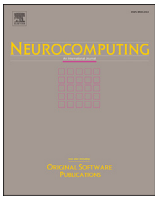
Neurocomputing 659 (2026) 131845

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



WP-CrackNet: A collaborative adversarial learning framework for end-to-end weakly-supervised road crack detection

Nachuan Ma[a], Zhengfei Song[a], Qiang Hu[a], Xiaoyu Tang[b], Chengxi Zhang[c], Rui Fan[a,*], Lihua Xie[d]

[a] College of Electronic and Information Engineering, Shanghai Research Institute for Intelligent Autonomous Systems, Tongji University, Shanghai, China
[b] School of Electronics and Information Engineering, and Xingzhi College, South China Normal University, Shanwei, China
[c] School of Internet of Things Engineering, Jiangnan University, Wuxi, China
[d] School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, Singapore

ARTICLE INFO

Communicated by L. Zhu

Keywords:
Weakly-supervised
Road cracks
Deep learning
Computer vision

ABSTRACT

Road crack detection is essential for intelligent infrastructure maintenance in smart cities. To reduce reliance on costly pixel-level annotations, we propose WP-CrackNet, an end-to-end weakly-supervised method that trains with only image-level labels for pixel-wise crack detection. WP-CrackNet integrates three components: a classifier generating class activation maps (CAMs), a reconstructor measuring feature inferability, and a detector producing pixel-wise road crack detection results. During training, the classifier and reconstructor alternate in adversarial learning to encourage crack CAMs to cover complete crack regions, while the detector learns from pseudo labels derived from post-processed crack CAMs. This mutual feedback among the three components improves learning stability and detection accuracy. To further boost detection performance, we design a path-aware attention module (PAAM) that fuses high-level semantics from the classifier with low-level structural cues from the reconstructor by modeling spatial and channel-wise dependencies. Additionally, a center-enhanced CAM consistency module (CECCM) is proposed to refine crack CAMs using center Gaussian weighting and consistency constraints, enabling better pseudo-label generation. We create three image-level datasets and extensive experiments show that WP-CrackNet achieves comparable results to supervised methods and outperforms existing weakly-supervised methods, significantly advancing scalable road inspection. The source code package and datasets are available at https://mias.group/WP-CrackNet/.

1. Introduction

Road cracks, typically appearing as narrow lines or curves on road surfaces, serve as early indicators of structural deterioration in civil infrastructure. Although subtle, road cracks can progressively compromise the integrity of roadways and pose significant safety hazards. For instance, in 2020, deteriorated road conditions contributed to 12.6 % of traffic accidents in the UK [1]. Therefore, regular inspection and timely maintenance are crucial to reducing risks and extending the service life of networks [2,3]. At present, manual visual inspection remains the primary approach for crack detection [4], but it is costly, labor-intensive, time-consuming, and highly subjective—especially for extensive highway networks exceeding 100,000 km in countries like China and the US [5]. These limitations underscore the urgent need for automated systems capable of efficiently and objectively analyzing road conditions [6]. As shown in Fig. 1, collection devices—such as car cameras, smartphones, and surveillance cameras—offer a promising solution by continuously collecting road surface images and uploading the data to cloud-based platforms for automated road crack detection [7]. The detection results can provide actionable insights for infrastructure administrators, thereby facilitating more timely and data-driven maintenance decisions.

Recent advancements in deep learning, particularly Convolutional Neural Networks (CNNs) and Transformer-based networks, have significantly enhanced automated road crack detection, which are generally categorized into three types: (1) image classification networks that distinguish between crack and non-crack images [8–11]; (2) object

* Corresponding author.
Email addresses: 2111481@tongji.edu.cn (N. Ma), 2151094@tongji.edu.cn (Z. Song), 2252974@tongji.edu.cn (Q. Hu), tangxy@scnu.edu.cn (X. Tang), cxzhang@jiangnan.edu.cn (C. Zhang), rui.fan@ieee.org (R. Fan), elhxie@ntu.edu.sg (L. Xie).

https://doi.org/10.1016/j.neucom.2025.131845
Received 29 August 2025; Received in revised form 10 October 2025; Accepted 15 October 2025
Available online 16 October 2025

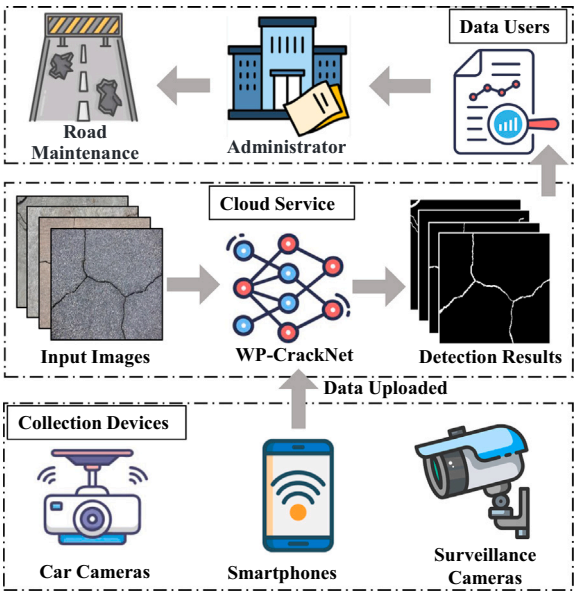


Fig. 1. Road inspection for intelligent infrastructure maintenance with the proposed WP-CrackNet.

detection networks that identify cracks at the instance level (location and class) [12–14]; and (3) semantic segmentation networks that provide pixel-level crack detection and have emerged as the dominant approach in this field [15–19]. However, the training of supervised semantic segmentation networks relies heavily on manually annotated datasets, which are costly and time-consuming to construct. This reliance limits the scalability of road surface perception, as the massive amounts of collected data are difficult to efficiently annotate at the pixel level, thereby hindering the widespread deployment in large-scale infrastructure maintenance.

To overcome this limitation, we propose a novel ***W**eakly-supervised **P**ixel-wise **C**rack Detection **Net**work (WP-CrackNet)* via adversarial mutual learning. Unlike prior methods that rely on offline pseudo-label generation [20–22] or hand-crafted cues [23], WP-CrackNet adopts an end-to-end training strategy that jointly optimizes three synergistic components—a classifier, a reconstructor, and a detector—using only image-level labels. This design simplifies the training pipeline and promotes a more stable learning process. The classifier generates CAMs to localize discriminative regions, while the reconstructor measures the inferability between road and crack features to enhance structural understanding and encourage crack CAMs to fully cover crack regions (inspired by Ref. [22]). The detector then produces pixel-wise road crack predictions based on pseudo labels derived from post-processed crack CAMs. Through alternating adversarial training between the classifier and reconstructor, whose outputs feed into the detector, WP-CrackNet achieves stable and mutually reinforcing optimization. Furthermore, we design the PAAM to effectively integrate high-level semantic information with low-level structural cues from the classifier and reconstructor to improve detection performance. Additionally, we introduce the CECCM to refine crack CAMs using center Gaussian weighting and consistency constraints for better generation of pseudo labels. Experimental evaluations on three crack datasets demonstrate that WP-CrackNet not only achieves detection performance comparable to supervised methods but also surpasses SoTA general and road crack detection-specific weakly-supervised methods. The main contributions can be summarized as follows:

1. We propose WP-CrackNet, a novel end-to-end weakly-supervised pixel-wise crack detection method trained using only image-level labels.
2. We utilize an adversarial mutual learning strategy for three components which improves learning stability and road crack detection performance.
3. We design a path-aware attention module to effectively integrate semantic context and structural cues and a center-enhanced CAM consistency module to refine crack CAMs for better generation of pseudo labels.
4. We validate WP-CrackNet on three self-created datasets, showing comparable performance to supervised methods and outperforming other weakly-supervised methods.

The remainder of this paper is organized as follows: Section 2 reviews related work. Section 3 details the proposed methodology. Section 4 presents implementation details, ablation studies, and comparative experiments. Finally, Section 5 concludes the paper.

2. Related work

2.1. Deep learning-based road crack detection methods

Supervised methods based on CNNs and Transformers have been extensively developed for road crack detection, employing architectures such as FCN [24], SegNet [25], Deeplab [26], U-Net [15], *etc*. Among them, Zou et al. [27] proposed Deepcrack, by fusing features from different scales of SegNet [28] to obtain hierarchical details, subsequently resulting in accurate pixel-wise road crack detection results. Similarly, Yang et al. [29] proposed a feature pyramid-based hierarchical boosting network (FPHBN), introducing a side network on the HED network [30] to learn hierarchical feature information for road crack detection. Another Deepcrack version [31] incorporated a side-output layer into VGG-16 [32], adopting guided filtering and conditional random field techniques to achieve improved road crack detection performance. Tao et al. [33] proposed a novel convolutional-transformer network to combine both local and global information extracted from road crack images. In addition, a boundary awareness module was designed to capture boundary details of road cracks to refine crack detection results. Nevertheless, the methods mentioned above are data-driven, and training them relies on massive pixel-level human-annotated labels. The process of obtaining such fine labels is significantly time-consuming and laborious. To reduce this burden, label-efficient methods have been explored. Despite their potential, these methods still face notable limitations. The unsupervised method in [34] struggles to detect fine cracks due to their subtle visual characteristics. The semi-supervised method in [35] remains sensitive to the quality and quantity of labeled data, with performance degrading significantly when labels are scarce. The weakly-supervised method in [23] relies on manually crafted image processing techniques to generate pseudo labels from CAMs offline, leading to unstable outcomes and suboptimal performance.

2.2. Weakly-supervised semantic segmentation methods

Weakly-supervised semantic segmentation (WSSS) reduces reliance on costly pixel-level annotations by using weaker supervision such as scribbles [36], bounding boxes [37], and image-level labels [38–40]. Among these, image-level labels are favored for their simplicity and scalability but provide only class presence without spatial information, making precise localization difficult. To tackle this, researchers tend to leverage CAMs [41] to highlight discriminative regions learned by classifiers and use refined CAMs as pseudo labels to train a segmentation network. However, CAMs typically focus on the most distinctive parts, failing to cover the entire target class region. To explicitly expand the CAMs, methods adopting sub-category classification [38], cross-image relationships [39], contrastive learning [42], attention modules [43,44] and adversarial erasing mechanisms [20–22] have been proposed. For instance, Kweon et al. [20] used a pre-trained classifier to erase discriminative regions, enabling more precise CAM generation by encouraging

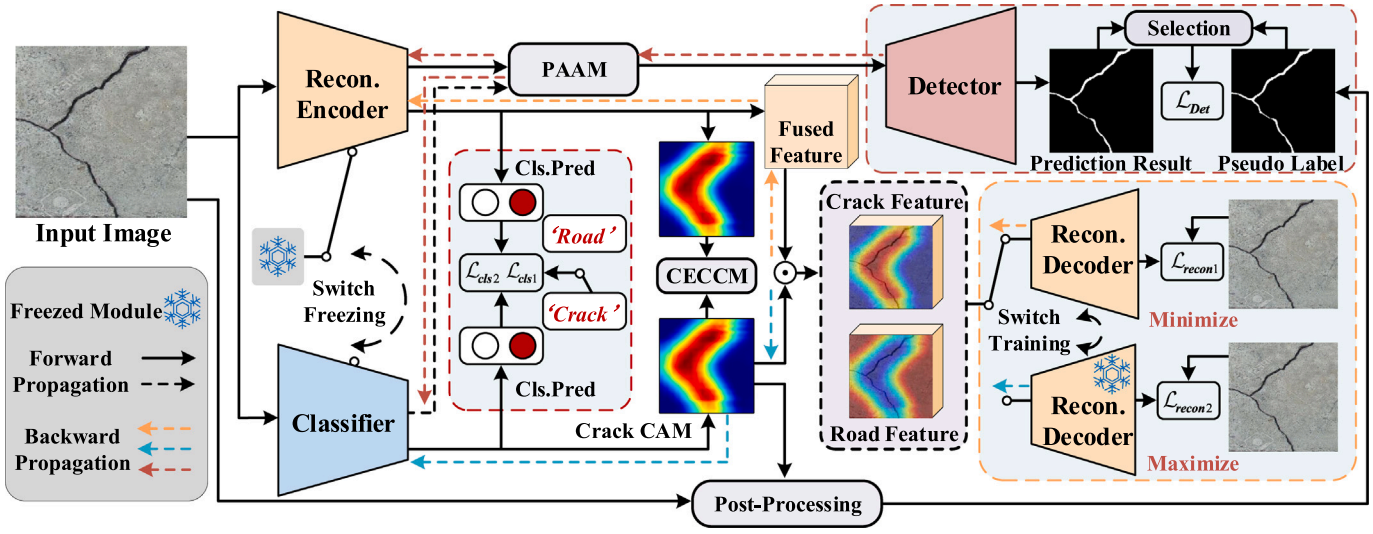


Fig. 2. The overall architecture of WP-CrackNet, consisting of a classifier, a reconstructor, and a detector. During the training phase, the classifier generates CAMs to localize discriminative regions, the reconstructor measures feature inferability, and the detector learns from pseudo labels derived from post-processed crack CAMs to produce pixel-wise road crack detection results.

activation of less discriminative regions. However, the usage of fixed classifier limits its performance. Yoon et al. [21] further introduced a triplet learning framework that relaxes reliance on the fixed classifier, enabling more flexible guidance of the erasing process and resulting in more complete CAMs. Kweon et al. [22] formulated WSSS as adversarial learning of the classifier and the reconstructor, using the reconstruction task to obtain an effective regularization of CAM generation. Inspired by the methods described above, we proposed WP-CrackNet for end-to-end weakly-supervised road crack detection. Unlike these methods relying on multi-stage processing, our method integrates feature extraction, crack localization, and segmentation into a unified network, further unlocking the potential of mutual adversarial learning. Also, PAAM and CECCM are designed to improve the quality of the generated crack CAMs and enhance the detection performance. The extensive experimental results demonstrate the superiority of WP-CrackNet over SoTA weakly-supervised semantic segmentation methods on the task of road crack detection.

3. Proposed methodology

3.1. Architecture overview

Let the road image training set be $\mathcal{I}^R = \{(\boldsymbol{I}_1^R, \boldsymbol{T}^R), \ldots, (\boldsymbol{I}_n^R, \boldsymbol{T}^R)\}$ and the crack image training set be $\mathcal{I}^C = \{(\boldsymbol{I}_1^C, \boldsymbol{T}^C), \ldots, (\boldsymbol{I}_m^C, \boldsymbol{T}_m^C)\}$, where $\boldsymbol{T}^R = [1, 0]$ denotes the non-crack class label and $\boldsymbol{T}^C = [0, 1]$ denotes the crack class label, respectively. Here, $\boldsymbol{I}_i^R$, $\boldsymbol{I}_i^C \in \mathbb{R}^{H \times W \times 3}$ denote the i-th road and crack image with H and W as the height and width. The overall architecture of WP-CrackNet is illustrated in Fig. 2, comprising a classifier Cls, a reconstructor Rec, and a detector Det. The classifier is trained via image-level supervision and generates crack CAMs to highlight crack regions within the crack image. The reconstructor is designed with an encoder–decoder architecture, aiming to assess the inferability of crack versus non-crack regions by reconstructing the input crack image. The detector is trained using pseudo labels derived from the post-processed crack CAMs to output refined pixel-wise crack detection results.

During the training phase, multi-layer fused feature maps from the encoder of reconstructor $Rec.E$ are decomposed into crack and road features based on crack CAMs, which are reconstructed separately. Inspired by Ref. [22], when crack CAMs fully cover all crack regions, the inferability between crack and road features becomes low. To leverage this, an adversarial scheme trains Rec to reconstruct one feature from the other, while Cls learns to generate crack CAMs that hinder this reconstruction. Simultaneously, Det refines its predictions by combining high-level semantic information from Cls with low-level structural cues from Rec, guided by the post-processed crack CAMs. This collaborative framework forms a reciprocal feedback loop among the three modules, improving training stability and detection performance under weak supervision.

3.2. Adversarial training of classifier and reconstructor

Given an input image $\boldsymbol{I} \in \{\boldsymbol{I}^R, \boldsymbol{I}^C\}$, the classifier and reconstructor output class predictions q and q^{rec} indicating the presence of cracks, along with CAMs $\boldsymbol{M}$ and $\boldsymbol{M}^{rec}$ that highlight discriminative regions. This process is formulated as:

$$\boldsymbol{M},\ q = Cls(\boldsymbol{I}), \quad \boldsymbol{M}^{rec},\ q^{rec} = Rec.E(\boldsymbol{I}). \tag{1}$$

In line with [45], ResNet38 [46] is employed as the backbone network for both Cls and $Rec.E$, followed by a 1×1 convolution layer serving as the classification head to generate CAMs. For input crack image $\boldsymbol{I}^C$, multi-layer fused feature map $\boldsymbol{Z}$ is obtained by passing it through $Rec.E$, and is decomposed into crack feature map $\boldsymbol{Z}_C$ and road feature map $\boldsymbol{Z}_R$ by using the corresponding crack CAM $\boldsymbol{M}_C$:

$$\boldsymbol{Z}_C = \boldsymbol{Z} \odot \boldsymbol{M}_C, \quad \boldsymbol{Z}_R = \boldsymbol{Z} \odot (1 - \boldsymbol{M}_C), \tag{2}$$

where $\odot$ denotes element-wise multiplication. Then, a switch training strategy is adopted, where Cls and $Rec.E$ are updated in turn. $\boldsymbol{Z}_C$ and $\boldsymbol{Z}_R$ are passed into the decoder of reconstructor $Rec.D$ (using a UNet-based network) to obtain corresponding reconstruction results:

$$\boldsymbol{O}_C = Rec.D(\boldsymbol{Z}_C), \quad \boldsymbol{O}_R = Rec.D(\boldsymbol{Z}_R). \tag{3}$$

When Cls is frozen, $Rec.E$ is trained to reconstruct one feature from the other. Conversely, with $Rec.E$ frozen, $\hat{\boldsymbol{O}}_C$ and $\hat{\boldsymbol{O}}_R$ are obtained, and Cls is trained to generate $\boldsymbol{M}_C$ that hinders the reconstruction of the original image.

Furthermore, considering the potential activation drift or instability in CAMs introduced by the adopted adversarial training scheme, along with the inherently slender and low-contrast nature of cracks, we propose a center-enhanced CAM consistency module (CECCM) to better guide the generation of $\boldsymbol{M}_C$. This module enhances spatial alignment between Cls and $Rec.E$ by applying Gaussian-based center weighting and

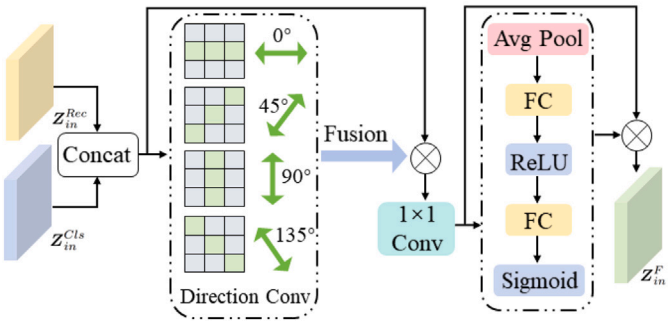


Fig. 3. The proposed PAAM, consisting of a spatial attention branch and a channel attention branch.

enforcing consistency between the center-enhanced crack CAMs. Given an input crack image, $\boldsymbol{M}_C$ and $\boldsymbol{M}_C^{rec}$ are obtained from Cls and $Rec.E$, respectively. Then, a center–enhancement operation is applied to both CAMs, guided by their spatial center of mass. Specifically, for a CAM M, we compute its normalized spatial center as:

$$\mu_x = \frac{\sum_{x,y} x \cdot \boldsymbol{M}(x,y)}{\sum_{x,y} \boldsymbol{M}(x,y)}, \quad \mu_y = \frac{\sum_{x,y} y \cdot \boldsymbol{M}(x,y)}{\sum_{x,y} \boldsymbol{M}(x,y)}. \tag{4}$$

Then, a spatial Gaussian prior $\boldsymbol{G} \in \mathbb{R}^{H \times W}$ centered at (μ_x, μ_y) is constructed:

$$\boldsymbol{G}(x,y) = \exp\left(-\frac{(x-\mu_x)^2 + (y-\mu_y)^2}{2\sigma^2}\right), \tag{5}$$

where σ controls the spread of the Gaussian. Then, the center-enhanced CAMs are obtained via element-wise multiplication:

$$\boldsymbol{M}_{CG} = \boldsymbol{M}_C \odot \boldsymbol{G}_C, \quad \boldsymbol{M}_{CG}^{rec} = \boldsymbol{M}_C^{rec} \odot \boldsymbol{G}_C^{rec}. \tag{6}$$

Finally, we impose a consistency loss between the two center-enhanced CAMs to encourage Cls and $Rec.E$ to produce spatially aligned and structure-consistent activations.

3.3. Iterative training of detector

Given an input crack image $\boldsymbol{I}^C$, we empirically select the output feature map from an intermediate residual block in both Cls and $Rec.E$ as the input to Det. This layer leverages dilated convolutions to extract abundant semantic representations while maintaining a relatively high spatial resolution. The obtained feature maps are denoted as $\boldsymbol{Z}_{in}^{Cls}$ and $\boldsymbol{Z}_{in}^{Rec}$, which provide high-level semantic information and low-level structural cues for the training of Det.

To effectively fuse these features, we concatenate them as $\boldsymbol{Z}_{in}^{Con}$ and feed the result into a novel attention mechanism named Path-Aware Attention Module (PAAM), which enhances the discriminative capability of the fused representation, especially by adaptively emphasizing crack-relevant regions through modeling both spatial and channel-wise dependencies. The structure of PAAM is illustrated in Fig. 3, consisting of a spatial attention branch and a channel attention branch. To capture the directional characteristics of cracks, we apply directional convolutions $D_\theta(\cdot)$ along four orientations $\theta \in \{0°, 45°, 90°, 135°\}$. For each direction, we compute the absolute response:

$$R_\theta = |D_\theta(\boldsymbol{Z}_{in}^{Con})|, \tag{7}$$

and the spatial attention map $A_{spatial}$ is obtained by aggregating the directional responses followed by a sigmoid activation δ:

$$A_{spatial} = \delta(R_{0°} + R_{45°} + R_{90°} + R_{135°}). \tag{8}$$

The spatial attention map serves as a soft mask that highlights potential crack paths across the feature map. We perform path-weighted fusion by applying this attention to the input feature via element-wise multiplication, followed by a convolution block with 1×1 kernel size to reduce computational cost and refine the fused features:

$$\boldsymbol{Z}_{in}^{spatial} = ReLU(BN(W_{1\times1}(\boldsymbol{Z}_{in}^{Con} \odot A_{spatial}))). \tag{9}$$

Then, $\boldsymbol{Z}_{in}^{spatial}$ is fed into the channel attention branch, which models inter-channel dependencies by emphasizing informative feature channels to better capture complementary cues such as texture, edges, and context for crack detection. Following the squeeze-and-excitation paradigm, we apply a global context aggregation using adaptive average pooling to obtain a channel-wise descriptor:

$$\boldsymbol{s} = AvgPool(\boldsymbol{Z}_{in}^{spatial}) \in \mathbb{R}^{C\times1\times1}. \tag{10}$$

The descriptor is then reshaped into a vector and passed through two fully connected layers, interleaved with ReLU and sigmoid activations, to generate the channel attention map:

$$A_{channel} = \delta(Fc(ReLU(Fc(s_{flatten})))). \tag{11}$$

Finally, the reshaped channel attention map is element-wise multiplied with $\boldsymbol{Z}_{in}^{spatial}$ to produce $\boldsymbol{Z}_{in}^{F}$, which serves as the input to Det (composed of a series of convolution and transposed convolution layers).

To enable the network to learn the mapping from input crack image to pixel-wise road crack detection results $\boldsymbol{Y}_{out}$, we train Det using the pseudo label $\boldsymbol{Y}_{pse}$, which is obtained by post-processing the corresponding crack CAM with dense Conditional Random Fields (denseCRF) [47]. The denseCRF algorithm refines the coarse CAM by modeling long-range dependencies between pixels based on both spatial proximity and color similarity. It encourages pixels with similar appearances and close spatial distances to be assigned the same label, which is particularly effective for road crack detection where cracks are often thin and low-contrast.

Since the quality of crack CAMs evolves during training, the pseudo labels are dynamically updated, making the learning process inherently iterative. To further improve training robustness, we introduce a selection mechanism that filters out pseudo labels that are entirely empty (*i.e., all-black masks*), thereby avoiding supervision from uninformative samples and enhancing the quality of the learning process.

3.4. Loss function

For the training of WP-CrackNet, the total loss function is defined as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{Rec} + \mathcal{L}_{Cls} + \mathcal{L}_{Det}, \tag{12}$$

where $\mathcal{L}_{Rec}$, $\mathcal{L}_{Cls}$, and $\mathcal{L}_{Det}$ denote the training losses for Rec, Cls and Det. During training, $\mathcal{L}_{Rec}$ and $\mathcal{L}_{Cls}$ are alternately optimized with the other module frozen, and the parameters of Det are consistently updated with $\mathcal{L}_{Det}$ to guide precise road crack detection.

3.4.1. Reconstructor loss

Given an input image $\boldsymbol{I}$, we first use the standard binary cross-entropy (BCE) loss between the class prediction q^{rec} and image-level ground-truth label $\boldsymbol{T} \in \{\boldsymbol{T}^R, \boldsymbol{T}^C\}$ to facilitate the learning of classification ability, denoted as:

$$\mathcal{L}_{cls1} = -\left[\boldsymbol{T}\log\delta(q^{rec}) + (1-\boldsymbol{T})\log(1-\delta(q^{rec}))\right]. \tag{13}$$

As mentioned above, we train Rec to reconstruct one feature from the other. To ensure consistency, we minimize the difference between the reconstructed result and the input crack image within the opposite class region. Specifically, for crack and road features, we minimize the

Table 1
Ablation study results for pixel-wise crack detection performance to investigate the impact of integrating outputs of *Cls* and *Rec.E*, and to validate the effectiveness of the proposed CECCM and PAAM on the DeepCrack dataset [31]. The symbol ✓indicates the used module for the training of the detection branch.

Cls	*Det*	*Rec.E*	CECCM	PAAM	Precision (%)↑	Recall (%)↑	Accuracy (%)↑	F1-Score (%)↑	IoU (%)↑
✓	✓				90.111	68.308	98.305	77.709	63.545
✓	✓	✓			83.920	74.032	98.263	78.666	64.835
✓	✓	✓	✓		83.828	78.341	98.409	80.992	68.055
✓	✓	✓	✓	✓	82.868	80.682	98.443	**81.760**	**69.148**

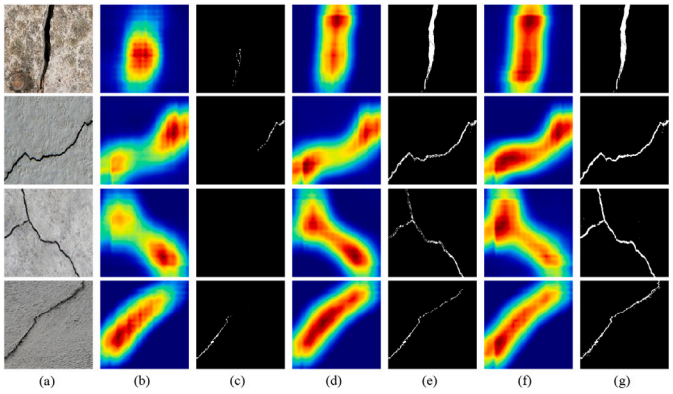


Fig. 4. Ablation study results illustrating crack CAMs (b, d, f) obtained under different training strategies—classifier only, adversarial training, and adversarial training with CECCM—and their corresponding pseudo labels (c, e, g) produced via denseCRF, with (a) showing the input image.

following losses:

$$\mathcal{L}^{C}_{recon1} = |(1 - \boldsymbol{M}^{rec}_{C}) \odot (\boldsymbol{I}_C - \boldsymbol{O}_C)|_1, \tag{14}$$

$$\mathcal{L}^{R}_{recon1} = |\boldsymbol{M}^{rec}_{C} \odot (\boldsymbol{I}_C - \boldsymbol{O}_R)|_1, \tag{15}$$

where $|\cdot|_1$ represents L1 loss. Therefore, the final reconstructor loss is formulated as follows:

$$\mathcal{L}_{Rec} = \lambda_{c1}\mathcal{L}_{cls1} + \lambda^{C}_{r1}\mathcal{L}^{C}_{recon1} + \lambda^{R}_{r1}\mathcal{L}^{R}_{recon1}, \tag{16}$$

where λ_{c1}, λ^{C}_{r1}, and λ^{R}_{r1} are weighting parameters used to harmonize these losses, and we have $\mathcal{L}_{recon1} = \lambda^{C}_{r1}\mathcal{L}^{C}_{recon1} + \lambda^{R}_{r1}\mathcal{L}^{R}_{recon1}$.

3.4.2. Classifier loss

Given an input image $\boldsymbol{I}$, similarly, the BCE loss is employed to facilitate the learning of classification ability for *Cls*:

$$\mathcal{L}_{cls2} = -\left[\boldsymbol{T}\log\delta(q) + (1-\boldsymbol{T})\log(1-\delta(q))\right]. \tag{17}$$

For input crack images, the objective of *Cls* is to generate crack CAMs that hinder *Rec* from reconstructing the original image. To this end, we minimize the similarity between the reconstruction results and the input image on crack region, with a similar constraint also applied to the road region:

$$\mathcal{L}^{C}_{recon2} = -|(1 - \boldsymbol{M}_C) \odot (\boldsymbol{I}_C - \hat{\boldsymbol{O}}_C)|_1, \tag{18}$$

$$\mathcal{L}^{R}_{recon2} = -|\boldsymbol{M}_C \odot (\boldsymbol{I}_C - \hat{\boldsymbol{O}}_R)|_1. \tag{19}$$

Furthermore, a consistency loss between center-enhanced crack CAMs from *Cls* and *Rec* is designed to enforce spatial alignment and enhance the focus on central crack regions, thereby facilitating more accurate and robust crack localization, which is denoted as:

$$\mathcal{L}_{CEC} = \frac{1}{HW}\sum|\boldsymbol{M}_{CG} - \boldsymbol{M}^{rec}_{CG}|. \tag{20}$$

Therefore, the final classifier loss is formulated as follows:

$$\mathcal{L}_{Cls} = \lambda_{c2}\mathcal{L}_{cls2} + \lambda^{C}_{r2}\mathcal{L}^{C}_{recon2} + \lambda^{R}_{r2}\mathcal{L}^{R}_{recon2} + \lambda_{cec}\mathcal{L}_{CEC}, \tag{21}$$

where λ_{c2}, λ^{C}_{r2}, λ^{R}_{r2}, and λ_{cec} are weighting parameters used to harmonize these losses. We have $\mathcal{L}_{recon2} = -\lambda^{C}_{r2}\mathcal{L}^{C}_{recon2} - \lambda^{R}_{r2}\mathcal{L}^{R}_{recon2}$.

3.4.3. Detector loss

Given input crack image I_C, as illustrated in 3.3, we can obtain road crack detection results $\boldsymbol{Y}_{out}$ by passing through *Cls*, *Rec*, and *Det*. The detector is trained using the corresponding pseudo label $\boldsymbol{Y}_{pse}$, and the detector loss is denoted as:

$$\mathcal{L}_{Det} = -\frac{1}{HW}\sum\left[\boldsymbol{Y}_{pse}\log\delta(\boldsymbol{Y}_{out}) + (1-\boldsymbol{Y}_{pse})\log\left(1-\delta(\boldsymbol{Y}_{out})\right)\right]. \tag{22}$$

4. Experimental results

4.1. Datasets

The **Crack500** [29] dataset consists of 500 high-resolution road images (2000 × 1500 pixels) with pixel-level annotations. Each image is divided into 16 non-overlapping regions, and those with over 1000 crack pixels are retained, yielding 1896 training, 348 validation, and 1124 test images. This dataset includes four crack types and poses challenges such as shadows, occlusions, and varying lighting. For our experiments, supervised methods are trained on the original dataset with all images resized to 256×256 pixels for consistency. Additionally, we create a new training set by combining 756 undamaged road images cropped from the original dataset with 1896 crack images, and assign image-level labels to facilitate the training of weakly-supervised methods.

The **DeepCrack** [31] dataset consists of 537 pixel-level annotated images of cracks on concrete and asphalt surfaces across various scenes and scales. Each image is captured at a resolution of 544×384 pixels and split into 300 training and 237 test images. Similarly, we train supervised methods on the original dataset with the size of 256 × 256 pixels and create a new training set for weakly-supervised methods by combining 296 undamaged road images (cropped and rescaled from the original dataset) with 300 crack images, all labeled at the image level.

The **CFD** [48] dataset consists of 118 high-quality images of concrete surfaces with cracks ranging from 1 mm to 3 mm in width, each annotated at the pixel level with a resolution of 480 × 320 pixels. The dataset features diverse illumination conditions, increasing the difficulty of accurate crack detection. For experiments, 70 images are used for training and 48 for testing, all resized to 256 × 256 for the training of supervised methods. To support weakly-supervised methods and account for the subtle nature of cracks in this dataset, we enlarge each image in the training set and divide it into a 3 × 3 grid of patches. This results in a new training set consisting of 268 crack images and 256 undamaged images, all annotated with image-level labels.

Table 2
Quantitative experimental results of pixel-wise crack detection performance on the Crack500 dataset [29].

Training strategy	Methods	Precision (%)↑	Recall (%)↑	Accuracy (%)↑	F1-Score (%)↑	IoU (%)↑
Specific supervised	Deepcrack19 [31]	57.581	86.733	95.687	69.213	52.920
	Deepcrack18 [27]	70.919	70.356	96.731	70.636	54.603
	Crack-Att [49]	66.497	70.883	96.376	68.620	52.230
	HrSegNet [50]	67.652	74.092	96.572	**70.726**	**54.710**
	CT-crackseg [33]	62.158	72.126	96.472	66.772	50.119
Weakly-supervised	AEFT [21]	68.406	26.986	95.222	38.703	23.995
	OC-CSE [20]	62.014	26.897	94.993	37.520	23.092
	VWL [51]	85.821	4.019	94.598	7.6782	3.992
	ACR [22]	6.580	75.302	38.859	12.103	6.441
	WS-SCS [23]	72.904	38.411	95.759	50.314	33.613
	WP-CrackNet	71.812	55.579	96.298	**62.661**	**45.625**
Unsupervised	UP-CrackNet [34]	65.377	58.609	95.484	61.808	44.726

Table 3
Quantitative experimental results of pixel-wise crack detection performance on the DeepCrack dataset [31].

Training strategy	Methods	Precision (%)↑	Recall (%)↑	Accuracy (%)↑	F1-Score (%)↑	IoU (%)↑
Specific supervised	Deepcrack19 [31]	88.785	68.923	97.756	77.603	63.403
	Deepcrack18 [27]	71.720	88.403	98.350	79.192	65.552
	Crack-Att [49]	89.967	69.327	98.339	78.310	64.352
	HrSegNet [50]	82.492	80.337	98.412	81.400	68.634
	CT-crackseg [33]	85.381	78.838	98.501	**81.979**	**69.461**
Weakly-supervised	AEFT [21]	81.899	68.112	97.969	74.372	59.199
	OC-CSE [20]	89.786	66.131	98.209	76.164	61.504
	VWL [51]	86.356	56.650	97.738	68.418	51.996
	ACR [22]	88.339	66.424	98.168	75.829	61.069
	WS-SCS [23]	98.499	30.722	96.952	46.836	30.579
	WP-CrackNet	82.868	80.682	98.443	**81.760**	**69.148**
Unsupervised	UP-CrackNet [34]	63.412	88.852	98.049	74.006	58.738

Table 4
Quantitative experimental results of pixel-wise crack detection performance on the CFD dataset [48].

Training strategy	Methods	Precision (%)↑	Recall (%)↑	Accuracy (%)↑	F1-Score (%)↑	IoU (%)↑
Specific supervised	Deepcrack19 [31]	20.892	88.142	94.372	33.778	20.321
	Deepcrack18 [27]	46.231	69.159	98.188	55.417	38.329
	Crack-Att [49]	70.086	43.530	98.778	53.705	36.710
	HrSegNet [50]	29.782	43.474	98.322	35.348	21.469
	CT-crackseg [33]	60.907	54.916	98.692	**57.757**	**40.604**
Weakly-supervised	AEFT [21]	93.233	4.008	98.432	7.685	3.996
	OC-CSE [20]	81.154	6.397	98.452	11.860	6.304
	VWL [51]	94.161	3.746	98.429	7.206	3.738
	ACR [22]	80.268	4.217	98.423	8.013	4.174
	WS-SCS [23]	77.764	2.485	98.400	4.816	2.468
	WP-CrackNet	62.262	49.682	98.690	**55.265**	**38.184**
Unsupervised	UP-CrackNet [34]	10.978	63.785	90.987	18.731	10.333

4.2. Implementation details

All experiments are conducted on a single NVIDIA RTX 4090 GPU, with each model trained for 200 epochs. The initial learning rate is set to 0.001 and adjusted dynamically using the polynomial decay policy. To enhance generalization, standard data augmentation techniques such as random cropping, resizing, and horizontal flipping are applied to the input images. Following the empirical settings suggested in [22], we configure the loss weights as follows: for $\mathcal{L}_{Rec}$, we set $\lambda_{c1} = 1$ and $\lambda_{r1}^{C} = \lambda_{r1}^{R} = 0.5$; for $\mathcal{L}_{Cls}$, we use $\lambda_{c2} = 1$, $\lambda_{r2}^{C} = 0.8$, $\lambda_{r2}^{R} = 0.3$, and $\lambda_{cec} = 0.5$.

During inference on the CFD dataset, each image is divided into a 3×3 grid of patches, which are resized and classified by Cls to filter out background regions. Only patches predicted to contain cracks are passed to the detection module, and final result is obtained by merging the outputs from these selected patches. For fair comparison, all weakly-supervised methods for comparison are processed using denseCRF refinement following CAM generation. Additionally, they are trained using the same Rec and Det architecture as our method. For evaluation, we adopt a comprehensive set of metrics, including precision, recall, accuracy, IoU, and F1-score, to quantitatively assess the detection performance of WP-CrackNet against existing methods. In addition, model parameters and frames per second (FPS) are used to evaluate the model complexity and processing speed.

4.3. Ablation study

To assess the impact of integrating both high-level semantic information and low-level structural cues from Cls and $Rec.E$, and to evaluate the effectiveness of the proposed CECCM and PAAM, we conduct an ablation study on the DeepCrack dataset [31]. The quantitative experimental results are presented in Table 1. The first row shows crack detection results using only Cls and Det, while the last row represents the process of obtaining fused feature map through Cls and Det, which is then enhanced by PAAM, with crack CAMs further improved by the CECCM for better generation of pseudo labels. The experimental results

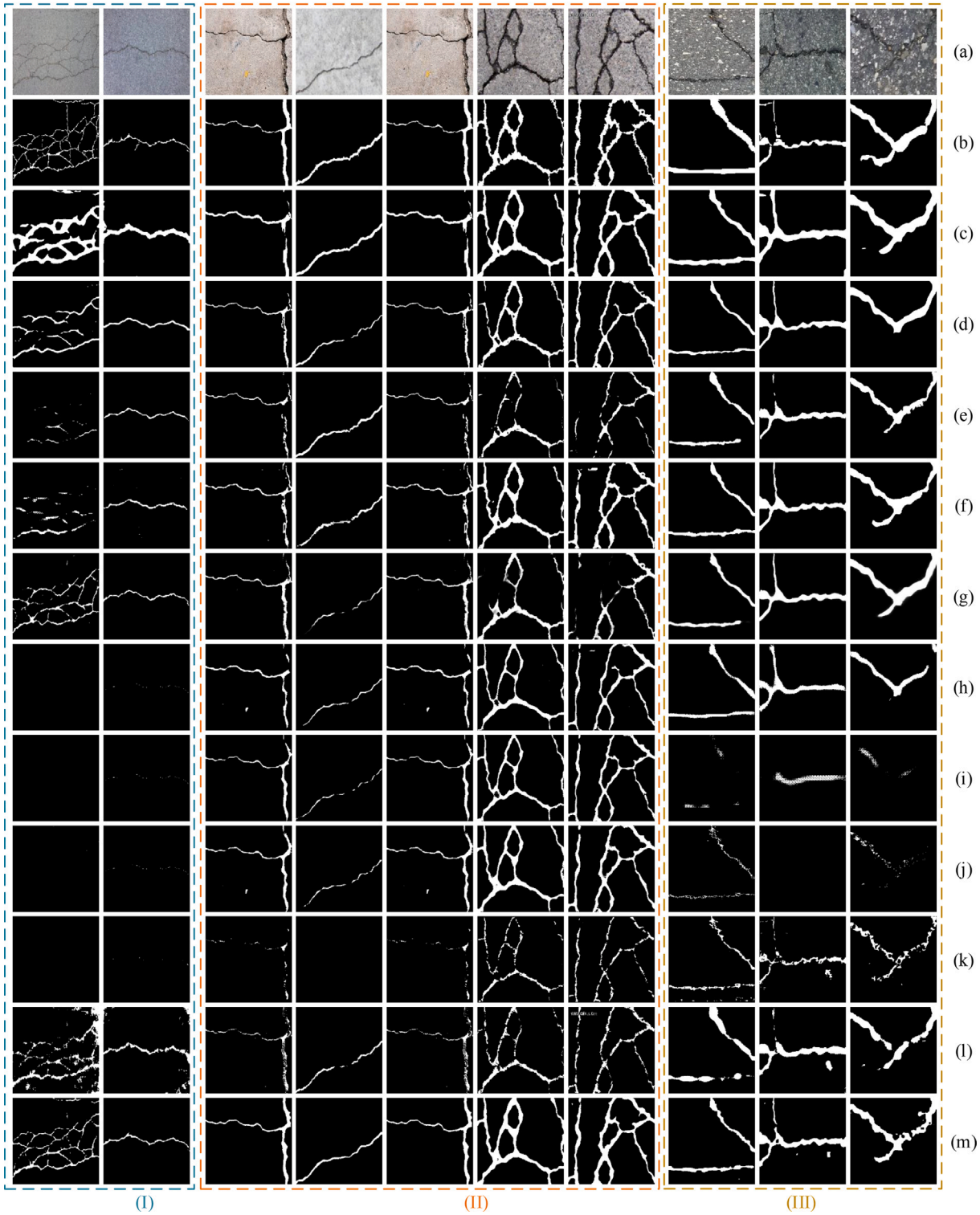


Fig. 5. Examples of experimental results on the (I) CFD [48], (II) DeepCrack [31], and (III) Crack500 [29] datasets: (a) Input images; (b) Ground Truth; (c) Deepcrack19 [31]; (d) Deepcrack18 [27]; (e) Crack-Att [49]; (f) HrSegNet [50]; (g) CT-crackseg [33]; (h) AEFT [21]; (i) OC-CSE [20]; (j) ACR [22]; (k) WS-SCS [23]; (l) UP-CrackNet [34]; (m) WP-CrackNet.

that integrate all these modules attain the best detection performance, demonstrating the effectiveness of the proposed CECCM and PAAM.

In addition, qualitative experiments are conducted to visually validate the effectiveness of the adopted adversarial training strategy and the proposed CECCM. The results in Fig. 4 indicate that the adversarial training strategy enables crack CAMs to fully cover crack regions, while the proposed CECCM enhances the generation and spatial aggregation of crack CAMs. Furthermore, with the aid of denseCRF, high-quality pseudo labels can be derived for the training of pixel-wise road crack detection. In Fig. 4, (a) is the input image; (b), (d), and (f) are crack CAMs obtained under different training settings (classifier only, adversarial training, and adversarial training with CECCM, respectively); and

(c), (e), and (g) are their corresponding pseudo labels after applying denseCRF.

4.4. Comparison with SoTA methods

To validate the effectiveness of our proposed WP-CrackNet, we conduct a comprehensive comparison against four SoTA general weakly-supervised semantic methods, five supervised methods, one weakly-supervised method, and one unsupervised method specifically designed for road crack detection. The evaluation is performed on the Crack500 [29], Deepcrack [31], and CFD [48] datasets. Quantitative and qualitative results are presented in Tables 2–4 and Fig. 5, respectively. The results clearly indicate that WP-CrackNet achieves detection performance comparable to road crack detection-specific supervised methods while utilizing only image-level labels, outperforming other weakly-supervised methods as well as the unsupervised method.

Specifically, across the three datasets, our proposed WP-CrackNet demonstrates an improvement in IoU of 12.012 % – −41.633 %, 7.644 % – −38.569 %, and 31.880 %– −35.716 % compared to other SoTA general and road crack detection-specific weakly-supervised methods. These results stem from WP-CrackNet's online and iterative pseudo-label generation with a selection mechanism, which reduces noise and enhances label reliability, as well as the incorporation of CECCM and PAAM. CECCM enables more precise crack CAMs, while PAAM effectively fuses high-level semantics from the classifier with low-level structural cues from the reconstructor by modeling spatial and channel-wise dependencies. Furthermore, compared with the road crack detection-specific method WS-SCS, WP-CrackNet reduces reliance on hand-crafted cues and adopts an end-to-end joint optimization strategy, leading to greater adaptability and robustness.

When compared to five supervised methods tailored for road crack detection, WP-CrackNet exhibits only a marginal IoU drop of 0.313 % and 2.420 % compared to CT-crackSeg [33] and outperforms other supervised methods on the DeepCrack and CFD datasets. On the Crack500 dataset, it experiences an IoU decrease of 4.494 %–9.085 % relative to these supervised methods. The relatively lower performance on the Crack500 dataset can be attributed primarily to two factors: (1) Higher scene variability and diverse crack morphologies: Crack500 contains four crack types with variations in widths, lengths, and branching patterns, captured on different road materials under diverse conditions. Real-world challenges such as shadows, occlusions, and lighting variations, together with complex background textures, increase intra-class variability and make crack boundary localization more difficult when only image-level labels are available; (2) More pixel-level labels for supervised methods: Crack500 contains a larger number of finely annotated pixel-level labels than the other datasets, allowing supervised networks to learn precise geometric priors and handle small or partially occluded cracks more effectively. In contrast, WP-CrackNet relies on implicit localization through image-level cues, which limits its boundary accuracy in these cases. Nevertheless, considering the substantial reliance of supervised techniques on extensive pixel-level manual annotations, our proposed WP-CrackNet, which achieves comparable detection results using only image-level labels, offers significant potential to enhance the scalability of road defect detection.

Furthermore, in comparison to the unsupervised method UP-CrackNet [34], WP-CrackNet achieves improvements in IoU by 0.899 %, 10.410 %, and 27.851 % on the Crack500, DeepCrack, and CFD datasets, respectively. The results align with the intuitive expectation that leveraging image-level label information generally leads to better detection performance than methods that do not utilize any label information. Notably, for datasets such as CFD, where cracks are thin and not obvious, our proposed WP-CrackNet exhibits superior accuracy in crack boundary prediction while effectively suppressing noise in the detection results.

Table 5 reports model parameters and processing efficiency of WP-CrackNet and representative supervised road crack detection methods on an NVIDIA RTX3090 using 256×256 inputs. For WP-CrackNet, in the inference phase, only the trained $Rec.E$, Cls, Det, and the PAAM are required for processing the test data. Experimental results show that WP-CrackNet has a slightly larger number of parameters compared with these supervised methods, and its FPS ranks at a moderate level. The relatively larger parameter size of WP-CrackNet is mainly attributed to the multi-module collaborative training strategy designed to achieve weakly supervised road crack detection. Considering that WP-CrackNet requires only image-level annotations while delivering detection performance comparable to supervised methods, it remains highly practical. In future iterations, techniques such as model pruning and knowledge distillation will be employed to compress the parameter size, along with hardware-friendly designs, to make the model suitable for deployment on edge devices (such as drones) for real-time road crack detection tasks.

Table 5
Quantitative experimental results in terms of model parameters and processing efficiency.

Methods	Model parameters (M)↓	FPS (on RTX3090)↑
Deepcrack18 [27]	30.905	59.175
Deepcrack19 [31]	14.720	287.888
SCCDNet [52]	31.705	143.585
Crack-Att [49]	45.804	64.710
HrSegNet [50]	9.641	285.635
CDLN [53]	19.151	67.984
LECSFormer [54]	16.528	96.881
CT-crackseg [33]	22.882	36.890
WP-CrackNet	86.458	73.613

5. Conclusion

In this paper, we propose WP-CrackNet, an innovative end-to-end weakly-supervised road crack detection method that leverages image-level labels to reduce reliance on costly pixel-level annotations, greatly enhancing the scalability of road inspection. WP-CrackNet consists of three components: a classifier that creates CAMs, a reconstructor that assesses the inferability between road and crack features, and a detector that outputs pixel-wise road crack detection results. The classifier and reconstructor are trained adversarially in turns, while the detector is trained with pseudo labels derived from post-processed crack CAMs. Our designed PAAM effectively fuses high-level semantics from the classifier and low-level structural cues from the reconstructor by modeling spatial and channel-wise dependencies, improving the detection performance. Additionally, the proposed CECCM improves the quality of crack CAMs through center Gaussian weighting and consistency constraints, optimizing pseudo-label generation. Extensive experiments conducted on three datasets demonstrate the effectiveness of WP-CrackNet and its superiority over SoTA general and road crack detection-specific weakly-supervised methods in detecting road cracks by only using image-level labels.

Future work will focus on compressing and accelerating WP-CrackNet through model pruning and knowledge distillation, enabling not only cloud-based detection but also real-time deployment on edge devices. Additionally, we plan to incorporate a small amount of fine-grained pixel-level annotations for joint training, aiming to further improve detection performance and enhance domain adaptation capabilities.

CRediT authorship contribution statement

Nachuan Ma: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Zhengfei Song:** Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation. **Qiang Hu:** Visualization, Validation, Methodology, Investigation,

Formal analysis, Data curation. **Xiaoyu Tang:** Writing – review & editing, Methodology, Investigation. **Chengxi Zhang:** Writing – review & editing, Methodology, Investigation. **Rui Fan:** Writing – review & editing, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization. **Lihua Xie:** Writing – review & editing, Supervision, Project administration, Methodology, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research was supported by the National Natural Science Foundation of China under Grant 62473288, the Fundamental Research Funds for the Central Universities, NIO University Programme (NIO UP), and the Xiaomi Young Talent Program.

Data availability

Data will be made available upon request.

References

[1] 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. 4 (4) (2022) tdac026.
[2] R. Fan, et al., Road crack detection using deep convolutional neural network and adaptive thresholding, in: 2019 IEEE Intelligent Vehicles Symposium (IV), IEEE, 2019, pp. 474–479.
[3] C.-W. Liu, et al., These maps are made by propagation: adapting deep stereo networks to road scenarios with decisive disparity diffusion, IEEE Transactions on Image Processing, 2025.
[4] R. Fan, M. Liu, Road damage detection based on unsupervised disparity map segmentation, IEEE Trans. Intell. Transp. Syst. 21 (11) (2020) 4906–4911.
[5] R. Fan, et al., Pothole detection based on disparity transformation and road surface modeling, IEEE Trans. Image Process. 29 (2019) 897–908.
[6] R. Fan, et al., Rethinking road surface 3-D reconstruction and pothole detection: from perspective transformation to disparity map segmentation, IEEE Trans. Cybern. 52 (7) (2022) 5799–5808.
[7] Y. Yuan, et al., EcRD: edge-cloud computing framework for smart road damage detection and warning, IEEE Internet Things J. 8 (16) (2020) 12734–12747.
[8] A. Krizhevsky, et al., ImageNet classification with deep convolutional neural networks, Commun. ACM 60 (6) (2017) 84–90.
[9] J. Fan, et al., Deep convolutional neural networks for road crack detection: qualitative and quantitative comparisons, in: 2021 IEEE International Conference on Imaging Systems and Techniques (IST), IEEE, 2021, pp. 1–6.
[10] Y. Hou, et al., MobileCrack: object classification in asphalt pavements using an adaptive lightweight deep learning, J. Transp. Eng. Part B Pavements 147 (1) (2021) 04020092.
[11] S. Guo, et al., LIX: implicitly infusing spatial geometric prior knowledge into visual semantic segmentation for autonomous driving, IEEE Transactions on Image Processing, 2025.
[12] M.W. Khan, et al., Real-time road damage detection using an optimized YOLOv9s-fusion in IoT infrastructure, IEEE Internet Things J. 12 (11) (2025).
[13] M.W. Khan, et al., Real-time road damage detection and infrastructure evaluation leveraging unmanned aerial vehicles and tiny machine learning, IEEE Internet Things J. 11 (12) (2024).
[14] C.-K. Tsung, et al., HPPH: Computer vision-based service for high-performance pavement health recognition, IEEE Internet Things J. 12 (11) (2025).
[15] J. Huyan, et al., CrackU-Net: a novel deep convolutional neural network for pixelwise pavement crack detection, Struct. Control Health Monit. 27 (8) (2020) e2551.
[16] T. Zhang, et al., ECSNet: an accelerated real-time image segmentation CNN architecture for pavement crack detection, IEEE Trans. Intell. Transp. Syst. 24 (12) (2023).
[17] G. Zhu, et al., A lightweight encoder–decoder network for automatic pavement crack detection, Comput.-Aided Civ. Infrastruct. Eng. 39 (12) (2024) 1743–1765.
[18] Z. Song, et al., Robust and real-time road crack detection through collaborative dual-branch learning on robotic sensing platform, in: 2025 IEEE International Conference on Real-Time Computing and Robotics (RCAR), IEEE, 2025, pp. 55–60.
[19] N. Ma, et al., Self-derived multi-modal knowledge distillation for real-time road crack detection, IEEE Sens. J. (2025).
[20] H. Kweon, et al., Unlocking the potential of ordinary classifier: class-specific adversarial erasing framework for weakly supervised semantic segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 6994–7003.
[21] S.-H. Yoon, et al., Adversarial erasing framework via triplet with gated pyramid pooling layer for weakly supervised semantic segmentation, in: European Conference on Computer Vision, Springer, 2022, pp. 326–344.
[22] H. Kweon, et al., Weakly supervised semantic segmentation via adversarial learning of classifier and reconstructor, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 11329–11339.
[23] J. König, et al., Weakly-supervised surface crack segmentation by generating pseudo-labels using localization with a classifier and thresholding, IEEE Trans. Intell. Transp. Syst. 23 (12) (2022) 24083–24094.
[24] C.V. Dung, et al., Autonomous concrete crack detection using deep fully convolutional neural network, Autom. Constr. 99 (2019) 52–58.
[25] T. Chen, et al., Pavement crack detection and recognition using the architecture of SegNet, J. Ind. Inf. Integr. 18 (2020) 100144.
[26] X. Sun, et al., DMA-Net: DeepLab with multi-scale attention for pavement crack segmentation, IEEE Trans. Intell. Transp. Syst. 23 (10) (2022) 18392–18403.
[27] Q. Zou, et al., DeepCrack: learning hierarchical convolutional features for crack detection, IEEE Trans. Image Process. 28 (3) (2018) 1498–1512.
[28] V. Badrinarayanan, et al., SegNet: a deep convolutional encoder-decoder architecture for image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 39 (12) (2017) 2481–2495.
[29] F. Yang, et al., Feature pyramid and hierarchical boosting network for pavement crack detection, IEEE Trans. Intell. Transp. Syst. 21 (4) (2019) 1525–1535.
[30] S. Xie, Z. Tu, Holistically-nested edge detection, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), IEEE, 2015, pp. 1395–1403.
[31] Y. Liu, et al., DeepCrack: a deep hierarchical feature learning architecture for crack segmentation, Neurocomputing 338 (2019) 139–153.
[32] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556, 2014.
[33] H. Tao, et al., A convolutional-transformer network for crack segmentation with boundary awareness, in: 2023 IEEE International Conference on Image Processing (ICIP), IEEE, 2023, pp. 86–90.
[34] N. Ma, et al., Up-CrackNet: unsupervised pixel-wise road crack detection via adversarial image restoration, IEEE Trans. Intell. Transp. Syst. 25 (10) (2024).
[35] X. Liu, K. Wu, X. Cai, W. Huang, Semi-supervised semantic segmentation using cross-consistency training for pavement crack detection, Road Mater. Pavement Des. 25 (6) (2024) 1368–1380.
[36] X. Zhang, et al., Scribble hides class: promoting scribble-based weakly-supervised semantic segmentation with its class label, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, 2024, pp. 7332–7340.
[37] A. Khoreva, et al., Simple does it: weakly supervised instance and semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 876–885.
[38] Y.-T. Chang, et al., Weakly-supervised semantic segmentation via sub-category exploration, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 8991–9000.
[39] J. Fan, et al., CIAN: cross-image affinity net for weakly supervised semantic segmentation, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, 2020, pp. 10762–10769.
[40] F. Zhang, et al., Complementary patch for weakly supervised semantic segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 7242–7251.
[41] B. Zhou, et al., Learning deep features for discriminative localization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2921–2929.
[42] J. Xie, et al., C2AM: contrastive learning of class-agnostic activation map for weakly supervised object localization and semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 989–998.
[43] Y. Wang, et al., Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 12275–12284.
[44] T. Wu, et al., Embedded discriminative attention mechanism for weakly supervised semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 16765–16774.
[45] X. Zhang, et al., Adversarial complementary learning for weakly supervised object localization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1325–1334.
[46] Z. Wu, C. Shen, A. Van Den Hengel, Wider or deeper: revisiting the ResNet model for visual recognition, Pattern Recognit. 90 (2019) 119–133.
[47] P. Krähenbühl, V. Koltun, Efficient inference in fully connected CRFs with Gaussian edge potentials, Advances in Neural Information Processing Systems 24, 2011 109–117
[48] Y. Shi, et al., Automatic road crack detection using random structured forests, IEEE Trans. Intell. Transp. Syst. 17 (12) (2016) 3434–3445.
[49] N. Xu, et al., Crack-Att Net: crack detection based on improved U-Net with parallel attention, Multimed. Tools Appl. 82 (27) (2023) 42465–42484.
[50] Y. Li, et al., Real-time high-resolution neural network with semantic guidance for crack segmentation, Autom. Constr. 156 (2023) 105112.
[51] L. Ru, et al., Weakly-supervised semantic segmentation with visual words learning and hybrid pooling, Int. J. Comput. Vis. 130 (4) (2022) 1127–1144.
[52] H. Li, et al., SCCDNet: a pixel-level crack segmentation network, Appl. Sci. 11 (11) (2021) 5074.
[53] P. Manjunatha, et al., CrackDenseLinkNet: a deep convolutional neural network for semantic segmentation of cracks on concrete surface images, Struct. Health Monit. 23 (2) (2024) 796–817.
[54] J. Chen, et al., Refined crack detection via LECSFormer for autonomous road inspection vehicles, IEEE Trans. Intell. Veh. 8 (3) (2022).

Author biography

Nachuan Ma received the B.E. degree in Electrical Engineering and Automation from China University of Mining and Technology in 2019, and the M.Sc. degree in Electronic Engineering from The Chinese University of Hong Kong in 2020. He worked as a Research Assistant at the Department of Electronic and Electrical Engineering, Southern University of Science and Technology, from 2020 to 2021. He is pursuing his Ph.D. degree at Tongji University with a research focus on computer vision and deep learning.

Zhengfei Song is currently pursuing his B.E. degree at Tongji University with a research focus on computer vision techniques for autonomous driving.

Qiang Hu is currently pursuing his B.E. degree at Tongji University with a research focus on computer vision techniques for autonomous driving.

Xiaoyu Tang received a B.S. degree from South China Normal University, Guangzhou, China, in 2003 and an M.S. degree from Sun Yat-sen University, Guangzhou, China, in 2011. He is currently pursuing a Ph.D. degree at South China Normal University. Mr. Tang works as an associate professor, master supervisor, and deputy dean at Xingzhi College, South China Normal University. His research interests include image processing and intelligent control, artificial intelligence, the Internet of Things, and educational informatization.

Chengxi Zhang received B.S. and M.S. degrees from Harbin Institute of Technology, China, in 2012 and 2015; and Ph.D. degree from Shanghai Jiao Tong University, China, in 2019. He is now an Associate Professor at Jiangnan University, China. His interests are space engineering, robotic systems & control. He is a Committee Member of CAAYAC, a Distinguished Reviewer for Intelligence & Robotics, 2024. He is an Associate Editor of Frontiers in Aerospace Engineering, on the Editorial Board of Aerospace Systems and Astrodynamics.

Rui Fan received the B.Eng. degree in Automation from the Harbin Institute of Technology in 2015 and the Ph.D. degree in Electrical and Electronic Engineering from the University of Bristol in 2018. He worked as a Research Associate at the Hong Kong University of Science and Technology from 2018 to 2020 and a Postdoctoral Scholar-Employee at the University of California San Diego between 2020 and 2021. He began his faculty career as a Full Research Professor in the College of Electronics & Information Engineering at Tongji University in 2021. He was promoted to Full Professor in 2022 and attained tenure in 2024, both in the same college and at the Shanghai Research Institute for Intelligent Autonomous Systems. His research interests include computer vision, deep learning, and robotics, with a specific focus on humanoid visual perception under the two-streams hypothesis. Prof. Fan served as an associate editor for ICRA'23/25 and IROS'23/24, an area chair for ICIP'24, and a senior program committee member for AAAI'23/24/25/26. He organized several impactful workshops and special sessions in conjunction with WACV'21, ICIP'21/22/23, ICCV'21, ECCV'22, and ICCV'25. He was honored by being included in the Stanford University List of Top 2% Scientists Worldwide between 2022 and 2024, recognized on the Forbes China List of 100 Outstanding Overseas Returnees in 2023, acknowledged as one of Xiaomi Young Talents in 2023, and awarded the Shanghai Science & Technology 35 Under 35 honor in 2024 as its youngest recipient.

Lihua Xie received the Ph.D. degree in electrical engineering from the University of Newcastle, Australia, in 1992. Since 1992, he has been with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, where he is currently a professor and Director, Center for Advanced Robotics Technology Innovation. He served as the Head of Division of Control and Instrumentation and Co-Director, Delta-NTU Corporate Lab for Cyber-Physical Systems. He held teaching appointments in the Department of Automatic Control, Nanjing University of Science and Technology from 1986 to 1989. Dr Xie's research interests include robust control and estimation, networked control systems, multi-agent networks, and unmanned systems. He is an Editor-in-Chief for Unmanned Systems and Associate Editor for IEEE Control System Magazine. He has served as Editor of IET Book Series in Control and Associate Editor of a number of journals including IEEE Transactions on Automatic Control, Automatica, IEEE Transactions on Control Systems Technology, IEEE Transactions on Network Control Systems, and IEEE Transactions on Circuits and Systems-II. He was an IEEE Distinguished Lecturer (Jan 2012–Dec 2014). Dr Xie is Fellow of Academy of Engineering Singapore, IEEE, and IFAC.