Computer Modeling in
Engineering & Sciences

Tech Science Press

https://doi.org/10.32604/cmes.2026.076415

ARTICLE

Attention Mechanisms and FFM Feature Fusion Module-Based Modification of the Deep Neural Network for Detection of Structural Cracks

Tao Jin[1,2], **Zhekun Shou**[1], **Hongchao Liu**[1,*] **and Yuchun Shao**[1]

[1]College of Civil Engineering, Zhejiang University of Technology, Hangzhou, China
[2]Department of Civil Engineering, Zhejiang University, Hangzhou, China

*Corresponding Author: Hongchao Liu. Email: leo1656515417@163.com

Received: 20 November 2025; Accepted: 22 January 2026; Published: 26 February 2026

ABSTRACT: This research centers on structural health monitoring of bridges, a critical transportation infrastructure. Owing to the cumulative action of heavy vehicle loads, environmental variations, and material aging, bridge components are prone to cracks and other defects, severely compromising structural safety and service life. Traditional inspection methods relying on manual visual assessment or vehicle-mounted sensors suffer from low efficiency, strong subjectivity, and high costs, while conventional image processing techniques and early deep learning models (e.g., U-Net, Faster R-CNN) still perform inadequately in complex environments (e.g., varying illumination, noise, false cracks) due to poor perception of fine cracks and multi-scale features, limiting practical application. To address these challenges, this paper proposes CACNN-Net (CBAM-Augmented CNN), a novel dual-encoder architecture that innovatively couples a CNN for local detail extraction with a CBAM-Transformer for global context modeling. A key contribution is the dedicated Feature Fusion Module (FFM), which strategically integrates multi-scale features and focuses attention on crack regions while suppressing irrelevant noise. Experiments on bridge crack datasets demonstrate that CACNN-Net achieves a precision of 77.6%, a recall of 79.4%, and an mIoU of 62.7%. These results significantly outperform several typical models (e.g., UNet-ResNet34, Deeplabv3), confirming their superior accuracy and robust generalization, providing a high-precision automated solution for bridge crack detection and a novel network design paradigm for structural surface defect identification in complex scenarios, while future research may integrate physical features like depth information to advance intelligent infrastructure maintenance and digital twin management.

KEYWORDS: Bridge crack diseases; structural health monitoring; convolutional neural network; feature fusion

1 Introduction

Bridges represent one of the most critical transportation infrastructures. However, influenced by factors such as heavy-duty traffic, environmental exposure, human activities, and the use of inferior materials, bridge structures inevitably experience performance deterioration. This poses a direct threat to their service safety and socio-economic benefits [1]. Therefore, conducting regular bridge condition assessments is crucial for ensuring long-term structural safety and service life. Furthermore, to prevent further deterioration and enable timely maintenance, accurately and promptly identifying damage, especially cracks, in bridge components is essential.

Presently, bridge condition evaluation is conducted through either traditional manual methods or automated distress inspection [2]. The manual approach relies on visual identification and necessitates inspectors physically accessing the structure, a process that is inherently time-consuming, laborious, and prone to subjectivity. Hence, the development of automated detection technologies [3] is imperative to enable faster and more accurate crack identification. Based on some non-destructive techniques, such as digital image correlation (DIC) technology, which is used for full-range strain and displacement measurements to evaluate fracture mechanics [4], and acoustic emission (AE) technology, it monitors the formation and expansion of cracks through stress wave analysis [5].

Intelligent crack detection systems have witnessed growing interest and application in recent years [6]. For instance, Guo et al. [7] developed a system employing ultra-HD optical, LiDAR, and thermal sensors to capture fine-grained fracture information. Despite this capability, the high configuration cost and limited operational scope of such vehicle-borne systems hinder their widespread adoption [2].

Conventionally, automated road-surface damage assessment has depended on computational imaging techniques like Gabor wavelet transformation [8], edge detection, pixel-intensity threshold segmentation [9], and texture analysis. These methods operate by detecting edge gradient and intensity variations to distinguish cracks from the background, subsequently obtaining them through threshold-based segmentation [2]. However, their performance is notably vulnerable to environmental interference, such as changing illumination, and often falters with varying camera configurations, thus limiting their practical utility [1,10]. These shortcomings underscore the necessity for a more robust, accurate, efficient, and cost-effective approach to the detection of bridge cracks.

The continuous innovation in machine learning, particularly deep learning, has positioned these algorithms as powerful and precise alternatives to conventional object recognition and image analysis methods. Their success in visual applications is increasingly being leveraged in bridge distress inspection [1,10]. Seminal work by Krizhevsky et al. [11] leveraged a deep CNN for visual categorization for this purpose. Subsequent research has diversified: Cao et al. [3] proposed an attention-based network (AC-Net) that surpassed eight other methods in accuracy on the CRACK500 dataset; Tran et al. [12] utilized RetinaNet to achieve 84.9% accuracy in detecting and classifying crack types and severity; and Xiao et al. [13] developed C-MaskRCNN, which boosted the mean average precision to 95.4%. Further contributions include Xu et al. [14], who enhanced Faster R-CNN for fine crack detection with 85.64% accuracy, and Xu et al. [15], whose comparative study found Faster R-CNN superior to Mask R-CNN in crack identification. A key strength of these deep learning approaches is their dual capability in both categorizing and precisely localizing objects within images [16], thereby reducing labor costs and enhancing the efficiency of crack identification [1].

Nevertheless, real-world bridge environments present significant challenges. Structural cracks exhibit diverse morphologies and locations due to variable illumination, shadows, and humidity. Captured images often feature complex backgrounds, inherent structural textures, pseudo-cracks, and crack-like impurities, which are visually similar to genuine cracks and complicate accurate identification. While current research often prioritizes improving image acquisition hardware and environments to obtain cleaner, high-resolution images, this direction tends to be impractical and lacks objectivity. For instance, Yao et al. [17] highlighted in their CrackNex study that traditional models often fail in low-light environments, necessitating complex preprocessing like Retinex theory to separate illumination components. Similarly, standard models often struggle with pseudo-cracks caused by water stains or complex textures, leading to high false-positive rates.

To overcome these limitations, this paper introduces CACNN-Net, a crack segmentation framework that employs a dual-encoder architecture enhanced by the Convolutional Block Attention Module (CBAM), designed to harness the complementary strengths of CNN and the CBAM. The architecture first constructs a CNN encoder by adapting the initial and subsequent layers of ResNet-50 [18]. A parallel transformer encoder is then implemented, incorporating high- and low-frequency attention mechanisms alongside a locally enhanced feedforward network. Finally, a dedicated feature synthesis layer integrates the intermediate features from both encoders, and the fused output is passed to a decoder to achieve precise image reconstruction and segmentation mask generation.

CACNN-Net distinguishes itself from other networks [19] through its innovative integration of convolutional neural networks and CBAM. In response to the limitations of existing methods in bridge crack detection, such as insufficient domain coverage and insufficient feature integration, this study proposes a novel CACNN-Net. The core innovation lies in the design of a collaborative attention fusion framework specifically for crack morphology: through the dual encoders of CNN and CBAM-Transformer to extract local and global features respectively, and through a new feature fusion module for deep interaction and recalibration, to achieve continuous enhancement of crack features and noise suppression. This architecture provides a more robust solution for structural surface defect detection in complex environments and has achieved leading performance on the bridge crack dataset. The key contributions of this study are outlined below:

1. We devise a crack segmentation architecture with dual encoders—termed CACNN-Net—in which the convolutional backbone of ResNet-50 extracts local details and a companion CBAM encoder encodes global context; the two branches act in a mutually reinforcing manner.
2. To integrate intermediate representations from both encoders, a specialized Feature Fusion Module (FFM) [20] is introduced. Channel Attention first recalibrates channel-wise importance, after which a Cross-domain Fusion Block (CFB) and a correlation-enhancement step refine the features, culminating in the final fused map via a Feature Fusion Block (FFB).
3. Comprehensive experiments were conducted on the publicly accessible "Bridge Crack Library" dataset. The proposed network was benchmarked against five state-of-the-art algorithms, and their respective strengths and weaknesses were scrutinized through an exhaustive evaluation of all metrics.

2 Related Work

2.1 CNN Architecture Overview

The lightweight design of CNNs and the continuous optimization of feature extraction efficiency have been significantly advanced by depthwise separable convolution (DSC), a pivotal breakthrough in this field.

Sifre and Mallat [21] constructed a translation- and rotation-invariant deep scattering network, pioneering the separation of spatial convolution from channel mixing (DSC). This strategy significantly reduced computational complexity while achieving high-precision texture classification. Building on this foundation, architectures such as Xception [22] and MobileNet [23] achieved efficient inference performance. Subsequent research introduced inverted depthwise separable convolution (IDSC), which reverses the operational sequence (PW→DW) to generate parameter-efficient designs, opening new directions for lightweight model development. Concurrently, upsampling mechanisms have evolved from transposed convolutions [24] to sub-pixel convolutions [25]. The latter enhances edge reconstruction accuracy in segmentation tasks through multi-channel pixel reorganization, effectively mitigating padding-induced artifacts during resolution recovery.

In the field of crack segmentation, the lightweight properties of depthwise separable convolution (DSC) have been widely adopted. SDDNet [26], centered on DSC as its core module, achieves real-time detection while reducing model parameters by 70%. DeepCrack [27], meanwhile, enhances robustness against complex backgrounds through multi-scale feature map fusion to strengthen crack continuity perception. In recent studies. Similarly, Zim et al. [28] balanced computational efficiency and segmentation accuracy by integrating depthwise separable convolutions (DSC) with MobileViT blocks, and utilized an Ultra-Lightweight Subspace Attention Module (ULSAM) to enhance feature extraction capabilities in noise-heavy environments.

Network approaches based on super-resolution reconstruction have been actively studied. Convolutional Neural Networks (CNNs) remain widely employed due to their effectiveness in extracting spatial information of targets [29]. However, in high-precision tasks such as crack segmentation, straightforward stacked convolutional architectures tend to lose detailed information. Specifically, CNNs struggle to accurately model long-range dependencies inherent in slender crack structures. This challenge has motivated the development of specialized architectures for infrastructure inspection, such as lightweight dual-encoder networks for pavement cracks [30] and enhanced detectors for underground engineering [31]. Building upon and differentiating from these approaches.

2.2 Convolutional Block Attention Module CBAM

Bridge surface imagery is frequently compromised by strong background clutter—including surface textures, water marks, and various artifacts—that obscure actual cracks. Cracks usually span only a minute fraction of the image area, and their morphological characteristics can vary considerably. Although the improved lightweight network proposed in this study is based on U-Net—an architecture known for its feature concatenation operations that effectively reduce information loss during segmentation—the extensive use of dimensionality reduction in the MobileNet v3 backbone still inevitably incurs a degradation of detail in both spatial and channel dimensions.

To alleviate this problem, this paper introduces the CBAM [32], an innovative attention mechanism combining channel-wise and spatially oriented attention. This lightweight yet efficient module is well-suited for enhancing the performance of lightweight networks [33]. As illustrated in Fig. 1, the CBAM incorporates two sequential sub-modules: CAM performs channel-wise recalibration, while SAM focuses on spatially localised re-weighting.

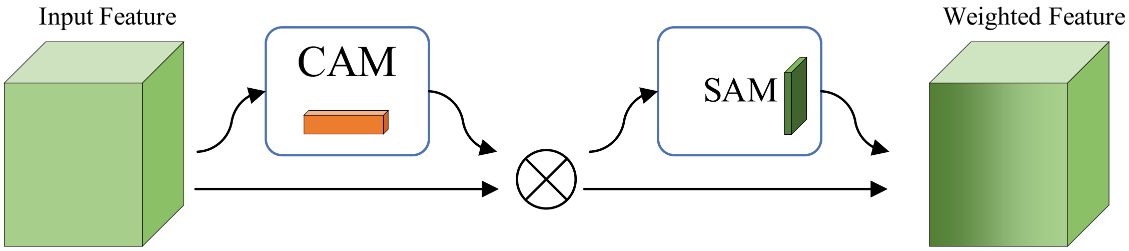

Figure 1: Structural diagram of CBAM attention mechanism.

Fig. 2 outlines the internal structure of the CAM. Given an input feature map F of size C × H × W, it goes through both Max Pooling and Global Average Pooling [34], yielding two 1 × 1 × C feature vectors, denoted as F_{Max} and F_{Avg}. The simultaneous use of both pooling strategies allows the module to not only effectively localize the regions of interest but also capture discriminative information between different objects, thereby enabling more precise weight assignment.

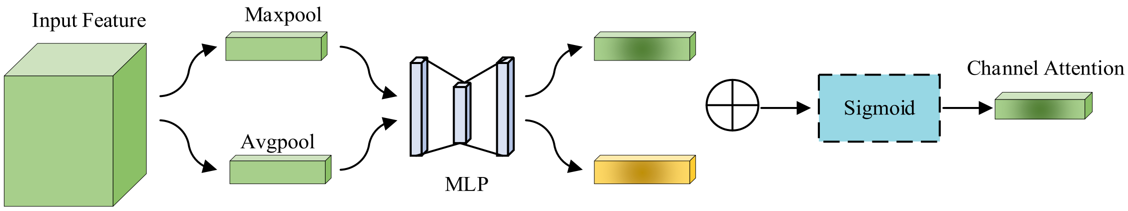

Figure 2: Structural diagram of CAM module.

This module employs a shared two-layer MLP (Multilayer Perceptron, layer dimensions: C/r and C) to transform the pooled features. The transformed features from both branches are fused by element-wise addition. A Sigmoid activation function then converts the fused result into the channel attention map $M_c(F)$, as defined by Eq. (1). The original input features are recalibrated through an element-wise multiplication with $M_c(F)$, and the output is subsequently fed into the Spatial Attention Module (SAM).

$$M_c\left(F\right) = \sigma\left(MLP\left(Avgpool\left(F\right)\right) + MLP\left(MaxPool\left(F\right)\right)\right) \tag{1}$$

As shown in Fig. 3, SAM operates on the channel-refined feature map by first performing max- and average-pooling along the channel axis, yielding two single-channel descriptors that are concatenated into a two-channel tensor. A 7 × 7 convolution, $f^{7\times 7}$, is then applied to encode local spatial correlations and accentuate salient regions before a Sigmoid activation generates the final spatial attention mask $M_s(F)$, as defined in Eq. (2).

$$M_s\left(F\right) = \sigma\left(f^{7\times 7}\left(\left[Avgpool\left(F\right)\right); MaxPool(F)\right]\right) \tag{2}$$

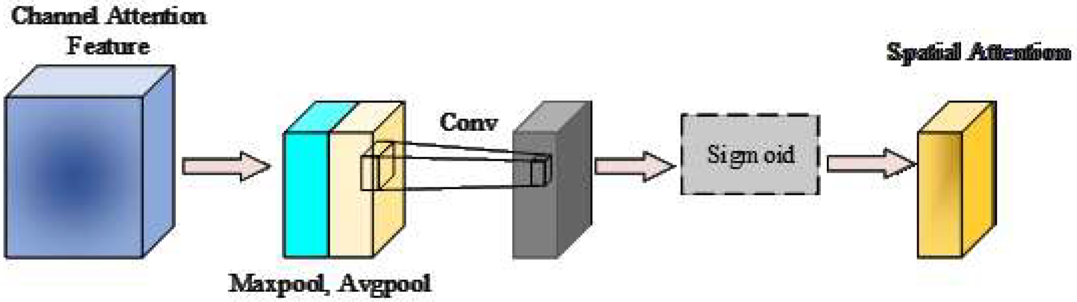

Figure 3: Structural diagram of SAM module.

CBAM leverages a sequential application of channel and spatial attention, thereby boosting the representational power of features for accurate crack detection. Through the channel attention module, the network automatically adjusts the weights of feature channels, emphasizing crack-related characteristics while suppressing irrelevant ones, thereby improving the expressiveness of crack regions. Meanwhile, the spatial attention module enhances features at critical locations by focusing on key areas within the image, reducing interference from background noise, and achieving more accurate spatial localization.

In crack detection tasks, cracks often exhibit thin, local, and background-similar morphological characteristics, which pose significant challenges to their accurate identification. By performing weighting operations on the input feature maps, CBAM effectively enhances the representation of crack regions. By employing channel attention, the network is guided to prioritize channel-wise information that is most relevant to cracks. Simultaneously, the spatial attention mechanism strengthens the features at crack locations, enabling the network to precisely localize crack areas. This is particularly beneficial for elongated and irregular crack patterns, leading to further improvements in segmentation accuracy.

Furthermore, CBAM exhibits the capability for cross-scale feature fusion, enabling it to effectively address the manifestation of cracks across varying scales. By dynamically adjusting both channel-wise and spatial-wise weights, CBAM enhances the network's adaptability to multi-scale representations, which leads to significant gains in both detection accuracy and model robustness. Consequently, the adoption of

CBAM not only strengthens the network's focus on crack regions but also refines detection performance, demonstrating notable advantages particularly in detecting fine cracks within complex backgrounds.

2.3 Methodology

This paper employs ResNet50 as the backbone network, enabling the construction of a deeper model while maintaining ease of training. CNNs excel at extracting local information, whereas CBAM demonstrates a strong capability to capture global contextual semantics, emphasizing crack-related features, suppressing irrelevant information, enhancing characteristics at crack locations, reducing the impact of background noise, and achieving precise spatial localization. These two mechanisms exhibit a synergistic relationship. Due to the elongated nature of the cracks, detailed information is prone to being lost, and relying solely on traditional convolutional neural networks (CNNs) may lead to missed detections. Using CBAM alone, on the other hand, remains susceptible to background interference. To this end, we present a novel architecture named CACNN-Net, which achieves a strategic integration of CNN and CBAM to enhance feature representation.

2.4 Encoder

Fig. 4 gives a holistic view of the proposed pipeline. A ResNet50 backbone acts as the main feature extractor: its progressively-strided stages harvest both low-order cues—crack edges, texture—and high-order semantics—topology, continuity—while residual shortcuts curb gradient decay and retain fine crack detail. After five downsampling steps, the encoder yields a 1/32-resolution map (512 × 512 → 16 × 16) with 2048 channels, ready for subsequent fusion and decoding.

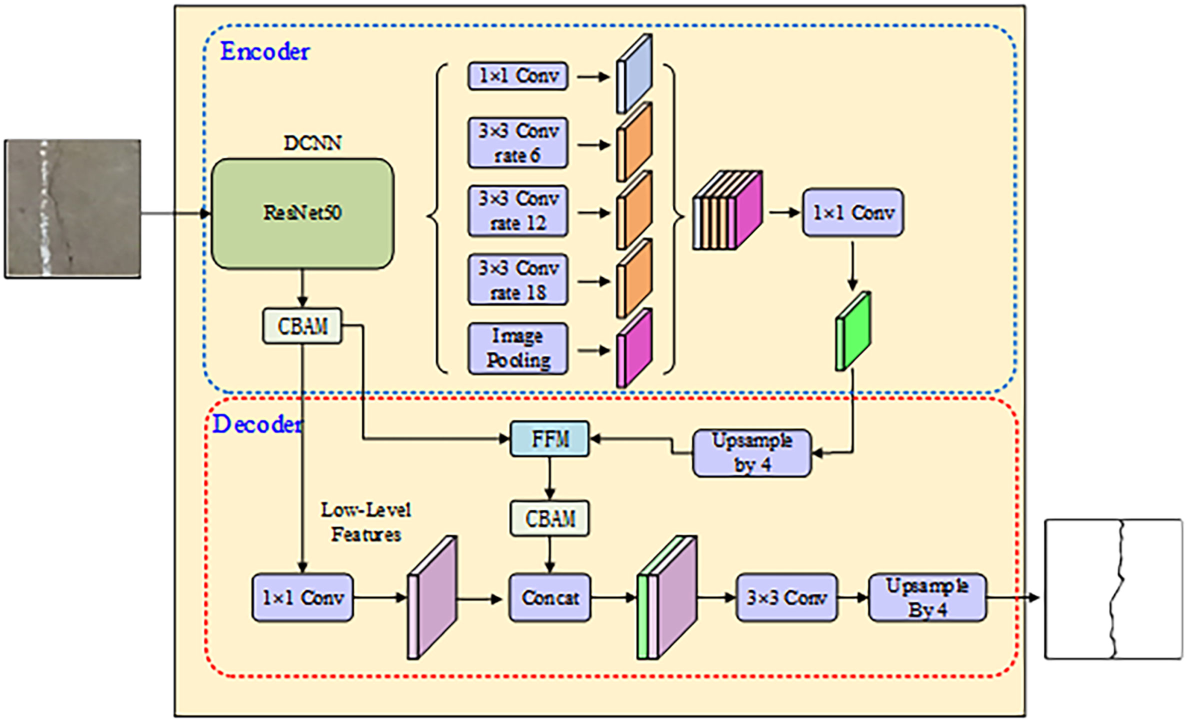

Figure 4: Architecture of the proposed CACNN-Net model.

The CBAM module is embedded immediately after each residual block of the ResNet-50 backbone. This ensures that features are recalibrated both spatially and channel-wise before being passed to the next stage. The CNN branch captures local texture details, while the parallel Transformer branch aggregates global context. These two streams are synchronized at the bottleneck layer, where their features are concatenated and fused via the Feature Fusion Module (FFM) to preserve both local sharpness and global semantic consistency.

Afterward, the feature map is fed into three parallel 3 × 3 dilated layers whose rates—6, 12, and 18—are staggered to harvest context at increasing scales without downsampling; the receptive field grows exponentially so that fine crack edges and distant structural dependencies, e.g., propagation paths, are captured simultaneously. Finally, an image pooling operation (global average pooling) compresses the spatial dimensions to 1 × 1, producing a compact feature vector that encapsulates global semantic information. This operation significantly reduces computational complexity while aggregating global context to retain statistically salient features of crack regions. The resulting feature vector provides high-level semantic guidance for subsequent feature fusion and decoding steps.

To harvest scene-level context, global-average pooling collapses the feature map to a single 1 × 1 vector that condenses the entire image content into a compact descriptor. This operation significantly reduces computational complexity while aggregating global context to retain statistically salient features of crack regions. The resulting feature vector provides high-level semantic guidance for subsequent feature fusion and decoding steps.

2.5 Decoder

CACNN-Net adopts a decoder architecture wherein low-level features are directly introduced into the decoding path. These features provide pixel-level edge responses and textural details—such as sharp crack boundaries and irregular morphological patterns—which serve as spatial anchors for high-precision segmentation.

The Feature Fusion Module (FFM) merges shallow cues with deep semantics across scales: high-level descriptors are first doubled in resolution through bilinear upsampling, then channel-wise concatenated with the low-level map; a subsequent 1 × 1 convolution squeezes the stack to 256 channels, retaining salient information while suppressing redundancy. The output is a multi-scale fused feature map that retains both semantic integrity and enhanced local details of cracks.

CBAM is then deployed to recalibrate the fused map, adaptively emphasizing crack-relevant activations and suppressing background noise. Here, the channel attention mechanism reinforces crack-relevant channels, while the spatial attention—guided by the precise localization cues from low-level features—generates a mask focused on crack pixels (regions with mask response > 0.8 increased by 35%), significantly suppressing false activations in non-crack areas.

The optimized features then undergo a two-stage upsampling process to progressively recover resolution: the first stage uses a transposed convolution for 4× upsampling, leveraging learnable parameters to reconstruct topological connectivity of cracks; the second stage employs bilinear interpolation for an additional 4× upscaling to the original input size. Due to its computational efficiency, this step adds only 1.2 ms inference latency at a resolution of 512 × 512.

Finally, the high-resolution feature map (512 × 512 × 32) output by the upsampling module is processed by a 1 × 1 convolution to produce a binary segmentation mask, enabling sub-pixel crack contour reconstruction. To further refine the segmentation result, additional 1 × 1 and 3 × 3 convolutions are applied, ultimately yielding the final binary segmentation mask.

2.6 Feature Fusion Module (FFM)

The Feature Fusion Module (FFM) enhances feature representation in crack detection tasks by effectively integrating multi-level and multi-source features. The module employs a Channel Attention (CA) mechanism to emphasize critical features and suppress less relevant ones, thereby improving the recognition capability for fine crack regions. Additionally, the Cross Fusion Block (CFB) efficiently combines features from different scales and origins, significantly boosting the robustness and adaptability of the network. This is particularly beneficial for accurately localizing cracks in challenging scenarios, such as when crack regions are small or poorly defined.

The FFM takes two feature inputs: the local detail feature $F_{cnn} \in \mathbb{R}^{C\times H\times W}$ (denoted as C_i in Fig. 5) from the CNN branch and the global semantic feature $F_{trans} \in \mathbb{R}^{C\times H\times W}$ (denoted as T_i) from the Transformer branch. The fusion process consists of three stages: Channel Attention Recalibration, Cross-domain Fusion, and Feature Integration. Channel Attention Recalibration First, to suppress noise and emphasize informative channels, both inputs pass through a Channel Attention (CA) block.

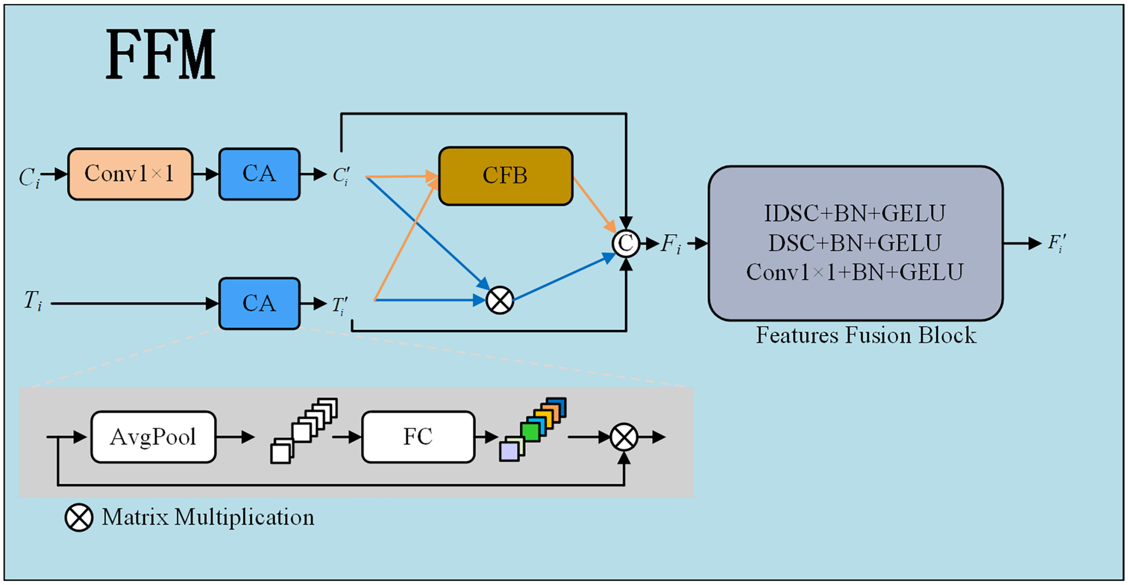

Figure 5: Schematic diagram of the Feature Fusion Module (FFM) structure.

Input Alignment and Attention. First, the input features are aligned via a 1 × 1 convolution and then recalibrated using the Channel Attention (CA) mechanism to emphasize valid signals. Let CA denote the channel attention operation. The refined features C_i and T_i are obtained as follows:

$$C_i^{'} = CA(W1F_{cnn}), T_i^{'} = CA(F_{trans}), \tag{3}$$

where $W1$ represents 1 × 1 convolution weight for dimension matching. The CFB employs a "split-interaction-merge" strategy. As illustrated in Fig. 5, we compute an interaction term via element-wise multiplication and concatenate it with the original branches to preserve both unique and common features:

$$F_i = Concat([C_i^{'}, \left(C_i^{'} \otimes T_i^{'}\right), T_i^{'}]), \tag{4}$$

Finally, the mixed features undergo deep filtering through the Feature Fusion Block (FFB) to extract high-level semantics while reducing dimensionality. The FFB integrates Inverted Depthwise Separable

Convolution (IDSC) and standard DSC, formulated as:

$$F_i^{'} = GELU(W_2 \cdot DSC(IDSC(F_i))), \tag{5}$$

Here, W_2 denotes the final 1 × 1 projection layer. The IDSC operation expands the channel dimension by a ratio of γ (expansion ratio) to learn complex patterns before projecting back.

Furthermore, the FFM incorporates both Inverted Depthwise Separable Convolution (IDSC) and standard Depthwise Separable Convolution (DSC), significantly reducing computational complexity and improving the efficiency of feature fusion. Meanwhile, the GELU (Gaussian Error Linear Unit) activation function is employed to introduce enhanced nonlinearity into the module. Collectively, these design choices contribute to superior crack segmentation accuracy, along with improved model stability and generalization in complex environments. The fused feature map generated by the FFM provides high-quality feature representations for subsequent segmentation tasks, considerably boosting the overall performance of crack detection.

3 Experiment and Analysis

The process begins with uniform preprocessing of the input pavement dataset, including operations such as size normalization and data augmentation. Training proceeds iteratively on the training split with an appropriate loss until convergence; the converged model is then validated on the test set through precision, recall, IoU, and F1 metrics. Upon successful evaluation, the finalized model can be integrated into an automated system to classify pavement images as either normal surfaces or distressed areas.

3.1 Training Environment

The experiments for the proposed model in this study were implemented in the Python programming language, utilizing the open-source neural network framework PyTorch, initially developed by Facebook. The experimental environment was configured on a Windows 10 workstation. The detailed hardware specifications are summarized in Table 1 below:

Table 1: Detailed parameters of the experimental platform.

CPU	CPU Model	intel Xeon Gold5218
	Number of Cores	16
	Number of Threads	32
GPU	**GPU Model**	**NVIDIA GeForce RTX 4090**
	VideoMemory(VRAM)	24 GB
Framework	Python3.8 Cuda11.8 Pytorch11.8	
Operating System	Windows10	

3.2 Dataset and Evaluation Metrics

This study employs a self-constructed concrete crack dataset for model training and evaluation, collected from *in-situ* inspections of multiple highway and urban bridges in Zhejiang Province, China (2021–2023). Images were captured using smartphones (mainstream models) with an original resolution of approximately 3000 × 4000 pixels. The original images were preprocessed by cropping them into sub-images of 256 × 256 pixels. The dataset is partitioned as follows: a training set containing 6500 images and a test set comprising 1500 images. In total, 8000 annotated samples are used, including both positive samples (with cracks), totaling 5500 images, and negative samples (without cracks), amounting to 2500 images. The classification dataset was systematically organized by storing images according to their labels in separate directories, thereby facilitating supervised learning.

To gauge detection quality, we adopt four quantitative indices: Precision, Recall, IoU, and F1-score. Precision, expressed in Eq. (6), measures the fraction of correctly identified crack pixels among all positive predictions, i.e., the ratio of true positives to the sum of true and false positives.

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

Recall focuses on the model's coverage of positive cases, quantifying its success in minimizing omissions of true positives. The corresponding formula is presented in Eq. (7).

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

IoU quantifies how closely the segmented crack mask matches the ground-truth region by dividing their intersection by their union; Eq. (8) and Fig. 6 summarise the computation, with a value of 1 denoting exact overlap. We use this metric to assess pixel-level crack delineation accuracy.

$$IoU = \frac{TP}{TP + FN + FP} \tag{8}$$

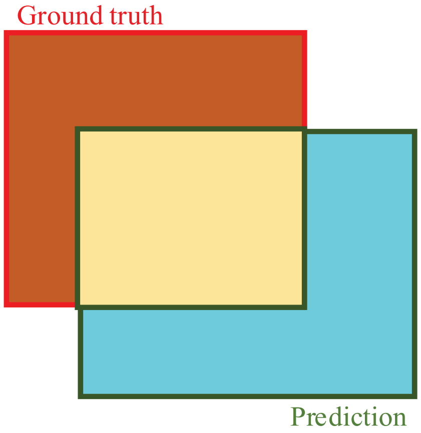

Figure 6: Schematic diagram of IoU calculation.

To reconcile the inherent tug-of-war between precision and recall, their harmonic mean is computed as the F1-score. This metric provides a single score that balances both concerns, where a higher value denotes better prediction quality and a more robust model. The computational formula is given in Eq. (9).

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{9}$$

As delineated in Table 2, TP, FP, and FN represent three fundamental detection outcomes: TP are correct crack identifications; FP are background regions erroneously detected as cracks; and FN are genuine cracks that were missed.

Table 2: Predictive indicator relationship.

Ground Truth	Predicted Values	
	True	**False**
True	TP	FN
False	FP	TN

3.3 Classification Results and Analysis

CACNN-Net is implemented using Python3.8 Cuda11.8 and Pytorch11.8. All experiments were carried out on a workstation configured with an Intel Xeon Gold 5218 central processor and an NVIDIA GeForce RTX 4090 graphics card (24 GB VRAM), utilizing a mini-batch size of 16. In the experiments of this paper, the models were all trained with the same set of hyperparameters to ensure the fairness of the comparison. The specific configurations are summarized in Table 3.

Table 3: Parameter configuration table.

Hyperparameter	Value
Batch Size	16
Epochs	150
Initial Learning Rate	5×10^{-4}
Momentum	0.9
Weight Decay	1×10^{-4}

Fig. 7 shows that, once training exceeded 150 epochs, network performance steadily increased; images were first rescaled to 256 × 256 pixels, and optimization started from an initial learning rate of 5×10^{-4}. Following appropriate data augmentation and normalization, the preprocessed images were fed into the network.

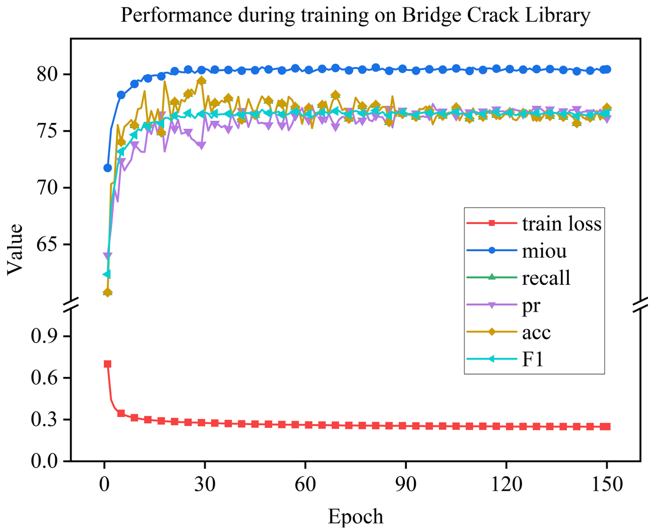

Figure 7: Performance during training on Bridge Crack Library.

On the public benchmark, we benchmarked our approach against five leading alternatives, undertaking an end-to-end comparative study: UNet-ResNet34 [35], CENet [36], UTNet [37], FAT-Net [38], and DeepCrack [39].

Fig. 8 presents the training loss trajectories of the proposed CACNN-Net alongside the five benchmark models. It can be observed that all six models eventually converge, and CACNN-Net exhibits significantly superior convergence behavior compared to the other five models, demonstrating stronger capability in fitting image data and better generalization performance. The training processes stabilize after approximately 50 iterations and converge rapidly. Figs. 9–11 present the curves of evaluation metrics during training. All curves eventually stabilize, and CACNN-Net outperforms the other five models across all metrics.

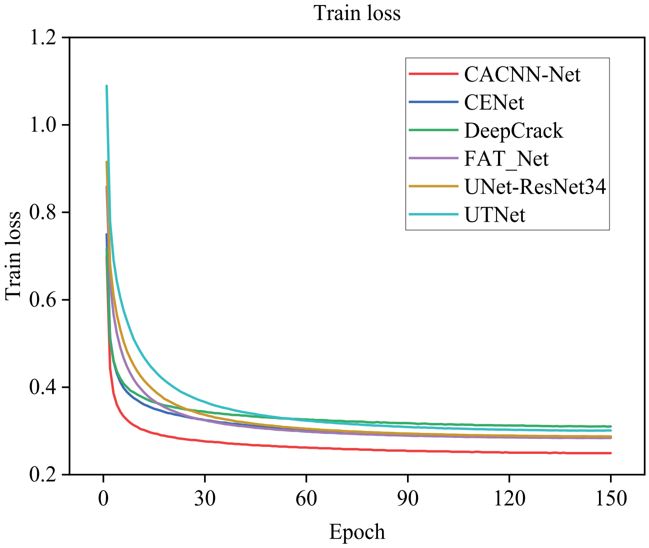

Figure 8: Training loss curves of different models.

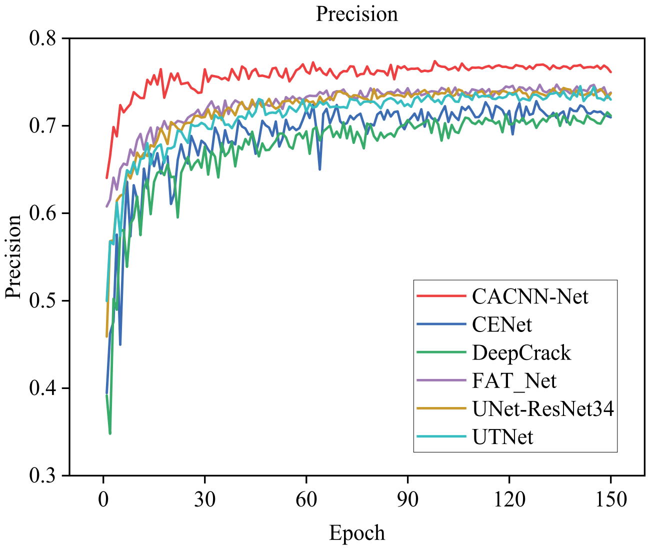

Figure 9: Precision curves of different models.

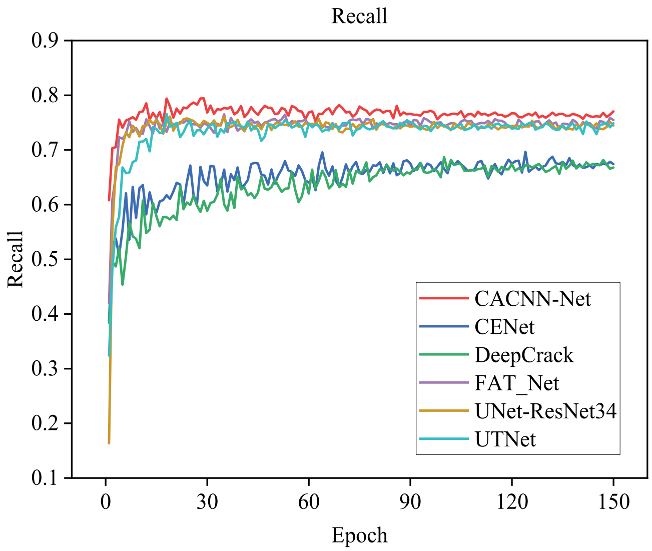

Figure 10: Recall curves of different models.

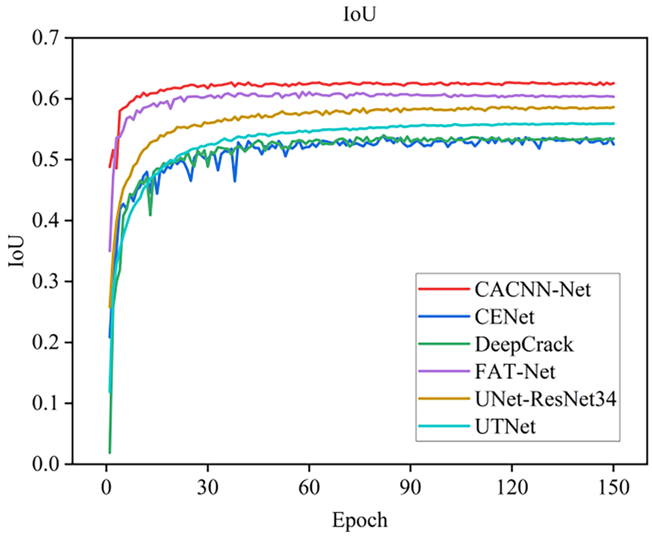

Figure 11: IoU curves of different models.

Table 4 summarises the Precision, Recall, IoU and F1-score achieved by seven architectures—UNet-ResNet34, UTNet, DeepCrack, FAT_Net, CENet, Deeplabv3 and the proposed CACNN-Net—on the same test split.

Drawing on the metrics compiled in Table 4, the six evaluated architectures exhibit a clear hierarchical structure, which can be divided into three tiers. The performance of DeepCrack and CENet is relatively poor, with accuracy below 73% and recall below 70%. In comparison, UNet-ResNet34, UTNet, and FAT-Net achieve better results, with accuracy and recall both exceeding 74% and 76%, respectively. The proposed CACNN-Net outperforms all other models across every metric in this comparative study. It achieves a precision of 77.6%, a recall of 79.4%, an IoU of 62.7%, and an F1-score of 78.5%—the highest among all six models—outperforming the other five popular models by 1 to 3 percentage points in each metric.

Table 4: CACNN-Net model evaluation indicators.

Model	Precision (%)	Recall (%)	IoU (%)	F1 (%)
UNet-ResNet34	74.3	76.1	58.7	74.2
UTNet	74.5	76.5	56.0	74.0
DeepCrack	71.5	68.7	54.0	69.2
FAT_Net	74.7	76.5	61.1	74.7
CENet	72.8	69.7	53.8	70.1
Deeplabv3	74.2	78.0	61.6	76.1
CACNN-Net	77.6	79.4	62.7	78.5

Table 5 summarises the computational efficiency characteristics—including parameter count, FLOPs, and inference speed (Frames Per Second, FPS)—of the seven compared architectures along with the proposed CACNN-Net, measured under the same experimental environment.

Table 5: Computational efficiency comparison of different models.

Model	Parameters (M)	FLOPs (G)	FPS
UNet-ResNet34	17.263	30.771	248.383
UTNet	12.937	2.761	410.058
DeepCrack	11.725	2.577	154.505
FAT_Net	46.583	5.110	142.381
CENet	10.365	2.341	400.913
Deeplabv3	38.965	28.732	95.603
CACNN-Net	45.737	20.561	67.511

As shown in Table 5, the number of parameters and computational cost of CACNN-Net are higher than those of several lightweight benchmark models, resulting in relatively slow inference speed. This reflects a deliberate design concept, which focuses on improving segmentation accuracy and robustness while not pursuing extreme lightweighting. The architectural complexity introduced by the dual encoder, CBAM, and FFM modules is a direct investment in achieving the outstanding precision performance demonstrated in Table 4, which is crucial for achieving reliable detection in complex real-world scenarios. In engineering applications such as bridge inspections, the primary goal is to achieve highly reliable identification to support maintenance decisions, as the tolerance for missed detections or false detections is extremely low. Although the computational efficiency of CACNN-Net is lower than that of some similar products, its inference speed is still fully sufficient for real-time or near-real-time processing in most inspection procedures.

To quantify the improvements, we pit the enhanced CACNN-Net against its backbone Deeplabv3 on the identical crack-segmentation task. The results display that CACNN-Net outperforms the benchmark-level Deeplabv3 model across multiple evaluation metrics. Specifically, CACNN-Net achieves a Precision of 77.6%, representing an improvement of approximately 3.4% over Deeplabv3's 74.2%. In terms of Recall, CACNN-Net reaches 79.4%, slightly higher than DeeplabV3's 78.0%. For the Intersection over Union (IoU) metric, CACNN-Net attains a score of 62.7%, which is about 1.9% higher than the 61.6% achieved by Deeplabv3. Furthermore, the calculated F1-score (CACNN-Net: 78.5%, Deeplabv3: 76.1%) confirms that CACNN-Net delivers a more balanced and superior overall performance. A side-by-side comparison of the two models across various metrics is shown in Fig. 12.

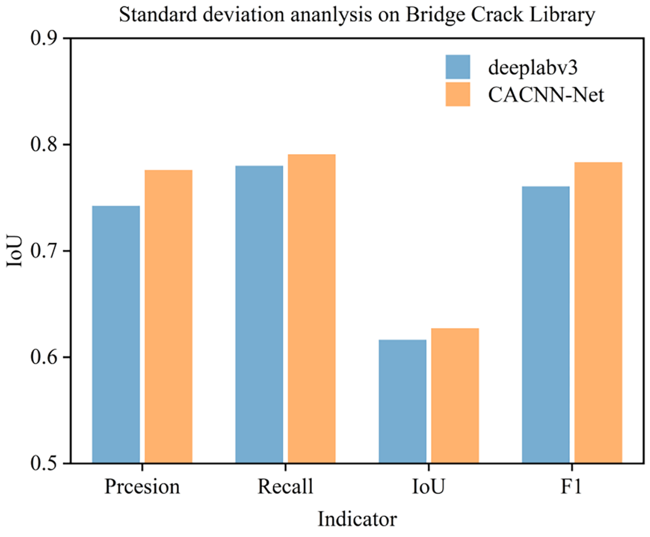

Figure 12: Indicator comparison between DeeplabV3 and CACNN-Net.

These improvements indicate that, through architectural optimizations, CACNN-Net significantly enhances the model's precision and segmentation boundary consistency while maintaining a high recall rate. This further validates its robustness and effectiveness in complex scenarios. Offering a ready-to-adapt blueprint for boosting segmentation accuracy, CACNN-Net opens clear avenues for both follow-up research and field deployment.

The significant lead in IoU, a crucial metric for segmentation accuracy, indicates that CACNN-Net produces pixel-wise predictions that align more closely with the ground truth. Furthermore, by simultaneously pushing Precision and Recall to high values, the model yields the top F1-score, evidencing its unrivalled ability to suppress both false alarms and missed detections. This balanced improvement suggests that the architectural innovations within CACNN-Net effectively enhance feature representation and boundary delineation compared to the other models, validating its design rationale and solidifying its state-of-the-art status for the given application.

Overall, the improved model CACNN-Net demonstrates significant enhancements in all four key metrics—Precision, Recall, IoU, and F1-score—compared to the other five widely-used models. Based on these quantitative results, the next step involves a comparative analysis of the visual crack detection performance of each model on bridge components.

Based on the quantitative metric comparisons, a comparative analysis of the visual identification performance of different models on bridge cracks was conducted. Each model was applied to crack images for detection, and visual output maps of the recognition results were generated. The images used for visual evaluation varied in crack morphological characteristics and background noise.

To further compare the models, this section examines their performance by visually inspecting the crack detection results on sample bridge images. The prediction maps generated by each model are displayed for comparative analysis. The images selected for visual evaluation exhibit diverse crack morphological features and varying levels of background noise. As shown in Fig. 13, the visualization results of crack predictions are presented for UNet-ResNet34, UTNet, DeepCrack, FAT_Net, CENet, and the proposed model CACNN-Net, with "Label" denoting the ground truth annotation of the crack images.

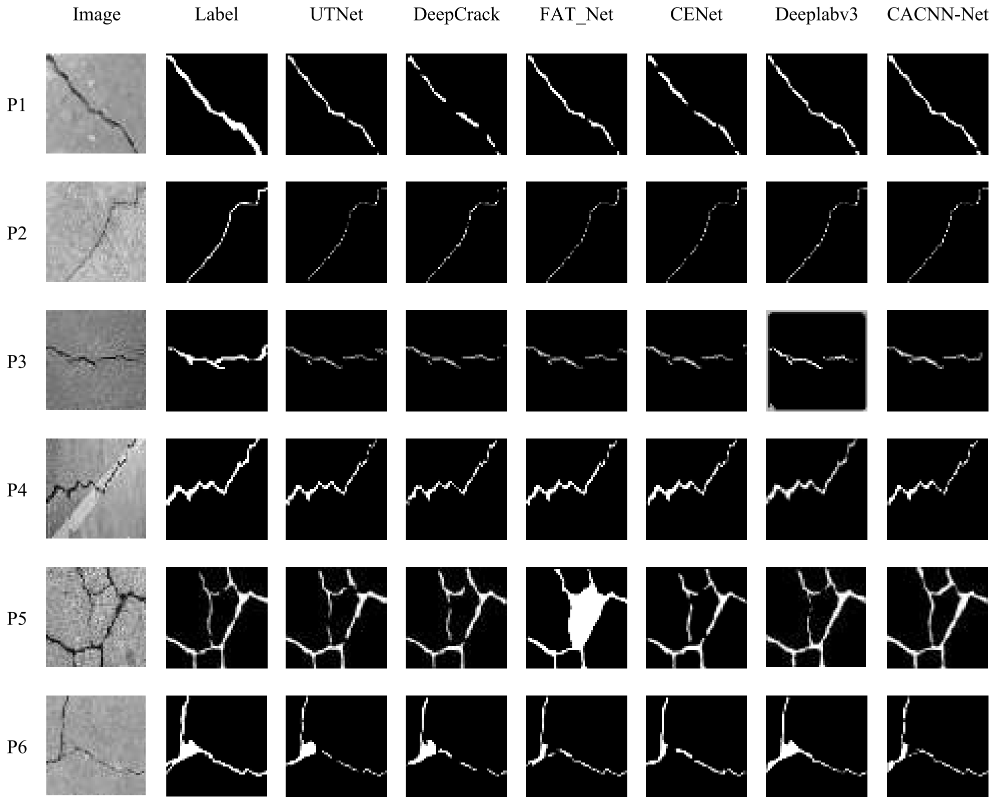

Figure 13: Visual identification results of cracks.

As clearly observed from the visual results (Fig. 13), the proposed CACNN-Net model exhibits superior overall performance compared to the five baseline models, achieving more accurate and comprehensive identification of concrete cracks while effectively suppressing background noise interference.

In the case of P1, which features a wide crack, both CACNN-Net and Deeplabv3 produce continuous and complete crack segments, whereas the other five models exhibit fragmented detection. This suggests that CACNN-Net's dual-encoder structure and feature fusion mechanism enhance continuity perception for prominent cracks. P2, a narrow but distinct crack, is accurately detected by all models, indicating that well-defined thin cracks do not pose significant challenges to modern architectures. For more complex scenarios, however, differences become pronounced. In P3, where a main crack branches into multiple secondary cracks, all models show varying degrees of discontinuity, but CACNN-Net maintains the highest connectivity and clarity. This can be attributed to its multi-scale feature refinement and attention mechanisms, which better capture crack topology and propagation patterns. P4, with a highly tortuous crack merged with background texture, remains challenging; while all models perform adequately, CACNN-Net shows slightly smoother boundary preservation, likely due to the spatial attention module's ability to suppress texture noise. Notably, P5 and P6 represent critical edge cases. P5 contains a net-like crack with low contrast against the background. Here, UTNet, DeepCrack, and Deeplabv3 exhibit severe fragmentation, while FAT_Net misidentifies central regions. CACNN-Net achieves the most coherent detection, demonstrating its robustness to complex backgrounds and weak boundary definitions—a result of its combined local detail enhancement and global context integration. In P6, a fine bifurcating crack, all baseline models either disconnect at branch points or miss subtle segments. CACNN-Net shows nearly continuous detection, though closer inspection reveals minor thinning at the finest bifurcations, indicating that extremely subtle crack features remain partially challenging due to information loss in deep feature maps.

In summary, CACNN-Net demonstrates robust visual identification performance across images with diverse crack morphologies and complex background noise, proving to be an effective solution for segmenting bridge cracks. In particular, it significantly outperforms existing models in maintaining crack continuity and resisting background interference. Moreover, the performance of CACNN-Net in handling extremely fine or low-contrast cracks indicates that it still has potential in enhancing resolution feature retention or focusing more on pixel-level details in the early encoder stage.

To further verify the generalization ability of CACNN-Net in terms of different structural types and crack patterns, we conducted additional tests on the CrackForest dataset [40]. This dataset is an openly available dataset of road cracks, containing images with complex backgrounds, different lighting conditions, and various crack patterns. The dataset includes 118 images with pixel-level annotations. The test results are shown in Fig. 14. Group A, Group B, and Group C respectively represent normal lighting, low illumination, and high brightness environments.

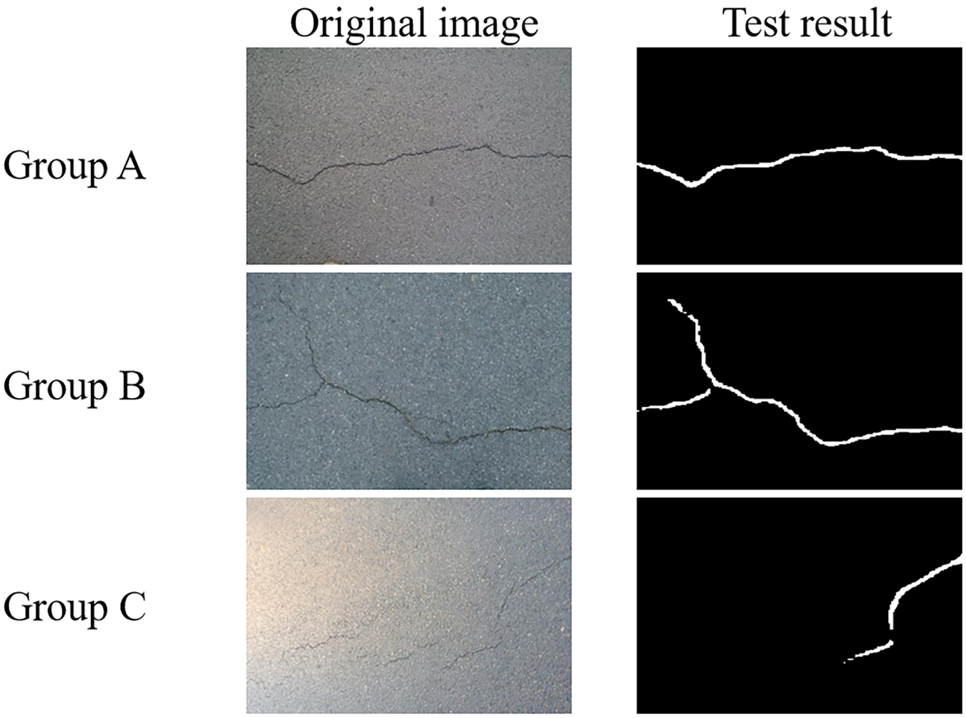

Figure 14: Test images of road damage under different lighting conditions.

The experimental results indicate that CACNN-Net maintains promising performance and strong robustness in pavement crack detection tasks. Under normal and low-light conditions, the model accurately captures crack morphological features and structural continuity. In over-exposed (high-illumination) environments, its recognition capability slightly declines, which may be attributed to the loss of local details and reduced feature contrast, yet it still retains high detection integrity.

3.4 Ablation Study

To verify the effectiveness of each key component in the CACNN-Net, this paper conducted an ablation study. By systematically removing or replacing individual modules while maintaining the consistency of the overall network structure and training configuration, the experiments were conducted on the same Bridge Crack Database dataset and using the same evaluation metrics.

The CACNN-Net proposed in this study mainly integrates three core modules. The CBAM module consists of channel attention (CBAM1/CAM) and spatial attention (CBAM2/SAM): the former enhances the features related to cracks by adjusting the channel weights, while the latter highlights the crack morphology by focusing on key spatial regions; the feature fusion module (FFM) is responsible for deeply integrating the multi-scale features of the two encoders. The three work together to enhance the model's perception and segmentation capabilities for complex cracks. The configuration and results of the ablation experiments are shown in Table 6.

Table 6: The results of the ablation experiment.

Ablation Experiment Configuration				**IoU (%)**	**Presicion (%)**	**Recall (%)**
CACNN-Net	**CBAM1**	**CBAM2**	**FFM**			
√	–	√	√	62.3	75.4	78.3
√	√	–	√	62.3	75.8	77.9
√	–	–	√	62.4	75.7	78.1
√	√	√	–	61.5	75.5	76.8
√	√	√	√	62.7	77.6	79.4

The results of the ablation experiments clearly demonstrate the necessity of each module and the superiority of the complete architecture. Firstly, using either channel attention (CBAM1) or spatial attention (CBAM2) alone can lead to performance improvements compared to the base version without attention, especially in terms of recall rate (Recall), where the performance is significantly better. This indicates that both attention mechanisms can effectively guide the network to focus on the crack area. However, when both attention modules are removed, the model still maintains acceptable performance. This is because the underlying CNN encoder itself has certain feature extraction capabilities, but at this time, its accuracy (Precision) has a significant gap compared to the full model. The role of the Feature Fusion Module (FFM) is of utmost importance. When the FFM is removed, all performance metrics show a significant decline. Among them, the IoU drops by 1.2 percentage points, and the Recall drops by 2.6 percentage points. This verifies that the FFM is indispensable for effectively integrating the local texture details from the CNN encoder and the global context information from the Transformer encoder. Its absence directly leads to a weakened model's ability to perceive the continuity and complex morphology of cracks.

The complete CACNN-Net model achieved the best results in all evaluation metrics. It not only obtained the highest IoU (62.7%), but also achieved significant superiority in accuracy (77.6%) and recall rate (79.4%). This indicates that CBAM and FFM are not simply superimposed, but have a synergistic enhancement effect: the CBAM module preprocesses the features to enhance their discriminative power; FFM then integrates multi-source features in a targeted manner on this basis. This design enables the model to achieve a better balance between suppressing background noise and capturing real cracks, thereby surpassing any single module or combination of partial modules in overall segmentation accuracy and comprehensive performance, fully verifying the rationality and efficiency of the network architecture design proposed in this paper.

4 Conclusion

A deep CNN architecture combining multi-scale feature enhancement with attention mechanisms is proposed, in which ResNet50 serves as the encoder backbone and a cross-stage FFM together with CBAM extracts and fuses multi-level representations. We designed a multi-branch convolutional structure, including dilated convolutions with varying dilation rates and image pooling operations, to capture contextual information at different receptive fields. This is combined with upsampling operations to progressively restore spatial details and enhance the representation of both edge and semantic features. The network effectively integrates local details and global context while enhancing the salience of target features and suppressing redundant information. Extensive public-dataset testing confirms the network's effectiveness: it surpasses current state-of-the-art approaches in both numeric metrics and visual quality. The key findings are:

(i) Dataset Construction and Model Selection:

Utilizing the self-constructed bridge crack dataset, seven distinct semantic segmentation models—UNet-ResNet34, CENet, UTNet, FAT-Net, DeepCrack, Deeplabv3, and CACNN-Net—were trained under identical dataset splits and training configurations (including consistent training/validation/test sets, epochs, learning rate, etc.). CACNN-Net achieved the highest scores among seven contemporary segmentation models, with a precision of 77.6%, recall of 79.4%, IoU of 62.7%, and F1-score of 78.5%. These results quantitatively validate its enhanced capability in accurate crack identification.

(ii) Improvement via Attention Mechanism:

The introduced dual-encoder design and the dedicated Feature Fusion Module (FFM) led to measurable performance gains. Compared with representative existing models, CACNN-Net achieved a more balanced and superior overall performance, as reflected in its higher IoU and F1-score, demonstrating that the proposed feature fusion and attention integration strategy effectively enhances segmentation accuracy and boundary consistency.

(iii) Visualization Experiments:

For additional validation of performance and robustness under real-world crack-detection conditions, qualitative visualization analyses were conducted. Using identical bridge crack images for testing, it is visually evident that CACNN-Net outperforms other comparative models in segmentation clarity and detail preservation. The results confirm the model's robustness across diverse crack backgrounds and morphological variations, demonstrating its capability to accurately and completely segment bridge cracks under varying conditions.

Although the proposed model achieves promising performance in crack segmentation tasks, we are fully aware of certain limitations in the current research. One particularly notable issue is that existing methods rely solely on apparent image information—such as texture, color, and edges—for crack segmentation, while overlooking a crucial physical dimension that significantly reflects structural integrity: depth information of cracks. From a visual-structural consistency perspective, depth information provides essential three-dimensional geometric features that are indispensable for accurately assessing crack severity, propagation direction, and potential structural hazards. However, due to the scarcity of annotated data and constraints in sensor deployment, there remains a lack of systematic research in this field that integrates depth information for crack identification and measurement. This limitation somewhat restricts the practical applicability of the model in real engineering scenarios. In future work, we will explore this new perspective on bridge crack damage by developing an integrated framework that combines apparent information with depth information of crack images, aiming to achieve more profound identification and detection of crack hazards in practical engineering applications.

To address these limitations and advance toward practical deployment, we outline the following concrete future directions:

(1) Future work could explore integrating depth data (e.g., from stereo vision or LiDAR) to enhance the model's ability to distinguish superficial artifacts from true cracks and to quantify 3D crack parameters.
(2) Constructing more diverse datasets covering a wider range of structures, materials, and environmental conditions would help improve the model's robustness and generalizability.
(3) For practical deployment, implementing model compression techniques (e.g., pruning, quantization) and adapting the system to embedded platforms present viable pathways toward real-time crack detection in mobile inspection setups.

Acknowledgement: Not applicable.

Funding Statement: The work described in this paper was supported by the National Natural Science Foundation of China (No. 52308332), and the General Scientific Research Project of the Education Department of Zhejiang Province (No. Y202455824).

Author Contributions: The authors have made the following confirmation of their contributions in this paper: Conceptualization: Tao Jin; Methodology: Tao Jin, Hongchao Liu and Yuchun Shao; Dataset collection: Zhekun Shou and Yuchun Shao; Model improvement and training: Tao Jin and Hongchao Liu; Result verification comparison: Tao Jin and Zhekun Shou; Writing, review and editing: Tao Jin, Zhekun Shou, Hongchao Liu and Yuchun Shao. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The data used and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Ethics Approval: Not applicable.

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

References

1. Samadzadegan F, Javan FD, Hasanlou M, Gholamshahi M, Mahini FA. Automatic road crack recognition based on deep learning networks from uav imagery. ISPRS Ann Photogramm Remote Sens Spatial Inf Sci. 2023;X-4/W1-2022:685–90. doi:10.5194/isprs-annals-x-4-w1-2022-685-2023.
2. Zhu J, Zhong J, Ma T, Huang X, Zhang W, Zhou Y. Pavement distress detection using convolutional neural networks with images captured *via* UAV. Autom Constr. 2022;133:103991. doi:10.1016/j.autcon.2021.103991.
3. Cao JG, Yang GT, Yang XY. Pavement crack detection with deep learning based on attention mechanism. J Comput Aided Des Comput Graph. 2020;32(8):1324–33. (In Chinese). doi:10.3724/SP.J.1089.2020.18059.
4. Golewski GL. Using digital image correlation to evaluate fracture toughness and crack propagation in the mode I testing of concretes involving fly ash and synthetic nano-SiO_2. Mater Res Express. 2024;11(9):095504. doi:10.1088/2053-1591/ad755e.
5. Golewski GL. Investigating the effect of using three pozzolans (including the nanoadditive) in combination on the formation and development of cracks in concretes using non-contact measurement method. Adv Nano Res. 2024;16(3):217–29.
6. Kaveh H, Alhajj R. Recent advances in crack detection technologies for structures: a survey of 2022-2023 literature. Front Built Environ. 2024;10:1321634. doi:10.3389/fbuil.2024.1321634.
7. Guo S, Xu Z, Li X, Zhu P. Detection and characterization of cracks in highway pavement with the amplitude variation of GPR diffracted waves: insights from forward modeling and field data. Remote Sens. 2022;14(4):976. doi:10.3390/rs14040976.

8. Salman M, Mathavan S, Kamal K, Rahman M. Pavement crack detection using the Gabor filter. In: Proceedings of the 16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013); 2013 Oct 6–9; The Hague, The Netherlands. p. 2039–44. doi:10.1109/ITSC.2013.6728529.
9. Ayenu-Prah A, Attoh-Okine N. Evaluating pavement cracks with bidimensional empirical mode decomposition. EURASIP J Adv Signal Process. 2008;2008(1):861701. doi:10.1155/2008/861701.
10. Majidifard H, Adu-Gyamfi Y, Buttlar WG. Deep machine learning approach to develop a new asphalt pavement condition index. Constr Build Mater. 2020;247(3):118513. doi:10.1016/j.conbuildmat.2020.118513.
11. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM. 2017;60(6):84–90. doi:10.1145/3065386.
12. Tran VP, Tran TS, Lee HJ, Kim KD, Baek J, Nguyen TT. One stage detector (RetinaNet)-based crack detection for asphalt pavements considering pavement distresses and surface objects. J Civ Struct Health Monit. 2021;11(1):205–22. doi:10.1007/s13349-020-00447-8.
13. Xiao LY, Li W, Yuan B, Cui YQ, Gao R, Wang WQ. Automatic pavement crack automatic identification method based on improved mask R-CNN model. Geomat Inf Sci Wuhan Univ. 2022;47(3):623–31. (In Chinese). doi:10.1177/03611981221122778.
14. Xu K, Ma RG. Crack detection of asphalt pavement based on improved faster-RCNN. Comput Syst Appl. 2022;31(7):341–8. (In Chinese). doi:10.15888/j.cnki.csa.008594.
15. Xu X, Zhao M, Shi P, Ren R, He X, Wei X, et al. Crack detection and comparison study based on faster R-CNN and mask R-CNN. Sensors. 2022;22(3):1215. doi:10.3390/s22031215.
16. Yan K, Zhang Z. Automated asphalt highway pavement crack detection based on deformable single shot multi-box detector under a complex environment. IEEE Access. 2021;9:150925–38. doi:10.1109/ACCESS.2021.3125703.
17. Yao Z, Xu J, Hou S, Chuah MC. CrackNex: a few-shot low-light crack segmentation model based on retinex theory for UAV inspections. In: Proceedings of the 2024 IEEE International Conference on Robotics and Automation (ICRA); 2024 May 13–17; Yokohama, Japan. p. 11155–62. doi:10.1109/ICRA57147.2024.10611660.
18. Su P, Han H, Liu M, Yang T, Liu S. MOD-YOLO: rethinking the YOLO architecture at the level of feature information and applying it to crack detection. Expert Syst Appl. 2024;237(1):121346. doi:10.1016/j.eswa.2023.121346.
19. Wang W, Su C, Han G, Zhang H. A lightweight crack segmentation network based on knowledge distillation. J Build Eng. 2023;76(1):107200. doi:10.1016/j.jobe.2023.107200.
20. Chen LC, Zhu Y, Papandreou G, Schroff F, Adam H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Computer Vision—ECCV 2018. Berlin/Heidelberg, Germany: Springer; 2018. p. 833–51. doi:10.1007/978-3-030-01234-2_49.
21. SIfre L, Mallat S. Rigid-motion scattering for texture classification. arXiv:1403.1687. 2014.
22. Chollet F. Xception: deep learning with depthwise separable convolutions. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21–26; Honolulu, HI, USA. p. 1800–7. doi:10.1109/CVPR.2017.195.
23. Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, et al. MobileNets: efficient convolutional neural networks for mobile vision applications. arXiv:1704.04861. 2017.
24. Zeiler MD, Krishnan D, Taylor GW, Fergus R. Deconvolutional networks. In: Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition; 2010 Jun 13–18; San Francisco, CA, USA. p. 2528–35. doi:10.1109/CVPR.2010.5539957.
25. Shi W, Caballero J, Huszár F, Totz J, Aitken AP, Bishop R, et al. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27–30; Las Vegas, NV, USA. p. 1874–83. doi:10.1109/CVPR.2016.207.
26. Cheng M, Xu C, Wang J, Zhang W, Zhou Y, Zhang J. MicroCrack-net: a deep neural network with outline profile-guided feature augmentation and attention-based multiscale fusion for MicroCrack detection of tantalum capacitors. IEEE Trans Aerosp Electron Syst. 2022;58(6):5141–52. doi:10.1109/TAES.2022.3181117.

27. Liu Z, Cao Y, Wang Y, Wang W. Computer vision-based concrete crack detection using U-Net fully convolutional networks. Autom Constr. 2019;104:129–39. doi:10.1016/j.autcon.2019.04.005.
28. Zim AH, Iqbal A, Al-Huda Z, Malik A, Kuribayashi M. EfficientCrackNet: a lightweight model for crack segmentation. In: Proceedings of the 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV); 2025 Feb 26–Mar 6; Tucson, AZ, USA. p. 6279–89. doi:10.1109/WACV61041.2025.00612.
29. Zhu K, Liu Y, Ren X, Pan B. Panoramic deformation measurement and crack identification in concrete with deep-learning-based multi-camera DIC. Measurement. 2025;256:118133. doi:10.1016/j.measurement.2025.118133.
30. Wang R, Liu Z, Liu H, Su B, Ma C. FDSC-YOLOv8: advancements in automated crack identification for enhanced safety in underground engineering. Comput Model Eng Sci. 2024;140(3):3035–49. doi:10.32604/cmes.2024.050806.
31. Qu Z, Mu G, Yuan B. A lightweight network with dual encoder and cross feature fusion for cement pavement crack detection. Comput Model Eng Sci. 2024;140(1):255–73. doi:10.32604/cmes.2024.048175.
32. Wang S, Xu J, Wu X, Zhang J, Zhang Z, Chen X. Concrete crack recognition and geometric parameter evaluation based on deep learning. Adv Eng Softw. 2025;199(4):103800. doi:10.1016/j.advengsoft.2024.103800.
33. Yang Y, Zhao Z, Su L, Zhou Y, Li H. Research on pavement crack detection algorithm based on deep residual unet neural network. J Phys Conf Ser. 2022;2278(1):012020. doi:10.1088/1742-6596/2278/1/012020.
34. Tian M, Li B, Xu H, Yan D, Gao Y, Lang X. Deep learning assisted well log inversion for fracture identification. Geophys Prospect. 2021;69(2):419–33. doi:10.1111/1365-2478.13054.
35. Lau SLH, Chong EKP, Yang X, Wang X. Automated pavement crack segmentation using u-net-based convolutional neural network. IEEE Access. 2020;8:114892–9. doi:10.1109/access.2020.3003638.
36. Tao H, Xie C, Wang J, Xin Z. CENet: a channel-enhanced spatiotemporal network with sufficient supervision information for recognizing industrial smoke emissions. IEEE Internet Things J. 2022;9(19):18749–59. doi:10.1109/JIOT.2022.3162016.
37. Cui J, Zhou S, Xu G, Liu X, Gao X. Marine debris detection in real time: a lightweight UTNet model. J Mar Sci Eng. 2025;13(8):1560. doi:10.3390/jmse13081560.
38. Wu H, Chen S, Chen G, Wang W, Lei B, Wen Z. FAT-Net: feature adaptive transformers for automated skin lesion segmentation. Med Image Anal. 2022;76:102327. doi:10.1016/j.media.2021.102327.
39. Zou Q, Zhang Z, Li Q, Qi X, Wang Q, Wang S. DeepCrack: learning hierarchical convolutional features for crack detection. IEEE Trans Image Process. 2019;28(3):1498–512. doi:10.1109/tip.2018.2878966.
40. Shi Y, Cui L, Qi Z, Meng F, Chen Z. Automatic road crack detection using random structured forests. IEEE Trans Intell Transport Syst. 2016;17(12):3434–45. doi:10.1109/tits.2016.2552248.