted mask,  $\hat{\partial M}$ , and the boundary pixel set of the ground-truth mask,  $\partial M_{gt}$ . The boundary extraction is implemented via a convolutional difference operation: the mask is convolved with an all-ones kernel, and pixels where the convolution result is strictly less than the sum of the kernel elements, while the original mask value remains 1, are identified as boundary pixels. Subsequently, a  $3 \times 3$  max-pooling operation is applied to perform a 1-to-2 pixel dilation, thereby increasing the tolerance for boundary matching:

$$\text{Boundary IoU} = \frac{|\hat{\partial M} \cap \partial M_{gt}| + \varepsilon}{|\hat{\partial M} \cup \partial M_{gt}| + \varepsilon} \quad (22)$$

Traditional IoU and Dice metrics treat the internal and boundary regions of a mask equally; consequently, when the crack width is substantial, the correct prediction of internal pixels often obscures precision defects at the boundaries. Boundary IoU strictly confines its evaluation focus to the mask boundaries, enabling it to more sensitively reveal discrepancies in the segmentation quality of crack contours. Compared to regional IoU, it provides a more rigorous quantitative perspective that better aligns with practical engineering applications.

### 3.3. Model Performance Evaluation

#### 3.3.1. Qualitative Visual Analysis

Figure 5 illustrates a visual comparison of two representative samples across various intermediate stages. From left to right, these include the original image, ground truth mask, YOLO boxes overlay, RoI mask overlay (green), global UNet overlay (blue), and the final fusion mask overlay (red).

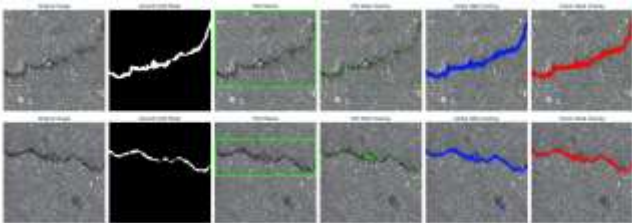
The operational mechanism of the two-stage fusion framework and the complementary nature of its constituent branches are distinctly observable. In the first-row sample, where the crack exhibits an arcuate trajectory with non-uniform width, the YOLO bounding boxes accurately localize the primary crack region, thereby furnishing an effective spatial prior for the RoI branch. Within these YOLO-delineated local regions, the RoI branch (green mask) executes a fine-grained segmentation of the crack contour, yielding sharply defined edges.

Conversely, while the global U-Net branch (blue mask) successfully captures the overall crack trajectory from a macroscopic perspective, its segmentation output suffers from noticeably overly broad boundaries. This corroborates the analysis in the introduction regarding the inadequate capacity of global segmentation models to capture local details. The fusion mask (red) synergizes the local precision of the RoI branch with the continuous integrity of the global branch. By significantly narrowing the predicted width while maintaining the completeness of the main crack backbone, the fusion

result achieves a high degree of concordance with the ground truth.

In the second row of samples, the crack extends to the edge of the image and exhibits a slender morphology, accompanied by minor aggregate noise interference in the background. The YOLO bounding box successfully localizes the main crack region. Within the Region of Interest (RoI), the RoI branch achieves relatively precise fine-line segmentation, effectively suppressing the interference from background aggregates. Although the global U-Net branch (blue mask) roughly captures the location of the crack, several small-area noise prediction regions emerge beneath it, reflecting the global branch's propensity for false positives in complex background areas. During the fusion stage, the global probability down-weighting strategy ( $w_{global} < 1$ ) effectively mitigates this false-positive noise. The final fused mask (red) successfully eliminates the noise while preserving the complete prediction of the main crack. Following a small connected component filtering process ( $A_{min} = 10$  pixels), the final result is substantially cleaner, devoid of obvious isolated noise artifacts.

These two sets of visualized samples substantiate the core design philosophy of the fusion framework: the localization prior provided by the YOLO detector effectively constrains the segmentation scope, thereby preventing the divergence of the global segmentation; the fine-grained segmentation from the RoI branch supplies high-quality local masks; the global branch guarantees the overall continuity of the cracks; and the adaptive probability fusion achieves an optimal balance between precision and recall across both branches.



**Figure 5. Comparison of typical sample visualization results of two-stage fusion reasoning**

### 3.3.2. Quantitative Result Analysis

To systematically validate the contribution of each core module, the following three sets of ablation experiments were designed by progressively integrating these modules. Comparative training was conducted

under identical training configurations and datasets, utilizing the fused inference results across five metrics as the basis for evaluation. To ensure fair comparison, all variants in the experiments employed the exact same Crack500 dataset splits and training hyperparameters, differing solely in their model architectures.

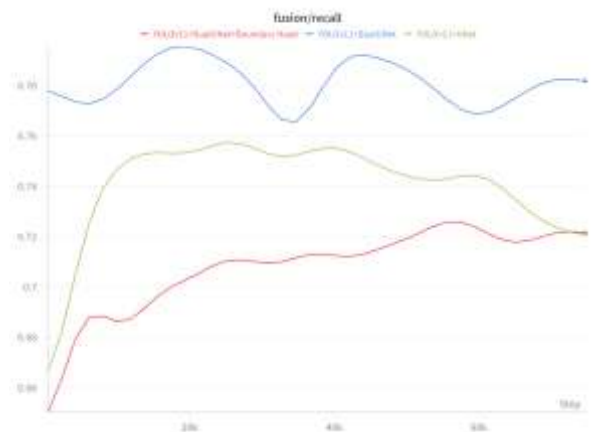
**Baseline Model (YOLO + U-Net):** A loosely coupled two-stage baseline method, consisting of a standard YOLO detector and a conventional single-output U-Net, serves as the comparative starting point. The U-Net lacks an RoI branch and relies exclusively on the global-image branch output for segmentation prediction. The two models were trained independently.

**Dual-Branch Model (YOLO + DualUNet):** Building upon the baseline, the U-Net is upgraded to the proposed DualUNet dual-branch architecture. This introduces a shared ResNet34 encoder, an RoI branch, and an RoI Head module, facilitating the joint training of both the global-image and RoI branches, alongside the fusion inference of their probability maps.

**Boundary Head Mechanism (YOLO + DualUNet + Boundary Head):** Expanding upon the dual-branch model, a Boundary Head supervised training mechanism is further incorporated. By applying specific supervisory signals to the mask boundary regions during the training process, this mechanism guides the network to more precisely delineate crack boundary contours while simultaneously optimizing the overall segmentation quality.

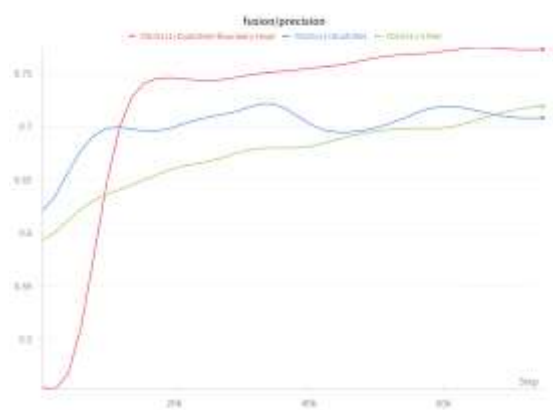
**Recall:** As illustrated in Figure 6, the dual-branch model maintains the highest recall throughout the entire training cycle, ultimately converging to approximately 0.79, which significantly outperforms the baseline model. The recall of the boundary-head model converges to roughly 0.72, lower than that of the dual-branch variant. This means that the DualUNet architecture improves the coverage capacity for fine crack regions by combining the global and RoI branches in a way that works together. While the introduction of the boundary head bolsters the precision of boundary delineation, it simultaneously induces the network to favor more conservative predictions in high-confidence boundary regions. This results in a marginal decrease in recall, despite yielding a substantial improvement in edge precision (a 21% increase in Boundary IoU). In practical engineering contexts, the

binarization threshold  $\tau$  can be optimized via Optuna to moderately relax the prediction criteria while preserving edge precision, thereby recovering a portion of the false-negative (missed detection) regions.



**Figure 6. Convergence curve of model recall**

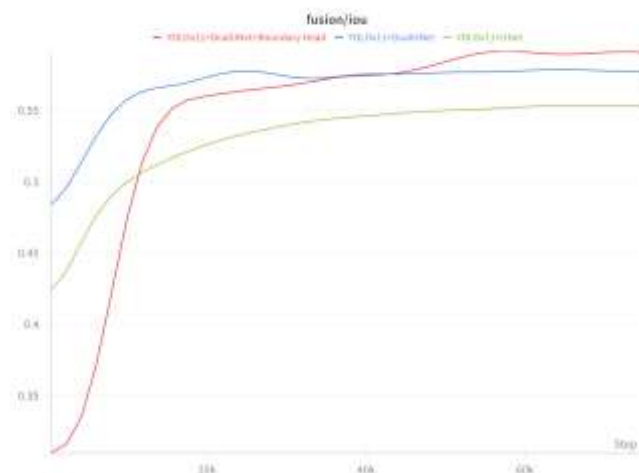
**Precision:** As depicted in Figure 7, the boundary-head model achieves the highest final precision, reaching 0.75. The precision scores of the dual-branch model and the baseline model are comparable, with both ultimately converging around 0.70 to 0.71. This demonstrates that the boundary-head supervision mechanism plays a constructive role in elevating prediction confidence and suppressing false positives, thereby enabling the network to predict genuine crack regions with greater accuracy.



**Figure 7. Convergence curve of model precision**

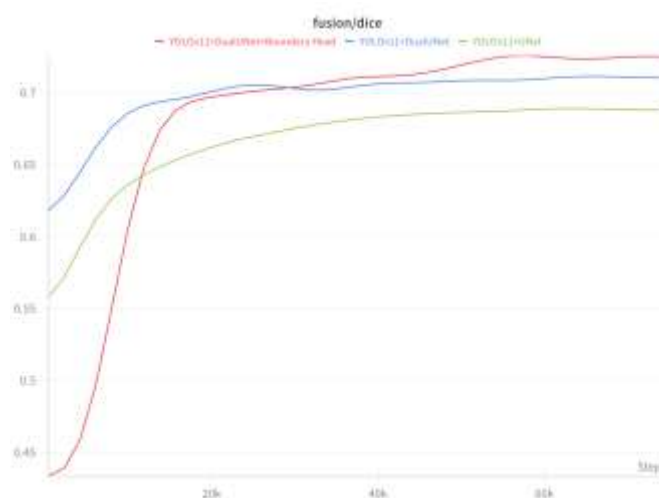
**Intersection over Union (IoU):** As illustrated in Figure 8, the final IoU values across the three models are comparable. Both the dual-branch and boundary-head variants achieve approximately 0.58, slightly outperforming the baseline model. The dual-branch and boundary-head models are largely

equivalent in terms of the regional overlap metric, demonstrating consistent improvements over the baseline. This validates the systematic enhancement of overall segmentation quality facilitated by the dual-branch architecture.



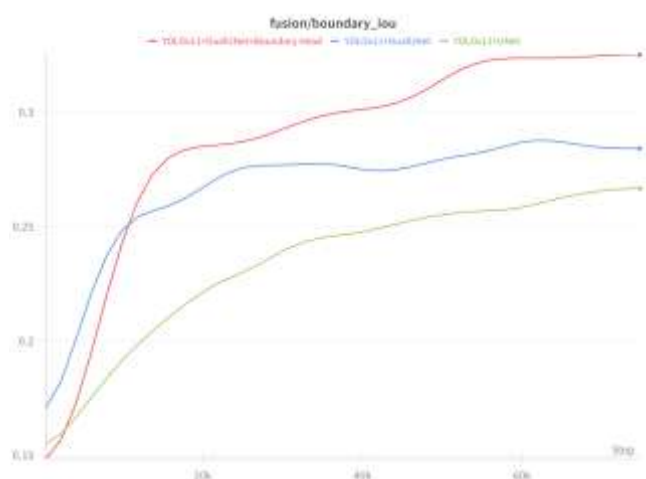
**Figure 8. Variation curve of model regional IoU**

**Dice Coefficient:** As shown in Figure 9, the boundary-head model achieves the highest final Dice coefficient at approximately 0.71, followed by the dual-branch model at roughly 0.70, and the baseline model at approximately 0.69. Although the numerical differences are marginal, the boundary-head model's leading position in the comprehensive F1 score aligns with its superior precision, reflecting the positive contribution of boundary supervision to the overall segmentation quality.



**Figure 9. Convergence curve of the model Dice coefficient**

Boundary Intersection over Union (Boundary IoU): As depicted in Figure 10, the boundary-head model leads with a final value of approximately 0.34, followed by the dual-branch model at roughly 0.28, while the baseline model ranks lowest at approximately 0.27. The boundary-head model exhibits a Boundary IoU improvement of approximately 26% over the baseline model and 21% over the dual-branch model; this constitutes the most significant relative enhancement among all five evaluated metrics. This result robustly substantiates the targeted contribution of the boundary-head supervision mechanism to crack boundary precision. By explicitly enforcing supervision on the boundary regions during the training phase, the model achieves a substantial, highly specific improvement in boundary segmentation accuracy during testing.



**Figure 10. Convergence curve of model Boundary IoU**

Table 1 clearly quantifies the specific contributions of each core module to the overall segmentation performance. Compared to the baseline model (YOLOv11+UNet), the recall of the dual-branch model (YOLOv11+DualUNet) exhibits a substantial leap from 0.75 to 0.79. This demonstrates that the synergistic strategy—integrating global semantic constraints with local Region of Interest (RoI) feature enhancement—effectively compensates for the perceptual blind spots inherent in single-model architectures, thereby significantly mitigating the false-negative rate for fine cracks.

Furthermore, following the introduction of the boundary head supervision mechanism to the dual-branch model, the recall undergoes a minor regression to

0.72, a consequence of the model's more stringent and conservative predictions within edge regions. Nevertheless, this configuration attains global optima in both precision (elevated to 0.75) and the Dice coefficient (elevated to 0.71). The Boundary IoU metric, which evaluates edge alignment, surges to 0.34, yielding substantial relative gains of approximately 26% and 21% over the baseline and dual-branch models, respectively.

In conclusion, the dual-branch model successfully secures the macroscopic coverage of minute cracks, whereas the boundary head mechanism facilitates precise contour delineation at the microscopic, pixel level. These two mechanisms are highly complementary; through the further optimization of post-processing strategies such as the binarization threshold, this fusion framework is capable of achieving a robust equilibrium between "high coverage" and "precise edges" in practical engineering applications characterized by complex backgrounds.

**Table 1. Comparison of segmentation performance of models at different stages in the ablation study on the validation set**

Model	Precision	Recall	Dice	Iou	Boundary Iou
YOLOv11+UNet	0.70	0.75	0.69	0.57	0.27
YOLOv11+DualUNet	0.71	0.79	0.70	0.58	0.28
YOLOv11+DualUNet+Boundary Head	0.75	0.72	0.71	0.58	0.34

## 4. Conclusion

To address the technical bottlenecks of existing methods concerning the missed detection of fine targets, insufficient pixel-level segmentation accuracy, and coarse crack boundary delineation, this paper systematically designs and implements a two-stage collaborative recognition framework characterized by "macroscopic localization guiding local fine segmentation." This framework achieves comprehensive performance surpassing the comparative baselines on the Crack500 dataset.

Regarding the improvements to the first-stage detector, to tackle the structural issue where the standard

YOLO series Feature Pyramid Network (FPN) architecture exhibits a perceptual blind spot for fine cracks at the P3 scale, a P2 small-object detection layer is introduced at the tail end of the feature self-fusion (i.e., the top of the FPN). This addition enhances the detection resolution for exceptionally fine cracks. Furthermore, a C2PSA channel attention module is integrated at the C2 feature level. This integration bolsters the model's representational capacity for ultra-fine targets such as narrow cracks, elevating the detector's perceptual downsampling scale to 1/4, thereby enabling the effective detection of crack targets with a width of merely around 4 pixels. The improved YOLOv11 provides a high-confidence, high-localization-accuracy Region of Interest (RoI) spatial prior for the second-stage segmentation task, ensuring the reliability of the two-stage collaborative pipeline from the source of information quality.

In the second stage, the core advantage of the proposed dual-branch DualUNet architecture lies in achieving efficient synergy between full-image macroscopic semantics and local RoI fine segmentation via a shared backbone, significantly enhancing parameter utilization and generalization capabilities. The global branch is responsible for providing global semantic constraints to prevent the RoI branch from losing contextual information due to local focusing. Meanwhile, the specially designed MSSE module within the RoI branch captures the directional geometric features of cracks through a multi-directional strip convolution structure, thereby substantially amplifying the segmentation response to boundary and width variations. Confronted with the extreme imbalance between positive and negative samples (approximately 1:100), a joint loss function comprising Weighted Binary Cross-Entropy (BCE) and Dice Loss is adopted to guarantee the stability of the crack pixel recall rate from the dual perspectives of gradient contribution and regional overlap.

Experimental results demonstrate that the dual-branch architecture increases the recall rate from 0.75 to 0.79, achieving a notable improvement of 5.3%. The boundary head supervision mechanism elevates precision from 0.70 to 0.75, representing a 7.1% increase. Moreover, the Boundary Intersection over Union (Boundary IoU) substantially surges from 0.27 to

0.34, yielding a proportional enhancement of approximately 26%. These two categories of technical improvements form a complementary relationship in the performance dimension, collectively realizing a systematic leap in the overall performance of the framework. This provides a novel perspective and a reliable technical paradigm for the intelligent inspection and structural safety assessment of bridges, laying a foundation for the future realization of real-time, automated bridge health monitoring.

Future work will further explore the following directions: conducting cross-domain validation on larger-scale, multi-scenario bridge crack datasets to fortify the model's generalization capability; and integrating 3D point clouds with multi-modal information to achieve cross-modal crack detection and comprehensive structural defect evaluation.

## Declarations

### *Author Contributions*

#### **The following statements should be used:**

Conceptualization, Z.W. and J.S.; methodology, Z.W.; software, Z.W. and Y.L.; validation, Z.W., H.C. and J.Z.; formal analysis, Z.W. and Y.L.; investigation, H.C. and J.Z.; resources, J.S.; data curation, Z.W.; writing—original draft preparation, Z.W.; writing—review and editing, J.S. and Y.L.; visualization, Z.W. and H.C.; supervision, J.S.; project administration, J.S. All authors have read and agreed to the published version of the manuscript.

### *Data Availability Statement*

• Data available in a publicly accessible repository that does not issue DOIs: Publicly available datasets were analyzed in this study. This data can be found here: [Yaan-Wang/Pavement-Defect-Datasets](#). The Crack500 dataset is publicly available and can be accessed through its corresponding open-source repositories.

### *Funding*

This research was funded by Research on Key Technologies of Digital Supervision, Testing and Its Management & Control Platform, grant number 2023-21.

### *Acknowledgements*

The authors would like to acknowledge the technical and organizational support provided by the Hunan Provincial Administration of Quality and Safety Supervision for Transportation Construction and Hunan CCCC Jingwei Information Technology Co., Ltd.

### Conflicts of Interest

The author declares that there is no conflict of interests regarding the publication of this manuscript. In addition, the ethical issues, including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, and redundancies, have been completely observed by the authors.

### References

- [1] Otsu, N. (1979). A threshold selection method from gray-level histograms. *Automatica*, 11(285-296).
- [2] Cha, Y. J., Choi, W., & Büyüköztürk, O. (2017). Deep learning-based crack damage detection using convolutional neural networks. *Computer-Aided Civil and Infrastructure Engineering*, 32(5), 361-378. <https://doi.org/10.1111/mice.12263>
- [3] Koch, C., Georgieva, K., Kasireddy, V., Akinci, B., & Fieguth, P. (2015). A review on computer vision based defect detection and condition assessment of concrete and asphalt civil infrastructure. *Advanced engineering informatics*, 29(2), 196-210. <https://doi.org/10.1016/j.aei.2015.01.008>
- [4] Huyan, J., Li, W., Tighe, S., Xu, Z., & Zhai, J. (2020). CrackU-net: A novel deep convolutional neural network for pixelwise pavement crack detection. *Structural Control and Health Monitoring*, 27(8), e2551. <https://doi.org/10.1002/stc.2551>
- [5] Liu, H., Miao, X., Mertz, C., Xu, C., & Kong, H. (2021). Crackformer: Transformer network for fine-grained crack detection. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 3783-3792).
- [6] Liu, Y., & Yeoh, J. K. (2021). Robust pixel-wise concrete crack segmentation and properties retrieval using image patches. *Automation in Construction*, 123, 103535. <https://doi.org/10.1016/j.autcon.2020.103535>
- [7] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 779-788).
- [8] Sun, S., Liu, W., & Cui, R. (2022, July). YOLO based bridge surface defect detection using decoupled prediction. In *2022 7th Asia-Pacific Conference on Intelligent Robot Systems (ACIRS)* (pp. 117-122). IEEE. 10.1109/ACIRS55390.2022.9845546
- [9] Terven, J., Córdova-Esparza, D. M., & Romero-González, J. A. (2023). A comprehensive review of yolo architectures in computer vision: From yolov1 to yolov8 and yolo-nas. *Machine learning and knowledge extraction*, 5(4), 1680-1716.
- [10] Xu, W., Li, H., Li, G., Ji, Y., Xu, J., & Zang, Z. (2025). Improved YOLOv8n-based bridge crack detection algorithm under complex background conditions. *Scientific Reports*, 15(1), 13074. <https://doi.org/10.1038/s41598-025-97842-2>
- [11] LI, J., MENG, X., HU, L., BAO, Y., & ZHAO, S. (2025). Bridge small target crack detection based on improved YOLOv8. *Journal of Tsinghua University (Science and Technology)*, 65(7), 1260-1271. 10.16511/j.cnki.qhdxxb.2025.26.023
- [12] Ren, W., & Zhong, Z. (2025). Building construction crack detection with BCCD YOLO enhanced feature fusion and attention mechanisms. *Scientific Reports*, 15(1), 23167. <https://doi.org/10.1038/s41598-025-05665-y>
- [13] Wang, N., Huang, S., Liu, X., Wang, Z., Liu, Y., & Gao, Z. (2025). MRA-YOLOv8: a network enhancing feature extraction ability for photovoltaic cell defects. *Sensors*, 25(5), 1542.
- [14] Xu, T., Zhang, G., Ruan, Y., Xu, H., Lu, R., & Lin, J. (2025). An improved YOLOv8 by fusing a coordinate attention mechanism and a bidirectional feature pyramid network for identifying power repair vehicles in the cable terminal field. *International Journal of Parallel, Emergent and Distributed Systems*, 1-18. <https://doi.org/10.1080/17445760.2025.2518139>
- [15] Khanam, R., & Hussain, M. (2024). Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725*. <https://doi.org/10.48550/arXiv.2410.17725>
- [16] Song, Y., Xing, L., Song, Y., & Li, J. (2025). Real-Time Detection and Monitoring of Structural Cracks Using ConcreteCrack. <https://doi.org/10.21203/rs.3.rs-7767819/v1>
- [17] Zhang, R., Guan, C., Fang, Y., Duan, Y., & Sui, X. (2026). A Two-Stage Concrete Crack Segmentation Method Based on the Improved YOLOv11 and Segment Anything Model. *Buildings*, 16(4), 794. <https://doi.org/10.3390/buildings16040794>
- [18] Gao, X., Cao, C., & Yi, X. (2025). Using the improved YOLOv11 model to enhance computer vision

- applications for building crack detection algorithms. *Scientific Reports*, 15(1), 38843. <https://doi.org/10.1038/s41598-025-22160-6>
- [19] Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention* (pp. 234-241). Cham: Springer international publishing. [https://doi.org/10.1007/978-3-319-24574-4\\_28](https://doi.org/10.1007/978-3-319-24574-4_28)
- [20] Yang, F., Zhang, L., Yu, S., Prokhorov, D., Mei, X., & Ling, H. (2019). Feature pyramid and hierarchical boosting network for pavement crack detection. *IEEE transactions on intelligent transportation systems*, 21(4), 1525-1535. 10.1109/TITS.2019.2910595
- [21] Yang, A., Chen, S., Yao, K., Huang, X., Wang, Y., Li, J., & Chen, Y. (2025, August). A Multi-scale Dilated Convolution Model with Edge Optimization for Crack Detection. In *Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data* (pp. 231-245). Singapore: Springer Nature Singapore. [https://doi.org/10.1007/978-981-95-5719-6\\_15](https://doi.org/10.1007/978-981-95-5719-6_15)
- [22] Liu, H., Miao, X., Mertz, C., Xu, C., & Kong, H. (2021). Crackformer: Transformer network for fine-grained crack detection. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 3783-3792).
- [23] Chen, S., Feng, Z., Xiao, G., Chen, X., Gao, C., Zhao, M., & Yu, H. (2024). Pavement crack detection based on the improved Swin-Unet model. *Buildings*, 14(5), 1442.
- [24] Xu, C., Zhang, Q., Mei, L., Chang, X., Ye, Z., Wang, J., ... & Yang, W. (2023). Cross-attention-guided feature alignment network for road crack detection. *ISPRS International Journal of Geo-Information*, 12(9), 382.
- [25] Du Nguyen, Q., & Thai, H. T. (2023). Crack segmentation of imbalanced data: The role of loss functions. *Engineering Structures*, 297, 116988. <https://doi.org/10.1016/j.engstruct.2023.116988>
- [26] Yang, E., Tang, Y., Zhang, A. A., Wang, K. C., & Qiu, Y. (2023). Policy gradient-based focal loss to reduce false negative errors of convolutional neural networks for pavement crack segmentation. *Journal of Infrastructure Systems*, 29(1), 04023002. <https://doi.org/10.1061/JITSE4.ISENG-215>
- [27] Chen, J. (2022, November). Optimized hybrid focal margin loss for crack segmentation. In *2022 International Conference on Digital Image Computing: Techniques and Applications (DICTA)* (pp. 1-7). IEEE. 10.1109/DICTA56598.2022.10034608
- [28] Chen, X., Shi, Y., & Pang, J. (2025). SECrackSeg: a high-accuracy crack segmentation network based on proposed UNet with SAM2 S-Adapter and edge-aware attention. *Sensors*, 25(9), 2642.
- [29] Rajput, V. (2021). Robustness of different loss functions and their impact on networks learning capability. *arXiv preprint arXiv:2110.08322*. <https://doi.org/10.48550/arXiv.2110.08322>
- [30] Yan, J., Wang, H., Yan, M., Diao, W., Sun, X., & Li, H. (2019). IoU-adaptive deformable R-CNN: Make full use of IoU for multi-class object detection in remote sensing imagery. *Remote Sensing*, 11(3), 286.
- [31] He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision* (pp. 2961-2969). Zhang, C., Chen, X., Liu, P., He, B., Li, W., & Song, T. (2024). Automated detection and segmentation of tunnel defects and objects using YOLOv8-CM. *Tunnelling and Underground Space Technology*, 150, 105857. <https://doi.org/10.1016/j.tust.2024.105857>
- [32] Xu, Y., Yan, S., Qi, Y., Ding, Z., & Zhang, D. (2025). CDIF-Net: cross-dimensional interactive fusion network with dual-branch attention for pavement crack segmentation. *Measurement Science and Technology*, 36(9), 095404. 10.1088/1361-6501/adfb9e
- [33] Zhang, L., Liao, Y., Wang, G., Chen, J., & Wang, H. (2022). A multi-scale contextual information enhancement network for crack segmentation. *Applied Sciences*, 12(21), 11135.
- [34] Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7132-7141).
- [35] Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019, July). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 2623-2631). <https://doi.org/10.1145/3292500.3330701>

## 参考文献:

- [1] Otsu, N. (1979). 基于灰度直方图的阈值选择方法。《Automatica》, 11, 285–296。
- [2] Cha, Y. J., Choi, W., & Büyüköztürk, O. (2017). 基于深度学习的裂缝损伤检测方法——卷积神经网络的应用。《Computer-Aided Civil and Infrastructure Engineering》, 32(5), 361–378。  
<https://doi.org/10.1111/mice.12263>
- [3] Koch, C., Georgieva, K., Kasireddy, V., Akinci, B., & Fieguth, P. (2015). 基于计算机视觉的混凝土与沥青基础设施缺陷检测与状况评估综述。《Advanced Engineering Informatics》, 29(2), 196–210。  
<https://doi.org/10.1016/j.aei.2015.01.008>
- [4] Huyan, J., Li, W., Tighe, S., Xu, Z., & Zhai, J. (2020). CrackU-net: 一种用于路面裂缝像素级检测的新型深度卷积神经网络。《Structural Control and Health Monitoring》, 27(8), e2551。  
<https://doi.org/10.1002/stc.2551>
- [5] Liu, H., Miao, X., Mertz, C., Xu, C., & Kong, H. (2021). Crackformer: 用于细粒度裂缝检测的Transformer 网络。载于 IEEE/CVF 国际计算机视觉会议论文集 (pp. 3783–3792)。
- [6] Liu, Y., & Yeoh, J. K. (2021). 基于图像块的鲁棒混凝土裂缝像素级分割与特征提取。《Automation in Construction》, 123, 103535。  
<https://doi.org/10.1016/j.autcon.2020.103535>
- [7] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You Only Look Once: 统一的实时目标检测方法。载于 IEEE 计算机视觉与模式识别会议论文集 (pp. 779–788)。
- [8] Sun, S., Liu, W., & Cui, R. (2022). 基于 YOLO 的桥梁表面缺陷检测 (采用解耦预测方法)。载于 2022 年第七届亚太智能机器人系统会议 (ACIRS) (pp. 117–122)。IEEE。  
10.1109/ACIRS55390.2022.9845546
- [9] Terven, J., Córdova-Esparza, D. M., & Romero-González, J. A. (2023). 计算机视觉中 YOLO 架构的综合综述: 从 YOLOv1 到 YOLOv8 及 YOLO-NAS。《Machine Learning and Knowledge Extraction》, 5(4), 1680–1716。
- [10] Xu, W., Li, H., Li, G., Ji, Y., Xu, J., & Zang, Z. (2025). 复杂背景条件下基于改进 YOLOv8n 的桥梁裂缝检测算法。《Scientific Reports》, 15(1), 13074。  
<https://doi.org/10.1038/s41598-025-97842-2>
- [11] Li, J., Meng, X., Hu, L., Bao, Y., & Zhao, S. (2025). 基于改进 YOLOv8 的桥梁小目标裂缝检测。《清华大学学报 (自然科学版)》, 65(7), 1260–1271。  
10.16511/j.cnki.qhdxxb.2025.26.023
- [12] Ren, W., & Zhong, Z. (2025). 基于 BCCD YOLO 及增强特征融合与注意力机制的建筑裂缝检测。《Scientific Reports》, 15(1), 23167。  
<https://doi.org/10.1038/s41598-025-05665-y>
- [13] Wang, N., Huang, S., Liu, X., Wang, Z., Liu, Y., & Gao, Z. (2025). MRA-YOLOv8: 一种增强光伏电池缺陷特征提取能力的网络。《Sensors》, 25(5), 1542。
- [14] Xu, T., Zhang, G., Ruan, Y., Xu, H., Lu, R., & Lin, J. (2025). 融合坐标注意力机制与双向特征金字塔网络的改进 YOLOv8 用于电缆终端场景中电力抢修车辆识别。《International Journal of Parallel, Emergent and Distributed Systems》, 1–18。  
<https://doi.org/10.1080/17445760.2025.2518139>
- [15] Khanam, R., & Hussain, M. (2024). YOLOv11: 关键架构改进综述。arXiv 预印本 arXiv:2410.17725。  
<https://doi.org/10.48550/arXiv.2410.17725>
- [16] Song, Y., Xing, L., Song, Y., & Li, J. (2025). 基于 ConcreteCrack 的结构裂缝实时检测与监测。  
<https://doi.org/10.21203/rs.3.rs-7767819/v1>
- [17] Zhang, R., Guan, C., Fang, Y., Duan, Y., & Sui, X. (2026). 基于改进 YOLOv11 与 Segment Anything 模型的两阶段混凝土裂缝分割方法。《Buildings》, 16(4), 794。  
<https://doi.org/10.3390/buildings16040794>
- [18] Gao, X., Cao, C., & Yi, X. (2025). 基于改进 YOLOv11 模型的建筑裂缝检测计算机视觉应用研究。《Scientific Reports》, 15(1), 38843。  
<https://doi.org/10.1038/s41598-025-22160-6>
- [19] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: 用于生物医学图像分割的卷积神经网络。载于国际医学图像计算与计算机辅助干预会议 (pp. 234–241)。Springer。  
[https://doi.org/10.1007/978-3-319-24574-4\\_28](https://doi.org/10.1007/978-3-319-24574-4_28)
- [20] Yang, F., Zhang, L., Yu, S., Prokhorov, D., Mei, X., & Ling, H. (2019). 用于路面裂缝检测的特征金字塔与分层提升网络。《IEEE Transactions on Intelligent Transportation Systems》, 21(4), 1525–1535。
- [21] Yang, A., Chen, S., Yao, K., Huang, X., Wang, Y., Li, J., & Chen, Y. (2025). 一种基于多尺度空洞卷积与边缘优化的裂缝检测模型。载于亚太 Web 与 Web 时代信息管理联合国际会议 (pp. 231–245)。Springer。

- [22] Liu, H., Miao, X., Mertz, C., Xu, C., & Kong, H. (2021). Crackformer: 用于精细裂缝检测的 Transformer 网络。
- [23] Chen, S., Feng, Z., Xiao, G., Chen, X., Gao, C., Zhao, M., & Yu, H. (2024). 基于改进 Swin-Unet 模型的路面裂缝检测。《Buildings》, 14(5), 1442。
- [24] Xu, C., Zhang, Q., Mei, L., Chang, X., Ye, Z., Wang, J., ... & Yang, W. (2023). 基于交叉注意力引导特征对齐的道路裂缝检测网络。《ISPRS International Journal of Geo-Information》, 12(9), 382。
- [25] Du Nguyen, Q., & Thai, H. T. (2023). 不平衡数据下裂缝分割: 损失函数的作用。《Engineering Structures》, 297, 116988。
- [26] Yang, E., Tang, Y., Zhang, A. A., Wang, K. C., & Qiu, Y. (2023). 基于策略梯度的焦点损失函数用于减少卷积神经网络在路面裂缝分割中的漏检。《Journal of Infrastructure Systems》, 29(1), 04023002。
- [27] Chen, J. (2022). 用于裂缝分割的优化混合焦点边界损失函数。载于国际数字图像计算会议 (DICTA) (pp. 1–7)。IEEE。
- [28] Chen, X., Shi, Y., & Pang, J. (2025). SECrackSeg: 基于改进 UNet 与 SAM2 S-Adapter 及边缘感知注意力的高精度裂缝分割网络。《Sensors》, 25(9), 2642。
- [29] Rajput, V. (2021). 不同损失函数的鲁棒性及其对网络学习能力的影响。arXiv 预印本。
- [30] Yan, J., Wang, H., Yan, M., Diao, W., Sun, X., & Li, H. (2019). IoU 自适应可变形 R-CNN: 在遥感图像多类别目标检测中的应用。《Remote Sensing》, 11(3), 286。
- [31] He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask R-CNN。载于 IEEE 国际计算机视觉会议论文集 (pp. 2961–2969)。
- [32] Xu, Y., Yan, S., Qi, Y., Ding, Z., & Zhang, D. (2025). CDIF-Net: 一种基于双分支注意力的跨维交互融合网络用于路面裂缝分割。《Measurement Science and Technology》, 36(9), 095404。
- [33] Zhang, L., Liao, Y., Wang, G., Chen, J., & Wang, H. (2022). 一种多尺度上下文信息增强裂缝分割网络。《Applied Sciences》, 12(21), 11135。
- [34] Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-Excitation 网络。载于 IEEE 计算机视觉与模式识别会议论文集 (pp. 7132–7141)。
- [35] Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: 新一代超参数优化框架。载于 ACM SIGKDD 知识发现与数据挖掘会议 (pp. 2623–2631)。  
<https://doi.org/10.1145/3292500.3330701>

### Manuscript Information

Word count: 10,317 words (excluding references).

### Peer-Review Record

Fast-track status: Not fast-tracked.

First-round reviews received: 3 reports.

Revision cycles completed: 3 rounds.

Final version submitted: April 10, 2026

### Disclaimer / Publisher's Note

The statements, opinions, and data contained in this article are solely those of the authors and do not necessarily represent the views of the *Journal of Hunan University (Natural Sciences)* or its editorial team. The journal and its editors disclaim any responsibility for injury to persons or property resulting from any ideas, methods, instructions, or products referred to in the content of this article.