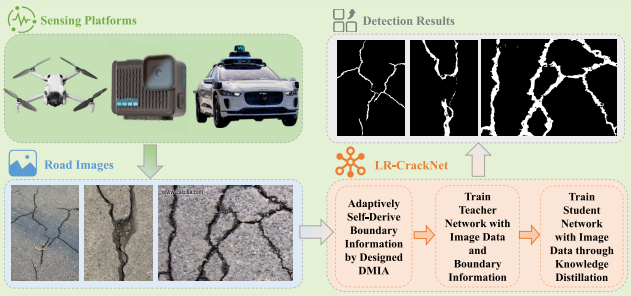
# Self-Derived Multimodal Knowledge Distillation for Real-Time Road Crack Detection

Nachuan Ma, Qiang Hu, Zhengfei Song, Sicen Guo, and Rui Fan, *Senior Member, IEEE*

***Abstract*—Road crack detection is vital for intelligent transportation and infrastructure maintenance. In recent years, deep learning-based methods have emerged for automated road crack detection with enhanced performance. Nevertheless, existing methods fail to balance detection performance, model complexity, and processing efficiency, hindering their deployment in autonomous road inspection systems. To fill this gap, we propose a novel lightweight network for real-time road crack detection, called the Lightweight Real-time Pixel-wise Road Crack Detection Network (LR-CrackNet). It employs a teacher–student (TS) network architecture, where the teacher network leverages both image data and adaptively self-derived boundary information from pixel-level annotations, while the student network relies solely on image input. The training process focuses on distilling the teacher network's detection capabilities into the student network. Both networks share a U-shaped segmentation design, integrating novel residual dual-layer depthwise convolution (RDLDC) blocks and a fast Transformer block for efficient integration of detailed hierarchical spatial information and global long-range contextual information. Furthermore, an attention-enhanced discriminator network is designed, which aims to improve detection performance by enforcing global consistency and providing pixel-level feedback. Comprehensive experimental results demonstrate that LR-CrackNet achieves state-of-the-art (SoTA) detection performance on the UDTIRI-Crack, DeepCrack, and CamCrack789 datasets and surpasses existing publicly available algorithms in processing efficiency, with 324.40 frames per second (FPS) on a single NVIDIA RTX3090. The source code package is available at https://mias.group/LR-CrackNet/**




***Index Terms*— Civil infrastructure maintenance, computer vision, deep learning, image processing, road crack.**

## I. Introduction

ROAD cracks can signal potential structural deterioration in urban transportation networks. If left unaddressed, they may propagate into more severe defects, substantially threatening infrastructure reliability and driving safety [1], [2]. For instance, in the United States, poor road conditions impose an annual average cost of $324 per driver in vehicle repairs [3]. In the U.K., poor road surfaces contributed to 12.6% of all traffic accidents in 2020 [4]. Therefore, regular and systematic road inspections are crucial for minimizing the risks associated with structural deterioration and traffic accidents [5]. However, current road crack detection still relies on manual inspection performed by certified engineers, the process of which is labor-intensive, expensive, and fraught with safety hazards [6], [7]. In addition, the inspection results are subjective, constrained by the inspector's experience and judgment [8], [9]. Therefore, the development of automated road crack detection methods is imperative.

With the rise of deep learning, researchers have increasingly adopted convolutional neural networks (CNNs) and Transformer-based models for road crack detection, aiming to enhance accuracy and robustness while reducing reliance on handcrafted features [10]. These methods are typically categorized into: 1) image classification networks that distinguish crack from noncrack images; 2) object detection networks that localize and classify crack instances; and 3) semantic segmentation networks that perform pixel-wise crack detection and have become the mainstream approach. For instance, Deepcrack [11] enhanced the fully convolutional network (FCN) [12] by integrating side-output layers and employing conditional random fields alongside guided filtering for improved road crack detection results. To support practical deployment, ECSNet [13] integrated small kernel convolutions

Received 11 August 2025; revised 17 August 2025; accepted 19 August 2025. Date of publication 16 September 2025; date of current version 16 October 2025. This work was supported in part by the Fundamental Research Funds for the Central Universities and in part by the Xiaomi Young Talents Program. The associate editor coordinating the review of this article and approving it for publication was Prof. Haidong Shao. *(Nachuan Ma, Qiang Hu, and Zhengfei Song contributed equally to this work.) (Corresponding author: Rui Fan.)*

Nachuan Ma, Qiang Hu, Zhengfei Song, and Sicen Guo are with the College of Electronic and Information Engineering, Tongji University, Shanghai 201804, China (e-mail: 2111481@tongji.edu.cn; 2252974@tongji.edu.cn; 2151094@tongji.edu.cn; guosicen@tongji.edu.cn).

Rui Fan is with the College of Electronic and Information Engineering, Shanghai Institute of Intelligent Science and Technology, Shanghai Research Institute for Intelligent Autonomous Systems, Shanghai Key Laboratory of Intelligent Autonomous Systems, State Key Laboratory of Autonomous Intelligent Unmanned Systems, and Frontiers Science Center for Intelligent Autonomous Systems (Ministry of Education), Tongji University, Shanghai 201804, China (e-mail: rui.fan@ieee.org).

This article has supplementary downloadable material available at https://doi.org/10.1109/JSEN.2025.3602294, provided by the authors.

Digital Object Identifier 10.1109/JSEN.2025.3602294

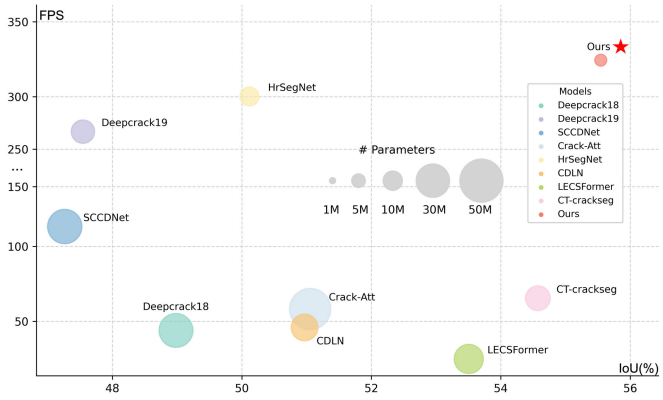



Fig. 1. Comparison results of eight road crack detection-specific methods and the proposed LR-CrackNet in terms of detection performance (IoU), model complexity (parameters), and processing efficiency (FPS).

and parallel max-pooling to enhance model efficiency while maintaining detection performance.

However, as shown in Fig. 1, none of the existing public methods for road crack detection can balance detection performance, model complexity, and processing efficiency. To fill this gap, we propose a novel **L**ightweight **R**eal-time Pixel-wise Road **C**rack Detection **Net**work (LR-CrackNet) via self-derived multimodal knowledge distillation. Specifically, a teacher–student (TS) network architecture is employed, and the training process focuses on distilling the teacher network's detection capabilities into the student network. The teacher network is pretrained using both image data and boundary information adaptively self-derived by a designed dynamic morphological iteration algorithm (DMIA), while the student network learns solely on image input. Both teacher and student networks share a U-shaped segmentation design, where novel RDLDC blocks and a novel fast Transformer block are proposed to efficiently integrate detailed hierarchical spatial information with global long-range contextual information for obtaining precise and robust crack detection results. Furthermore, during the training process, we propose a novel attention-enhanced discriminator network, which can improve the detection performance from high-dimensional global and local supervision via adversarial training. We validate the effectiveness of our proposed method on three public road crack datasets. Extensive experiments show that LR-CrackNet achieves state-of-the-art (SoTA) detection performance and surpasses existing methods in processing efficiency. The main contributions are as follows.

1) We propose LR-CrackNet, a novel lightweight network for real-time pixel-wise road crack detection via self-derived multimodal knowledge distillation.
2) We employ a TS framework, where the teacher utilizes both image and self-derived boundary information, while the student learns solely from images, enabling effective knowledge transfer.
3) We design U-shaped segmentation networks, integrating novel RDLDC blocks and a fast Transformer block to efficiently combine hierarchical spatial details with global contextual information.
4) We design an attention-enhanced discriminator network, which can enhance detection performance by enforcing global consistency and providing pixel-level feedback.

## II. Related Works

Recent advances in deep learning have led to the extensive use of CNNs and Transformer-based models for pixel-wise road crack detection. For instance, [14] proposed another Deepcrack, which fuses multiscale features from SegNet [15] to learn hierarchical information for improved road crack detection results. DMA-Net [16] added a multiscale attention module into the decoder of Deeplabv3+ [17] for generating attention masks and dynamically modulating weights across feature maps, thereby enhancing road crack detection performance. [18] combined Swin-Transformer [19] blocks with MLP layers, which can capture long-range dependencies for better feature representation of crack areas. However, to achieve higher detection accuracy, these methods inevitably introduce structural redundancy, which increases model complexity and computational inefficiency.

To facilitate the practical deployment of road crack detection models, lightweight CNN-based methods have been developed. For instance, [20] proposed RHACrackNet, using a hybrid attention block and integrating residual blocks into deeper encoder layers to maintain feature extraction with fewer parameters. Zhou et al. [21] introduced a split exchange convolution (SEConv) module, which splits feature maps into high and low-resolution parts, filtering redundant information and enhancing feature reuse. Li et al. [22] proposed HrSegNet, combining high-resolution and semantic paths with controlled channel capacity and a two-stage segmentation head to balance detection performance and efficiency.

However, lightweight methods often compromise model depth or structure, potentially trading detection performance for faster inference. To address this, some scholars have explored using knowledge distillation in road crack detection. For instance, [23] proposed LPCD-MSMD, a cascaded U-Net with extra branches capturing fine features, transferring knowledge via multiscale semantic map distillation to a student network with fewer layers and compressed channels. Similarly, [24] leveraged feature distillation to train a lightweight student network with fewer layers than the teacher network. However, relying solely on single RGB images limits the teacher network's ability to capture cracks' geometric characteristics, restricting the student's performance. Therefore, we propose LR-CrackNet, a novel lightweight real-time pixel-wise road crack detection method based on self-derived multimodal knowledge distillation. The extensive experiments reveal the superiority of LR-CrackNet in detection performance and processing speed with a small number of model parameters.

## III. Methodology

This section provides an in-depth overview of the proposed LR-CrackNet. As illustrated in Fig. 2, we adopt a TS architecture to facilitate knowledge transfer through feature distillation and logits distillation training strategies. During the training process, the teacher network is pretrained with multimodal inputs, including both image data and corresponding boundary information, which is adaptively self-derived by the designed DMIA, whereas the student network only receives image data. By constraining intermediate feature representations and detection results, such a design enables the student network to comprehensively learn the crack detection and boundary sensing capabilities of the teacher network. To ensure feature alignment, both teacher and student networks share the same

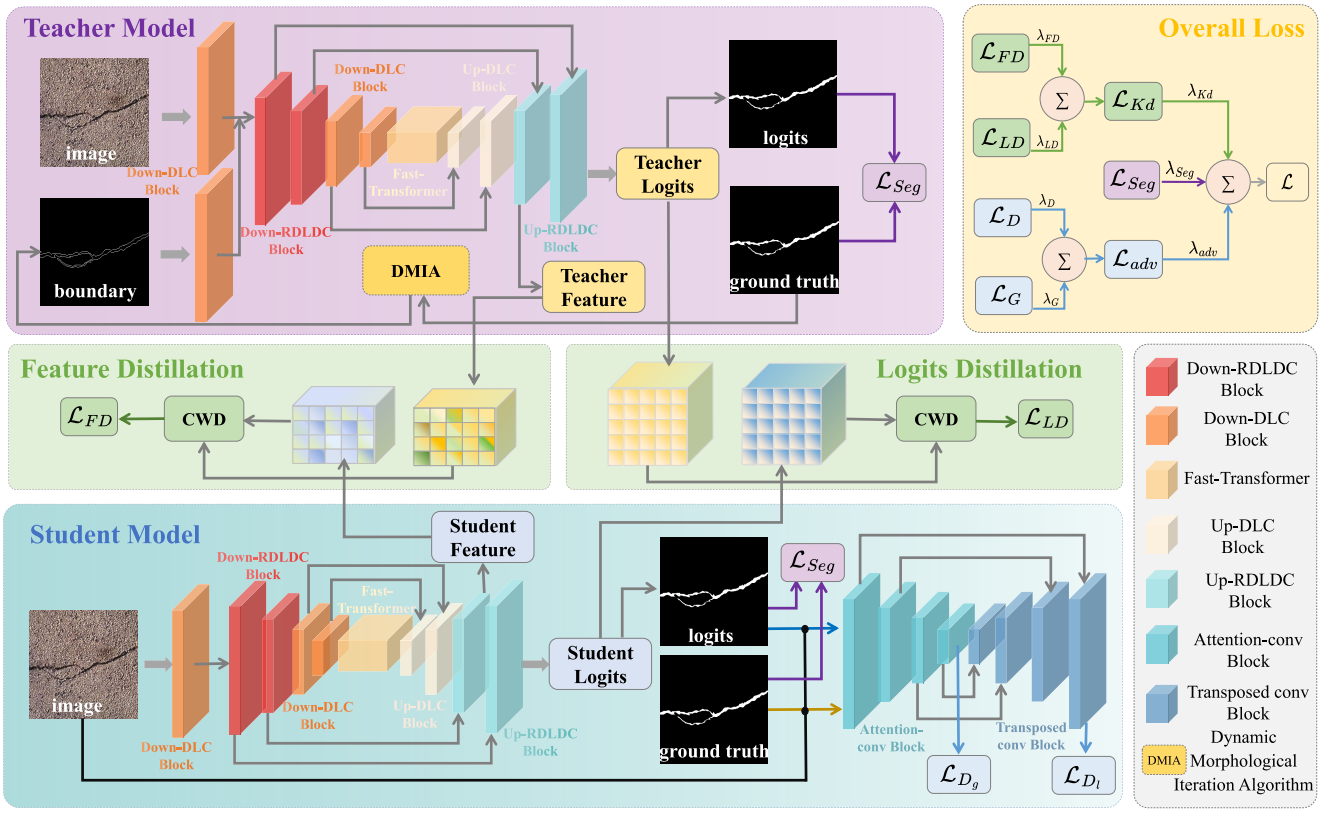



Fig. 2. Overall architecture of LR-CrackNet.

U-shaped segmentation network. Furthermore, to enhance the detection performance of the student network, we propose a novel attention-enhanced discriminator network to provide high-dimensional global and local supervision via adversarial training. During the inference phase, only the U-shaped segmentation network of the trained student network is employed to output pixel-wise road crack detection results, thus avoiding increased model complexity and computational overhead. The remainder of this section describes the proposed U-shaped segmentation network, attention-enhanced discriminator network, DMIA, and the designed loss functions.

### A. U-Shaped Segmentation Network

The proposed U-shaped segmentation network employs an encoder–decoder architecture. The encoder comprises dual-layer convolutional (DLC) blocks—two in the teacher network (for initial feature extraction from the input image and boundary map) and one in the student network (for initial feature extraction from the input image)—followed by two Down-RDLDC and two Down-DLC blocks for hierarchical feature extraction. The decoder comprises two Up-DLC and two Up-RDLDC blocks to map the extracted feature maps to pixel-wise road crack detection results. Skip connections between the encoder and the decoder facilitate the integration of low- and high-level features. In addition, we integrated the proposed Fast-Transformer block between the encoder and the decoder, enabling the network to simultaneously learn detailed hierarchical spatial and global long-range contextual information while improving overall processing efficiency.

Specifically, each DLC block comprises two consecutive `Conv-BatchNorm-ReLU` blocks, enabling more comprehensive information extraction compared to a single block. The proposed RDLDC enhances the DLC by improving both model performance and processing efficiency. A cross-connection structure between convolutional layers is designed to facilitate feature reuse, allowing the network to better capture and leverage multiscale information, thereby enhancing its feature representation capability. Furthermore, depthwise (DW) convolution layers are introduced to replace standard convolution layers, enabling independent convolution operations for each input channel. This approach effectively reduces parameters and computational costs while preventing information interference between channels.

The proposed Fast-Transformer block is composed of an attention embedding module (AEM), a residual feature mixer module (RFMM), and an MLP. In detail, as illustrated in Fig. 3(a), the core component of AEM is a channel-wise spatial attention module. Compared with the standard patch embedding module, this design can enhance the network's ability to perceive image structures, enabling it to capture global context more effectively while maintaining spatial structure awareness and adaptively focusing on discriminative regions. The formulation of AEM can be expressed as

$$F_e = F \cdot \sigma\left(\mathrm{Conv}_1 * \mathrm{Concat}\left(\mathrm{CWAP}\left(F\right), \mathrm{CWMP}\left(F\right)\right)\right) \quad (1)$$

$$\mathrm{CWAP}\left(F\right)(i,j) = \frac{1}{C}\sum_{c=1}^{C} F(c,i,j) \quad (2)$$

$$\mathrm{CWMP}\left(F\right)(i,j) = \max_{c=1}^{C} F(c,i,j) \quad (3)$$

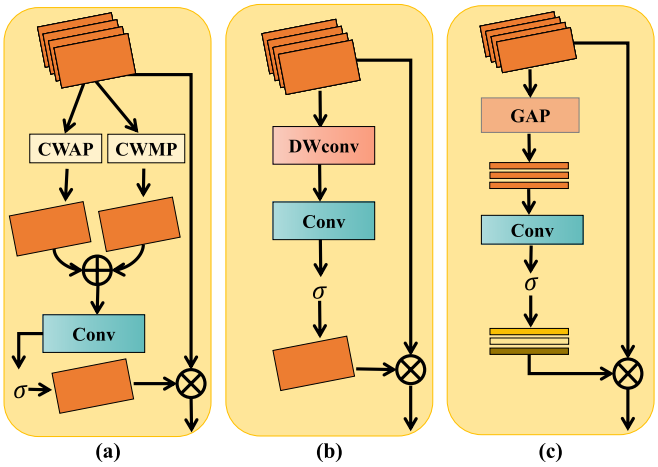



Fig. 3. Proposed (a) channel-wise spatial attention module. (b) Spatial interaction module. (c) Channel interaction module.

where $F$ denotes the input feature map, $F_e$ denotes the embedded output, Concat denotes concatenation along the channel dimension, $\text{Conv}_1*$ represents the standard convolution operation with $7 \times 7$ kernel size, and $\sigma$ denotes the sigmoid function. Notably, CWMP and CWAP represent channel-wise max-pooling and average pooling operations, respectively.

Then, to enhance feature extraction while maintaining processing efficiency, RFMM splits the embedded feature map $F_e$ into two channel-wise segments: $F_{e1}$ and $F_{e2}$. The former undergoes lightweight spatial and channel interaction operations [illustrated in Fig. 3(b) and (c)] to extract context-enhanced features, which are then element-wise multiplied with the latter to modulate and refine feature representation. This design facilitates selective information aggregation, which can improve the network's expressive power while reducing computational complexity compared to other attention-based modules such as [25]. A residual connection is further introduced to retain the original information flow, enhancing training stability and convergence. The formulation of RFMM is defined as

$$F_r = F_e + \text{DWConv} * \left(\left(\text{DWConv} * F''_{e1}\right) \cdot F_{e2}\right) \quad (4)$$

$$F''_{e1} = F'_{e1} \cdot \sigma\left(\text{Conv}_2 * \left(\text{GAP}\left(F'_{e1}\right)\right)\right) \quad (5)$$

$$F'_{e1} = F_{e1} \cdot \left(\text{Conv}_2 * \left(\text{DWConv} * \left(F_e\right)\right)\right) \quad (6)$$

where $F_r$ denotes the refined feature map through RFMM, DWConv$*$ denotes DW convolution operation with $3\times3$ kernel size, $\text{Conv}_2*$ represents the standard convolution operation with $1 \times 1$ kernel size, and GAP represents global average pooling operations.

Thereafter, the refined feature map $F_r$ is passed through a lightweight MLP, which serves to further enhance the network's feature representation ability and avoid overfitting by modeling nonlinear dependencies across channels. The structure of MLP comprises `Conv-GELU-Dropout-Conv-Dropout`.

### B. Attention-Enhanced Discriminator

As illustrated in Fig. 2, the proposed attention-enhanced discriminator adopts an encoder–decoder architecture. The encoder consists of four Attention-Conv (AC) blocks, while the decoder is composed of four standard transposed convolution blocks. The discriminator receives two sets of concatenated data: one is the predicted pixel-wise crack detection results $O$ conditioned on the input image $I$ and another is the ground-truth map $Y$ conditioned on $I$. During training, the discriminator is optimized to distinguish between real ($Y$ conditioned on $I$) and fake ($O$ conditioned on $I$) inputs from both global (over the whole concatenated data) and local (per-pixel) aspects, while the U-shaped segmentation network is simultaneously trained to generate $O$ that can fool the discriminator. This adversarial design helps to improve the crack detection performance of the U-shaped segmentation network in a higher-order way.

Specifically, in the encoder part, each ac block is composed of `Conv-BatchNorm-ReLU-directional-spatial attention (DSA)`. We innovatively design a DSA module, aiming to strengthen the representation of informative features from both directional and spatial perspectives, thereby enhancing the discriminator's discriminative ability. DSA first divides the input feature map into multiple subgroups along the channel dimension for group-wise processing, allowing each subgroup to extract features independently. This reduces intergroup interference and promotes more diverse and discriminative representations. Then, adaptive average pooling is applied along vertical and horizontal directions to each subgroup to capture directional global context. The resulting features are concatenated and fused via a convolution operation to enable localized cross-channel interaction. The process is formulated as follows:

$$F_d = \text{Conv}_2 * \text{Concat}\left(\text{AgPool}_v\left(F_g\right), \text{AgPool}_h\left(F_g\right)\right) \quad (7)$$

$$\text{AgPool}_v\left(F_g\right)(i,j) = \frac{1}{k_v}\sum_{m=1}^{k_v} F_g\left(i \cdot s_v + m, j\right) \quad (8)$$

$$\text{AgPool}_h\left(F_g\right)(i,j) = \frac{1}{k_h}\sum_{n=1}^{k_h} F_g\left(i, j \cdot s_h + n\right) \quad (9)$$

where $F_g$ denotes the input feature map after subgrouping, $F_d$ denotes the output feature map after directional-attention processing, $\text{AgPool}_v$ and $\text{AgPool}_h$ represent vertical and horizontal adaptive average pooling, respectively, $k_v$ and $k_h$ denote the size of pooling window in the vertical and horizontal directions, respectively, and $s_v$ and $s_h$ denote vertical and horizontal step lengths, respectively.

To further enhance spatial structure modeling, DSA incorporates the channel-wise spatial attention (introduced in Section III-A) in a parallel branch. The designed directional and spatial attention modules are jointly applied to each subgroup, and their outputs are fused via element-wise multiplication to adaptively reweight feature responses, thereby facilitating effective multiscale contextual integration. Subsequently, DSA incorporates the global average pooling operation (introduced in Section III-A) on the fused feature map after group normalization to capture global dependencies. Furthermore, a cross-connection strategy is employed, allowing $F_g$ to be involved in the reweighting of feature responses, which can reuse shallow feature information and alleviate the loss of important details during informative feature extraction.

### C. Dynamic Morphological Iteration Algorithm

Inspired by [26], we propose a novel DMIA, which can adaptively extract boundary maps from corresponding

ground-truth maps. These boundary maps are integrated with image data as a multimodal signal for the training process of the teacher network, thereby enhancing its sensitivity to crack boundaries. Such boundary-awareness is then effectively transferred to the student network via the designed knowledge distillation training strategy. Notably, compared to [26], the proposed DMIA can dynamically adjust its morphological parameters in response to variations in crack width. This adaptive mechanism enables the algorithm to accommodate a wide range of crack patterns and environmental conditions, thereby ensuring more precise and robust boundary delineation.

After preprocessing, DMIA estimates an adaptive kernel radius based on the input ground-truth map $Y$. Specifically, a small elliptical structuring element is iteratively applied to erode $Y$ until the foreground region vanishes. The number of erosion iterations required is then used to determine a suitable kernel radius through proportional scaling. Based on this radius, a diamond-shaped structuring element is constructed, where the included pixels satisfy a predefined Manhattan distance constraint. This structuring element is subsequently used to dilate $Y$, and the boundary map is obtained by subtracting $Y$ from the dilated result.

### D. Loss Function

For the training of LR-CrackNet, the total loss can be formulated as follows:

$$\mathcal{L} = \lambda_{Kd}\mathcal{L}_{Kd} + \lambda_{\text{Seg}}\mathcal{L}_{\text{Seg}} + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}} \tag{10}$$

where $\lambda_{Kd}$, $\lambda_{\text{Seg}}$, and $\lambda_{\text{adv}}$ are weighting parameters that harmonize the designed knowledge distillation loss $\mathcal{L}_{Kd}$, segmentation loss $\mathcal{L}_{\text{Seg}}$, and adversarial loss $\mathcal{L}_{\text{adv}}$, respectively.

*1) Knowledge Distillation Loss:* As stated above, we adopt two distillation training strategies to facilitate knowledge transfer in the designed TS architecture. Thus, $\lambda_{Kd}$ consists of two parts

$$\mathcal{L}_{Kd} = \lambda_{FD}\mathcal{L}_{FD} + \lambda_{LD}\mathcal{L}_{LD} \tag{11}$$

where $\lambda_{FD}$ and $\lambda_{LD}$ denote weighting parameters that harmonize feature distillation loss $\mathcal{L}_{FD}$ and logits distillation loss $\mathcal{L}_{LD}$, respectively. In detail, we adopt a channel-wise divergence (CWD) loss function to constrain intermediate feature representations (from the third upsampling block of the U-shaped segmentation network) and detection results between the teacher network and the student network. $\mathcal{L}_{FD}$ and $\mathcal{L}_{LD}$ are defined as

$$\mathcal{L}_{FD} = \frac{1}{N \cdot C}\sum_{n=1}^{N}\sum_{c=1}^{C} KL\left(\phi\left(T_{n,c}^{\text{up3}}\right) \| \phi\left(S_{n,c}^{\text{up3}}\right)\right) \tag{12}$$

$$\mathcal{L}_{LD} = \frac{1}{N \cdot C}\sum_{n=1}^{N}\sum_{c=1}^{C} KL\left(\phi\left(T_{n,c}^{\text{logits}}\right) \| \phi\left(S_{n,c}^{\text{logits}}\right)\right) \tag{13}$$

$$KL\left(P \| Q\right) = \sum_{i=1}^{H \cdot W} P_i \cdot \left(\log P_i - \log Q_i\right) \tag{14}$$

$$\phi\left(T\right)_i = \frac{e^{T_i}}{\sum_{j=1}^{H \cdot W} e^{T_j}}, \quad i = 1, 2, \ldots, H \cdot W \tag{15}$$

where $N$ and $C$ denote the batch size and the number of feature channels, respectively, $H$ and $W$ denote the height and width of the feature map, respectively, $i$ is used to index the spatial positions, and $j$ is used as the summation index over all spatial positions in the Softmax operation $\phi$. Compared with traditional pixel-level knowledge distillation losses, the proposed $\mathcal{L}_{Kd}$ enables more fine-grained alignment of semantic information at the channel level, which helps improve the detection performance of the student model.

*2) Segmentation Loss:* To effectively guide the training process of the student network, the designed segmentation loss combines the binary cross-entropy (BCE) loss and dice loss, denoted as

$$\mathcal{L}_{\text{seg}} = \lambda_{\text{bce}}\mathcal{L}_{\text{BCE}} + \lambda_{\text{dice}}\mathcal{L}_{\text{Dice}} \tag{16}$$

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{HW}\sum_{i,j}\left[Y_{ij}\log O_{ij} + \left(1 - Y_{ij}\right)\log\left(1 - O_{ij}\right)\right] \tag{17}$$

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2\sum_{i,j} O_{ij}Y_{ij} + \varepsilon}{\sum_{i,j} O_{ij} + \sum_{i,j} Y_{ij} + \varepsilon} \tag{18}$$

where $\varepsilon$ is a small constant to avoid division by zero. The combined $\mathcal{L}_{\text{seg}}$ takes advantage of both BCE and dice losses; the former ensures stable pixel-wise supervision, whereas the latter improves region-level consistency, especially for detecting thin and sparse cracks. This design facilitates effective learning in challenging scenarios where cracks are fragmented and cover only a small area of the image.

*3) Discriminator Loss:* Furthermore, to encourage the proposed network to produce crack detection outputs $O$ that more closely align with the ground-truth map $Y$, we introduce an adversarial loss $\mathcal{L}_{\text{adv}}$ to provide both global structural supervision and local per-pixel feedback. The adversarial loss $\mathcal{L}_{\text{adv}}$ is defined as

$$\mathcal{L}_{\text{adv}} = \lambda_D\mathcal{L}_D + \lambda_G\mathcal{L}_G \tag{19}$$

$$\mathcal{L}_D = \lambda_{D_g}\mathcal{L}_{D_g} + \lambda_{D_l}\mathcal{L}_{D_l} \tag{20}$$

where $\lambda_D$, $\lambda_G$, $\lambda_{D_g}$, and $\lambda_{D_l}$ denote weighting parameters that harmonize discriminator loss $\mathcal{L}_D$, generator loss $\mathcal{L}_G$, global discriminator loss $\mathcal{L}_{D_g}$, and local discriminator loss $\mathcal{L}_{D_l}$, respectively. In detail, these loss functions are defined as follows:

$$\begin{aligned}\mathcal{L}_{D_g} = -\big(&\mathbb{E}_{Y,I}\left[\log D_{en}\left(Y, I\right)\right] \\ &+\mathbb{E}_{O,I}\left[\log\left(1 - D_{en}\left(O, I\right)\right)\right]\big)\end{aligned}$$

$$\begin{aligned}\mathcal{L}_{D_l} = -\big(&\mathbb{E}_{Y,I}\left[\log D\left(Y, I\right)_{i,j}\right] \\ &+\mathbb{E}_{O,I}\left[\log\left(1 - D\left(O, I\right)_{i,j}\right)\right]\big)\end{aligned}$$

$$\begin{aligned}\mathcal{L}_G = -\big(&\mathbb{E}_{O,I}\left[\log D_{en}\left(O, I\right)\right] \\ &+\mathbb{E}_{O,I}\left[\log D\left(O, I\right)_{i,j}\right]\big)\end{aligned}$$

where $D_{en}$ represents the encoder part of the designed attention-enhanced discriminator and $D(Y, I)_{i,j}$ denotes the discriminator decision at pixel $(i, j)$.

## IV. Experiments

### A. Datasets

The **UDTIRI-Crack** [27], [28] is a high-quality integrated dataset consisting of 2500 road images (resolution:

$320 \times 320$) sourced from seven public datasets, which include five crack types (alligator, transverse, longitudinal, multifurcate, and pit) under various road materials (concrete, asphalt, etc.) and noise factors (oil spots, zebra crossing markings, etc.). These images have been annotated at the pixel level. The UDTIRI-Crack dataset is divided into three subsets, with 1500 images for training, 400 images for validation, and the remaining 600 images for testing.

The **DeepCrack** [11] dataset contains 537 road images (resolution: $544 \times 384$ pixels) with multiscale and multiscene cracks. These images have also been annotated at the pixel level. The DeepCrack dataset is randomly separated into three subsets, with 210 images for training, 90 images for evaluation, and the remaining 237 images for testing.

The **CamCrack789** [20] dataset contains 789 road images (resolution: $640 \times 480$) with four types of road cracks (common, intersecting, block, and alligator) under various noise factors (water stains, leaves, debris, etc.). These images have been annotated at the pixel level. The CamCrack789 dataset is randomly separated into three subsets, with 328 images for training, 218 images for evaluation, and the remaining 243 images used for testing.

### B. Implementation Details

All experiments are conducted on a single NVIDIA RTX3090. We use a progressive training strategy with warm-up and learning rate decay to stabilize LR-CrackNet's training. First, the teacher network is trained for 50 epochs with image data, and the learning rate (lr) is set as $1e-3$ to learn basic structural features. Then, boundary information is added, and lr is reduced to $1e-4$ to support stable multimodal learning with 100 training epochs. Finally, the student network is trained for 200 epochs with lr $= 1e-4$ via knowledge distillation and adversarial training. During training, the model is evaluated on the validation set every 5 epochs, and the parameters of the best validation performance are saved. In our implementation, according to experimental experience, we set $\{\lambda_{Kd}, \lambda_{\text{Seg}}, \lambda_{\text{adv}}, \lambda_{FD}, \lambda_{D_g}, \lambda_{D_l}\} = 1$, $\{\lambda_{\text{bce}}, \lambda_{\text{dice}}, \lambda_G\} = 0.5$, $\lambda_D = 0.4$, and $\lambda_{LD}$ as 1.5. To ensure fair comparison, we adopt precision, recall, accuracy, intersection over union (IoU), $F1$-score, and average IoU (AIoU) as evaluation metrics. In addition, model parameters and FPS are used to evaluate the model complexity and processing speed.

### C. Ablation Study

To assess the contribution of our proposed loss design, we perform a series of ablation studies on the DeepCrack [11] and CamCrack789 [20] datasets, as summarized in Table I. The experimental results clearly indicate that each component of the designed losses plays a meaningful role in enhancing the model's detection performance. Among all configurations, the complete version of utilizing all losses achieves the best detection performance, demonstrating the effectiveness of our overall design. These findings highlight the advantages of the proposed multimodal distillation architecture and the designed adversarial training strategy.

Furthermore, we conduct ablation studies on the DeepCrack [11] and CamCrack789 [20] datasets to examine how the integration of the designed RDLDC and fast Transformer (Fast-T) blocks affects the tradeoff between detection performance and processing efficiency in the proposed U-shaped segmentation framework. The quantitative results are presented in Table II. The first row corresponds to a baseline U-Net [29] with base channels reduced from 64 to 32 for higher processing efficiency. The second and third rows denote variants without Fast-T and without RDLDC, respectively, while the fourth integrates another Transformer variant [30]. These configurations allow us to isolate and evaluate the individual contributions of Fast-T and RDLDC blocks to the overall model performance. Results confirm their effectiveness in efficiently capturing detailed hierarchical spatial and global contextual information, thereby improving detection without incurring computational overhead.

TABLE I
ABLATION STUDIES ON THE DEEPCRACK [11] AND CAMCRACK789 [20] DATASETS TO VALIDATE THE EFFECTIVENESS OF THE DESIGNED LOSS FUNCTIONS

| $\mathcal{L}_{Seg}$ | $\mathcal{L}_{D_g}$ | $\mathcal{L}_{D_l}$ | $\mathcal{L}_{FD}$ | $\mathcal{L}_{LD}$ | DeepCrack | | CamCrack789 | |
|---|---|---|---|---|---|---|---|---|
| | | | | | F1-Score(%)↑ | IoU(%)↑ | F1-Score(%)↑ | IoU(%)↑ |
| ✓ | | | | | 81.397 | 68.630 | 80.233 | 66.992 |
| ✓ | ✓ | | | | 82.285 | 69.902 | 81.640 | 68.976 |
| ✓ | ✓ | ✓ | | | 82.454 | 70.146 | 82.204 | 69.785 |
| ✓ | ✓ | ✓ | ✓ | | 82.990 | 70.925 | 82.812 | 70.666 |
| ✓ | ✓ | ✓ | ✓ | ✓ | **84.099** | **72.561** | **83.025** | **70.977** |

TABLE II
ABLATION STUDIES ON THE DEEPCRACK [11] AND CAMCRACK789 [20] DATASETS TO VALIDATE THE EFFECTIVENESS OF THE DESIGNED RDLDC AND FAST-TRANSFORMER BLOCKS

| Methods | FPS↑ | DeepCrack | | CamCrack789 | |
|---|---|---|---|---|---|
| | | F1-Score(%)↑ | IoU(%)↑ | F1-Score(%)↑ | IoU(%)↑ |
| Base* | 376.69 | 79.587 | 66.095 | 80.563 | 67.452 |
| Fast-T (w/o) | **443.83** | 80.856 | 67.864 | 81.559 | 68.860 |
| RDLDC (w/o) | 242.79 | 82.069 | 69.590 | 82.416 | 70.091 |
| Standard-T (w*) | 266.22 | 81.861 | 69.292 | 82.195 | 69.772 |
| LR-CrackNet (w) | 324.40 | **84.099** | **72.561** | **83.025** | **70.977** |

### D. Comparison With Other SoTA Methods

To validate the superiority of the proposed LR-CrackNet in terms of both detection performance and processing efficiency, we compare it with eight public road crack detection-specific algorithms (Deepcrack18 [14], Deepcrack19 [11], SCCD-Net [31], Crack-Att [32], HrSegNet [22], CDLN [33], LECSFormer [34], and CT-crackseg [26]) on the UDTIRI-Crack [27], DeepCrack [11], and CamCrack789 [20] datasets. On the one hand, quantitative and qualitative results regarding detection performance are summarized in Tables III, IV, and V and illustrated in Fig. 4. As shown in the results, LR-CrackNet consistently achieves the best detection performance across all three datasets. For instance, on the UDTIRI-Crack dataset, it reaches an $F1$-score of 71.419%, surpassing the second-best method by 0.807%. Similar performance gains are observed on the other two datasets, validating the robustness of LR-CrackNet under diverse road conditions.

Table VI reports model parameters and processing efficiency of LR-CrackNet and other methods on an NVIDIA RTX3090 and a 32GB AGX Orin using $320 \times 320$ inputs. LR-CrackNet attains the highest FPS on the RTX3090 and ranks second on the AGX Orin, slightly behind HrSegNet, yet surpassing others. Given its markedly better detection accuracy

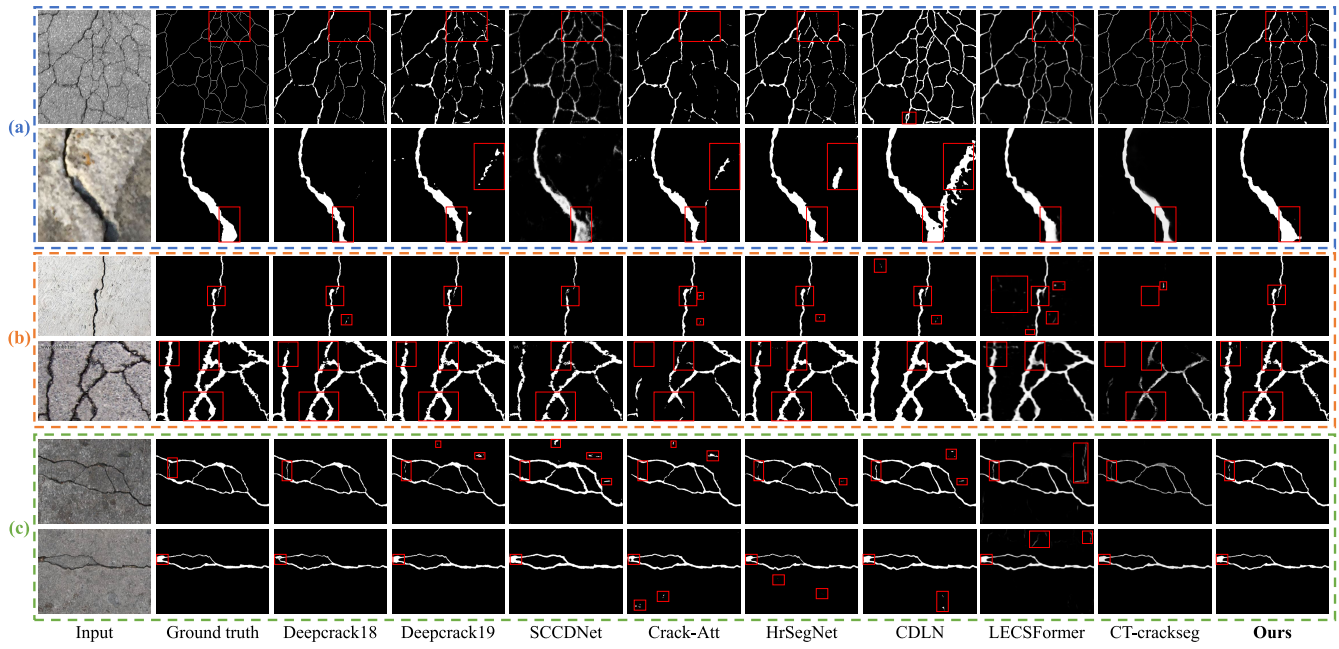



Fig. 4. Comparison results on the (a) UDTIRI-Crack [27], (b) Deepcrack [11], and (c) CamCrack789 [20] datasets.

TABLE III
QUANTITATIVE EXPERIMENTAL RESULTS OF DETECTION PERFORMANCE ON THE UDTIRI-CRACK DATASET [27]

| Methods | F1-Score (%)↑ | IoU (%)↑ | AIoU (%)↑ |
|---|---|---|---|
| Deepcrack18 [14] | 65.756 | 48.982 | 73.708 |
| Deepcrack19 [11] | 64.448 | 47.544 | 72.956 |
| SCCDNet [31] | 64.189 | 47.263 | 72.797 |
| Crack-Att [32] | 67.598 | 51.055 | 74.710 |
| HrSegNet [22] | 66.770 | 50.117 | 74.281 |
| CDLN [33] | 67.522 | 50.968 | 74.456 |
| LECSFormer [34] | 69.712 | 53.506 | 76.025 |
| CT-crackseg [26] | 70.612 | 54.573 | 76.574 |
| **ours** | **71.419** | **55.545** | **77.060** |

TABLE IV
QUANTITATIVE EXPERIMENTAL RESULTS OF PIXEL-WISE CRACK DETECTION PERFORMANCE ON THE DEEPCRACK DATASET [11]

| Methods | F1-Score (%)↑ | IoU (%)↑ | AIoU (%)↑ |
|---|---|---|---|
| Deepcrack18 [14] | 78.149 | 64.135 | 81.219 |
| Deepcrack19 [11] | 81.276 | 68.458 | 83.468 |
| SCCDNet [31] | 76.164 | 61.504 | 79.590 |
| Crack-Att [32] | 78.191 | 64.191 | 81.210 |
| HrSegNet [22] | 81.453 | 68.709 | 83.598 |
| CDLN [33] | 82.242 | 69.839 | 84.087 |
| LECSFormer [34] | 82.389 | 70.052 | 84.268 |
| CT-crackseg [26] | 80.825 | 67.821 | 83.158 |
| **ours** | **84.099** | **72.561** | **85.622** |

TABLE V
QUANTITATIVE EXPERIMENTAL RESULTS OF PIXEL-WISE CRACK DETECTION PERFORMANCE ON THE CAMCRACK789 DATASET [20]

| Methods | F1-Score (%)↑ | IoU (%)↑ | AIoU (%)↑ |
|---|---|---|---|
| Deepcrack18 [14] | 76.935 | 62.515 | 80.479 |
| Deepcrack19 [11] | 80.991 | 68.054 | 83.347 |
| SCCDNet [31] | 70.921 | 54.944 | 76.042 |
| Crack-Att [32] | 74.949 | 59.935 | 79.007 |
| HrSegNet [22] | 80.673 | 67.608 | 83.125 |
| CDLN [33] | 81.632 | 68.964 | 83.718 |
| LECSFormer [34] | 80.177 | 66.913 | 82.703 |
| CT-crackseg [26] | 79.569 | 66.070 | 82.334 |
| **ours** | **83.025** | **70.977** | **84.865** |

TABLE VI
QUANTITATIVE EXPERIMENTAL RESULTS IN TERMS OF MODEL PARAMETERS AND PROCESSING EFFICIENCY

| Methods | Parameters ($M$)↓ | FPS (on RTX3090)↑ | FPS (on AGX Orin)↑ |
|---|---|---|---|
| Deepcrack18 [14] | 30.905 | 43.935 | 2.762 |
| Deepcrack19 [11] | 14.720 | 276.695 | 27.030 |
| SCCDNet [31] | 31.705 | 113.317 | 9.365 |
| Crack-Att [32] | 45.804 | 58.257 | 5.847 |
| HrSegNet [22] | 9.641 | 300.15 | **51.726** |
| CDLN [33] | 19.151 | 45.934 | 13.857 |
| LECSFormer [34] | 16.528 | 65.460 | 12.062 |
| CT-crackseg [26] | 22.882 | 24.926 | 2.561 |
| **ours** | **3.856** | **324.40** | 46.419 |

than HrSegNet, LR-CrackNet remains highly competitive overall. The minor slowdown on AGX Orin is mainly due to the computational cost of attention modules, which enhance accuracy but are not yet optimized for edge devices. Future work will involve hardware-friendly designs and pruning to further cut latency. Combined with its accuracy and compactness, LR-CrackNet offers a strong balance of performance and efficiency, making it well-suited for real-time autonomous road inspection.

## V. CONCLUSION

Reliable road crack detection plays a vital role in ensuring road safety and enabling timely infrastructure maintenance. Despite advances in deep learning, current road crack detection

methods still struggle to balance detection performance, model complexity, and processing speed. In this article, we propose LR-CrackNet, a lightweight, robust network adopting a teacher–student architecture: the teacher exploits both images and boundary information self-derived from annotations, while the student uses only images. Knowledge distillation is realized via a shared U-shaped segmentation framework incorporating RDLDC blocks and a Fast-Transformer block to capture detailed hierarchical spatial details and global long-range contextual information efficiently. An attention-enhanced discriminator further enforces global consistency and provides local feedback to enhance detection performance. Extensive experiments on three public datasets demonstrate that LR-CrackNet achieves superior performance with fewer parameters and higher efficiency than publicly available methods. Future work will focus on building more diverse datasets and exploring two-stage frameworks to further improve detection reliability in complex real-world scenarios.

## REFERENCES


[1] R. Fan, X. Ai, and N. Dahnoun, "Road surface 3D reconstruction based on dense subpixel disparity map estimation," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 3025–3035, Jun. 2018.

[2] R. Fan, H. Wang, Y. Wang, M. Liu, and I. Pitas, "Graph attention layer evolves semantic segmentation for road pothole detection: A benchmark and algorithms," *IEEE Trans. Image Process.*, vol. 30, pp. 8144–8154, 2021.

[3] J. Nehme, "About long-term pavement performance," 2013 Report Card for America's Infrastructure, American Society of Civil Engineers, Reston, VA, USA, 2013. [Online]. Available: https://www.fhwa.dot.gov/research/tfhrc/programs/infrastructure/pavements/ltpp/

[4] *Road Safety Statistics*, Dept. Transp., London, U.K., 2022.

[5] R. Fan, U. Ozgunalp, B. Hosking, M. Liu, and I. Pitas, "Pothole detection based on disparity transformation and road surface modeling," *IEEE Trans. Image Process.*, vol. 29, pp. 897–908, 2020.

[6] N. Ma, R. Fan, and L. Xie, "UP-CrackNet: Unsupervised pixel-wise road crack detection via adversarial image restoration," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 10, pp. 13926–13936, Oct. 2024.

[7] R. Fan and M. Liu, "Road damage detection based on unsupervised disparity map segmentation," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 11, pp. 4906–4911, Nov. 2020.

[8] R. Fan, U. Ozgunalp, Y. Wang, M. Liu, and I. Pitas, "Rethinking road surface 3-D reconstruction and pothole detection: From perspective transformation to disparity map segmentation," *IEEE Trans. Cybern.*, vol. 52, no. 7, pp. 5799–5808, Jul. 2022.

[9] C.-W. Liu, Y. Zhang, Q. Chen, I. Pitas, and R. Fan, "These maps are made by propagation: Adapting deep stereo networks to road scenarios with decisive disparity diffusion," *IEEE Trans. Image Process.*, vol. 34, pp. 1516–1528, 2025.

[10] N. Ma et al., "Computer vision for road imaging and pothole detection: A state-of-the-art review of systems and algorithms," *Transp. Saf. Environ.*, vol. 4, no. 4, Nov. 2022, Art. no. tdac026.

[11] Y. Liu, J. Yao, X. Lu, R. Xie, and L. Li, "DeepCrack: A deep hierarchical feature learning architecture for crack segmentation," *Neurocomputing*, vol. 338, pp. 139–153, Apr. 2019.

[12] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[13] T. Zhang, D. Wang, and Y. Lu, "ECSNet: An accelerated real-time image segmentation CNN architecture for pavement crack detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 12, pp. 15105–15112, Dec. 2023.

[14] Q. Zou, Z. Zhang, Q. Li, X. Qi, Q. Wang, and S. Wang, "DeepCrack: Learning hierarchical convolutional features for crack detection," *IEEE Trans. Image Process.*, vol. 28, no. 3, pp. 1498–1512, Mar. 2019.

[15] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder–decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[16] X. Sun, Y. Xie, L. Jiang, Y. Cao, and B. Liu, "DMA-Net: DeepLab with multi-scale attention for pavement crack segmentation," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 10, pp. 18392–18403, Oct. 2022.

[17] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder–decoder with Atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 833–851.

[18] F. Guo, Y. Qian, J. Liu, and H. Yu, "Pavement crack detection based on transformer network," *Autom. Construct.*, vol. 145, Jan. 2023, Art. no. 104646.

[19] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.

[20] G. Zhu et al., "A lightweight encoder–decoder network for automatic pavement crack detection," *Comput.-Aided Civil Infrastruct. Eng.*, vol. 39, no. 12, pp. 1743–1765, Jun. 2024.

[21] Q. Zhou, Z. Qu, and F.-R. Ju, "A lightweight network for crack detection with split exchange convolution and multi-scale features fusion," *IEEE Trans. Intell. Vehicles*, vol. 8, no. 3, pp. 2296–2306, Mar. 2023.

[22] Y. Li, R. Ma, H. Liu, and G. Cheng, "Real-time high-resolution neural network with semantic guidance for crack segmentation," *Autom. Construct.*, vol. 156, Dec. 2023, Art. no. 105112.

[23] X. Wang, Z. Mao, Z. Liang, and J. Shen, "Multi-scale semantic map distillation for lightweight pavement crack detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 10, pp. 15081–15093, Oct. 2024.

[24] J. Zhang et al., "Crack segmentation-guided measurement with lightweight distillation network on edge device," *Comput.-Aided Civil Infrastruct. Eng.*, vol. 40, no. 16, pp. 2269–2286, Jun. 2025.

[25] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 3–19.

[26] H. Tao, B. Liu, J. Cui, and H. Zhang, "A convolutional-transformer network for crack segmentation with boundary awareness," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2023, pp. 86–90.

[27] N. Ma et al., "Vehicular road crack detection with deep learning: A new online benchmark for comprehensive evaluation of existing algorithms," 2025, *arXiv:2503.18082*.

[28] S. Guo et al., "UDTIRI: An online open-source intelligent road inspection benchmark suite," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 8, pp. 9920–9931, Aug. 2024.

[29] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, 2015, pp. 234–241.

[30] T. Zhang et al., "CAS-ViT: Convolutional additive self-attention vision transformers for efficient mobile applications," 2024, *arXiv:2408.03703*.

[31] H. Li et al., "SCCDNet: A pixel-level crack segmentation network," *Appl. Sci.*, vol. 11, no. 11, p. 5074, May 2021.

[32] N. Xu, L. He, and Q. Li, "Crack-Att Net: Crack detection based on improved U-Net with parallel attention," *Multimedia Tools Appl.*, vol. 82, no. 27, pp. 42465–42484, Nov. 2023.

[33] P. Manjunatha, S. F. Masri, A. Nakano, and L. C. Wellford, "Crack-DenseLinkNet: A deep convolutional neural network for semantic segmentation of cracks on concrete surface images," *Struct. Health Monitor.*, vol. 23, no. 2, pp. 796–817, Mar. 2024.

[34] J. Chen, N. Zhao, R. Zhang, L. Chen, K. Huang, and Z. Qiu, "Refined crack detection via LECSFormer for autonomous road inspection vehicles," *IEEE Trans. Intell. Vehicles*, vol. 8, no. 3, pp. 2049–2061, Mar. 2023.