Contents lists available at ScienceDirect

Structures

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

Pixel-level bridge crack detection and 3D localization: An improved hybrid approach combining segmentation and 3D reconstruction

Yang Fang

Huazhong University of Science and Technology, Wuhan, China

ARTICLE INFO

Keywords:
Bridge inspection
Crack detection
3D reconstruction
Unmanned Aerial Vehicle (UAV)
Computer vision

ABSTRACT

Unmanned Aerial Vehicles (UAVs)-based bridge pier inspection faces two major challenges: (1) *accurately and efficiently detecting fine cracks, and* (2) *reconstructing high-fidelity 3D models of bridge piers that often lacks sufficient surface features for recovering true geometry and correctly mapping cracks*. To address these issues, this study presents an improved hybrid approach for pixel-level bridge crack detection and 3D localization. First, a YOLOv11-CSD model is proposed to localize cracks. Next, a U-Net-SR network is proposed to segment fine cracks within the identified regions. An improved CasMVSNet architecture is then developed for accurate depth estimation. Finally, OpenMVS is employed to generate high-fidelity 3D reconstruction and achieve precise crack mapping. Case studies demonstrate that the YOLO-CSD model achieves a recall of 0.7247 and a mAP@0.5 of 0.8196; the U-Net-SR model attains a recall of 0.8054 and a mIoU of 0.7502; the improved CasMVSNet yields an overall error of 0.344; and the 3D crack mapping achieves an average recall of 0.9607 and mIoU of 0.7601. The resulting high-resolution 3D crack models enable precise localization and geometric characterization.

1. Introduction

With the rapid expansion of bridge infrastructure worldwide and the continued aging of existing structures, a growing number of bridges are reaching or exceeding their intended service life [1,2]. Over time, structural components such as bridge piers are increasingly exposed to environmental stressors, material fatigue, and cumulative traffic loads [3]. These factors accelerate the development of common deterioration phenomena, including cracks, surface spalling, and reinforcement corrosion. Failure to detect and address such defects in a timely manner may lead to progressive structural degradation and pose serious safety risks [4]. Therefore, regular and reliable inspection of bridge components, particularly load-bearing elements like piers, has become a critical component of infrastructure management and public safety assurance [5].

Over the past decade, 2D image-based crack detection has advanced significantly, driven by progress in Deep Learning (DL), computer vision, and the image acquisition technologies such as unmanned aerial vehicles (UAVs) [6–9]. Existing studies have shown promising results, achieving high accuracy in identifying surface defects across various materials. Nevertheless, the crack assessment in bridge piers continues to face certain persistent challenges.

The first challenge lies in the reliable detection of fine cracks, especially in UAV-acquired images where such defects are prevalent [10]. These microcracks often exhibit very low contrast relative to the surrounding concrete, making them susceptible to occlusion by image noise, illumination variability, and intricate surface textures. While state-of-the-art segmentation algorithms have improved detection performance, consistently identifying fine cracks in real-world scenarios remains a significant unresolved issue.

Secondly, the non-planar geometry of bridge piers—characterized by curvature, surface roughness, and localized pitting—introduces significant projective distortion in 2D imagery [11]. Cracks appearing on these irregular surfaces complicate the accurate measurement of their true physical dimensions and orientations. Moreover, UAV- and telephoto-based imaging techniques, while effective at capturing localized detail, often produce narrow fields of view with limited spatial context. This makes it difficult to determine the global position of detected cracks on the structure, which is crucial for damage mapping and long-term maintenance planning.

Recent efforts have explored 3D reconstruction techniques to correct 2D distortion and improve crack localization [12,13]. By projecting cracks onto a reconstructed 3D surface model, these methods can mitigate perspective effects and improve spatial accuracy. However, current approaches frequently suffer from incomplete reconstructions (e.g., holes or missing geometry) and low computational efficiency, limiting

E-mail address: y_fang@hust.edu.cn.

https://doi.org/10.1016/j.istruc.2025.110611
Received 11 July 2025; Received in revised form 12 October 2025; Accepted 2 November 2025
Available online 24 November 2025

their scalability in practical inspection workflows.

This study aims to address two key research gaps in the crack inspection of bridge piers: (1) *how to more accurately and efficiently detect fine cracks.* (2) *how to reconstruct high-fidelity 3D models of bridge piers, which often lack sufficient surface features for recovering true geometry and correctly mapping cracks.* To this end, this study proposes an improved hybrid approach for pixel-level bridge crack detection and 3D localization. The approach comprises four key components: (1) YOLO-CSD is proposed to efficiently extract crack regions of interest (ROIs) from UAV images, thereby reducing the computational load for subsequent semantic segmentation. (2) U-Net-SR is developed to accurately segment fine cracks within the identified ROIs. (3) An improved CasMVSNet is introduced for high-quality depth estimation from UAV-captured images. And (4) OpenMVS is employed to reconstruct detailed 3D bridge pier models and map crack textures, resulting in a final mesh with embedded crack semantic information.

The rest of this paper is structured as follows: Section 2 reviews related work on crack detection and 3D reconstruction of bridge piers. Section 3 details the proposed methodology. Section 4 presents the experimental validation, and Section 5 discusses the results, key contributions and limitations.

2. Related work

2.1. Crack detection

Crack detection plays a vital role in the maintenance of civil infrastructure, aiding in the prevention of further structural deterioration and ensuring public safety. Over the years, numerous automated crack detection approaches have been developed, evolving from traditional techniques to advanced deep learning models.

Traditional methods primarily include threshold-based [14–17], edge-based [18–22], and graph theory-based methods [23–27]. These perform adequately under ideal conditions but struggle with real-world variations in lighting, textures, and complex backgrounds, often requiring manual tuning [28,29]. To address these limitations, machine learning (ML)-based methods emerged, categorized into supervised (e.g. Support Vector Machine [30,31], Random Forest [32], and various ensemble models [33,34]) and unsupervised techniques (e.g. Principal Component Analysis [35,36], and clustering algorithms [37,38]). However, these shallow models rely on hand-crafted features, limiting their robustness and generalizability across diverse conditions.

With advancements in data acquisition, deep learning (DL)-based methods have become dominant, enabling end-to-end training and hierarchical feature learning. DL approaches for crack detection include image classification [39–41], object localization [42–44], and semantic segmentation [45,46], with segmentation gaining prominence for its pixel-level precision in structural health monitoring.

Convolutional Neural Networks (CNNs) form the backbone of many DL models, capturing spatial patterns through localized receptive fields and hierarchical layers. Key architectures include Fully Convolutional Networks (FCNs) for dense predictions [47–49], though they may overlook global context, leading to issues like the “all black” problem in imbalanced datasets [50]. SegNet adopts an encoder-decoder structure with pooling indices for efficiency [51–53], but lacks direct feature fusion, resulting in blurred boundaries and suboptimal performance in detecting fine crack details [54,55]. Enhancements such as dilated convolutions (also known as Atrous convolutions) [56–58] and Feature Pyramid Network (FPN) [59,60] expand receptive fields and retain spatial details. The DeepLab series, with Atrous Spatial Pyramid Pooling (ASPP), further improves fine-grained crack detection [61,62], though computational demands can hinder high-resolution tasks [63,64]. In contrast, U-Net and its variants [65], with symmetric encoder-decoder designs and skip connections, excel in segmenting small, thin structures under resource constraints [66–70].

Recent advances have introduced Transformer-based and generative models to overcome CNN limitations in global context modeling and data efficiency. Vision Transformers (ViT) leverage self-attention mechanisms for long-range dependencies, enabling better handling of complex crack patterns in varied environments [71,72]. Building on this framework, various ViT-based models, such as SegFormer and the Segment Anything Model (SAM), offer superior performance in accuracy and robustness for fine crack detection [73–75]. However, ViT models typically require a significant amount of data to achieve optimal performance, potentially leading to suboptimal results on smaller datasets. Moreover, their high computational costs, driven by larger parameter counts, result in substantial resource consumption compared to CNN-based methods [76]. To mitigate data deficiency challenges, diffusion models, which iteratively denoise images, have been adapted for crack segmentation, generating high-fidelity masks from partial annotations and addressing data scarcity in infrastructure monitoring [77–80]. These developments underscore the shift toward more robust, efficient, and adaptable methods for crack detection, paving the way for integrated systems in structural health monitoring.

2.2. 3D reconstruction of cracks on concrete bridge piers

Crack detection on bridge piers is typically conducted using telephoto lenses or UAV-based close-up imaging, which, while effective for capturing fine details, inherently lack broader spatial context. These imaging approaches, constrained by narrow fields of view, often fail to capture the spatial relationships among multiple cracks or between cracks and key structural features such as joints, edges, or reinforcement zones. Moreover, the complex, non-planar geometry of bridge piers, such as curvature, surface roughness, and concrete pitting, introduces significant projective distortions in two-dimensional imagery. As a result, such images may misrepresent the true spatial extent or orientation of cracks, undermining both qualitative interpretation and quantitative measurement.

To overcome these limitations, accurate 3D reconstruction of bridge piers has become increasingly essential in the domain of structural health monitoring. A high-fidelity 3D model not only restores the spatial integrity of surface imagery but also enables precise localization and metric analysis of surface defects. Two widely adopted approaches for 3D reconstruction in structural inspection are stereo vision-based and Structure-from-Motion (SfM)-based methods.

Stereo vision relies on the geometric principle of triangulation, using the disparity between two spatially separated cameras to estimate depth and reconstruct the 3D structure of a scene. For example, Kim et al. [81] proposed a stereo imaging framework combining a wide-range lens and a telephoto lens to achieve 3D reconstruction of cracks on concrete bridge piers. Deng et al. [82] integrated binocular visual simultaneous localization and mapping (VSLAM) with a DL-based crack segmentation algorithm, enabling accurate 3D representation and global localization of cracks from binocular video sequences. Feng et al. [12] employed a fusion of LiDAR-Inertial Odometry (LIO) and Visual-Inertial Odometry (VIO) SLAM techniques, combined with triangular meshing, to generate a textured point cloud and mesh model of the bridge pier. Crack semantic information extracted from individual images was then projected onto the 3D model using the corresponding camera projection matrices. Stereo vision-based methods are relatively underutilized in structural crack inspection. A key limitation lies in their dependence on a fixed camera baseline and local disparity estimation, which eliminates the need for global image registration. While this simplification allows for faster reconstruction, it often compromises accuracy. In structural health monitoring tasks, where high geometric precision is critical for reliable damage diagnosis and quantification, this trade-off renders stereo vision less suitable.

SfM-based methods are often favored for their superior accuracy in generating dense and geometrically reliable 3D models, despite being more computationally intensive. A typical workflow begins with the application of DL-based semantic segmentation algorithms to extract cracks from images. In parallel, 3D reconstruction is performed using software such as ContextCapture or COLMAP, which implement SfM and Multi-View Stereo (MVS) techniques to produce a detailed mesh model of the bridge piers. The estimated projection matrices are then used to accurately project the detected cracks onto the reconstructed 3D surface [83–85]. However, traditional SIFT-based SfM+MVS approaches rely solely on geometric depth estimation, which can suffer from textureless regions and feature mismatches, particularly on low-texture surfaces such as concrete bridge piers. In comparison, DL-based methods that extract learned depth features offer improved matching accuracy and efficiency, leading to more complete and accurate reconstructions. For instance, Qi et al. [13] proposed a 3D reconstruction framework that combines COLMAP with PatchMatchNet for dense point cloud generation. Following the reconstruction, a mesh model is created and its surface is unfolded onto a 2D plane. Detected crack information is then mapped onto this 2D surface and subsequently reprojected onto the 3D model. Yu et al. [86] introduced a Voxel Neural Radiance Fields (VNRF) approach to reconstruct the visual appearance of bridge pier structures in underwater environments. In this framework, cracks on the structural surface are represented as 3D voxels, enabling volumetric visualization and analysis of underwater damage. Recently, 3D Gaussian Splatting (3DGS) has received extensive attention in 3D reconstruction tasks as an emerging deep learning-based technique, though its application in bridge crack reconstruction remains relatively limited. Duan et al. [87] proposed a framework integrating the Segment Anything Model 2 (SAM 2) with 3DGS for high-quality defect segmentation and 3D reconstruction from monocular videos. This method first employs SAM 2 for defect segmentation and tracking, followed by SfM to generate sparse point clouds, and then utilizes 3DGS for efficient rendering.

3. Methodology

3.1. Pipeline

The pipeline of the proposed approach (Fig. 1) consists of the following steps: (1) The UAV captures images around the bridge pier. (2) YOLO-CSD is used to localize crack regions and remove irrelevant background information. (3) U-Net-SR performs semantic segmentation on the ROIs to obtain the crack regions. (4) COLMAP estimates the internal and external parameters of the images via SfM, and to reconstruct a sparse point cloud. (5) Improved CasMVSNet estimates depth maps from the UAV images and corresponding camera parameters. Finally, (6) OpenMVS generates a dense point cloud using the sparse point cloud, depth maps, and crack masks, resulting in a 3D mesh model of the bridge pier with embedded crack semantics.

3.2. Crack localization based on YOLOv11-CSD

As a state-of-the-art object detection algorithm, YOLOv11 adopts three core components: a backbone for feature extraction, a neck for feature fusion, and a head for final prediction. However, in bridge crack detection scenarios, challenges such as low image resolution, small and thin crack structures, and complex or low-contrast background conditions often hinder its ability to capture fine-grained structural details. To address these limitations, this section proposes an YOLOv11-CSD (Fig. 2) which introduces three architectural improvements to boost detection performance under such constraints.

(1) Space-to-Depth Convolution (SPD-Conv) module [88]. Traditional stride= 2 convolutions are effective in expanding the receptive field and capturing global features, but they tend to reduce spatial resolution and result in the loss of fine-grained information, particularly for small and thin cracks [89,90]. To address this issue, the SPD-Conv module is introduced to replace the second through fifth stride= 2 convolution layers in the YOLOv11 backbone. The module integrates a non-strided 3 × 3 convolution for preserving spatial details with a Space-to-Depth (SPD) operation that reduces spatial dimensions and maps them into the channel dimension. This structure retains rich local information, enhances the model's ability to detect small objects, and improves performance under complex background conditions.

(2) Context Augmentation Module (CAM) [91]. The original SPPF module in YOLOv11, which relies on fixed-scale pooling operations, lacks the capacity to adaptively capture contextual cues beyond the immediate receptive field [92]. To address this limitation, the CAM module is introduced as a replacement for Spatial Pyramid Pooling Fast (SPPF). CAM employs dilated convolutions with multiple dilation rates to aggregate multi-scale contextual information, enabling the model to better differentiate cracks from visually similar background patterns. Feature fusion in CAM is performed using channel-wise concatenation. By incorporating surrounding structural context into feature representation, CAM enhances the model's robustness against background interference

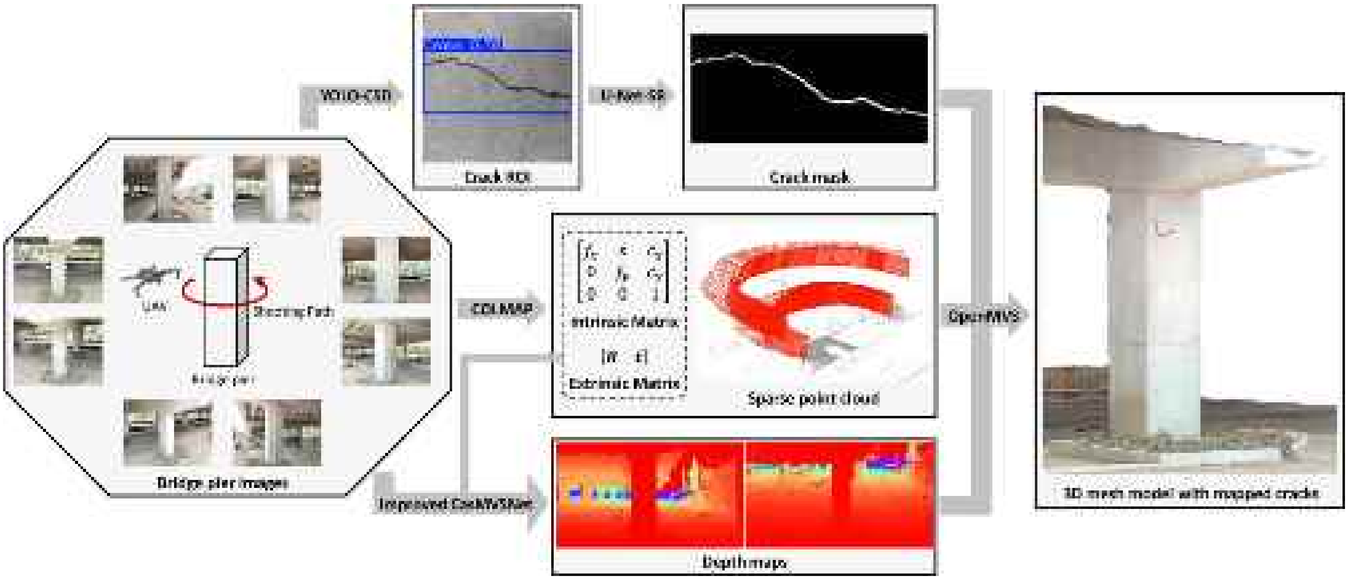


Fig. 1. Pixel-level bridge crack detection and 3D localization pipeline.

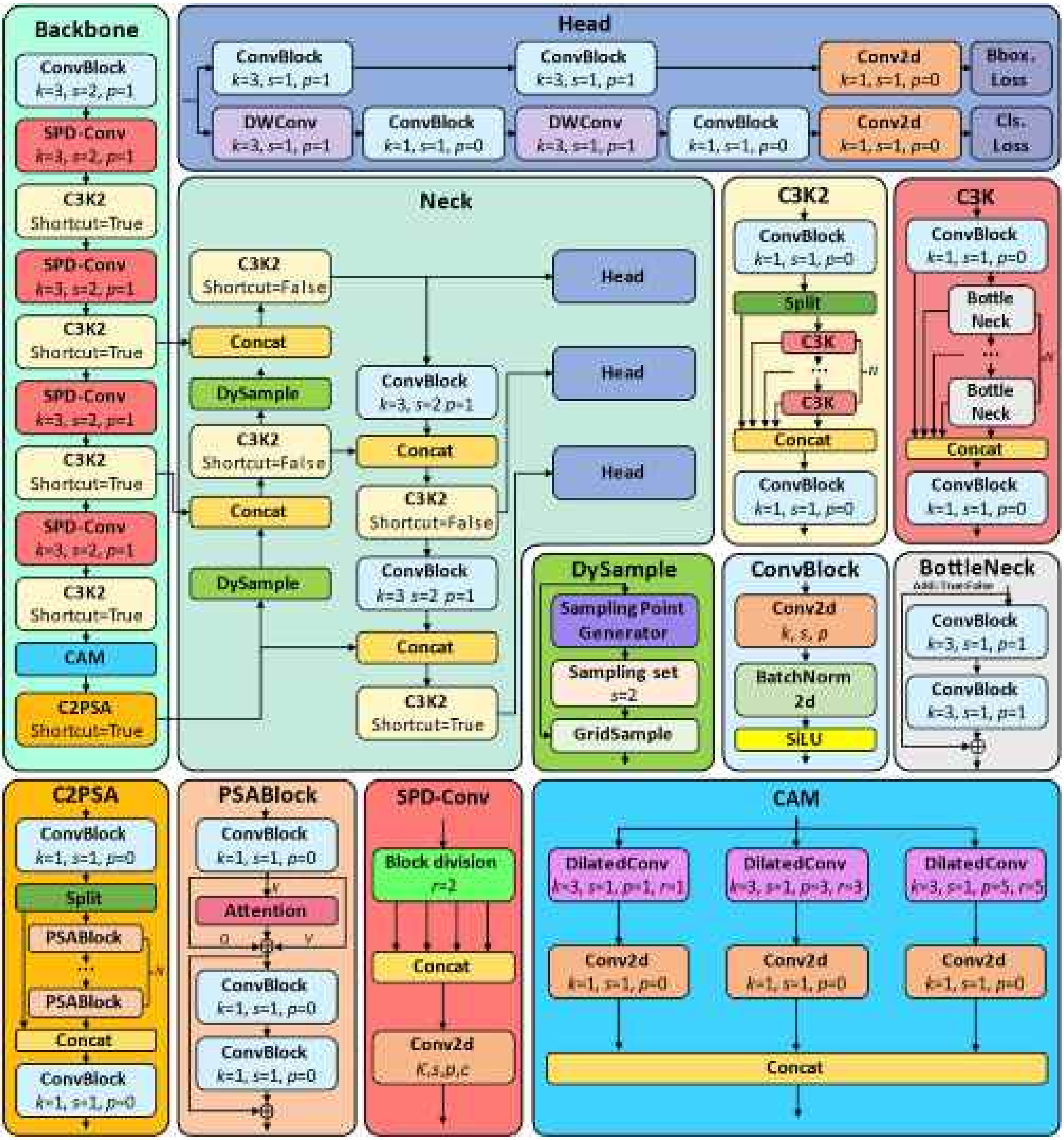


Fig. 2. YOLOv11-CSD architecture.

and significantly improves detection accuracy in real-world scenarios.

(3) Lightweight dynamic upsampling module (DySample) [93]. YOLOv11 employs computationally cheap nearest-neighbor interpolation for upsampling but lacks adaptability to image content. This compromises the balance between semantic consistency and fine-detail reconstruction, particularly for critical boundaries in tasks like crack detection [94]. DySample addresses this by reformulating upsampling as point-wise sampling. Instead of fixed rules or expensive dynamic convolutions, it generates content-aware resampling by learning sampling coordinates directly from feature maps. This enables flexible, detail-preserving reconstruction with minimal overhead.

3.3. *Crack segmentation based on U-Net SR*

The images captured by UAVs typically feature a wide field of view, with bridge cracks appearing small and displaying low contrast against the surrounding concrete surface. These characteristics present challenges for accurate crack segmentation. In this study, a U-Net-SR is proposed, with the architecture shown in Fig. 3.

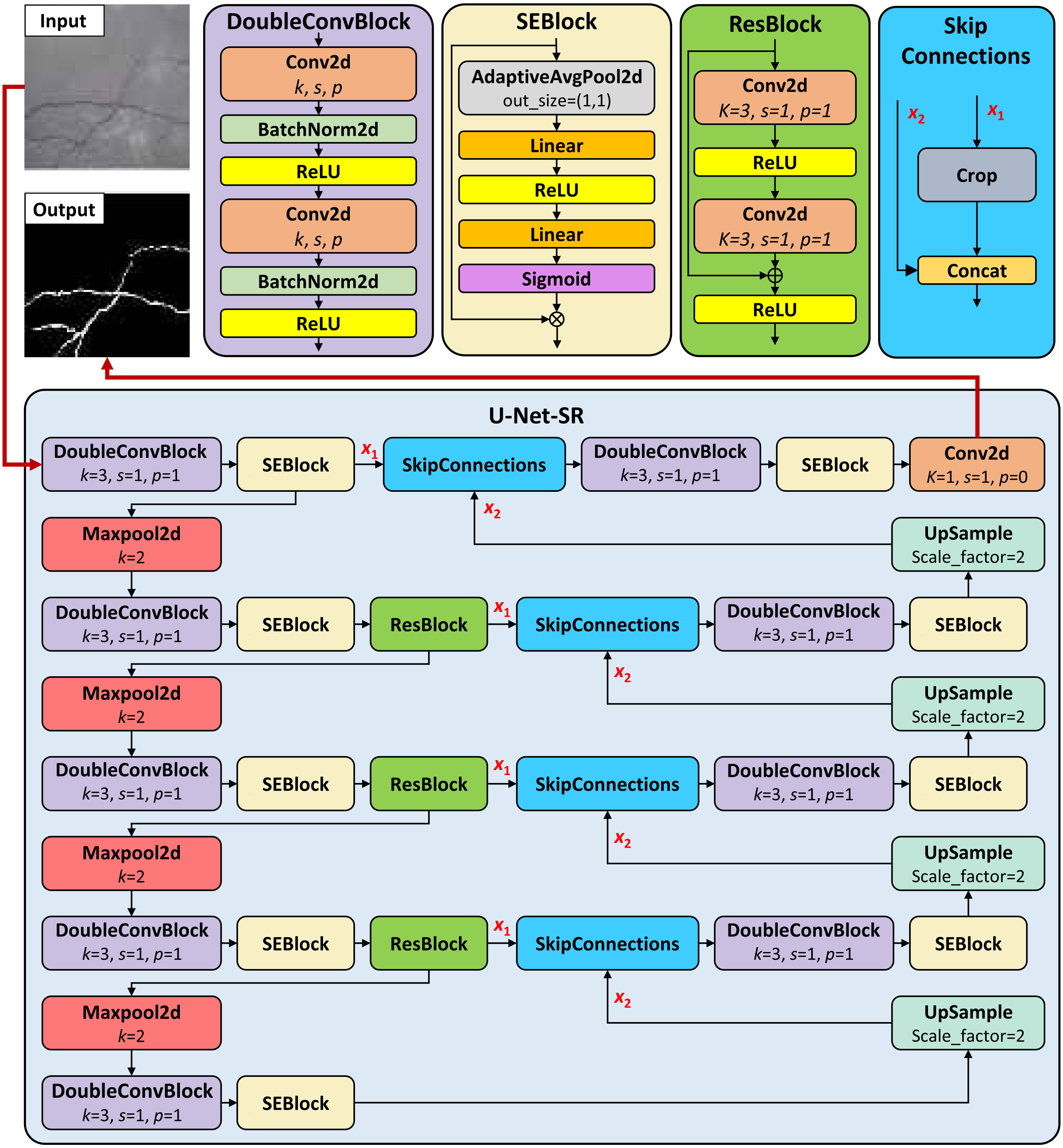


Fig. 3. U-Net-SR architecture.

(1) Squeeze-and-Excitation (SE) block [95]. Bridge crack images often exhibit varying object sizes and uneven feature distributions. By emphasizing important features and suppressing irrelevant ones, SE can improve the model's sensitivity to key structural details.

(2) Residual block [96]. As network depth increases, issues such as vanishing gradients and overfitting may arise, degrading the model's performance. To mitigate these problems, Residual Blocks are integrated into the encoder of the U-Net architecture, enabling increased network depth while maintaining stability.

Let the overall input image be denoted as $x \in \mathbb{R}^{H\times W\times C}$, where H, W, and C represent the height, width, and number of channels, respectively. Semantic segmentation models operate over the entire image domain $\Omega = \mathbb{R}^{H\times W}$, resulting in a computational complexity of $o(HWC)$. This full-resolution processing not only increases the computational burden but also heightens the risk of misclassifying irrelevant background regions.

The computational complexity of object detection algorithms is generally much lower than that of the semantic segmentation. Therefore, the proposed pipeline first leverages YOLO-CSD for global inference, followed by a more computationally intensive U-Net-SR for local segmentation. This hierarchical strategy significantly reduces overall computational cost by narrowing the scope of high-resolution analysis to only the most relevant regions.

Specifically, the semantic segmentation model restricts the search space to a union of localized regions $\cup_{i=1}^{n} R_i \subset \Omega$, where $R_i \in \mathbb{R}^{h\times w}$ corresponds to a candidate region predicted to contain cracks. The pipeline is formulated as a function composition $g(f(x))$, where $f(\cdot)$ denotes the coarse crack localization performed by YOLO-CSD, outputting n candidate bounding boxes $\{R_i\}_{i=1}^{n}$, and $g(\cdot)$ denotes the fin-grained segmentation performed by U-Net-SR on each extracted region R_i.

The overall computational complexity of the pipeline can thus be expressed as:

$$o(HWC_{\text{YOLOv11-CSD}}) + \sum_{i=1}^{n} o(hwC_{\text{U-Net-SR}}) \qquad (1)$$

By restricting the segmentation task to the regions of interest predicted by YOLO-CSD, the pipeline avoids redundant computations over non-informative background areas, thereby improving efficiency and reducing the likelihood of false positives.

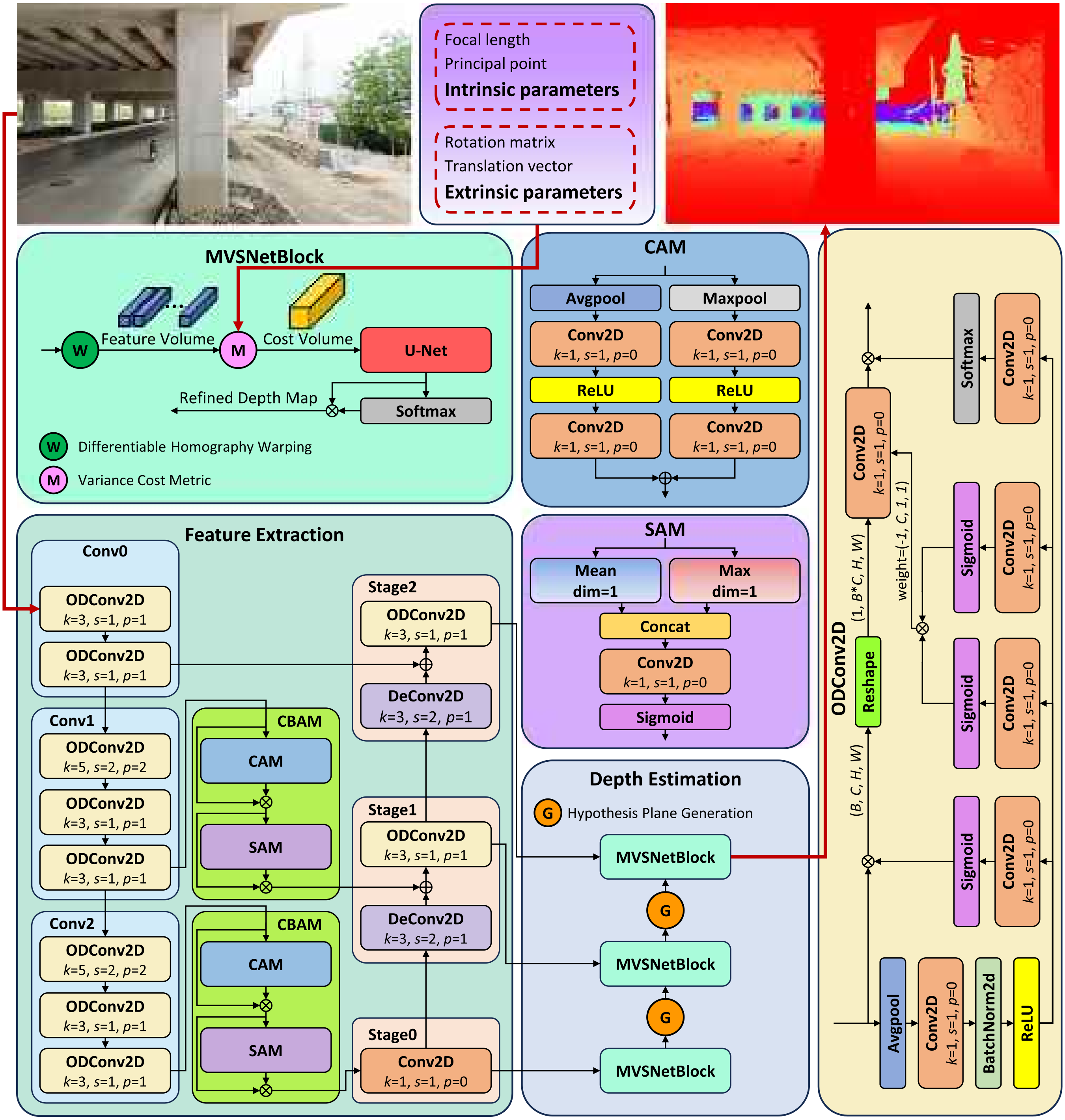


Fig. 4. Improved CasMVSNet architecture.

3.4. Depth estimation based on improved CasMVSNet

CasMVSNet, an improved version of the Multi-View Stereo Network (MVSNet), is a DL-based framework for MVS reconstruction [97]. It combines an FPN with a cascade optimization strategy to deliver accurate and efficient depth estimation from unstructured multi-view imagery. However, while FPN facilitates multi-scale feature extraction, it does not differentiate the relative importance of features at various scales. In practice, features from target reconstruction regions are typically more valuable, whereas those from background areas contribute less.

To enhance the feature extraction and representation capabilities of CasMVSNet, this study proposes an improved architecture (Fig. 4) incorporating two key modifications:

(1) Omni-Dimensional Dynamic Convolution (ODConv) [98]. ODConv dynamically adjusts convolutional weights via multi-dimensional attention (spatial, input/output channels, kernel space). By jointly optimizing four attention mechanisms within a single convolution layer, it significantly enhances feature representation capabilities without substantial computational cost.
(2) Convolutional Block Attention Module (CBAM) [99]. CBAM is a lightweight yet effective attention mechanism consisting of two submodules: the Channel Attention Module (CAM) and the Spatial Attention Module (SAM). By sequentially applying channel-wise and spatial attention, CBAM adaptively highlights informative features while suppressing less relevant ones.

3.5. 3D bridge pier reconstruction and crack texture mapping

This study leverages OpenMVS for 3D reconstruction and crack mapping. The pipeline (Fig. 5) comprises:

(1) Camera pose estimation and sparse reconstruction: Raw images are processed using COLMAP to estimate camera poses and generate a sparse point cloud through SfM.
(2) Dense point cloud generation: OpenMVS integrates the camera poses, sparse point cloud, and depth maps precomputed by the improved CasMVSNet to densify the reconstruction. The precomputed depth maps serve as initial estimates and are refined for each view using OpenMVS's *PatchMatch* function, which propagates depth hypotheses across neighboring pixels while enforcing multi-view consistency through photometric and geometric constraints. These refined depth maps are then fused into a consistent dense point cloud by back-projecting pixels into 3D space and filtering outliers based on visibility and reprojection error thresholds.
(3) Mesh reconstruction: The dense point cloud is converted into a triangulated mesh using OpenMVS's *ReconstructMesh* function, which applies Poisson surface reconstruction to preserve surface topology.
(4) Texture mapping of segmented cracks: Using the camera poses, OpenMVS's *TextureMesh* function projects segmented crack textures onto the mesh, blending multi-view images for seamless mapping.

4. Case study

4.1. Dataset description

In this study, a crack localization dataset and a crack segmentation dataset were constructed to train the YOLOv11-CSD and U-Net-SR models, respectively. The improved CasMVSNet was trained on DTU dataset [100] to support accurate 3D reconstruction. A high-resolution bridge pier image dataset was constructed to evaluate the 3D reconstruction performance of the proposed MVS-based pipeline for automated reconstruction and crack mapping of bridge piers.

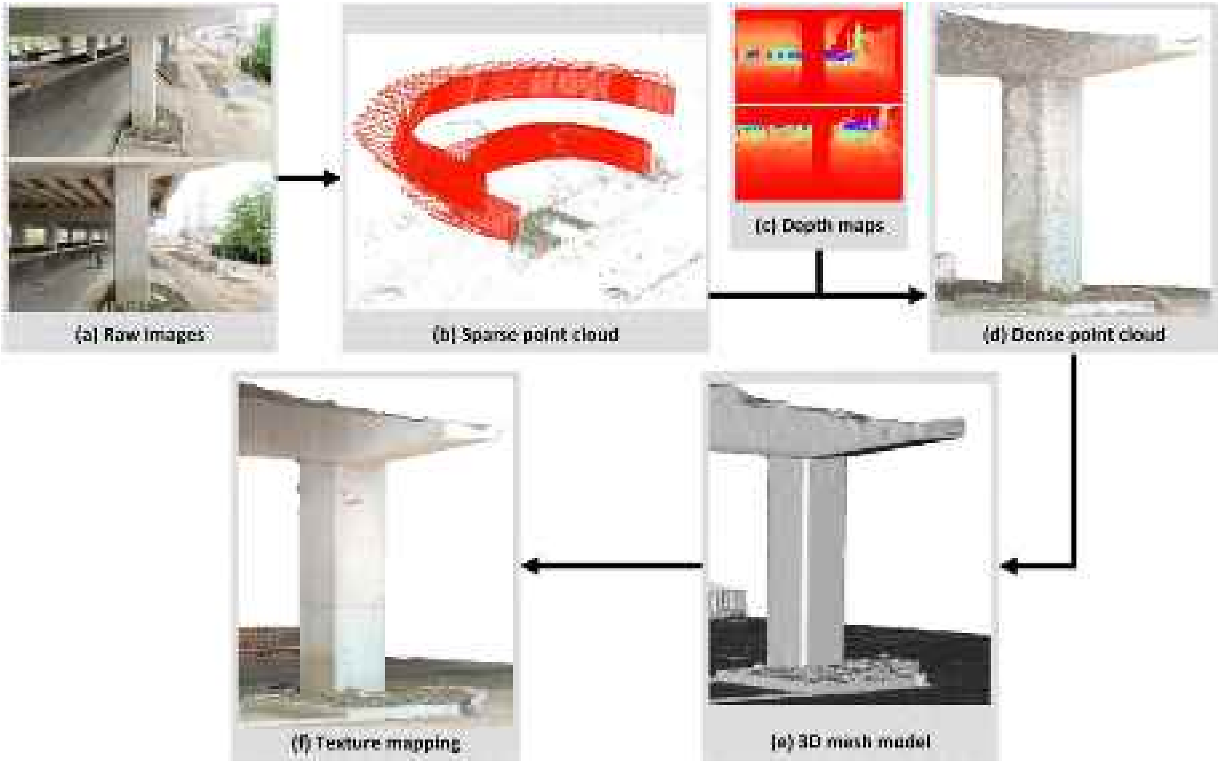


Fig. 5. Pipeline for 3D reconstruction and crack mapping of bridge pier.

4.1.1. Crack localization dataset

The crack localization dataset is derived from an open-source concrete crack dataset on the Roboflow platform, consisting of 5070 images [101]. The dataset is partitioned into 4196 training images, 353 validation images, and 521 test images. Fig. 6 presents several example images. During the training phase, the training set is augmented threefold, resulting in a total of 12,588 augmented training samples. Analysis of 100 randomly sampled images (Fig. 7(a) and (b)) reveals two key bounding box characteristics: (1) aspect-ratio statistics indicate numerous elongated crack boxes with pronounced width-height asymmetry. (2) a high frequency of low crack area ratios (<20 %) suggests an abundance of small-scale cracks in the dataset.

4.1.2. Crack segmentation dataset

The crack segmentation dataset comprises 11298 images aggregated from multiple public datasets (CFD, CRACK500, DeepCrack, Gaps, Volker, Rissbilder, and Noncrack). This comprehensive collection features diverse crack manifestations across various substrates, including concrete and asphalt surfaces. The dataset is partitioned into training (9040 images), validation (1129 images), and test sets (1129 images). Fig. 8 illustrates representative samples from these datasets.

4.1.3. High-resolution bridge pier image dataset

The high-resolution bridge pier image dataset was collected using a DJI Air 3 drone, which includes 147 4032 × 2268 images of a bridge pier from multiple angles during an infrastructure renovation project. A subset of these images is shown in Fig. 9.

4.2. Training setup

The computational platform consisted of an Intel(R) Xeon(R) Gold 6138 CPU paired with an NVIDIA RTX A5000 GPU.

For YOLO-CSD model training, input images are resized to 224 × 224 pixels and processed in batches of 32 over a total of 1000 epochs. Early stopping is implemented with a patience parameter of 20 epochs to prevent overfitting. The training process uses an initial learning rate of 0.01 and a weight decay coefficient of 0.005 for regularization.

For U-Net-SR model training, the input dimensions are resized to 224 × 224 pixels, with a batch size of 32. The model was trained for 1000 epochs using the Adam optimizer (momentum coefficient = 0.9) and an initial learning rate of 0.0001.

For the improved CasMVSNet, input images were resized to 1152 × 864 pixels. The model was trained for 16 epochs with a batch size of 4, a depth sampling number (Numdepth) of 192, and an initial learning rate of 0.001. The interval scale was set to 1.06. The FPN extracted multi-scale features at resolutions of 1/16, 1/4, and full scale relative to the input. During cascade processing, the depth hypothesis planes (Ndepths) were set to 48, 32, and 8 for each stage, with corresponding depth intervals (Depth_inter_r) of 4, 2, and 1.

4.3. Crack localization evaluation

4.3.1. Evaluation metrics

The performance of the crack localization model is evaluated using five metrics:

Precision (P_{obj}): The ratio of correctly localized bounding boxes (with IoU≥0.5) to all predicted positive boxes, reflecting detection purity.

$$P_{\mathrm{obj}} = \frac{\mathrm{TP}_{\mathrm{obj}}}{\mathrm{TP}_{\mathrm{obj}} + \mathrm{FP}_{\mathrm{obj}}} \tag{2}$$

Where, $\mathrm{TP}_{\mathrm{obj}}$ is the number of objects correctly detected as the target class, $\mathrm{FP}_{\mathrm{obj}}$ is the number of objects incorrectly detected as the target class.

Recall (R_{obj}): The ratio of ground-truth cracks correctly detected (with IoU≥0.5) to the total number of ground-truth instances, reflecting the model's ability to cover actual cracks.

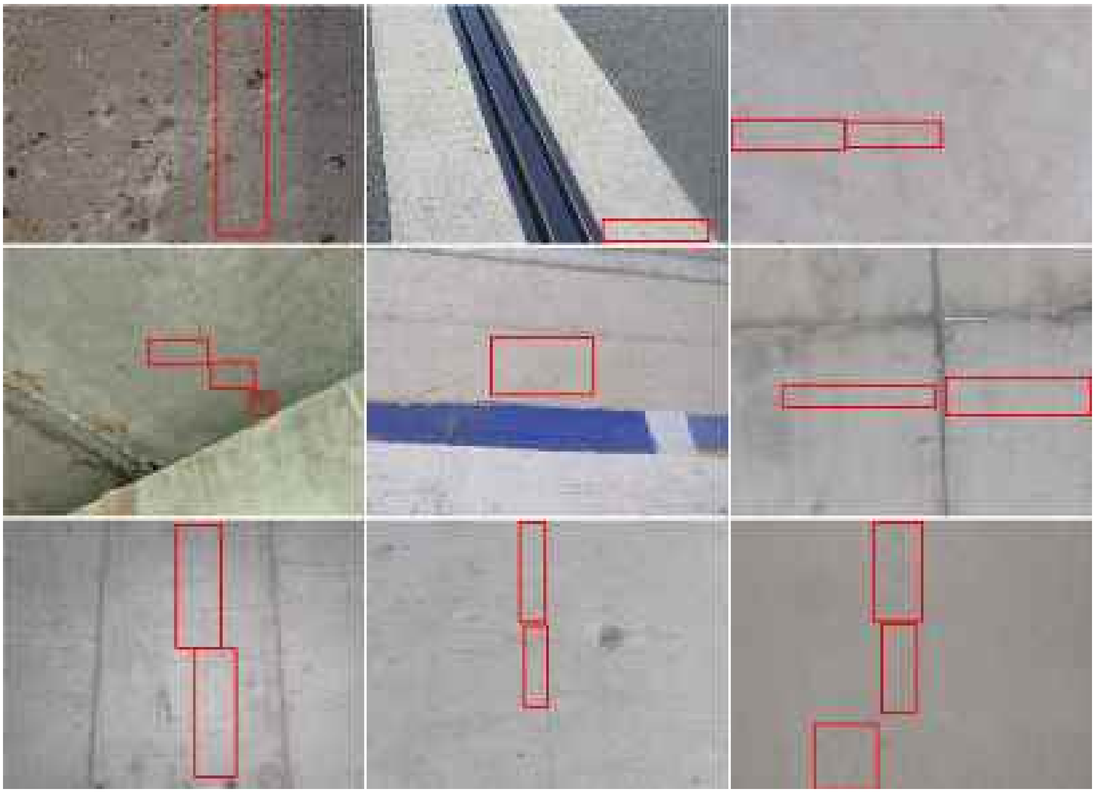
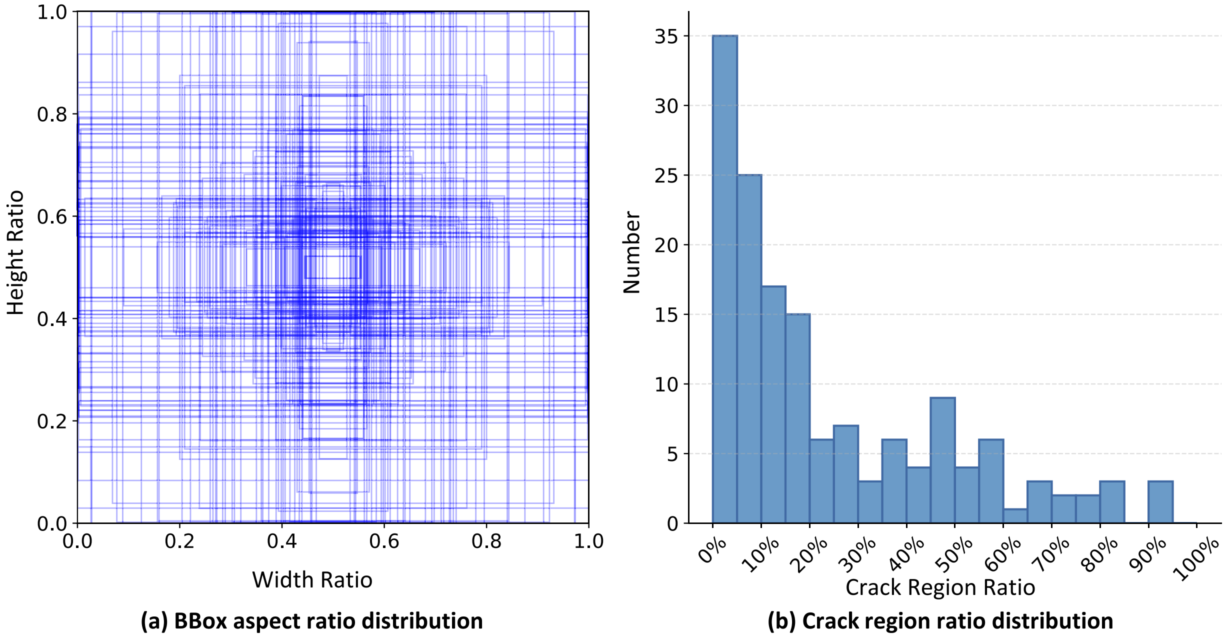


Fig. 6. Example images from the crack localization dataset.



(a) BBox aspect ratio distribution

(b) Crack region ratio distribution

Fig. 7. Crack localization dataset distribution.

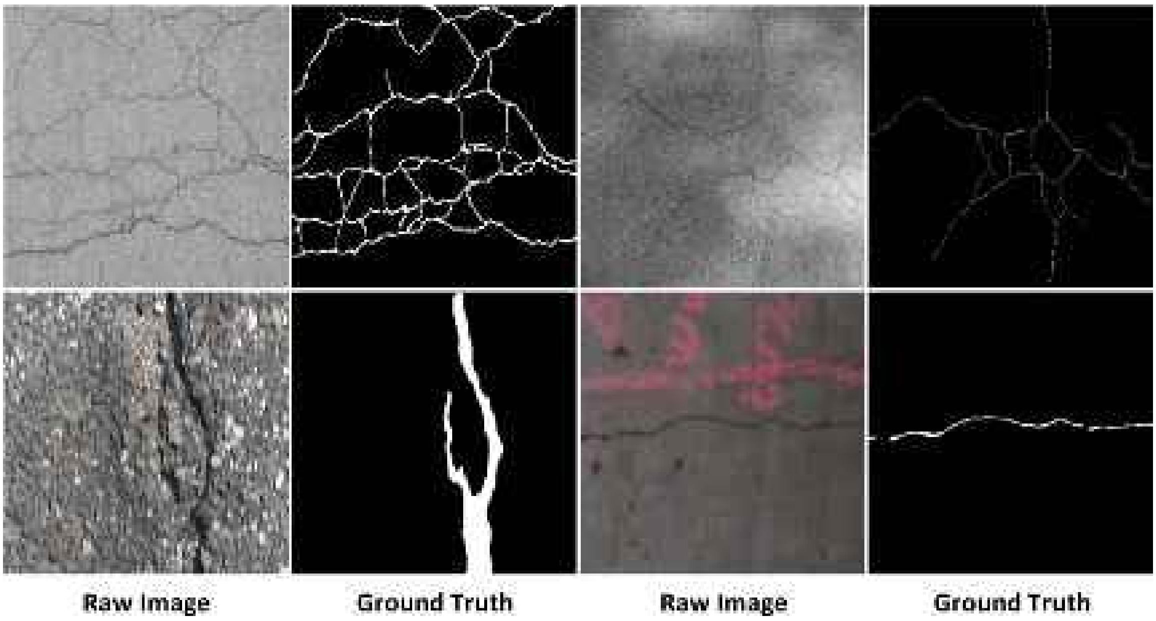


Fig. 8. Crack segmentation dataset.

$$R_{\mathrm{obj}} = \frac{\mathrm{TP}_{\mathrm{obj}}}{\mathrm{TP}_{\mathrm{obj}} + \mathrm{FN}_{\mathrm{obj}}} \tag{3}$$

Where, $\mathrm{FN}_{\mathrm{obj}}$ is the number of objects missed by the model.

F1-score (F_{obj}): The harmonic mean of precision and recall, balancing both above aspects.

$$F_{\mathrm{obj}} = \frac{2 \times P_{\mathrm{obj}} \times R_{\mathrm{obj}}}{P_{\mathrm{obj}} + R_{\mathrm{obj}}} \tag{4}$$

mAP@0.5: The mean Average Precision at an IoU threshold of 0.5, evaluating detection performance under a moderate localization criterion.

mAP@0.5:0.95: The mean Average Precision averaged over multiple IoU thresholds from 0.5 to 0.95 with a step size of 0.05, assessing the robustness and precision of bounding box localization under stricter criteria.

4.3.2. Ablation study

To comprehensively evaluate the impact of each improved module on YOLOv11-CSD performance, ablation studies were conducted on the crack localization dataset. The objective was to quantify the contribution of each module to detection accuracy and efficiency. Seven control groups were established, including the improved YOLOv11 model (Baseline) and various combinations of the three enhancements: CAM, SPD-Conv, and DySample. As shown in Table 1, the enhanced YOLOv11-CSD significantly reduces detail loss and improves small-object detection in complex scenes.

Fig. 9. subset of the high-resolution bridge pier image dataset.

Table 1
Ablation study results for YOLOv11-CSD.

Baseline	CAM	SPD-Conv	DySample	Params (M)	GFLOPs	FPS	P_{obj}	R_{obj}	F_{obj}	mAP@0.5	mAP@0.5: 0.95
√	×	×	×	2.6	6.4	256	0.8586	0.6659	0.7501	0.7525	0.5164
√	√	×	×	4.3	7.9	248	0.8684	0.6721	0.7577	0.7536	0.5216
√	×	√	×	3.3	12.5	240	0.8915	0.6704	0.7653	0.7672	0.5536
√	×	×	√	2.6	6.5	260	0.8611	0.6815	0.7607	0.7521	0.5179
√	√	√	×	5.1	14.0	232	0.8983	0.7125	0.7947	0.8070	0.5879
√	√	×	√	4.3	8.0	240	0.8926	0.6741	0.7681	0.7804	0.5592
√	×	√	√	3.4	12.6	244	0.8786	0.7173	0.7898	0.7862	0.5643
√	√	√	√	5.1	14.1	238	**0.9001**	**0.7247**	**0.8055**	**0.8196**	**0.5936**

4.3.3. *Comparative analysis*

To further evaluate the effectiveness of the proposed YOLOv11-CSD on the crack localization dataset, a comparative analysis was performed against several established models, including YOLOv5, YOLOv8, YOLOv10, YOLOv11, DINO-DETR-R50, and RT-DETR-R50. As illustrated in Fig. 10, the visualization results demonstrate that YOLOv11-CSD produces outputs closest to the ground truth. The blue segmentation lines precisely matching the crack's length, width, continuity, and fine details, exhibiting minimal omissions or noise.

The quantitative results in Table 2 confirm YOLOv11-CSD's effectiveness on datasets constaining a high proportion of small cracks. Notably, it achieves competitive or superior scores in key metrics (e.g., R_{obj} of 0.7247 and mAP@0.5 of 0.8196), closely matching or exceeding strong baselines like RT-DETR-R50, while maintaining superior efficiency (e.g., 238 FPS). Overall, YOLOv11-CSD demonstrates strong performance across evaluation criteria, effectively balancing high accuracy with real-time inference capabilities in challenging small-object detection.

4.4. *Crack segmentation evaluation*

4.4.1. *Evaluation metrics*

The crack segmentation model is evaluated using four metrics:

Precision (P_{pixel}): The ratio of correctly predicted crack pixels to all pixels classified as crack, indicating the purity of pixel-level segmentation.

$$P_{\text{pixel}} = \frac{\text{TP}_{\text{pixel}}}{\text{TP}_{\text{pixel}} + \text{FP}_{\text{pixel}}} \tag{5}$$

Where, TP_{pixel} is the number of true positive pixels, FP_{pixel} is the number of false positive pixels.

Recall (R_{pixel}): The ratio of correctly predicted crack pixels to the total number of actual crack pixels, reflecting the completeness of segmentation coverage.

$$R_{\text{pixel}} = \frac{\text{TP}_{\text{pixel}}}{\text{TP}_{\text{pixel}} + \text{FN}_{\text{pixel}}} \tag{6}$$

Where, FN_{pixel} is number of the false negative pixels.

F1-score (F_{pixel}): The harmonic mean of pixel-level precision and recall, offering a balanced evaluation.

$$F_{obj} = \frac{2 \times P_{\text{pixel}} \times R_{\text{pixel}}}{P_{\text{pixel}} + R_{\text{pixel}}} \tag{7}$$

mIoU: The mean IoU, computed by averaging the IoU across all samples or classes. A higher mIoU indicates more accurate and consistent segmentation performance.

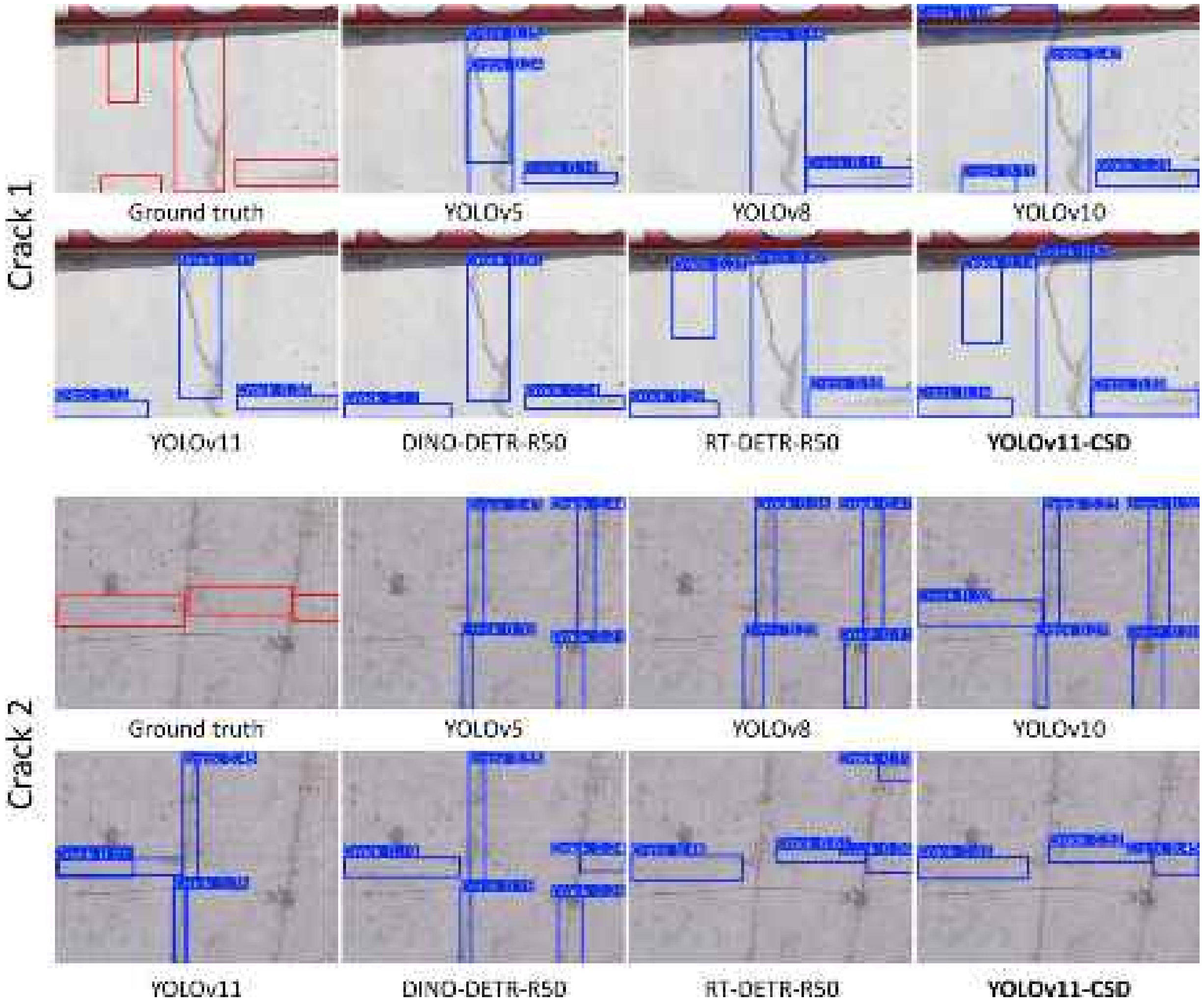


Fig. 10. Visual comparison of different models on the crack localization dataset.

Table 2
Performance comparison of different models on the crack localization dataset.

Method	**Params (M)**	**GFLOPs**	**FPS**	P_{obj}	R_{obj}	F_{obj}	**mAP@0.5**	**mAP@0.5–0.95**
YOLOv5	2.5	7.2	278	0.8522	0.6257	0.7216	0.7213	0.4955
YOLOv8	3.0	8.2	270	0.8646	0.6573	0.7468	0.7359	0.5054
YOLOv10	2.7	8.4	278	0.8574	0.6566	0.7437	0.7376	0.5081
YOLOv11	2.6	6.4	256	0.8586	0.6659	0.7501	0.7525	0.5164
DINO-DETR-R50	34.3	160.1	28	0.8966	0.7074	0.7908	0.7946	0.5754
RT-DETR-R50	31.5	81.9	115	**0.9024**	0.7206	0.8013	0.8168	**0.5946**
YOLOv11-CSD (Ours)	5.1	14.1	238	0.9001	**0.7247**	**0.8055**	**0.8196**	0.5936

4.4.2. *Ablation study*

To comprehensively evaluate the impact of each improved module on U-Net-SR performance, ablation studies were conducted on the crack segmentation dataset. The results of the ablation experiments are shown in Table 3. The experiment set up three control groups: the U-Net before optimization (Baseline), the U-Net with integrated SE module, and the U-Net with integrated residual block. The results show that combining both modules yields the highest performance gains, confirming their complementary effect in enhancing feature representation and training stability.

Table 3
Ablation study results for U-Net-SR.

Baseline	**SE**	**Residual Block**	**Params (M)**	**GFLOPs**	**FPS**	P_{pixel}	R_{pixel}	F_{pixel}	**mIoU**
√	×	×	31.0	41.9	26.4	0.8708	0.7637	0.8137	0.7168
√	√	×	31.2	41.9	23.5	0.8718	0.7806	0.8237	0.7413
√	×	√	31.1	41.9	25.1	**0.8826**	0.7945	0.8362	0.7446
√	√	√	31.3	41.9	22.6	0.8782	**0.8054**	**0.8402**	**0.7502**

4.4.3. Comparative analysis

To further evaluate the performance of the proposed U-Net-SR algorithm, experiments were conducted on the crack segmentation dataset using seven representative semantic segmentation networks as benchmarks: FCN, PSPNet, DeepLab-v3 + , U-Net, U-Net3 + , SegFormer-MiT-B5, and Mask2Former-R50.

As illustrated in Fig. 11, U-Net-SR produces segmentation outputs closely matching the ground truth, with white lines accurately capturing crack continuity, boundaries, and fine details, while exhibiting minimal omissions or noise. This performance is comparable to advanced models like SegFormer-MiT-B5 and Mask2Former-R50, without falling short in handling complex textures and narrow cracks. In contrast, earlier baselines such as FCN and PSPNet display more fragmentation and excessive noise, whereas U-Net-SR maintains robust pixel-level delineation, effectively reducing missed detections in fine crack regions.

As shown in Table 4, U-Net-SR achieves the highest scores in key metrics, including R_{pixel} of 0.8054, F_{pixel} of 0.8402, and mIoU of 0.7502. It outperforms baselines like SegFormer-MiT-B5 and U-Net3 + . U-Net-SR also maintains efficiency similar to U-Net, with comparable parameters (31.3 M) and GFLOPs (41.9), while achieving a competitive FPS of 22.6 that outperforms heavier models like U-Net3 + (11.3 FPS) and Mask2Former-R50 (18.2 FPS).

4.5. Bridge pier 3D reconstruction

4.5.1. Evaluation metrics

The bridge pier 3D reconstruction model is evaluated using three metrics:

Accuracy: Mean distance from each point in the predicted point cloud to its nearest corresponding point in the ground truth point cloud.

$$\text{Accuracy} = \frac{1}{|P|}\sum_{p\in P}\min_{q\in Q}\|p-q\|_2 \tag{8}$$

Where, Pdenotes the predicted point set; Q denotes the ground truth point set, $\|\cdot\|_2$ is the Euclidean distance.

Completeness: Mean distance from each point in the ground truth point cloud to its nearest point in the predicted point cloud.

$$\text{Completeness} = \frac{1}{|Q|}\sum_{q\in Q}\min_{p\in P}\|q-p\|_2 \tag{9}$$

Overall: Average of Accuracy and Completeness.

$$\text{Overall} = \frac{\text{Accuracy} + \text{Completeness}}{2} \tag{10}$$

4.5.2. Ablation study

To evaluate the impact of incorporating CBAM and ODConv into the proposed architecture, we conduct an ablation study on the DTU dataset, a widely used multi-view stereo benchmark, using three key performance metrics: Accuracy (Acc.), Completeness (Comp.), and Overall error, all measured in millimeters. For all metrics, lower values indicate better reconstruction quality.

As summarized in Table 5. the ablation results demonstrate that while CBAM alone improves completeness, it negatively impacts accuracy. However, when combined with ODConv, the model benefits from improved representation learning, leading to better geometric reconstruction performance overall. Specifically, the combination of CBAM and ODConv achieves the best performance across all three metrics: accuracy improves to 0.373 mm, completeness further improves to 0.315 mm, and the overall error is reduced to 0.344 mm.

This result indicates a synergistic effect between CBAM and ODConv, where CBAM enhances feature attention and ODConv provides more flexible and expressive convolutional kernels for geometry modeling. Although this combination increases the number of parameters (from 0.93 M to 1.06 M) and Video Random Access Memory (VRAM) usage (from 5.3 GB to 6.1 GB), the performance gains indicate a worthwhile trade-off in high-quality 3D reconstruction tasks. Fig. 12 displays depth maps generated by the improved CasMVSNet architecture.

4.5.3. Comparative analysis

The improved CasMVSNet is evaluated against two established baselines, MVSNet and CasMVSNet, on the DTU benchmark. As shown in Table 6, the improved CasMVSNet significantly outperforms MVSNet across all metrics while requiring substantially less VRAM usage. Compared to the original CasMVSNet, it achieves a lower overall error (0.344 mm vs. 0.375 mm), primarily due to a notable improvement in completeness (0.315 mm vs. 0.407 mm), while maintaining comparable accuracy. These results demonstrate that the proposed enhancements improve geometric completeness without compromising precision, with only a modest increase in model size and VRAM usage.

A qualitative comparison of dense point cloud reconstruction is conducted on the high-resolution bridge pier image dataset. Fig. 13 presents depth maps generated by COLMAP, MVSNet, CasMVSNet, and the proposed improved CasMVSNet, respectively. These depth maps are fed into OpenMVS to generate the final dense reconstructions, as shown in Fig. 14. The proposed approach produces significantly denser and more complete point clouds, especially in fine structural details such as the pier surface and ground plane, demonstrating its superior capability in recovering accurate and comprehensive 3D geometry.

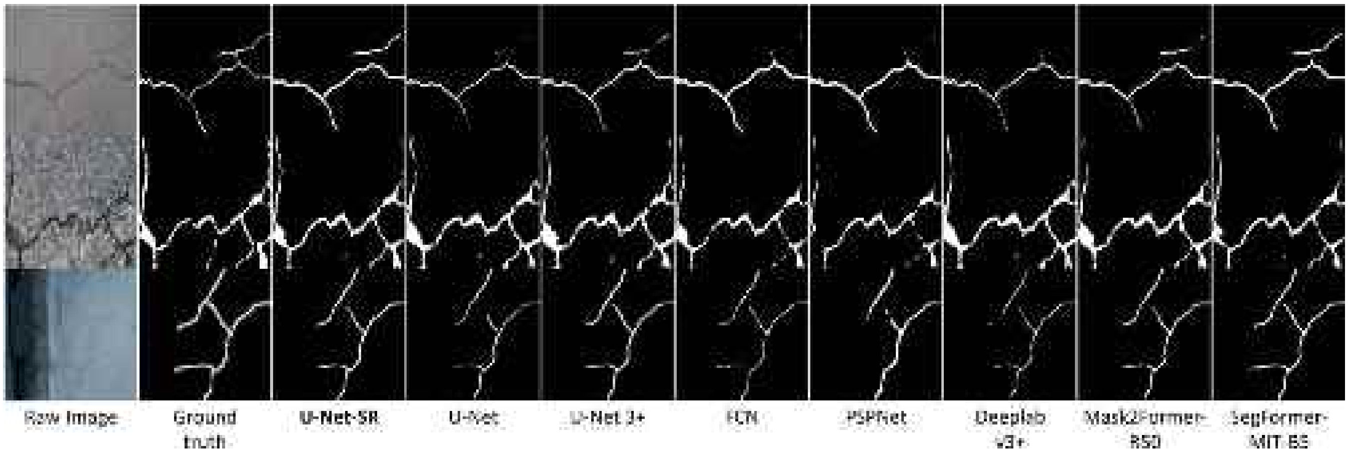


Fig. 11. Segmentation performance of different models.

Table 4
Performance comparison of different models on the crack segmentation dataset.

Method	Params (M)	GFLOPs	FPS	P_{pixel}	R_{pixel}	F_{pixel}	mIoU
FCN	47.1	37.9	24.7	0.8781	0.7723	0.8162	0.7277
PSPNet	46.6	34.2	25.3	0.8629	0.7193	0.7720	0.6805
DeepLab-v3 +	41.2	33.8	24.0	0.8740	0.7629	08079	0.7174
U-Net	31.0	41.9	26.4	0.8708	0.7637	0.8137	0.7168
U-Net3 +	27.0	153.1	11.3	0.8762	0.7832	0.8271	0.7326
SegFormer-MiT-B5	82.0	55.7	14.8	**0.8816**	0.7903	0.8335	0.7421
Mask2Former-R50	43.8	226.1	18.2	0.8724	0.7769	0.8219	0.7289
U-Net-SR (Ours)	31.3	41.9	22.6	0.8782	**0.8054**	**0.8402**	**0.7502**

Table 5
Ablation study results for the improved CasMVSNet.

CBAM	ODConv	Params (M)	VRAM (GB)	Runtime (s)	Acc. (mm)↓	Comp. (mm)↓	Overall (mm)↓
×	×	0.93	5.3	0.494	**0.325**	0.385	0.355
√	×	0.94	5.5	0.455	0.421	0.324	0.373
×	√	1.06	6.0	0.511	0.338	0.372	0.355
√	√	1.06	6.1	0.474	0.373	**0.315**	**0.344**

Note. All runtimes are measured as the average time per image over the entire test set, including data loading to GPU memory, pre-processing, model inference, and depth map post-processing.

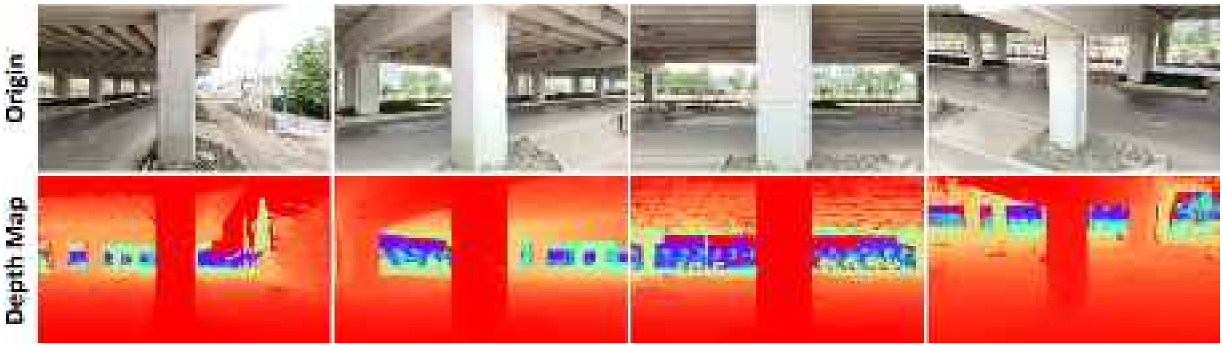


Fig. 12. Depth maps generated by the improved CasMVSNet.

Table 6
Performance comparison of different DL-based models on the DTU dataset.

Method	Depth map resolution (px)	Params (M)	VRAM (GB)	Runtime (s)	Acc. (mm)↓	Comp. (mm)↓	Overall (mm)↓
MVSNet [102]	640 × 512	0.68	11.3	1.207	0.469	0.627	0.548
CasMVSNet [97]	1152 × 864	0.93	5.3	0.494	**0.343**	0.407	0.375
Improved CasMVSNet (Ours)	1152 × 864	1.06	6.1	0.474	0.373	**0.315**	**0.344**

Note. All runtimes are measured as the average time per image over the entire test set, including data loading to GPU memory, pre-processing, model inference, and depth map post-processing.

4.6. 3D crack mapping

To quantitatively validate the improvement in 3D crack localization accuracy, three representative cracks are selected from the high-resolution bridge pier image dataset and the localization and segmentation quality are compared after projection onto the 3D mesh. The evaluation metrics include F1-score, Precision, Recall, and mIoU, computed by comparing the projected crack masks against ground-truth annotations in the 3D space. Four methods are benchmarked: Segformer, U-Net-SR, YOLOv11-CSD + SegFormer, and the proposed YOLOv11-CSD + U-Net-SR.

The results, summarized in Table 7, demonstrate that the hybrid approaches significantly outperform standalone segmentation models. For instance, the proposed YOLOv11-CSD + U-Net-SR achieves the highest average P_{pixel} (0.7762), R_{pixel} (0.9607), F_{pixel} (0.8584), and mIoU (0.7601) across the three cracks, indicating improved localization accuracy and completeness in the 3D domain. This enhancement is attributable to the initial ROI extraction by YOLOv11-CSD, which reduces background noise and focuses segmentation efforts on relevant regions, leading to more precise crack boundaries during 3D projection. Compared to standalone U-Net-SR, the hybrid method improves mIoU by an average of 5.3 %, confirming the second research objective of enhancing 3D crack localization through the integrated pipeline. Fig. 15 visualizes the 3D mapping results for these cracks segmented by the different models, further illustrating the superior alignment and detail preservation in the proposed approach.

5. Discussion

The proposed approach demonstrates a well-integrated and highly

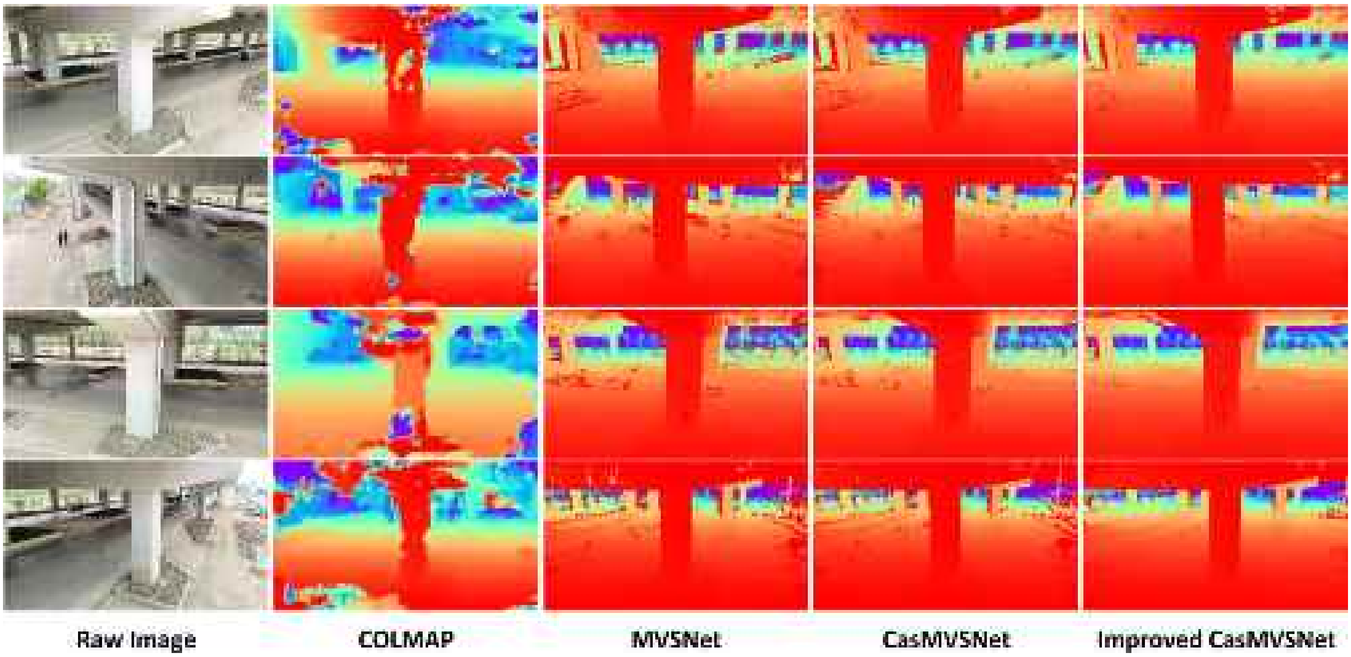


Fig. 13. Depth maps generated by different methods.

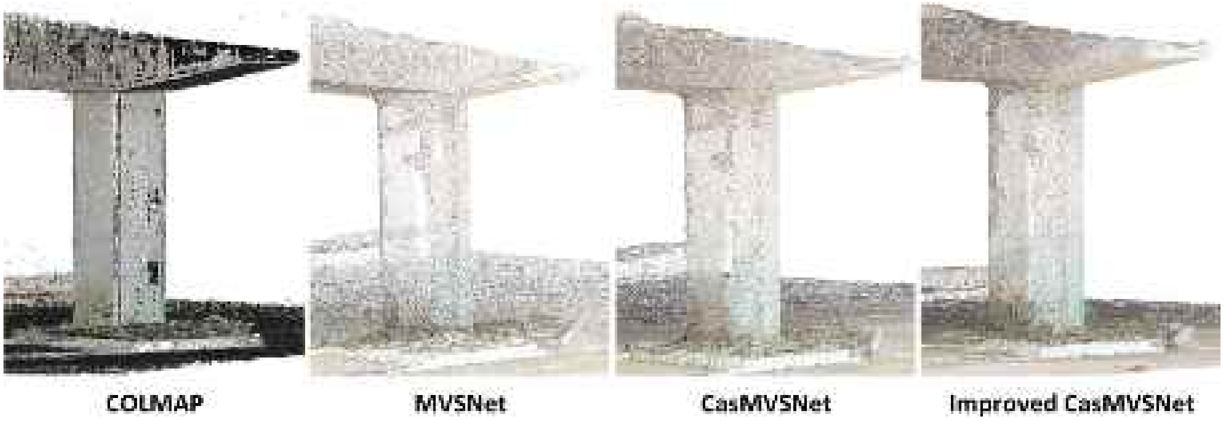


Fig. 14. Dense point clouds reconstructed using different methods.

Table 7
Quantitative evaluation of 3D crack mapping performance for different methods.

Method	Crack no.	P_{pixel}	R_{pixel}	F_{pixel}	mIoU
Segformer	1	0.7695	0.8649	0.8144	0.7046
	2	0.7036	**0.9877**	0.8217	0.6923
	3	0.6970	0.9803	0.8147	0.6861
U-Net-SR	1	0.7512	0.9529	0.8401	0.7299
	2	0.7368	0.9404	0.8262	0.7069
	3	0.7007	**0.9840**	0.8185	0.6905
Yolov11-CSD + SegFormer	1	0.7822	0.9570	0.8608	0.7717
	2	0.7441	0.9835	0.8472	0.7442
	3	0.7476	0.9709	0.8447	0.7458
Yolov11-CSD + U-Net-SR	1	**0.8064**	**0.9493**	**0.8720**	**0.7755**
	2	**0.7708**	0.9653	**0.8572**	**0.7561**
	3	**0.7515**	0.9674	**0.8459**	**0.7487**

effective pipeline for automated 3D reconstruction and crack mapping of bridge piers using UAV-acquired imagery.

This work innovates by synergistically integrating and optimizing established modules (e.g., SPD-Conv, CAM, DySample in YOLOv11-CSD; SE blocks in U-Net-SR; CBAM and ODConv in CasMVSNet) tailored for bridge crack detection—a domain characterized by small, low-contrast features on complex surfaces. This novel combination delivers enhanced small-object sensitivity (e.g., mAP@0.5 of 0.8196 in localization from Table 2, mIoU of 0.7502 in segmentation from Table 4) and geometric fidelity (e.g., an overall score of 0.344 from Table 6), thereby providing a scalable framework for structural health monitoring.

In the localization stage, YOLOv11-CSD, enhanced with SPD-Conv, CAM, and DySample, demonstrates superior performance in detecting small and low-contrast cracks. For segmentation, U-Net-SR incorporates squeeze-and-excitation blocks and residual connections to boost sensitivity to subtle features and training stability, enabling precise delineation of fine boundaries amid noise or textured backgrounds. This two-stage design minimizes computational overhead while sustaining high accuracy, ideal for processing high-resolution UAV data.

In 3D reconstruction, the improved CasMVSNet integrates CBAM and ODConv into feature extraction, markedly improving geometric fidelity and depth estimation. Paired with OpenMVS for dense reconstruction, the pipeline yields detailed, topologically consistent 3D mesh models with embedded crack semantics, facilitating advanced structural

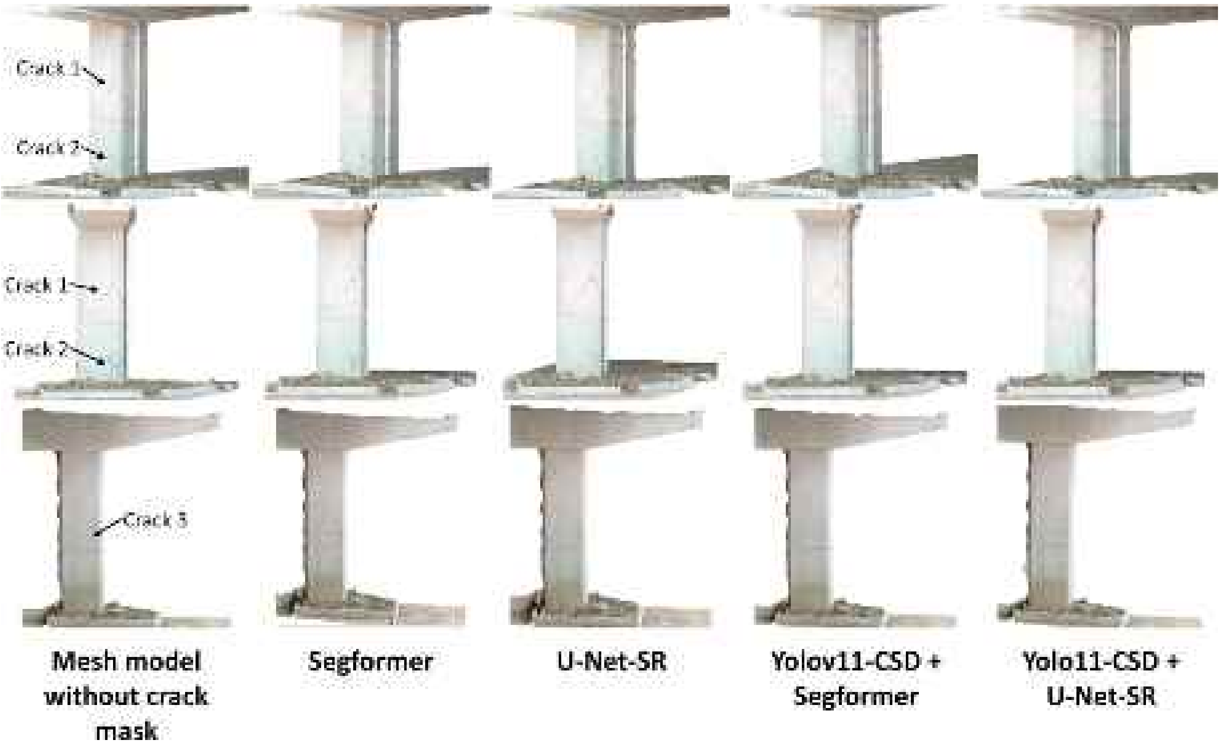


Fig. 15. 3D mapping of crack segmented by different models.

analysis.

A key strength of the proposed framework lies in its full-cycle automation, from UAV data acquisition to semantic 3D crack mapping, underscoring its practical applicability for large-scale infrastructure inspection. The visual and quantitative results demonstrate the pipeline's effectiveness in generating accurate, high-resolution models that retain both geometric and damage information, which is critical for informed maintenance planning and long-term structural health monitoring.

While the proposed approach demonstrates strong performance under daylight conditions, its effectiveness under low-light or nighttime scenarios remains limited. UAV-acquired imagery captured at night often suffers from increased noise, reduced contrast, and uneven illumination, which can obscure fine cracks and degrade both localization and segmentation accuracy. Although the current models are robust against common visual noise, their performance has not been extensively validated in nocturnal or poorly lit conditions. Future research should investigate integrating low-light enhancement techniques, such as deep learning-based denoising, or multi-spectral/infrared imaging. Adaptive exposure control during UAV data acquisition could also be explored to ensure consistent image quality across varying lighting environments.

6. Conclusion

This study presents an improved hybrid approach for pixel-level bridge crack detection and 3D localization. The proposed pipeline integrates crack detection, segmentation, depth estimation, and dense reconstruction into a unified framework tailored for UAV-acquired imagery. The YOLO-CSD model accurately extracts the crack ROIs, achieving a P_{obj} of 0.9001, R_{obj} of 0.7247, F_{obj} of 0.8055, mAP@0.5 of 0.8196, and mAP@0.5:0.95 of 0.5936. By filtering out irrelevant background information, it significantly reduces the computational burden of downstream semantic segmentation. Next, the U-Net-SR model demonstrates superior performance in segmenting fine cracks, achieving a P_{pixel} of 0.8782, R_{pixel} of 0.8054, F_{pixel} of 0.8402, and mIoU of 0.7502. An improved CasMVSNet architecture is introduced to generate high-quality depth maps from UAV images, yielding an accuracy of 0.373, completeness of 0.315, and overall error of 0.344. Finally, OpenMVS is employed to reconstruct detailed 3D mesh models with embedded crack semantic information. Quantitative evaluation of 3D crack mapping on representative cracks demonstrates that the hybrid YOLOv11-CSD + U-Net-SR method achieves the highest average P_{pixel} of 0.7762, R_{pixel} of 0.9607, F_{pixel} of 0.8584, and mIoU of 0.7601. Experimental validation confirms the effectiveness of the proposed approach in delivering precise, high-resolution crack mapping for structural health monitoring.

Overall, this approach shows strong potential to enhance automated bridge inspection workflows by reducing manual intervention and improving the reliability of condition assessments for aging infrastructure. Future work will focus on improving robustness under challenging environmental conditions and extending the framework to detect additional forms of structural deterioration beyond surface cracking.

CRediT authorship contribution statement

Yang Fang: Writing – review & editing, Writing – original draft, Validation, 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.

References

[1] Bridges, ASCE's 2025 Infrastructure Report Card | (2017). ⟨https://infrastructurereportcard.org/cat-item/bridges-infrastructure/⟩ (accessed June 18, 2025).

[2] Zhou X, Zhang X. Thoughts on the development of bridge technology in China. Engineering 2019;5:1120–30. https://doi.org/10.1016/j.eng.2019.10.001.

[3] Sun L, Shang Z, Xia Y, Bhowmick S, Nagarajaiah S. Review of bridge structural health monitoring aided by big data and artificial intelligence: from condition assessment to damage detection. J Struct Eng 2020;146:04020073. https://doi.org/10.1061/(ASCE)ST.1943-541X.0002535.

[4] Falowo OO, Makinde O, Akindureni Y. Degree of corrosion and other deterioration effects in highway bridge deck component using non-destructive self potential and vibration testing: case study from SW Nigeria. Discov Civ Eng 2025;2:8. https://doi.org/10.1007/s44290-025-00165-4.

[5] Koch C, Georgieva K, Kasireddy V, Akinci B, Fieguth P. A review on computer vision based defect detection and condition assessment of concrete and asphalt civil infrastructure. Adv Eng Inform 2015;29:196–210. https://doi.org/10.1016/j.aei.2015.01.008.

[6] Hüthwohl P, Lu R, Brilakis I. Multi-classifier for reinforced concrete bridge defects. Autom Constr 2019;105:102824. https://doi.org/10.1016/j.autcon.2019.04.019.

[7] Jeong E, Seo J, Wacker JP. UAV-aided bridge inspection protocol through machine learning with improved visibility images. Expert Syst Appl 2022;197:116791. https://doi.org/10.1016/j.eswa.2022.116791.

[8] Li R, Yu J, Li F, Yang R, Wang Y, Peng Z. Automatic bridge crack detection using Unmanned aerial vehicle and Faster R-CNN. Constr Build Mater 2023;362:129659. https://doi.org/10.1016/j.conbuildmat.2022.129659.

[9] Tran TS, Nguyen SD, Lee HJ, Tran VP. Advanced crack detection and segmentation on bridge decks using deep learning. Constr Build Mater 2023;400:132839. https://doi.org/10.1016/j.conbuildmat.2023.132839.

[10] Ding W, Yang H, Yu K, Shu J. Crack detection and quantification for concrete structures using UAV and transformer. Autom Constr 2023;152:104929. https://doi.org/10.1016/j.autcon.2023.104929.

[11] Liu Y-F, Nie X, Fan J-S, Liu X-G. Image-based crack assessment of bridge piers using unmanned aerial vehicles and three-dimensional scene reconstruction. ComputAided Civ Infrastruct Eng 2020;35:511–29. https://doi.org/10.1111/mice.12501.

[12] Feng C-Q, Li B-L, Liu Y-F, Zhang F, Yue Y, Fan J-S. Crack assessment using multi-sensor fusion simultaneous localization and mapping (SLAM) and image super-resolution for bridge inspection. Autom Constr 2023;155:105047. https://doi.org/10.1016/j.autcon.2023.105047.

[13] Qi Y, Lin P, Yang G, Liang T. Crack detection and 3D visualization of crack distribution for UAV-based bridge inspection using efficient approaches. Structures 2025;78:109075. https://doi.org/10.1016/j.istruc.2025.109075.

[14] Li Q, Liu X. Novel approach to pavement image segmentation based on neighboring difference histogram method. 2008 Congr Image Signal Process 2008:792–6. https://doi.org/10.1109/CISP.2008.13.

[15] Lakshmikantha MR, Prat PC, Ledesma A. Image Analysis for thE Quantification of A Developing Crack Network on A Drying Soil. Geotech Test J 2009;32:505–15. https://doi.org/10.1520/GTJ102216.

[16] Liu C, Tang C-S, Shi B, Suo W-B. Automatic quantification of crack patterns by image processing. Comput Geosci 2013;57:77–80. https://doi.org/10.1016/j.cageo.2013.04.008.

[17] Liang S, Jianchun X, Xun Z. An algorithm for concrete crack extraction and identification based on machine vision. IEEE Access 2018;6:28993–9002. https://doi.org/10.1109/ACCESS.2018.2844100.

[18] Subirats P, Dumoulin J, Legeay V, Barba D. Automation of pavement surface crack detection using the continuous wavelet transform. 2006 Int Conf Image Process 2006:3037–40. https://doi.org/10.1109/ICIP.2006.313007.

[19] Zhou J, Huang PS, Chiang F-P. Wavelet-based pavement distress detection and evaluation. OE 2006;45:027007. https://doi.org/10.1117/1.2172917.

[20] Wang KCP, Li Q, Gong W. Wavelet-Based Pavement Distress Image Edge Detection with À Trous Algorithm. Transp Res Rec 2007;2024:73–81. https://doi.org/10.3141/2024-09.

[21] Ayenu-Prah A, Attoh-Okine N. Evaluating Pavement Cracks with Bidimensional Empirical Mode Decomposition. EURASIP J Adv Signal Process 2008;2008:861701. https://doi.org/10.1155/2008/861701.

[22] Talab AMA, Huang Z, Xi F, HaiMing L. Detection crack in image using Otsu method and multiple filtering in image processing techniques. Optik 2016;127:1030–3. https://doi.org/10.1016/j.ijleo.2015.09.147.

[23] Li Q, Zou Q, Zhang D, Mao Q. FoSA: F* Seed-growing Approach for crack-line detection from pavement images. Image Vis Comput 2011;29:861–72. https://doi.org/10.1016/j.imavis.2011.10.003.

[24] Kaul V, Yezzi A, Tsai Y. Detecting Curves with Unknown Endpoints and Arbitrary Topology Using Minimal Paths. IEEE Trans Pattern Anal Mach Intell 2012;34:1952–65. https://doi.org/10.1109/TPAMI.2011.267.

[25] Amhaz R, Chambon S, Idier J, Baltazart V. Automatic Crack Detection on Two-Dimensional Pavement Images: An Algorithm Based on Minimal Path Selection. IEEE Trans Intell Transp Syst 2016;17:2718–29. https://doi.org/10.1109/TITS.2015.2477675.

[26] Zou Q, Li Q, Zhang F, Xiong Qian Wang Z, Wang Q. Path voting based pavement crack detection from laser range images. In: 2016 IEEE International Conference on Digital Signal Processing (DSP). Beijing, China: IEEE; 2016. p. 432–6. https://doi.org/10.1109/ICDSP.2016.7868594.

[27] Ouma YO, Hahn M. Wavelet-morphology based detection of incipient linear cracks in asphalt pavements from RGB camera imagery and classification using circular Radon transform. Adv Eng Inform 2016;30:481–99. https://doi.org/10.1016/j.aei.2016.06.003.

[28] Shi Y, Cui L, Qi Z, Meng F, Chen Z. Automatic Road Crack Detection Using Random Structured Forests. IEEE Trans Intell Transp Syst 2016;17:3434–45. https://doi.org/10.1109/TITS.2016.2552248.

[29] Pattnaik T, Kanungo P, Kar T, Sahoo PK. Multiple linear regression based illumination normalization for non-uniform light image thresholding. EPrime Adv Electr Eng Electron Energy 2024;7:100411. https://doi.org/10.1016/j.prime.2023.100411.

[30] Chen F-C, Jahanshahi MR, Wu R-T, Joffe C. A texture-Based Video Processing Methodology Using Bayesian Data Fusion for Autonomous Crack Detection on Metallic Surfaces. ComputAided Civ Infrastruct Eng 2017;32:271–87. https://doi.org/10.1111/mice.12256.

[31] Ai D, Jiang G, Siew Kei L, Li C. Automatic pixel-level pavement crack detection using information of multi-scale neighborhoods. IEEE Access 2018;6:24452–63. https://doi.org/10.1109/ACCESS.2018.2829347.

[32] Wang S, Liu X, Yang T, Wu X. Panoramic crack detection for steel beam based on structured random forests. IEEE Access 2018;6:16432–44. https://doi.org/10.1109/ACCESS.2018.2812141.

[33] Rodriguez-Lozano FJ, León-García F, Gámez-Granados JC, Palomares JM, Olivares J. Benefits of ensemble models in road pavement cracking classification. ComputAided Civ Infrastruct Eng 2020;35:1194–208. https://doi.org/10.1111/mice.12543.

[34] Peng X, Zhong X, Zhao C, Chen A, Zhang T. A UAV-based machine vision method for bridge crack recognition and width quantification through hybrid feature learning. Constr Build Mater 2021;299:123896. https://doi.org/10.1016/j.conbuildmat.2021.123896.

[35] Abdel-Qader I, Pashaie-Rad S, Abudayyeh O, Yehia S. PCA-Based algorithm for unsupervised bridge crack detection. Adv Eng Softw 2006;37:771–8. https://doi.org/10.1016/j.advengsoft.2006.06.002.

[36] Wang H, Xiong Z, Finn AM, Chaudhry Z. A context-driven approach to image-based crack detection. Mach Vis Appl 2016;27:1103–14. https://doi.org/10.1007/s00138-016-0779-1.

[37] Noh Y, Koo D, Kang Y-M, Park D, Lee D. Automatic crack detection on concrete images using segmentation via fuzzy C-means clustering. 2017 Int Conf Appl Syst Innov (ICASI) 2017:877–80. https://doi.org/10.1109/ICASI.2017.7988574.

[38] Fan X, Wu J, Shi P, Zhang X, Xie Y. A novel automatic dam crack detection algorithm based on local-global clustering. Multimed Tools Appl 2018;77:26581–99. https://doi.org/10.1007/s11042-018-5880-1.

[39] Zhang A, Wang KCP, Li B, Yang E, Dai X, Peng Y, Fei Y, Liu Y, Li JQ, Chen C. Automated pixel-level pavement crack detection on 3D asphalt surfaces using a deep-learning network. ComputAided Civ Infrastruct Eng 2017;32:805–19. https://doi.org/10.1111/mice.12297.

[40] Cha Y-J, Choi W, Büyüköztürk O. Deep Learning-Based Crack Damage Detection Using Convolutional Neural Networks. ComputAided Civ Infrastruct Eng 2017;32:361–78. https://doi.org/10.1111/mice.12263.

[41] Zhou S, Song W. Deep learning-based roadway crack classification using laser-scanned range images: A comparative study on hyperparameter selection. Autom Constr 2020;114:103171. https://doi.org/10.1016/j.autcon.2020.103171.

[42] Li C, Xu P, Niu L, Chen Y, Sheng L, Liu M. Tunnel crack detection using coarse-to-fine region localization and edge detection. WIREs Data Min Knowl Discov 2019;9:e1308. https://doi.org/10.1002/widm.1308.

[43] Majidifard H, Adu-Gyamfi Y, Buttlar WG. Deep machine learning approach to develop a new asphalt pavement condition index. Constr Build Mater 2020;247:118513. https://doi.org/10.1016/j.conbuildmat.2020.118513.

[44] Zhou Z, Zhang J, Gong C. Automatic detection method of tunnel lining multi-defects via an enhanced You Only Look Once network. ComputAided Civ Infrastruct Eng 2022;37:762–80. https://doi.org/10.1111/mice.12836.

[45] Qu Z, Wang C-Y, Wang S-Y, Ju F-R. A method of hierarchical feature fusion and connected attention architecture for pavement crack detection. IEEE Trans Intell Transp Syst 2022;23:16038–47. https://doi.org/10.1109/TITS.2022.3147669.

[46] Guo J-M, Markoni H, Lee J-D. BARNet: boundary aware refinement network for crack detection. IEEE Trans Intell Transp Syst 2022;23:7343–58. https://doi.org/10.1109/TITS.2021.3069135.

[47] Ye X-W, Jin T, Chen P-Y. Structural crack detection using deep learning–based fully convolutional networks. Adv Struct Eng 2019;22:3412–9. https://doi.org/10.1177/1369433219836292.

[48] Dung CV, Anh LD. Autonomous concrete crack detection using deep fully convolutional neural network. Autom Constr 2019;99:52–8. https://doi.org/10.1016/j.autcon.2018.11.028.

[49] Liu Y, Yao J, Lu X, Xie R, Li L. DeepCrack: A deep hierarchical feature learning architecture for crack segmentation. Neurocomputing 2019;338:139–53. https://doi.org/10.1016/j.neucom.2019.01.036.

[50] Zhang K, Zhang Y, Cheng H-D. CrackGAN: Pavement Crack Detection Using Partially Accurate Ground Truths Based on Generative Adversarial Learning. IEEE Trans Intell Transp Syst 2021;22:1306–19. https://doi.org/10.1109/TITS.2020.2990703.

[51] Schmugge SJ, Rice L, Lindberg J, Grizziy R, Joffey C, Shin MC. Crack Segmentation by Leveraging Multiple Frames of Varying Illumination. 2017 IEEE Winter Conf Appl Comput Vis (WACV) 2017:1045–53. https://doi.org/10.1109/WACV.2017.121.

[52] Zou Q, Zhang Z, Li Q, Qi X, Wang Q, Wang S. DeepCrack: learning hierarchical convolutional features for crack detection. IEEE Trans Image Process 2019;28:1498–512. https://doi.org/10.1109/TIP.2018.2878966.

[53] Bang S, Park S, Kim H, Kim H. Encoder–decoder network for pixel-level road crack detection in black-box images. ComputAided Civ Infrastruct Eng 2019;34:713–27. https://doi.org/10.1111/mice.12440.

[54] Li S, Zhao X, Zhou G. Automatic pixel-level multiple damage detection of concrete structure using fully convolutional network. ComputAided Civ Infrastruct Eng 2019;34:616–34. https://doi.org/10.1111/mice.12433.

[55] Yamanakkanavar N, Lee B. A novel M-SegNet with global attention CNN architecture for automatic segmentation of brain MRI. Comput Biol Med 2021;136:104761. https://doi.org/10.1016/j.compbiomed.2021.104761.

[56] Zhang J, Lu C, Wang J, Wang L, Yue X-G. Concrete Cracks Detection Based on FCN with Dilated Convolution. Appl Sci 2019;9:2686. https://doi.org/10.3390/app9132686.

[57] Jiang W, Liu M, Peng Y, Wu L, Wang Y. HDCB-Net: A Neural Network With the Hybrid Dilated Convolution for Pixel-Level Crack Detection on Concrete Bridges. IEEE Trans Ind Inform 2021;17:5485–94. https://doi.org/10.1109/TII.2020.3033170.

[58] Ye W, Deng S, Ren J, Xu X, Zhang K, Du W. Deep learning-based fast detection of apparent concrete crack in slab tracks with dilated convolution. Constr Build Mater 2022;329:127157. https://doi.org/10.1016/j.conbuildmat.2022.127157.

[59] Ni F, Zhang J, Chen Z. Pixel-level crack delineation in images with convolutional feature fusion. Struct Control Health Monit 2019;26:e2286. https://doi.org/10.1002/stc.2286.

[60] Yang F, Zhang L, Yu S, Prokhorov D, Mei X, Ling H. Feature Pyramid and Hierarchical Boosting Network for Pavement Crack Detection. IEEE Trans Intell Transp Syst 2020;21:1525–35. https://doi.org/10.1109/TITS.2019.2910595.

[61] Ji A, Xue X, Wang Y, Luo X, Xue W. An integrated approach to automatic pixel-level crack detection and quantification of asphalt pavement. Autom Constr 2020;114:103176. https://doi.org/10.1016/j.autcon.2020.103176.

[62] Sun X, Xie Y, Jiang L, Cao Y, Liu B. DMA-Net: DeepLab With Multi-Scale Attention for Pavement Crack Segmentation. IEEE Trans Intell Transp Syst 2022;23:18392–403. https://doi.org/10.1109/TITS.2022.3158670.

[63] Wang Y, Yang L, Liu X, Yan P. An improved semantic segmentation algorithm for high-resolution remote sensing images based on DeepLabv3+. Sci Rep 2024;14:9716. https://doi.org/10.1038/s41598-024-60375-1.

[64] Pan Z, Zhang X, Jiang Y, Li B, Golsanami N, Su H, Cai Y. High-precision segmentation and quantification of tunnel lining crack using an improved DeepLabV3+. Undergr Space 2025;22:96–109. https://doi.org/10.1016/j.undsp.2024.10.002.

[65] Ai D, Jiang G, Lam S-K, He P, Li C. Computer vision framework for crack detection of civil infrastructure—A review. Eng Appl Artif Intell 2023;117:105478. https://doi.org/10.1016/j.engappai.2022.105478.

[66] Liu Z, Cao Y, Wang Y, Wang W. Computer vision-based concrete crack detection using U-net fully convolutional networks. Autom Constr 2019;104:129–39. https://doi.org/10.1016/j.autcon.2019.04.005.

[67] Huyan J, Li W, Tighe S, Xu Z, Zhai J. CrackU-net: A novel deep convolutional neural network for pixelwise pavement crack detection. Struct Control Health Monit 2020;27:e2551. https://doi.org/10.1002/stc.2551.

[68] Ai D, Jiang G, Lam S-K, He P, Li C. Automatic pixel-wise detection of evolving cracks on rock surface in video data. Autom Constr 2020;119:103378. https://doi.org/10.1016/j.autcon.2020.103378.

[69] Shang J, Xu J, Zhang AA, Liu Y, Wang KCP, Ren D, Zhang H, Dong Z, He A. Automatic Pixel-level pavement sealed crack detection using Multi-fusion U-Net network. Measurement 2023;208:112475. https://doi.org/10.1016/j.measurement.2023.112475.

[70] Lu W, Qian M, Xia Y, Lu Y, Shen J, Fu Q, Lu Y. Crack _ PSTU: Crack detection based on the U-Net framework combined with Swin Transformer. Structures 2024;62:106241. https://doi.org/10.1016/j.istruc.2024.106241.

[71] Asadi Shamsabadi E, Xu C, Rao AS, Nguyen T, Ngo T, Dias-da-Costa D. Vision transformer-based autonomous crack detection on asphalt and concrete surfaces. Autom Constr 2022;140:104316. https://doi.org/10.1016/j.autcon.2022.104316.

[72] Wang Z, Leng Z, Zhang Z. A weakly-supervised transformer-based hybrid network with multi-attention for pavement crack detection. Constr Build Mater 2024;411:134134. https://doi.org/10.1016/j.conbuildmat.2023.134134.

[73] Su G, Qin Y, Xu H, Liang J. Automatic real-time crack detection using lightweight deep learning models. Eng Appl Artif Intell 2024;138:109340. https://doi.org/10.1016/j.engappai.2024.109340.

[74] Ge K, Wang C, Guo YT, Tang YS, Hu ZZ, Chen HB. Fine-tuning vision foundation model for crack segmentation in civil infrastructures. Constr Build Mater 2024;431:136573. https://doi.org/10.1016/j.conbuildmat.2024.136573.

[75] Ye G, Dai W, Tao J, Qu J, Zhu L, Jin Q. An improved transformer-based concrete crack classification method. Sci Rep 2024;14:6226. https://doi.org/10.1038/s41598-024-54835-x.

[76] Shi T, Luo H. Deep learning for automated detection and classification of crack severity level in concrete structures. Constr Build Mater 2025;472:140793. https://doi.org/10.1016/j.conbuildmat.2025.140793.

[77] Han C, Yang H, Ma T, Wang S, Zhao C, Yang Y. CrackDiffusion: a two-stage semantic segmentation framework for pavement crack combining unsupervised and supervised processes. Autom Constr 2024;160:105332. https://doi.org/10.1016/j.autcon.2024.105332.

[78] Lei Q, Zhong J, Dong M, Ota K. Faithful crack image synthesis from evolutionary pixel-level annotations via latent semantic diffusion model. Expert Syst Appl 2025;275:126986. https://doi.org/10.1016/j.eswa.2025.126986.

[79] Lin C, Liu R, Lin W, Zou Y, Wei X, Su Y. Underwater dam crack image enhancement and crack detection based on improved diffusion model and SDI-ASF-YOLO11. Constr Build Mater 2025;492:142861. https://doi.org/10.1016/j.conbuildmat.2025.142861.

[80] Qin S, Fang Y, Li Y, Li Z. Generation of Infrastructure Crack Images for Self-Supervision Training Based on Diffusion Model. IEEE Trans Intell Transp Syst 2025:1–17. https://doi.org/10.1109/TITS.2025.3597284.

[81] Kim H, Sim S-H, Spencer BF. Automated concrete crack evaluation using stereo vision with two different focal lengths. Autom Constr 2022;135:104136. https://doi.org/10.1016/j.autcon.2022.104136.

[82] Deng L, Sun T, Yang L, Cao R. Binocular video-based 3D reconstruction and length quantification of cracks in concrete structures. Autom Constr 2023;148:104743. https://doi.org/10.1016/j.autcon.2023.104743.

[83] Kalfarisi R, Wu ZY, Soh K. Crack Detection and Segmentation Using Deep Learning with 3D Reality Mesh Model for Quantitative Assessment and Integrated Visualization. J Comput Civ Eng 2020;34:04020010. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000890.

[84] Lin JJ, Ibrahim A, Sarwade S, Golparvar-Fard M. Bridge Inspection with Aerial Robots: Automating the Entire Pipeline of Visual Data Capture, 3D Mapping, Defect Detection, Analysis, and Reporting. J Comput Civ Eng 2021;35:04020064. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000954.

[85] Ni Y, Mao J, Wang H, Xi Z, Chen Z. Surface damage detection and localization for bridge visual inspection based on deep learning and 3D reconstruction. Struct Control Health Monit 2024;2024:9988793. https://doi.org/10.1155/2024/9988793.

[86] Yu Z, Shen Y, Zhang Y, Xiang Y. Automatic crack detection and 3D reconstruction of structural appearance using underwater wall-climbing robot. Autom Constr 2024;160:105322. https://doi.org/10.1016/j.autcon.2024.105322.

[87] Duan D, Wang Z, Xin Y, Ding Y. Defect segmentation and 3D reconstruction in concrete structures using SAM 2 and 3D Gaussian splatting. J Civ Struct Health Monit 2025. https://doi.org/10.1007/s13349-025-00993-z.

[88] Sunkara R, Luo T. No More Strided Convolutions or Pooling: A New CNN Building Block for Low-Resolution Images and Small Objects. In: Amini M-R, Canu S, Fischer A, Guns T, Kralj Novak P, Tsoumakas G, editors. Machine Learning and Knowledge Discovery in Databases. Cham: Springer Nature Switzerland; 2023. p. 443–59. https://doi.org/10.1007/978-3-031-26409-2_27.

[89] C. 1 Zou, S. 2 Yu, Y. 2 Yu, H. 2 Gu, X. 3 1 S.K.L. of R. Xu, zouchang@sia.cn (C.Z.), C.A. of S. Shenyang Institute of Automation, B. 100049 University of Chinese Academy of Sciences, zouchang@sia.cn (C.Z.), C.A. of S. Shenyang Institute of Automation, I. of D.-S. Science, C.A. of S. Engineering, Side-Scan Sonar Small Objects Detection Based on Improved YOLOv11, (2025). ⟨https://doi.org/10.3390/jmse13010162⟩.

[90] Wang C, Han Y, Yang C, Wu M, Chen Z, Yun L, Jin X. CF-YOLO for small target detection in drone imagery based on YOLOv11 algorithm. Sci Rep 2025;15:16741. https://doi.org/10.1038/s41598-025-99634-0.

[91] J. Xiao, T. Zhao, Y. Yao, Q. Yu, Y. Chen, CONTEXT AUGMENTATION AND FEATURE REFINEMENT NETWORK FOR TINY OBJECT DETECTION, in: 2021. ⟨https://openreview.net/forum?id=q2ZaVU6bEsT⟩ (accessed June 9, 2025).

[92] Wu C, Zhang S, Wang W, Wu Z, Yang S, Chen W. Computation and analysis of phenotypic parameters of Scylla paramamosain based on YOLOv11-DYPF keypoint detection. Aquac Eng 2025;111:102571. https://doi.org/10.1016/j.aquaeng.2025.102571.

[93] Liu W., Lu H., Fu H., Cao Z. 2023. (accessed June 10, 2025). Learning to Upsample by Learning to Sample, in: 2023: pp. 6027–6037. https://openaccess.thecvf.com/content/ICCV2023/html/Liu_Learning_to_Upsample_by_Learning_to_Sample_ICCV_2023_paper.html (accessed June 10, 2025)..

[94] Zou L, Chen A, Yang X, Sun Y. An improved method of AUD-YOLO for surface damage detection of wind turbine blades. Sci Rep 2025;15:5833. https://doi.org/10.1038/s41598-025-89864-7.

[95] Hu J., Shen L., Albanie S., Sun G., Wu E. 2019. Squeeze-and-Excitation Networks, (2019). https://doi.org/10.48550/arXiv.1709.01507.

[96] He K., Zhang X., Ren S., Sun J. 2015. Deep Residual Learning for Image Recognition, (2015). https://doi.org/10.48550/arXiv.1512.03385.

[97] Gu X., Fan Z., Zhu S., Dai Z., Tan F., Tan P. 2020. Cascade Cost Volume for High-Resolution Multi-View Stereo and Stereo Matching, in: 2020: pp. 2495–2504. https://openaccess.thecvf.com/content_CVPR_2020/html/Gu_Cascade_Cost_Volume_for_High-Resolution_Multi-View_Stereo_and_Stereo_Matching_CVPR_2020_paper.html (accessed June 16, 2025)..

[98] C. Li, A. Zhou, A. Yao, OMNI-DIMENSIONAL DYNAMIC CONVOLUTION, (2022).

[99] Woo S, Park J, Lee J-Y, Kweon IS. CBAM: Convolutional Block Attention Module. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y, editors. Computer Vision – ECCV 2018. Cham: Springer International Publishing; 2018. p. 3–19. https://doi.org/10.1007/978-3-030-01234-2_1.

[100] Aanæs H, Jensen RR, Vogiatzis G, Tola E, Dahl AB. Large-scale data for multiple-view stereopsis. Int J Comput Vis 2016;120:153–68. https://doi.org/10.1007/s11263-016-0902-9.

[101] Concrete Crack > Browse, Roboflow (n.d.). ⟨https://universe.roboflow.com/yolo-3oia3/concrete-crack-6loqq-sed2c⟩ (accessed June 15, 2025).

[102] Yao Y, Luo Z, Li S, Fang T, Quan L. MVSNet: depth inference for unstructured multi-view stereo. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y, editors. Computer Vision – ECCV 2018. Cham: Springer International Publishing; 2018. p. 785–801. https://doi.org/10.1007/978-3-030-01237-3_47.