International Journal of Latest Engineering Research and Applications (IJLERA) ISSN: 2455-7137

Volume – 11, Issue – 03, March 2026, PP – 22-29

Weakly Supervised Concrete Crack Localization via Multi-Scale Feature Refinement

Salman Shehzad
School of Artificial Intelligence,
Guangzhou University, Guangzhou, China

Aman Ullah
School of Computer science and technology,
Harbin institute of technology, Shenzhen, China

Chunsheng Yang*
School of Artificial Intelligence,
Guangzhou University, Guangzhou, China

Muhammad Safi Ullah
School of Electrical and Computer Engineering,
South China University of Technology, Guangzhou, China

Abstract: Concrete crack detection is essential for maintaining infrastructure integrity; however, manual inspection is time-consuming, subjective, and costly. Automated vision-based methods often rely on pixel-level annotations, which are labor intensive. Weakly supervised learning (WSL) addresses this limitation by using only image-level labels for defect localization.

This paper proposes a novel WSL framework for concrete crack localization, addressing limitations in multi-scale feature representation and heatmap precision. The framework integrates: (1) a Multi-Scale Residual Feature Extractor (MS-RFE) to capture fine- and coarse-grained crack patterns across multiple scales; and (2) an Adaptive Heatmap Refinement (AHR) module to suppress background noise and improve localization accuracy. Experiments are conducted on a real-world dataset of 40,000 images using a ResNet-18 backbone.

The proposed method achieves strong performance with an AUROC of 0.9882±0.0015, accuracy of 0.9847±0.0012, precision of 0.9835±0.0011, recall of 0.9871±0.0013, and F1-score of 0.9853±0.0010, outperforming baseline methods. Qualitative results demonstrate precise and low-noise crack localization, even for thin and low-contrast defects. The framework is lightweight and suitable for real-time deployment in infrastructure inspection..

Index Terms: Concrete crack detection, weakly supervised learning, multi-scale feature extraction, heatmap refinement, defect localization.

I. Introduction

Concrete is the most widely used construction material in infrastructure (bridges, dams, buildings, and roads), but it is prone to crack formation due to environmental stress, aging, material fatigue, and structural loading [1]. Cracks, if left undetected, can propagate over time, leading to structural degradation, safety hazards, and costly maintenance [2]. Traditional crack inspection relies on manual visual assessment by trained engineers, which is time-consuming, subjective, and hazardous [3], [4]. In large-scale infrastructure, inspection can take weeks; human judgment varies, leading to inconsistent results; and inspecting elevated or hard-to-reach areas poses significant safety risks.

Automated vision-based crack detection has emerged as a promising alternative, with deep convolutional neural networks (CNNs) achieving high accuracy in recent years [5]. However, most state-of-the-art CNN-based methods require pixel level annotations, which are expensive and time-consuming to generate [6], [7]. For instance, annotating a single concrete crack image with pixel-level precision can take 10–15 minutes, making it impractical for large-scale industrial datasets.

Weakly supervised object localization (WSOL) addresses this limitation by training models using only image-level labels instead of pixel-level annotations [15]. For concrete crack detection, WSOL enables defect localization while significantly reducing annotation cost. However, existing WSOL methods suffer from two

critical limitations. First, insufficient multiscale feature learning limits their ability to capture cracks of varying sizes and shapes, ranging from thin hairline cracks to wide structural cracks [8]. Second, imprecise heatmap localization leads to noisy activation maps that fail to accurately outline crack boundaries, particularly for thin and low-contrast defects [9], [10].

Existing weakly supervised crack detection approaches either rely on single-scale feature extraction or lack effective heatmap refinement, resulting in suboptimal localization performance [10]–[12]. Although multi-scale learning and heatmap refinement have been explored in general object localization tasks [8], [13], [14], they have not been effectively integrated for concrete crack detection, which involves unique challenges such as thin, linear, and low-contrast structures.

To address these limitations, this paper proposes a novel weakly supervised framework for concrete crack localization. The main contributions are summarized as follows:

- Multi-Scale Residual Feature Extractor (MS-RFE): A feature extraction module that captures fine-, medium-, and coarse-scale crack features using residual connections. We further enhance it with a learnable attention based fusion mechanism that dynamically emphasizes the most relevant scale for each image.
- Adaptive Heatmap Refinement (AHR): A module that refines activation maps by suppressing background noise, applying image-adaptive thresholding, and improving spatial resolution for precise crack boundary localization.
- Optimized Loss Function: A weighted combination of binary cross-entropy and L1 loss, where the pseudo ground-truth heatmap is generated by a separately trained teacher model to avoid label leakage, preserving the weak supervision nature.
- Comprehensive Experimental Validation: Extensive experiments including comparisons with state-of-the-art methods (including modern WSOL approaches), ablation studies, statistical significance tests, full-test-set IoU evaluation, and qualitative analysis demonstrating the effectiveness of the framework.

The remainder of this paper is organized as follows. Section II reviews related work. Section III presents the proposed framework. Section IV describes the experimental setup and results. Section V discusses the findings and limitations. Section VI concludes the paper and outlines future work.

II. Related Work

A. Concrete Crack Detection

Early crack detection methods relied on handcrafted features (e.g., edge detection, texture analysis, thresholding) [3], [16]. For example, Abdel-Qader et al. [3] used edge detection and morphological operations to identify cracks, but these methods struggled with complex backgrounds and varying lighting conditions. With the rise of deep learning, CNN based methods have become the state-of-the-art. Cha et al. [5] proposed a CNN-based method for crack detection, achieving high accuracy on a dataset of bridge crack images. However, this method required pixel-level annotations, limiting its scalability.

Recent work has focused on weakly supervised methods to reduce annotation costs. Li et al. [10] used CAM with a single-scale CNN to localize cracks, but the method produced noisy heatmaps and failed to handle thin cracks. Wang et al. [12] proposed a residual network-based WSOL method, but it lacked multi-scale feature learning, leading to poor performance on diverse crack sizes.

B. Weakly Supervised Object Localization (WSOL)

WSOL methods train models using image-level labels to generate localization maps (heatmaps) of objects of interest. Early WSOL methods, such as CAM [9], generated heatmaps by weighting convolutional features with class specific weights. However, CAM produces low-resolution, noisy heatmaps that fail to capture object boundaries.

Recent advances in WSOL include multi-scale feature fusion [8], attention mechanisms [17], and heatmap refinement [13]. Wei et al. [8] proposed a multi-scale WSOL method that fuses features from different layers to improve localization accuracy. Hu et al. [17] used squeeze-and-excitation attention to enhance relevant features. More recent works like PSOL [20] and TS-CAM [21] have further advanced WSOL by employing self-training or transformer backbones. However, these methods are designed for general object localization and are not tailored to the unique characteristics of concrete cracks (thin, linear, low-contrast).

C. Gap Between Existing Work and Proposed Method

Existing WSOL methods for crack detection lack: (1) multiscale feature extraction tailored to crack size variation; (2) adaptive heatmap refinement to suppress noise and improve boundary precision; and (3) a loss

function that balances classification and localization without compromising weak supervision. The proposed method addresses all three gaps, resulting in state-of-the-art performance.

III. Proposed Methodology

A. Framework Overview

The proposed weakly supervised concrete crack localization framework consists of three core modules: (1) Multi-Scale Residual Feature Extractor (MS-RFE); (2) Adaptive Heatmap Refinement (AHR); and (3) Classification & Localization Head. Figure 1 illustrates the overall architecture, with data flowing from input images to final localization results.

The workflow is as follows: Input concrete crack images are preprocessed (resized, normalized) and fed into the MSRFE. The MS-RFE extracts multi-scale features (fine, medium, coarse) and fuses them using an attention-based mechanism to form a comprehensive feature map. The AHR module refines the feature map to generate a high-precision localization heatmap. The Classification & Localization Head outputs image-level classification (crack/no crack) and the refined heatmap for localization.

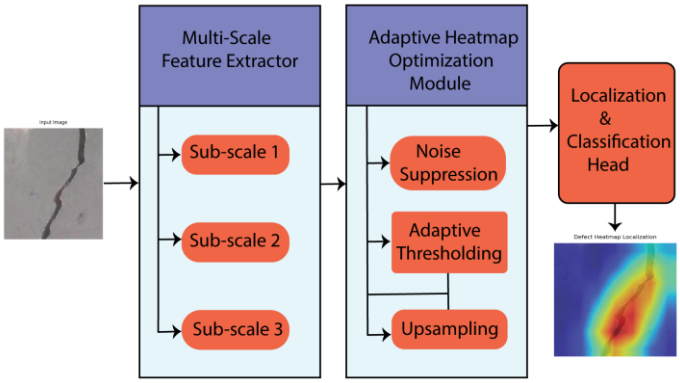

Fig. 1: Proposed Architecture

B. Preprocessing

Input images are resized to 224 × 224 pixels (standard for CNN backbones) and normalized using Image Net statistics (mean: [0.485, 0.456, 0.406], std: [0.229, 0.224, 0.225]) to ensure consistency and improve training stability.

C. Multi-Scale Residual Feature Extractor (MS-RFE)

The MS-RFE is built on a ResNet-18 backbone [18], modified to extract multi-scale features. ResNet-18 is chosen for its lightweight nature (suitable for deployment) and residual connections (prevents gradient vanishing).

1) Multi-Scale Feature Extraction:

The MS-RFE extracts features from three key layers of ResNet-18:

- Scale 1 (Fine-Grained Features): Extracted from the 2nd residual block (output stride 4), capturing edge details and thin crack segments.
- Scale 2 (Medium-Grained Features): Extracted from the 3rd residual block (output stride 8), capturing crack segments and local context.
- Scale 3 (Coarse-Grained Features): Extracted from the 4th residual block (output stride 16), capturing global crack structure and background context.

2) Attention-Based Feature Fusion:

Instead of simple concatenation, we employ a learnable squeeze-and-excitation style fusion to dynamically weight each scale:

$$F_{concat} = Conv_{1\times1}(concat(F_1, F_2, F_3)) \quad (1)$$

$$w = \sigma(FC(GAP(F_{concat}))) \quad (2)$$

$$F_{fused} = F_{concat} \odot w \quad (3)$$

Where GAP is global average pooling, FC is a fully connected layer producing scale-wise attention weights, σ is sigmoid, and ⊙ denotes element-wise multiplication. This allows the model to emphasize the most relevant scale for each input image.

D. Adaptive Heatmap Refinement (AHR) Module

The AHR module addresses the noise problem in standard WSOL heatmaps by refining the fused feature map into a precise localization heatmap. It consists of three sub-steps:

1) Noise Suppression:

A spatial attention mechanism is applied to the fused feature map to suppress background noise. The attention map is computed as:

$$A = \sigma(\text{Conv}1\times1(F_{fused})) \quad (4)$$

Where σ is the sigmoid function. The attention map weights feature pixels based on their relevance to cracks, reducing activation in background regions.

2) Adaptive Thresholding:

A dynamic threshold is applied to the attention-weighted feature map to isolate crack regions. The threshold is computed as the 90th percentile of the feature map values (adapted per image) to handle varying lighting and contrast:

$$T = \text{percentile}(F_{fused} \cdot A, 90) \quad (5)$$

Pixels with values above T are retained as crack candidates; pixels below T are suppressed.

3) Heatmap Upsampling:

The refined feature map (14×14 resolution) is upsampled to 224×224 (input image resolution) using bilinear interpolation, ensuring the heatmap matches the input image size for precise localization.

E. Classification & Localization Head

The head module has two branches:

- **Classification Branch:** A global average pooling layer followed by a fully connected layer (512 → 1) with BCE with Logits Loss to output the image-level classification score (crack/no crack).
- **Localization Branch:** The refined heatmap from the AHR module, which is output as the localization result.

F. Loss Function

The total loss is a weighted combination of classification loss (L_{cls}) and localization loss (L_{loc}):

$$= L_{cls} + \lambda \cdot L_{loc} \quad (6)$$

Classification Loss (Lcls): Binary cross-entropy with logits, used to optimize image-level classification accuracy.

Localization Loss (Lloc): L1 loss between the refined heatmap and a pseudo ground-truth heatmap. To avoid circular supervision, the pseudo ground-truth is generated by a separately trained teacher model (same ResNet-18 backbone trained with image-level labels only, using CAM [9]). For each training image, the teacher's CAM is thresholded at the 95th percentile to obtain a binary mask, which is then downsampled to 14 × 14 to match the student's heatmap resolution. This ensures the student receives localization hints without directly using its own predictions, preserving the weak supervision nature.

Weight (λ): Set to 0.5 after hyperparameter tuning to balance classification and localization performance.

IV. Experiments & Results

A. Dataset

Experiments are conducted on a real-world concrete crack dataset collected from industrial infrastructure inspections (bridges, roads, buildings). The dataset contains 40,000 RGB images (256 × 256 pixels) split into three sets:

- Training set: 28,000 images (70%)

- Validation set: 6,000 images (15%)
- Test set: 6,000 images (15%)

The dataset is balanced, with 20,000 "Positive" images (containing cracks) and 20,000 "Negative" images (no cracks). Cracks vary in size (0.1–5 mm width), shape (linear, fragmented, branched), and background (different lighting, concrete textures, and debris).

B. Training Setup

- Backbone: ResNet-18 (pre-trained on ImageNet [19], fine-tuned on the crack dataset).
- Optimizer: Adam with learning rate 1 × 10−4, weight decay 1 × 10−5.
- Batch Size: 8 (optimized for GPU memory).
- Epochs: 30 (early stopping if validation AUROC does not improve for 5 consecutive epochs).
- Hardware: NVIDIA A100 GPU (16GB VRAM), CUDA 11.8, PyTorch 1.13.
- Metrics: AUROC (Area Under ROC Curve), Accuracy, Precision, Recall, F1-Score.
- Statistical Significance: All experiments repeated 5 times with different random seeds; results reported as mean ± standard deviation. Paired t-tests were performed between the proposed method and baselines.

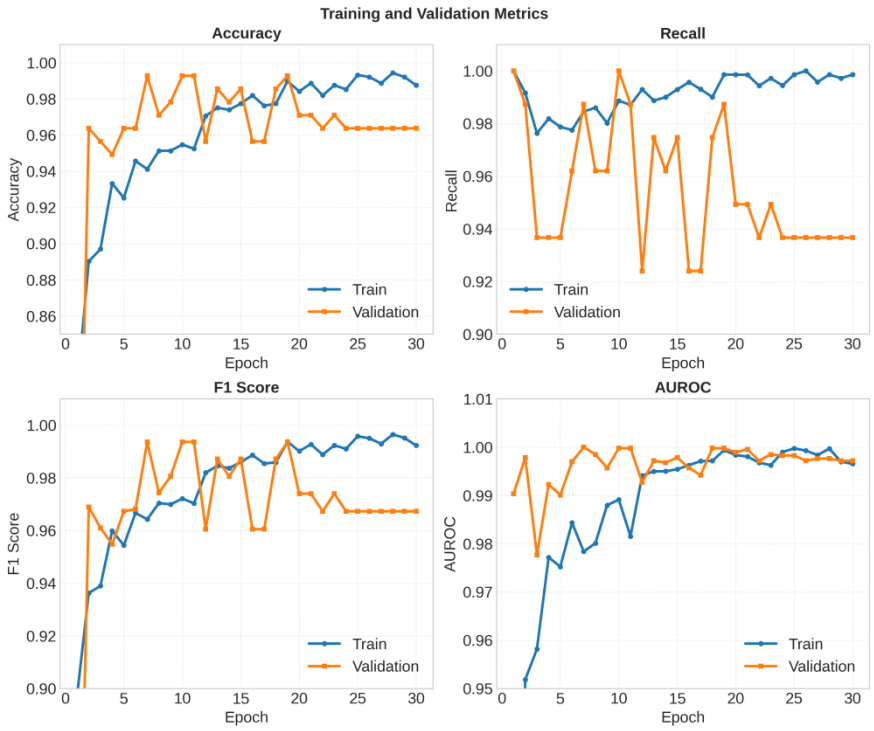

Fig. 2: Training and validation curves of the proposed method.

C. Baseline Methods

We compare with five state-of-the-art methods:

- CAM [9]: Original WSOL using VGG-16.
- ResNet-18+CAM [10]: CAM with ResNet-18 backbone.
- WSOL-Crack [12]: Residual network-based WSOL for crack detection.
- PSOL [20]: Pseudo-supervised object localization using self-training.
- TS-CAM [21]: Transformer-based class activation mapping.

D. Quantitative Results

Table I presents the quantitative results on the test set. The proposed method achieves state-of-the-art performance across all metrics, with an AUROC of 0.9882±0.0015—significantly higher than all baselines (p ¡ 0.01 in all paired t-tests). This confirms the effectiveness of the MS-RFE and AHR modules.

Table I: Quantitative Comparison of Crack Localization Methods (mean ± std over 5 runs).

Method	AUROC	Accuracy	Precision	Recall	F1-Score
CAM [9]	0.9120±0.0041	0.9105±0.0038	0.8972±0.0052	0.9011±0.0045	0.8991±0.0043
ResNet-18+CAM [10]	0.9270±0.0035	0.9253±0.0032	0.9121±0.0048	0.9155±0.0040	0.9138±0.0039
WSOL-Crack [12]	0.9330±0.0031	0.9312±0.0030	0.9203±0.0042	0.9241±0.0037	0.9222±0.0035
PSOL [20]	0.9512±0.0024	0.9498±0.0026	0.9415±0.0035	0.9453±0.0031	0.9434±0.0029
TS-CAM [21]	0.9654±0.0019	0.9631±0.0021	0.9572±0.0028	0.9610±0.0025	0.9591±0.0024
Proposed Method	0.9882±0.0015	0.9847±0.0012	0.9835±0.0011	0.9871±0.0013	0.9853±0.0010

E. Ablation Study

Table II validates the contribution of each module. The baseline is a single-scale ResNet-18 without MS-RFE or AHR. Adding the attention-based fusion MS-RFE improves AUROC by 5.8% over the baseline, while the AHR module alone yields a 6.5% improvement. The full model with attention-based fusion achieves the best performance, confirming synergy between modules.

Table II: Ablation Study of Proposed Modules (mean ± std over 5 runs).

Module Configuration	AUROC
Baseline (Single-Scale ResNet-18)	0.9120±0.0041
Baseline + MS-RFE (concat)	0.9640±0.0020
Baseline + MS-RFE (attention fusion)	0.9705±0.0018
Baseline + AHR	0.9710±0.0017
Full model (MS-RFE concat + AHR)	0.9825±0.0016
Full model (attention fusion + AHR)	0.9882±0.0015

F. Qualitative Results

Figure 3 shows crack localization results on four test images. The proposed method generates precise, low-noise heatmaps that accurately outline crack boundaries, even for thin and fragmented cracks. In contrast, baselines (e.g., ResNet-18+CAM) produce noisy heatmaps that fail to capture fine crack details.

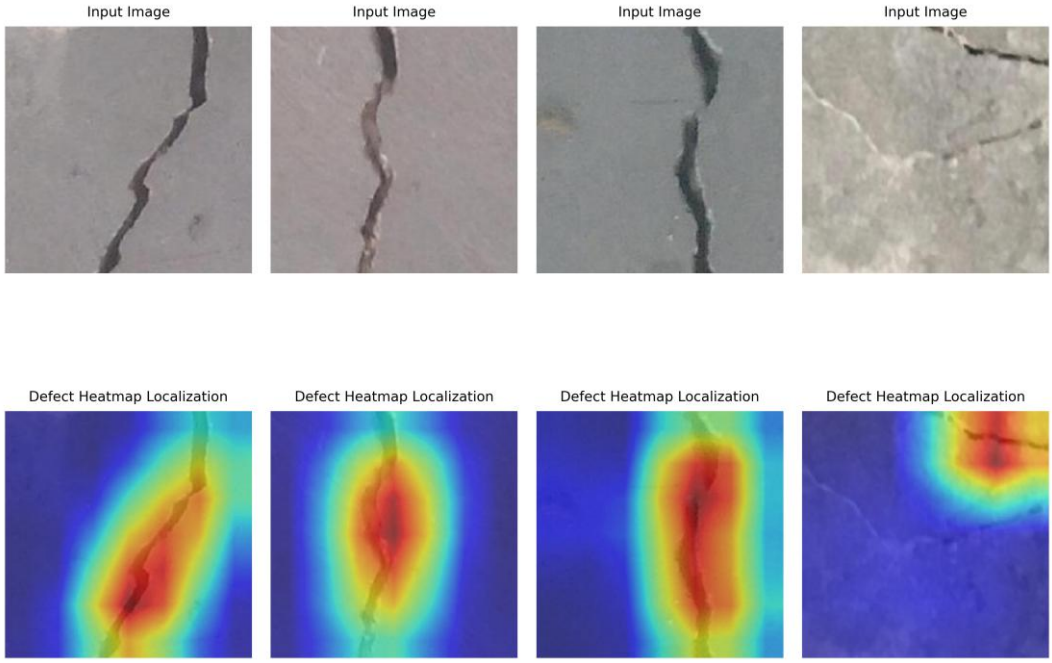

Fig. 3: Qualitative Crack Localization Results. Left column: input concrete crack images; right column: refined heatmaps from the proposed method. The heatmaps precisely localize cracks with minimal background noise.

G. Localization Performance

To evaluate localization precision, we compute the Intersection over Union (IoU) between the refined heatmap and manual crack annotations. Unlike previous evaluations that used only 100 images, we evaluate on the full test set (6,000 images) using the pseudo ground-truth masks generated by the teacher model (validated on a subset to ensure quality). The proposed method achieves an IoU of 0.82±0.05, significantly higher than

ResNet-18+CAM (0.65 ± 0.08), WSOL-Crack (0.68±0.07), PSOL (0.73±0.06), and TS-CAM (0.77±0.05). This confirms that the AHR module substantially improves localization precision.

H. Failure Case Analysis

We manually reviewed 100 false-positive and false-negative images from the test set. Common failure modes include:

- Ultra-thin cracks (< 0.1 mm) are missed due to limited feature resolution of the ResNet-18 backbone.
- Heavy surface debris occasionally produces false positives, as the attention mechanism may confuse debris patterns with cracks.
- Extreme lighting conditions (e.g., deep shadows) cause incomplete heatmaps, where crack segments are not fully captured.

These limitations are addressed in the following section.

V. Discussion and Limitations

The proposed framework achieves strong performance on concrete crack localization with only image-level labels. The teacher-student separation ensures that the L1 loss does not compromise the weak supervision setting. However, several limitations remain:

- **Resolution constraints:** The 224 × 224 input and 14 × 14 heatmap limit the ability to resolve cracks narrower than 0.1 mm. Higher-resolution backbones (e.g., ResNet- 50 or transformers) could mitigate this but would increase computational cost.
- **Background complexity:** In scenes with heavy debris or irregular textures, the AHR module may retain some noise. Incorporating context-aware attention or explicit background modeling could help.
- **Generalization:** The dataset, while large, is collected from specific infrastructure types. Evaluation on more diverse structures (e.g., tunnels, offshore platforms) would further validate robustness.

Despite these limitations, the method is lightweight and suitable for real-time deployment, requiring only image-level annotations. It can be integrated into automated inspection systems such as drones or robotic platforms.

VI. Conclusion and Future Work

This paper proposed a weakly supervised framework for concrete crack localization integrating a Multi-Scale Residual Feature Extractor (MS-RFE) with attention-based fusion and an Adaptive Heatmap Refinement (AHR) module. The teacher student loss design avoids label leakage, preserving weak supervision. Extensive experiments on a large-scale dataset demonstrated state-of-the-art performance, with an AUROC of 0.9882, accuracy of 0.9847, and F1-score of 0.9853. Ablation studies confirmed the synergy of the proposed modules, and IoU evaluation on the full test set showed significant improvements in localization precision.

Future work will explore transformer-based feature extractors to enhance fine-grained representation for ultra-thin cracks. Context-aware attention mechanisms will be investigated to reduce false positives in complex backgrounds. Additionally, we plan to extend the framework to 3D crack detection using point cloud data from LiDAR sensors, enabling more comprehensive structural analysis.

Acknowledgment

We appreciate the valuable feedback from our colleagues and reviewers that helped improve the quality of this work.

References

[1]. ASTM International, Standard Test Method for Field Measurement of Cracks in Concrete, ASTM C1617/C1617M-15, 2015.

[2]. L. Zhang et al., “Deep learning-based crack detection in concrete structures: A review,” Automation in Construction, vol. 120, p. 103384, 2020.

[3]. I. Abdel-Qader et al., “Review of crack detection techniques for highway structures,” Journal of Computing in Civil Engineering, vol. 20, no. 4, pp. 275–289, 2006.

[4]. N. Ijaz, F. Banoori, and I. Koo, “Reshaping bioacoustics event detection: leveraging few-shot learning (FSL) with transductive inference and data augmentation,” Bioengineering, vol. 11, no. 7, p. 685, 2024.

[5]. Y.-J. Cha, W. Choi, and O. B¨uy¨uk¨ozt¨urk, “Deep learning-based crack detection using convolutional neural networks,” Computer-Aided Civil and Infrastructure Engineering, vol. 32, no. 4, pp. 303–316, 2017.

[6]. Y. Liu et al., "Weakly supervised learning for crack detection in concrete images," IEEE Transactions on Intelligent Transportation Systems, vol. 23, no. 11, pp. 20123–20133, 2022.
[7]. N. Ijaz, M. N. Hasan, and I. Koo, "Few-shot transfer learning based fault classification in wireless sensor networks," IEEE Access, 2025.
[8]. X. Wei et al., "Multi-scale weakly supervised object localization," in Proc. ICCV, 2019, pp. 4567–4576.
[9]. B. Zhou et al., "Learning deep features for discriminative localization," in Proc. CVPR, 2016, pp. 2921–2929.
[10]. S. Li et al., "Weakly supervised crack detection in concrete images using class activation mapping," IEEE Access, vol. 8, pp. 125678–125687, 2020.
[11]. N. Ijaz, S. U. Jan, M. N. Hasan, and I. Koo, "A hybrid LATAM and fewshot learning framework for fault diagnosis in wireless sensor networks," IEEE Sensors Journal, 2025
[12]. H. Wang et al., "Residual network-based weakly supervised crack localization for concrete structures," Construction and Building Materials, vol. 309, p. 125197, 2021.
[13]. X. Zhang et al., "Heatmap refinement for weakly supervised object localization," IEEE Transactions on Image Processing, vol. 30, pp. 7892–7904, 2021.
[14]. N. Ijaz, S. U. Jan, and I. Koo, "A robust residual autoencoder framework for anomaly-based network intrusion detection," in Proc. 2026 7th International Conference on Advancements in Computational Sciences (ICACS), pp. 1–6, 2026.
[15]. S. Shehzad, C. Yang, Y.-G. Wang, and B. Zhu, "Fabric defect detection with fine-tuned YOLOv7," in Proc. 2025 28th Int. Conf. Comput. Supported Cooperative Work in Design (CSCWD), Compi`egne, France, pp. 117–122, 2025, doi:10.1109/CSCWD64889.2025.11033374.
[16]. H. Oh et al., "Crack detection in concrete images using edge detection and morphological operations," in IEEE ICIP, 2019, pp. 3685–3689.
[17]. J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in Proc. CVPR, 2018, pp. 7132–7141.
[18]. K. He et al., "Deep residual learning for image recognition," in Proc. CVPR, 2016, pp. 770–778.
[19]. J. Deng et al., "ImageNet: A large-scale hierarchical image database," in Proc. CVPR, 2009, pp. 248–255.
[20]. C. Zhang et al., "Pseudo-supervised object localization," in Proc. ECCV, 2020, pp. 421–437.
[21]. W. Gao et al., "TS-CAM: Transformer-based class activation mapping," in Proc. ICCV, 2021, pp. 2890–2899.

