Measurement 258 (2026) 119453

Contents lists available at ScienceDirect

Measurement

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

Depth-aware RGB-D concrete crack segmentation and quantification using progressive cross-modal attention

Yingjie Wu [a,b], Shaoqi Li [a,*], Yancheng Li [b,*]

[a] College of Civil Engineering, Nanjing Tech University, Nanjing 211816, China
[b] School of Civil and Environmental Engineering, University of Technology Sydney, Ultimo, NSW 2007, Australia

ARTICLE INFO

Keywords:
Crack segmentation
Crack quantification
Future fusion
Cross-modal
Deep learning

ABSTRACT

Cracks in infrastructures, such as concrete structures and pavements, pose significant risks to structural safety and durability. The development of crack geometry provides critical information of structural reliability hence there is need, recommended by standards, to precisely quantify the crack details, such as crack length, width or depth. Although deep learning has inspired automated crack detection, its capacities in profiling the crack geometry is still in doubt since most methods that rely solely on RGB images face challenges in field conditions with low contrast, surface contamination, and complex textures. Such conditions often result in blurred boundaries and unreliable geometric measurements, limiting their applicability in practice. To address these challenges, this study proposes a progressive Cross-Modal Fusion Transformer (CMF-Former) that integrates RGB and depth modalities through hierarchical representation and adaptive feature interaction. The network separately models RGB and depth representations to retain modality-specific features, and introduces a progressive cross-modal attention mechanism to adaptively fuse complementary information across semantic stages. A multi-scale decoder is used to further facilitate accurate crack localization and restoration. Additionally, a depth-assisted quantification method is developed by leveraging depth information to automatically estimate distance and spatial scale, enabling direct measurement of crack geometric features. Experimental results show that CMF-Former achieves a highest mIoU of 86.51%, outperforming other RGB-based and RGB-D based models. In addition to segmentation performance, the proposed RGB-D framework notably enhances geometric quantification. For crack width estimation, the proposed method achieved an average Root Mean Square Error (RMSE) of 1.167, representing a substantial improvement compared to other RGB-based methods. Moreover, the relative error rates for crack length and depth estimation are 2.19% and 6.188%, respectively, demonstrating improved accuracy in capturing crack morphology.

1. Introduction

Cracks are among the most common forms of structural damage in infrastructure, and can significantly compromise load-bearing capacity with induced deterioration, resulting in serviceability loss, structural failure, or even collapse [1–4]. Therefore, accurate detection and quantification of crack geometry, such as length, width, and depth, are essential for structural assessment and maintenance planning [5–8]. However, traditional inspection methods rely on manual procedures, which are time-consuming, labor-intensive, and prone to subjective errors [9–11].

Recent advancements in deep learning (DL) have substantially improved automated crack detection and quantification. In particular, segmentation-based approaches have gained increasing attention due to their ability to localize cracks at the pixel level [12–15]. In most existing workflows, crack regions are segmented from RGB images, and geometric features are subsequently estimated through manual pixel-to-metric conversions based on reference objects or known scene dimensions. A variety of segmentation models have been adapted to support this process. Convolutional neural networks (CNNs), such as FCN [16], SegNet [17], U-Net [18], PSPNet [19], DeepLabV3+ [20], Fast-SCNN [21], and SegNeXt [22], have been widely adopted to extract hierarchical features through encoder–decoder architectures. More recently, Transformer-based networks such as ViT [23], PVT [24], Swin Transformer [25], and SegFormer [26] have demonstrated improved structural perception by capturing long-range dependencies and global

* Corresponding authors.
E-mail addresses: shaoqi@njtech.edu.cn (S. Li), yancheng.li@uts.edu.au (Y. Li).

https://doi.org/10.1016/j.measurement.2025.119453
Received 16 June 2025; Received in revised form 5 October 2025; Accepted 21 October 2025
Available online 24 October 2025

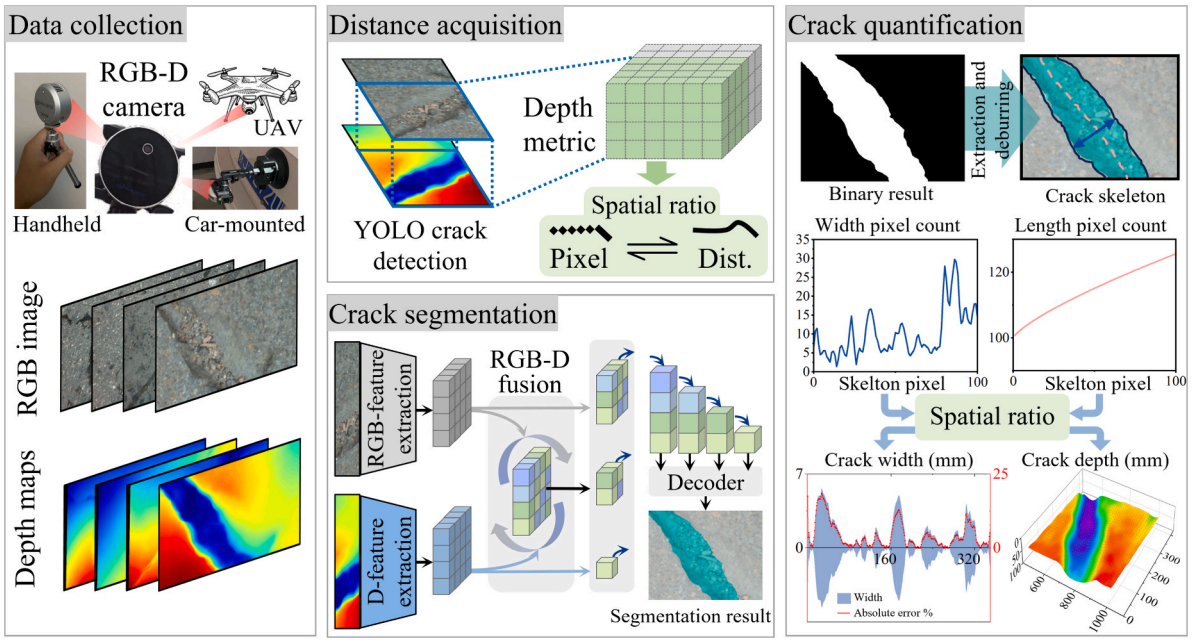

Fig. 1. Workflow of proposed method.

context.

While these approaches perform well in controlled environments, their reliance on RGB contexture feature can limit robustness under complex real-world conditions, reducing segmentation reliability and compromising the accuracy of downstream crack geometry quantification. Furthermore, the subsequent quantification process typically requires manual reference placement and is highly sensitive to imaging angles, which introduces human-induced errors and undermines the robustness and deployability of these methods in real-world scenarios [27,28].

With awareness of the limitations of RGB-based methods, several studies have investigated multi-modal fusion strategies for crack detection and quantification. One commonly explored approach involves combining RGB and thermal imagery to help distinguish cracks from complex backgrounds in low-contrast conditions [29–31]. However, thermal modality inherently lacks structural cues such as crack width variation and topology, which are critical for precise segmentation and meaningful quantification. In parallel, several studies have employed LiDAR-based methods to improve the precision and completeness of crack geometry measurement. For example, Zhang et al. [32] applied principal component analysis to 3D laser profiling data to identify crack-related points and estimated crack depth by computing their vertical distances to a fitted pavement surface. Pan et al. [33] reconstructed 3D pavement crack profiles, from which crack depth was obtained directly from the Z-coordinates of the point cloud. However, these methods rely heavily on raw geometric data capture by specialized equipment, without incorporating pixel-level segmentation and visual context, which can lead to inaccurate or unstable quantification. In addition, their complex workflows, expensive equipment, and intensive computation make them difficult to deploy in routine inspection scenarios.

Compared with thermal and LiDAR-based methods, RGB-D sensors simultaneously provide RGB features of surface appearance and depth features of geometric structure, enabling more comprehensive crack analysis. A few studies have also begun to explore RGB-D data for crack detection and quantification. For example, Kim et al. [34] used depth data to reconstruct a 3D pavement surface, applied RANSAC to fit a planar model, and mapped crack pixels onto this plane via coordinate transformation. Crack width was then measured by computing the 3D distance between edge pixels near the crack skeleton. Lin et al. [35] utilized the YOLACT++ algorithm to segment crack regions from RGB images and mapped the crack pixels into 3D space using depth data. A flat road surface was then fitted from the depth map, and crack width was measured by calculating the distance between crack edges projected onto this plane. However, these methods adopt two-stage frameworks that separates crack segmentation and geometric quantification. As the segmentation masks are typically generated from RGB-only inputs without incorporating depth information, they often lack geometric consistency, which in turn compromises the accuracy and reliability of subsequent crack geometry quantification.

The depth image provides precise information on surface discontinuity, such as pixel-wise elevation change, which is extremely valuable in tasks such as crack segmentation. Integration of the depth feature will certainly improve the existing RGB based crack detection/segmentation. However, few studies in the field have explored RGB-D fusion to jointly improve segmentation accuracy and geometric measurement. In the broader field of computer vision, several studies have attempted to leverage RGB-D fusion in tasks like indoor navigation and object recognition [36–38]. Although these works confirm the complementary nature of RGB and depth information, they are mostly designed for structured environments with distinct object boundaries and stable depth signals. In contrast, cracks often exhibit irregular shapes, low contrast, and appear within noisy backgrounds, which may limit the effectiveness of such designs when applied to this domain. Beyond general computer vision tasks, RGB-D fusion has also been explored in defect detection in industrial. For example, Wang et al. [39] proposed CLANet for RGB-D rail defect inspection, where RGB and depth features are fused at early stages using a multimodal attention block based on spatial refinement and feedback interaction. Zhang et al. [40] designed a dual-encoder RGB-D segmentation network for pore detection in bridge towers, where features from each stage of MobileNetV2 are extracted separately and fused at higher semantic levels by proposed FFEM modules. However, these methods rely on relatively rigid fusion strategies, where RGB and depth features are integrated at fixed stages with limited modality interaction.

As discussed above, although progress has been made with RGB-based, thermal, and LiDAR-assisted approaches, current methods still face the following challenges:

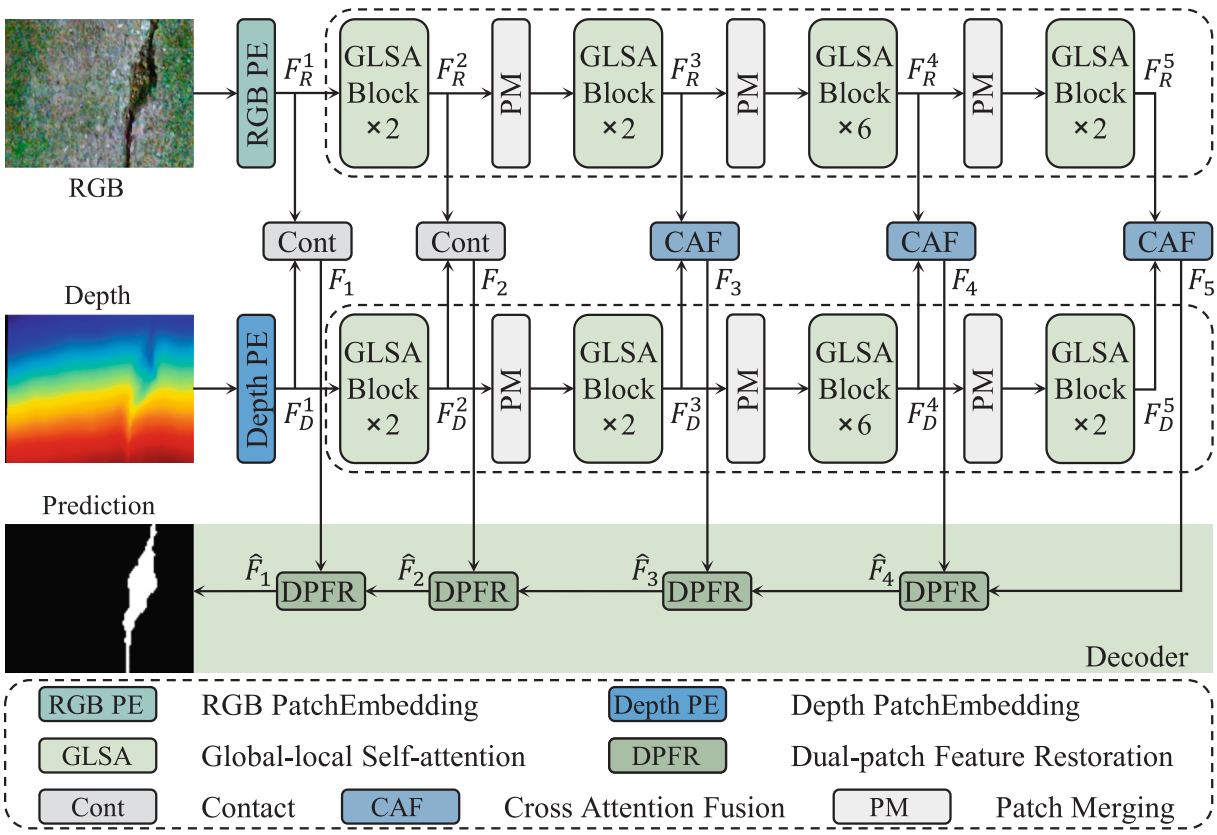

Fig. 2. Structure of CMF-Former ($\{F_R^i\}_{i=1}^5$ and $\{F_D^i\}_{i=1}^5$ represent the multi-scale RGB and depth feature maps extracted at each encoder stage. $\{F_i\}_{i=1}^5$ represents the fused cross-modal features obtained at different stages, while $\{\widehat{F}_I\}_{i=1}^4$ denotes the progressively restored multi-scale feature maps generated by the decoder).

- Insufficient robustness of RGB-only segmentation methods under complex conditions.
- Lack of geometric consistency due to the separation of segmentation and quantification stages.
- Limited adaptability of existing RGB-D fusion strategies, which remain relatively rigid.

These limitations collectively hinder accurate crack segmentation and reliable geometric quantification in real-world scenarios. To address these challenges, this paper proposes CMF-Former, a cross-modal fusion transformer that progressively integrates RGB and depth information for crack segmentation and geometry measurement. The main contributions are as follows:

1) A cross-modal feature fusion network for RGB-D crack segmentation with progressive multi-stage integration.
2) A progressive cross-attention fusion module that enables adaptive interaction between RGB and depth cues across semantic stages.
3) A depth-assisted quantification method for physical measurement of crack length, width, and depth.
4) Generalized RGB-D fusion designs applicable to both CNN and Transformer backbones.

2. Methodology

The workflow of the proposed method for crack detection and quantification is illustrated in Fig. 1. It consists of four stages: data collection, distance estimation, crack segmentation and crack quantification. An RGB-D sensor is employed to capture both RGB images and their corresponding depth maps of cracks. To obtain accurate depth-to-surface measurements, the RGB-D sensor is integrated with the YOLOv5 algorithm, enabling automated spatial distance estimation and subsequent calculation of the spatial ratio.

The collected RGB-D data are processed by the proposed CMF-Former, which performs multi-stage fusion of RGB and depth features through progressive cross-modal attention for precise pixel-level crack segmentation. Following segmentation, a depth-assisted quantification method is employed to measure the geometric properties of cracks, including length, width, and depth. Crack length and width are calculated based on skeleton extraction and contour analysis, and crack depth is obtained by mapping the segmented crack regions onto the corresponding depth map. This quantification process provides accurate geometric data of cracks, which is vital for safety evaluation, damage diagnosis, and maintenance planning of critical infrastructures.

3. Crack segmentation algorithm

3.1. CMF-Former

This section introduces CMF-Former, a Transformer-based network tailored for crack segmentation, which integrates RGB and depth features to improve the perception of fine-scale crack patterns under challenging field conditions. The structure of CMF-Former follows a hierarchical encoder-decoder framework, as illustrated in Fig. 2.

The encoder extracts multi-scale representations from RGB and depth inputs, capturing both surface texture and structural discontinuity. To preserve modality-specific characteristics, RGB and depth images are first processed through separate Patch Embedding modules (RGB-PE and Depth-PE), which utilize distinct convolutional operations suited to the different statistical properties of visual appearance and depth geometry. Subsequently, each modality is passed through a Global-Local Self-Attention (GLSA) block. This module is designed to capture long-range contextual dependencies while retaining fine-grained structural cues, which is critical for identifying thin,

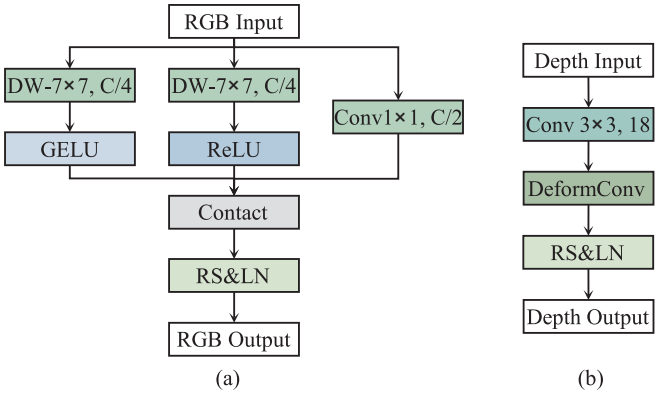

Fig. 3. Architecture of Patch Embedding module. (a) RGB Patch Embedding (RGB-PE); (b) Depth Patch Embedding (Depth-PE). (RS and LN denote *Reshape operation* and *Layer Normalization*, respectively.).

discontinuous, or branching cracks often observed in real-world structures.

In the first encoding stage, RGB and depth features are processed separately to preserve their individual advantages. RGB captures appearance cues such as color and texture, while depth provides overall geometric information. CAF module is applied from the second stage onward to enable bidirectional attention between RGB and depth features. This mechanism allows each modality to selectively attend complementary information from the counterpart, facilitating more effective feature alignment and fusion across semantic levels. For crack segmentation, this mechanism strengthens the representation of true crack contours, particularly when visual contrast is weak but geometric discontinuities are evident in the depth map. In the decoder, the proposed Dual-Patch Feature Restoration (DPFR) module integrates multi-scale encoder features through skip connections to progressively reconstruct high-resolution segmentation maps.

In addition, CMF-Former is a general Transformer-based architecture designed for RGB-D feature fusion. For other Transformer-based segmentation frameworks that operate on RGB-D data, GLSA block and CAF modules proposed in CMF-Former can serve as replacements of their attention mechanisms, facilitating more effective multi-modal feature interaction and extraction. The details of each module will be described in the following sections.

3.1.1. Patch Embedding module

Patch Embedding module converts RGB and depth inputs into token sequences for Transformer-based processing. To retain the distinct feature distributions of each modality, a dual-branch structure is introduced, where RGB and depth data are handled separately before fusion.

RGB-PE branch focuses on visual characteristics such as texture continuity and brightness variation, which are important for identifying thin and low-contrast cracks. It employs parallel 7×7 depth-wise convolutions (DWConv) to efficiently capture elongated crack patterns and spatial continuity along different directions. A parallel 1×1 convolution is used to retain localized appearance contrast and enhance channel-wise feature integration, as shown in Fig. 3(a). GELU [41] and ReLU [42] activations are applied to enhance non-linear representation capacity.

Depth-PE branch emphasizes structural features, including surface depressions and geometric discontinuities caused by cracks. As shown in Fig. 3(b), a deformable convolutional (DeformConv) [43] layer is used to enhance sensitivity to local geometric changes. By learning spatial offsets, it shifts the sampling pattern toward depth variations around crack edges, helping the network identify surface depressions and structural boundaries that are difficult to capture using RGB features alone.

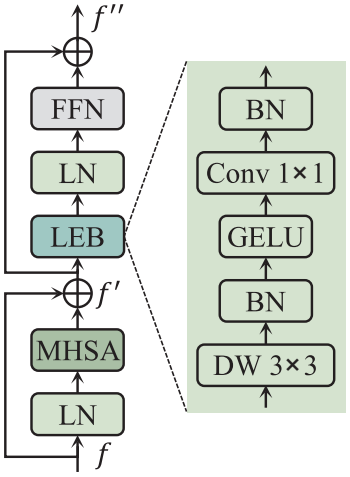

Fig. 4. Architecture of GLSA Block.

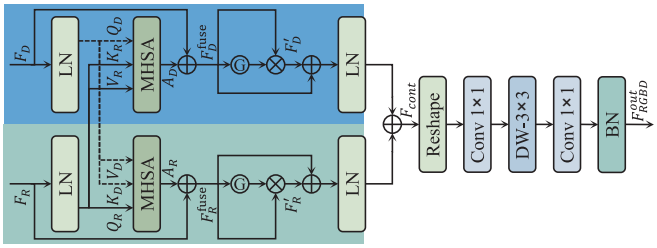

Fig. 5. Architecture of CAF module.

3.1.2. GLSA block

The output tokens from RGB-PE and Depth-PE branches are respectively passed into separate GLSA blocks to further refine modality-specific representations. As shown in Fig. 4, each GLSA block consists of a Multi-head Self-Attention (MHSA) module followed by a Local Enhancement Branch (LEB). MHSA module models long-range dependencies to associate fragmented crack regions and capture global continuity patterns. To compensate for the potential loss of local detail, the LEB is applied to the intermediate features and enhances spatial resolution through a lightweight convolutional structure, including a 3×3 DWConv, Batch Normalization (BN) [44], GELU activation, and a 1×1 convolution followed by another BN layer. This design enables each modality to retain global semantic context while reinforcing critical local cues, such as crack edge and contour discontinuities, which are essential for accurate segmentation under complex conditions. The output of LEB is passed through a LN layer and a Feed-Forward Network (FFN), followed by a second residual connection. The overall computation process of GLSA block is summarized in Eq. (1).

$$\begin{aligned} f' &= MHSA(LN(f)) + f \\ f'' &= FFN(LN(LEB(f'))) + f' \end{aligned} \tag{1}$$

3.1.3. CAF module

Following modality-specific encoding and global–local feature extraction through GLSA blocks, the RGB and depth features are prepared for cross-modal interaction. However, existing fusion methods often rely on early concatenation or late-stage addition, which lack adaptive alignment and tend to produce misaligned features or blurred crack boundaries. To address this, the proposed CAF module introduces bidirectional attention between RGB and depth streams, as shown in Fig. 5. Each modality generates queries and attends to the key-value embeddings of the counterpart, enabling dynamic and task-aware feature enhancement. RGB features incorporate geometric cues from depth to better localize cracks in weak-texture regions, while depth

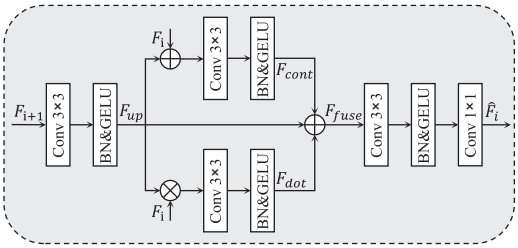

Fig. 6. Architecture of DPFR module.

features leverage appearance cues from RGB to suppress noise and restore incomplete boundaries. This design facilitates mutual guidance and effective feature selection, improving robustness in challenging inspection scenarios.

The detailed computation of CAF module is as follows: Let F_R and F_D denote the input features from the RGB and depth branches, respectively. To perform symmetric cross-attention, each modality is first normalized via LN and then projected into query embeddings (Q) using its own features, while key (K) and value (V) embeddings are derived from the other modality:

$$\begin{aligned} Q_R &= \widehat{F}_R W_Q^R,\ K_D = \widehat{F}_D W_K^D,\ V_D = \widehat{F}_D W_V^D, \\ Q_D &= \widehat{F}_D W_Q^D,\ K_R = \widehat{F}_R W_K^R,\ V_R = \widehat{F}_R W_V^R, \end{aligned} \tag{2}$$

where $\widehat{F}_R$ and $\widehat{F}_D$ represent the normalized RGB and depth features, respectively, and $W_Q^*, W_V^*, W_K^* \in R^{C\times C}$ are learnable projection matrices.

The attention outputs are computed by allowing each modality to attend informative regions of the other using symmetric cross-attention. These outputs are then fused with the original inputs through residual connections to preserve modality-specific information:

$$\begin{aligned} F_R^{fuse} &= F_R + Softmax\left(\frac{Q_R K_D^T}{\sqrt{d}}\right) V_D \\ F_D^{fuse} &= F_D + Softmax\left(\frac{Q_D K_R^T}{\sqrt{d}}\right) V_R \end{aligned} \tag{3}$$

To further emphasize salient features and suppress irrelevant ones, a channel-wise gating mechanism [45] is applied. Each fused feature map is passed through a linear layer and a sigmoid activation to generate a gating vector. The gated features F_R' and F_D' are obtained via element-wise multiplication, as follows:

$$\begin{aligned} F_R' &= \sigma\left(W_G^R F_R^{fuse}\right) \otimes F_R^{fuse} \\ F_D' &= \sigma\left(W_G^D F_D^{fuse}\right) \otimes F_D^{fuse} \end{aligned} \tag{4}$$

where σ denotes sigmoid activations, $\otimes$ denotes element-wise multiplication, and W_G^* are learnable weights.

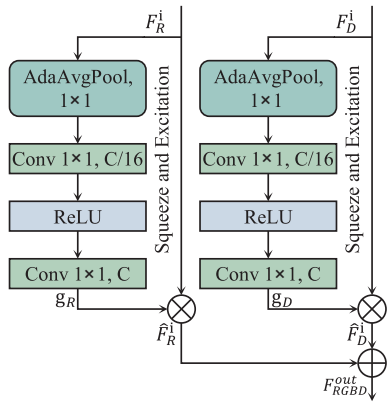

Fig. 8. Architecture of CFF module.

The gated features are concatenated with their corresponding residual-enhanced representations and subsequently passed through modality-specific projection layers, followed by LN to complete the integration. The resulting embeddings are reshaped and processed by a spatial refinement block consisting of a 1×1 convolution, a 3×3 DWConv, and a final 1×1 convolution, with BN applied to the output. The final fused representation is given by:

$$F_{RGBD}^{out} = BN(Conv_{1\times1}(DWConv_{3\times3}(Conv_{1\times1}(F_{cont})))) \tag{5}$$

This output preserves the spatial continuity and local geometric detail of crack patterns, enhancing robustness in real-world inspection scenes.

3.1.4. DPFR module

In the decoder, it is important to recover spatial detail from coarse high-level features, particularly when handling cracks characterized by narrow width, discontinuities, and low contrast. To this end, DPFR module is proposed to fuse upsampled high-level and low-level skip features via two complementary interaction paths: an element-wise multiplication branch that enhances spatial alignment, and a concatenation branch that promotes channel diversity and local context awareness. This dual-branch design facilitates the restoration of irregular and disconnected crack regions while preserving geometric continuity.

The architecture of DPFR is illustrated in Fig. 6, and the entire process can be defined as follows: Let F_{i+1} and F_i denote the upsampled high-level and the corresponding low-level feature, respectively. The high-level feature is first refined and then combined with the two branch outputs. The fused feature F_{fuse} is further enhanced through a convolutional refinement block to produce the final restored feature:

$$\widehat{F}_i = Conv_{1\times1}(GELU(BN(Conv_{3\times3}(F_{dot}, F_{cont}, F_{i+1})))) \tag{6}$$

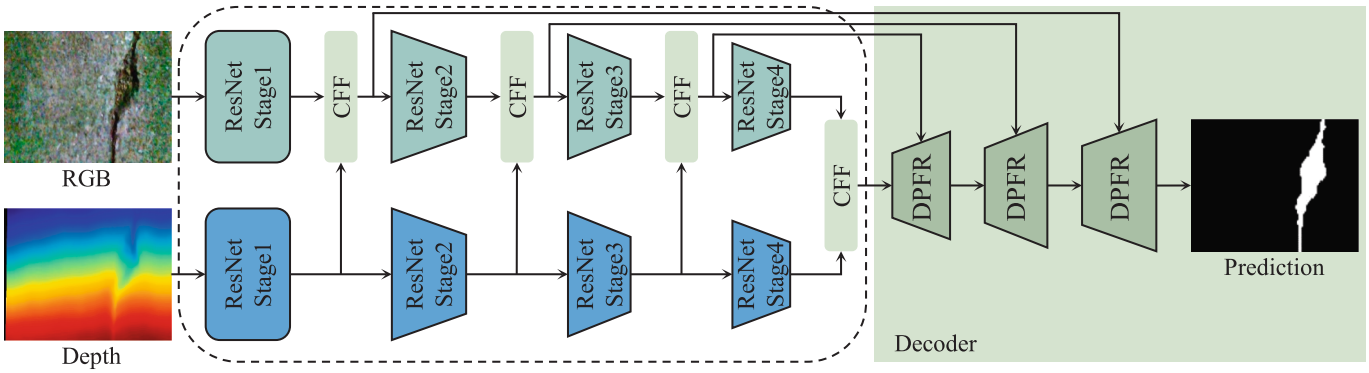

Fig. 7. Architecture of CNN-based feature fusion framework.

where F_{dot} and F_{cont} are the outputs of the multiplicative and concatenation branches. The restored feature improves spatial alignment and local structural continuity, which is essential for reconstructing thin and disconnected crack patterns in degraded visual conditions.

3.2. CNN-based feature fusion architecture

To assess the generalizability of the proposed fusion strategy beyond Transformer-based designs, it is incorporated into a CNN-based segmentation framework for comparative analysis. The architecture follows a standard encoder–decoder paradigm, where RGB and depth images are processed by two parallel ResNet-50 backbones to extract multi-level hierarchical features, as shown in Fig. 7. ResNet-50 is selected due to its effectiveness in feature representation and its training stability enabled by residual connections.

To integrate modality-specific features, a Cross-modal Feature Fusion (CFF) module is employed at each semantic level, as illustrated in Fig. 8. CFF module applies channel-wise attention independently to RGB and depth inputs, allowing informative cues such as edge contrast and crack boundaries in the RGB image, and surface depressions or geometric discontinuities in the depth map to be selectively enhanced prior to fusion. This design supports robust and geometry-aware feature integration within a convolutional framework, facilitating better structural representation for downstream decoding.

The module first applies Adaptive Averaging Pooling [46] to capture channel-wise global context for both modalities. The resulting channel-wise features are passed through two sequential 1×1 convolutions with ReLU and sigmoid activations to generate attention weights, which are used to reweight the original features via element-wise multiplication. The reweighted RGB features $\widehat{F}_R$ and depth features $\widehat{F}_D$ are then concatenated and fused through a final 1×1 convolution, followed by BN and ReLU activation:

$$F_{RGBD}^{out} = ReLU(BN(Conv([\widehat{F}_R; \widehat{F}_D]))) \tag{7}$$

The fused feature is then passed to the decoder for spatial reconstruction and crack prediction. CFF module provides a simple and effective way to adaptively combine RGB and depth features while maintaining compatibility with convolutional encoder–decoder architectures.

4. Crack quantification

This section outlines the procedure for crack quantification based on RGB-D data. The process begins with distance acquisition, where the YOLOv5 algorithm [47] is integrated with an RGB-D camera to estimate precise distances and derive spatial scaling factors in real time. Subsequently, coordinate transformation is applied to correct geometric distortions using camera calibration parameters, as detailed in Appendix A. Finally, the segmentation results obtained from CMF-Former are combined with the aligned depth map and the estimated spatial scale to compute physical crack characteristics, including length, width, and depth.

4.1. Distance acquisition

Accurately determining the distance is crucial for the quantification process, as it directly impacts the spatial ratio, thereby affecting the overall accuracy. However, the depth map captured by the RGB-D camera often contains excessive information from irrelevant background regions, making it unreliable to estimate distance based on the entire map. To address this challenge, YOLOv5 is employed to locate the region of interest via bounding boxes, owing to its maturity and stability, which enable reliable integration with the RGB-D setup. The distance for each predicted bounding box is estimated by averaging the depth values within a local window centered at the midpoint of the box. The detailed algorithmic process is provided as pseudocode in Appendix B.

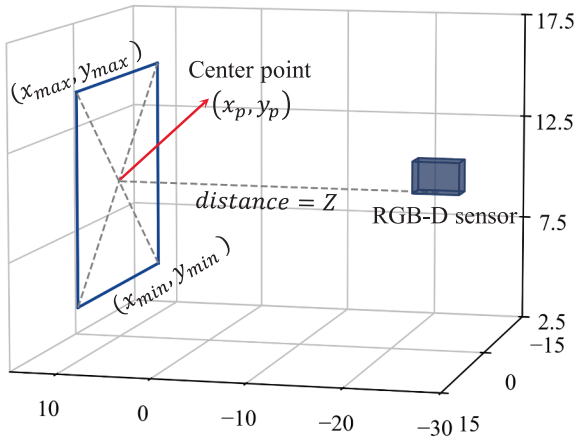

Fig. 9. Schematic for distance acquisition using RGB-D camera.

Fig. 9 illustrates the schematic for distance acquisition using the RGB-D camera. The coordinates of the center point $\left(x_p, y_p\right)$ are calculated as:

$$\begin{aligned} x_p &= \frac{x_{\max} + x_{\min}}{2} \\ y_p &= \frac{y_{\max} + y_{\min}}{2} \end{aligned} \tag{8}$$

where (x_{max}, y_{max}) and (x_{min}, y_{min}) denote the coordinates of the upper-left and lower-right corners of the bounding box, respectively. The distance estimation is then carried out based on the center point and its local neighborhood, following the method detailed in Appendix B.

4.2. Geometric quantification of cracks

To enable physical measurement of crack geometry following segmentation, a depth-assisted quantification framework is introduced. This framework incorporates crack skeletonization, contour extraction, and pixel-to-metric conversion based on depth information, providing estimates of crack length, width, and depth in real-world dimensions.

The process begins with skeleton extraction using the Zhang-Suen image thinning algorithm [48], which generates an initial medial axis of the crack. To suppress redundant branches and noise-induced artifacts, the skeleton is further refined using the Discrete Curve Evolution (DCE) method [49], which removes morphologically insignificant vertices while preserving the main crack trajectory. A comparison between the original and refined skeletons is shown in Fig. 10.

After skeletonization, crack contours are extracted using the Canny edge detector. For each skeleton point, a normal direction is constructed, and the crack width is calculated as the Euclidean distance between its two intersection points with the crack edges:

$$d = \sqrt{\left(x_i - x_j\right)^2 + \left(y_i - y_j\right)^2} \tag{9}$$

where (x_i, y_i) and $\left(x_j, y_j\right)$ are the coordinates of two intersection points.

The crack length is estimated by summing the pixel-wise distances between consecutive points:

$$L = \sum_{i=1}^{n-1} \sqrt{(x_{i+1} - x_i)^2 + (y_{i+1} - y_i)^2} \tag{10}$$

where n is the number of pixel points on the crack skeleton line, (x_i, y_i) and (x_{i+1}, y_{i+1}) are the coordinates of consecutive pixel points along the

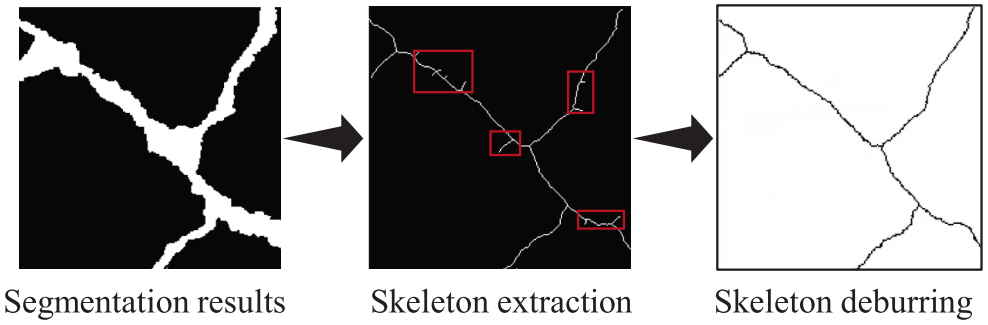

Fig. 10. Comparison of the original and DCE-refined crack skeletons.

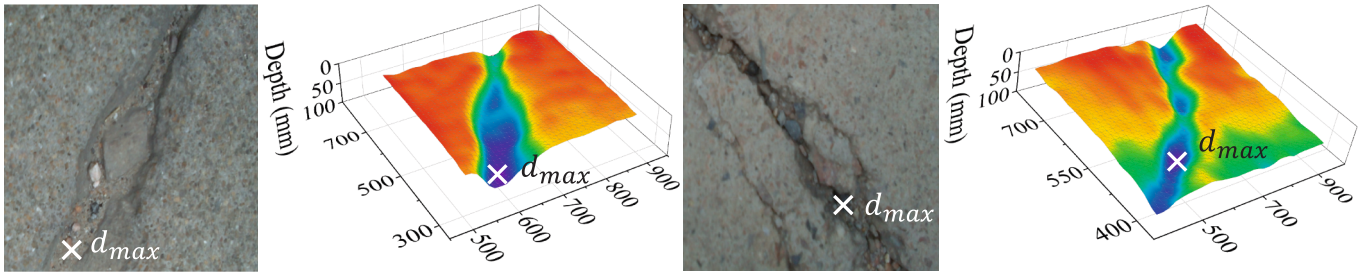

Fig. 11. Examples of 3D view of crack depth map.

Table 1
Specific parameters of RGB-D camera.

RGB camera	1920 × 1080 pixels, 30 FPS
RDB field of view	70° × 55° (±3°)
LiDAR camera	1024 × 768 pixels, 30 FPS
Recommended use range	0.2 m-9 *m*
LiDAR field of view	70° × 43° (±3°)

skeleton.

To convert these pixel-based measurements into real-world dimensions, the spatial ratio s between image pixels and physical length is computed using the RGB-D camera parameters:

$$s = d_p l / Pf \tag{11}$$

where d_p, l, P, and f denote the distance between the camera and the target structure, the size of the camera sensor, the image pixel resolution, and the focal length of camera, respectively.

Fig. 11 presents the crack RGB image and corresponding 3D depth map. The crack depth at a specific location is computed as the difference between the maximum depth value at that location and the distance from the camera to the intact surface of the structure, as follows:

$$D = d_{\max} - d_{surface} \tag{12}$$

where d_{max} denotes the maximum depth at the location, and $d_{surface}$ denotes the distance to the undamaged surface.

5. Datasets and evaluation metrics

5.1. Dataset description

In this study, the Intel RealSense L515 RGB-D camera is employed as

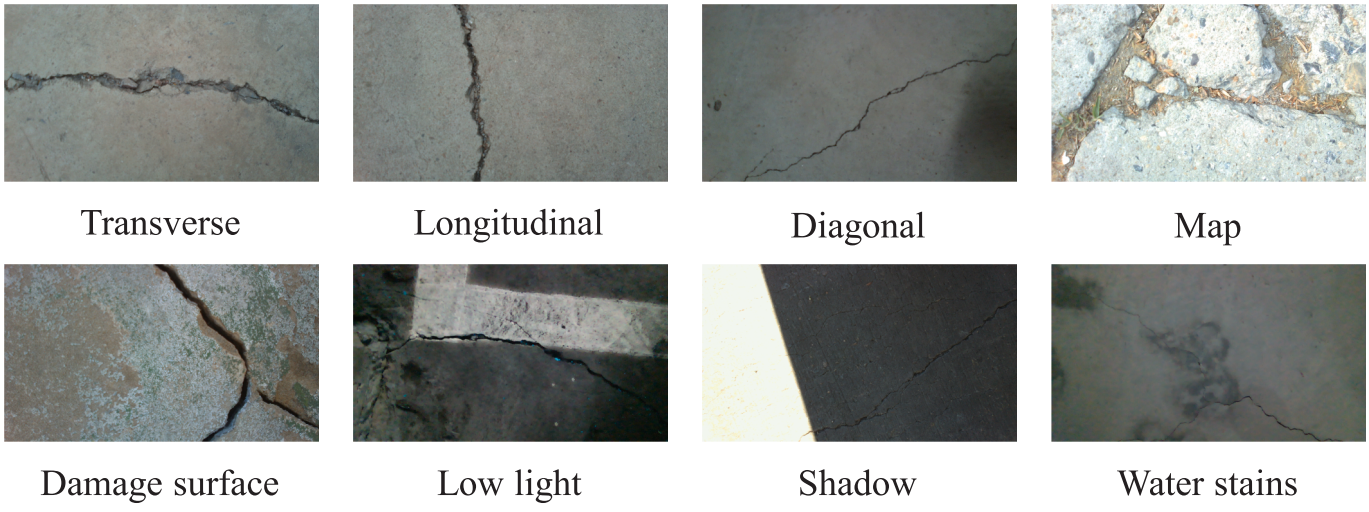

Fig. 12. Representative samples from RGB-D Crack dataset.

Table 2
Statistical distribution of crack types and background complexities in RGB-D crack dataset (The total percentage exceeds 100% because multiple crack types or background complexities may appear simultaneously in a single image).

Crack Types	Approx. Count	Percentage (%)
Diagonal	179	35.8
Longitudinal	168	33.6
Transverse	147	29.4
Low light	63	12.6
Shadow	53	10.6
Damage surface	48	9.6
Map	39	7.8
Water stains	28	5.6

the data acquisition device, which integrates an RGB camera and a LiDAR camera. The RGB camera captures images at a resolution of 1920 × 1080 pixels with a frame rate of 30 FPS. The detailed camera parameters are listed in Table 1. A dataset named RGB-D Crack was collected across various sections of the Nanjing Tech campus. To ensure reliable depth acquisition, images were captured under stable lighting conditions. Most outdoor samples were acquired during the early morning or late afternoon, while additional samples were obtained in indoor environments, thereby covering both natural and artificial illumination. The dataset includes diverse crack types such as transverse, longitudinal, and diagonal, etc. Representative examples are illustrated in Fig. 12, and the statistical distribution is summarized in Table 2.

The original dataset contained 500 pairs of RGB and corresponding depth images of concrete and pavement cracks. To increase data diversity, six augmentation techniques, including cropping, scaling, rotation, HSV contrast adjustment, blurring, and centrosymmetry [50], were applied, resulting in an expanded dataset of 3,500 image pairs. The dataset was subsequently divided into 2,450 pairs for training, 700 for validation, and 350 for testing, enabling comprehensive evaluation under diverse conditions.

In the L515 RGB-D camera, the RGB sensor and the LiDAR-based depth sensor are physically offset and operate at different resolutions, resulting in misalignment between the captured RGB and depth images. To address this issue, the Intel RealSense SDK is employed together with Python scripts [51] to perform image alignment. This process compensates for perspective and resolution differences, ensuring that the depth data is accurately aligned with the RGB images, which is an essential step for effective network training and accurate crack quantification.

5.2. Performance evaluation metrics

To evaluate the performance of the proposed network in crack segmentation, several standard metrics are used: Precision, Recall, Accuracy, F1-score, and mean Intersection over Union (mIoU). The evaluation metrics are defined as follows:

$$\begin{aligned} Precision &= \frac{TP}{TP+FP} \\ Recall &= \frac{TP}{TP+FN} \\ F1\ \text{-}\ score &= 2\times\frac{Precision\times Recall}{Precision+Recall} \\ mIoU &= \frac{1}{2}\sum_{i=0}^{1}\frac{TP}{FN+FP+TP} \end{aligned} \tag{13}$$

where TP represents the true positive samples that were correctly detected as cracks, FP represents the false positive samples that detected as non-cracks, and FN represents the false negative samples that were not detected as cracks.

Table 3
Various networks for comparison.

Model type	Method	Backbone
CNN-based	DeepLabV3+ [56]	ResNet-50
	U-Net [57]	ResNet-50
	PSPNet [58]	ResNet-50
Transformer-based	UperNet [59]	PVT [60]
	UperNet	ViT [61]
	UperNet	Swin [62]
	CMF-Former	

6. Results and discussion

6.1. Network implementation

To evaluate the effectiveness of the proposed cross-modal feature fusion strategy in crack segmentation, several representative semantic segmentation architectures, which are widely used in general-purpose vision tasks, were extended to support dual-modality input, incorporating both RGB images and depth maps. Each model was integrated with the proposed fusion modules to ensure consistent cross-modal interaction and maintain architectural comparability. Table 3 summarizes the employed networks, including their methodological categories and backbone architectures. CNN-based models adopt ResNet-50 as the backbone, while Transformer-based models employ diverse encoder designs.

In addition to these generic segmentation architectures, several crack-specific networks have also been included for comparison, such as CrackSegformer [52], DCCM [53], DBCNet [54], and ESSA-net [55], which were recently introduced for segmentation and quantification tasks. All networks were trained on a workstation equipped with two NVIDIA RTX 4090 GPUs, using 60,000 iterations and a mini-batch size of 8.

6.2. Quantitative evaluation of segmentation performance

To assess the impact of depth information, Table 4 presents the quantitative results of various networks. Among the RGB-only variants, CNN-based models achieved mIoU between 78.33 % and 79.12 %, while Transformer-based models yielded higher values, ranging from 81.04 % to 82.37 %. Crack-specific networks achieved further improvements compared with CNN baselines, with DBCNet reaching an mIoU of 83.80 %.

Nevertheless, RGB-D networks consistently outperformed not only their RGB-only counterparts but also the crack-specific architectures. When comparing RGB-only and RGB-D variants of the same backbone, the proposed fusion strategies yielded mIoU gains between 2.14 % and 3.25 %. The proposed CMF-Former achieved the best overall performance, with a Precision of 92.14 %, Recall of 89.96 %, F1-score of 90.57 %, and mIoU of 86.51 %, outperforming the best CNN-based model, DeepLabV3+, and the best Transformer-based model, Swin, by 4.14 % and 1.75 % in mIoU, respectively. These results demonstrate that explicit cross-modal fusion with depth information provides more substantial benefits than architecture-level specialization alone.

In addition to segmentation accuracy, Table 4 also reports computational efficiency. It can be observed that the introduction of depth information generally increases the computational complexity of all backbones, leading to higher FLOPs and reduced inference speed compared to their RGB-only counterparts. Despite this increase in computational burden, the proposed CMF-Former achieves the highest mIoU while using only 9.38 M parameters, which is considerably fewer than other Transformer-based models. However, the integration of depth information and progressive fusion substantially raises the computational cost to 483.63 GFLOPs and results in a relatively slow inference speed of 8.93 FPS.

To complement the quantitative results, Fig. 13 illustrates visual

Table 4
Quantitative accuracy and efficiency comparison of various networks (* denotes the network adopts the proposed feature fusion architecture, and computational complexity is calculated based on an input image size of 512 × 512).

Model type	Net index	Precision	Recall	F1-score	mIoU	Params (M)	GFLOPs	FPS
CNN-based	DeepLabV3+	90.84	88.43	88.52	79.12	40.35	93.21	48.81
	DeepLabV3+*	91.35	89.24	89.61	82.37	76.41	153.95	41.34
	U-Net	89.63	87.01	86.58	78.33	26.16	52.96	54.53
	U-Net *	90.38	87.41	87.08	81.04	62.22	113.1	47.56
	PSPNet	90.32	87.53	87.19	78.54	47.76	671.08	14.87
	PSPNet*	90.86	88.07	87.92	81.45	83.82	732.23	14.41
Transformer-based	PVT	90.77	87.67	88.93	81.23	36.89	146.82	28.11
	PVT*	91.43	88.42	89.64	83.37	85.03	243.17	23.2
	ViT	91.57	87.94	89.29	81.75	116.55	227.35	16.9
	ViT*	91.98	88.25	89.83	83.97	617.51	489.63	8.38
	Swin	91.65	88.29	89.61	82.24	48.47	182.57	26.22
	Swin*	91.93	89.7	90.12	84.76	135.89	365.73	17.28
Crack-Specific	CrackSegformer	87.81	87.36	86.24	80.2	47.14	71.6	39.38
	DCCM	89.06	88.36	87.13	81.08	26.04	44.94	41.12
	DBCNet	90.16	87.69	88.36	83.8	50.57	73.81	27.26
	USSA-net	88.76	88.53	81.23	82.76	26.09	462.72	9.62
	CMF-Former	**92.14**	**89.96**	**90.57**	**86.51**	**9.38**	483.63	8.93

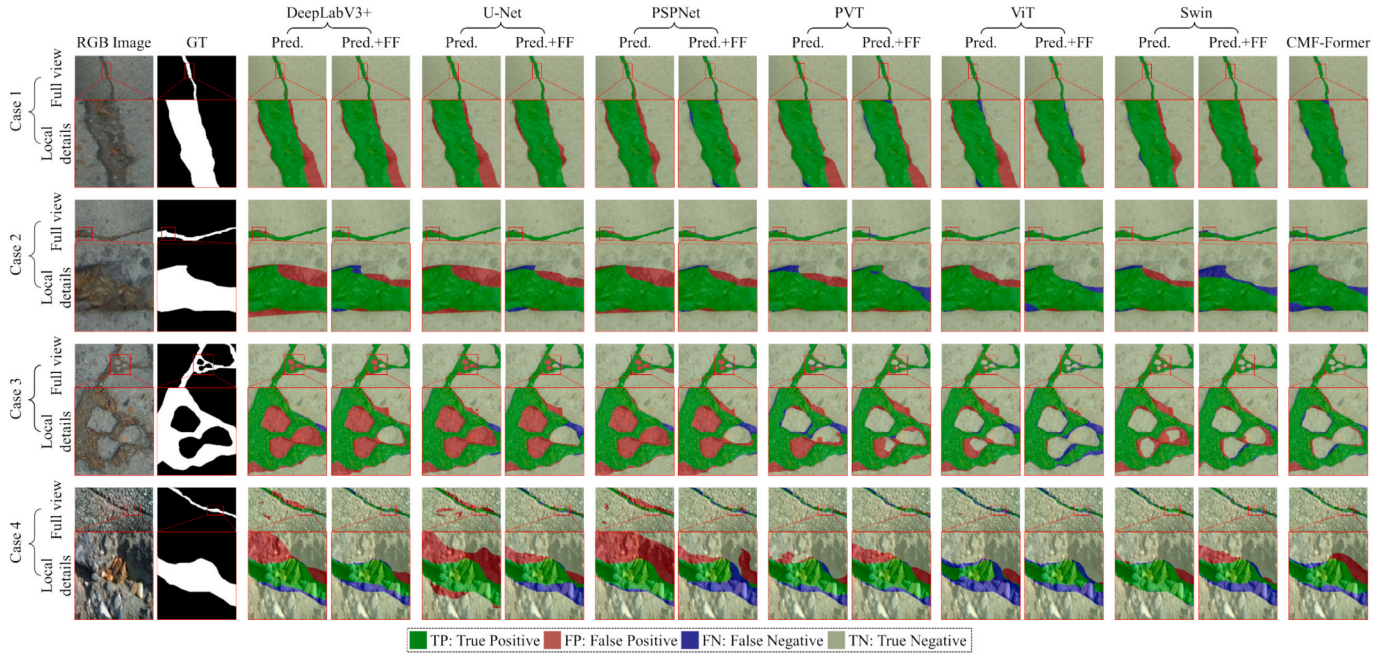

Fig. 13. Qualitative comparison of various networks using RGB features and RGB-D features ("Pred." denotes the baseline prediction, while "Pred. + FF" indicates prediction with future fusion.

comparisons between RGB-only and RGB-D variants of representative backbones, where red boxes highlight representative regions for qualitative analysis. Ground truth and predicted results are displayed for each method, with different colors indicating TP, FP, TN, and FN. The highlighted regions show that CNN-based models suffer from false positives in simple crack cases (Case 1 and 2), while Transformer-based models struggle with false negatives in complex backgrounds (Case 3 and 4). The addition of depth information consistently improves boundary localization and structural continuity, reducing both FP and FN regions. In particular, in Case 3, which presents a reticular crack pattern with complex connectivity, the depth map provides spatial continuity cues among crack branches, helping the model distinguish crack regions from textured backgrounds. In Case 4, where cracks closely resemble the surrounding material in color and texture, RGB-D networks demonstrate improved boundary delineation and reduced FP areas.

Furthermore, Fig. 14 provides qualitative comparisons between crack-specific networks and CMF-Former under diverse conditions, including longitudinal cracks, diagonal cracks, low-light environments, and surfaces with water stains. For longitudinal and diagonal cracks, all crack-specific models are generally capable of capturing the overall crack topology. However, CrackSegformer exhibits noticeable missing regions along thin crack branches, while the other algorithms tend to produce false detections near the crack boundaries. In low-light scenarios, where the contrast between crack and background pixels is substantially reduced, all models encounter difficulties in maintaining complete crack continuity, with CrackSegformer and DBCNet showing the most obvious missed detections and false detections. In the case of surfaces with water stains, the high visual similarity between stains and actual cracks poses a significant challenge. Consequently, most crack-specific networks exhibit substantial false detections and blurred boundaries, whereas CMF-Former achieves more accurate and coherent segmentation, successfully preserving the complete crack structure under these challenging conditions. These observations confirm that the integration of depth cues effectively supplements RGB features with geometric context, improving the detection of fine and low-contrast cracks in complex environments.

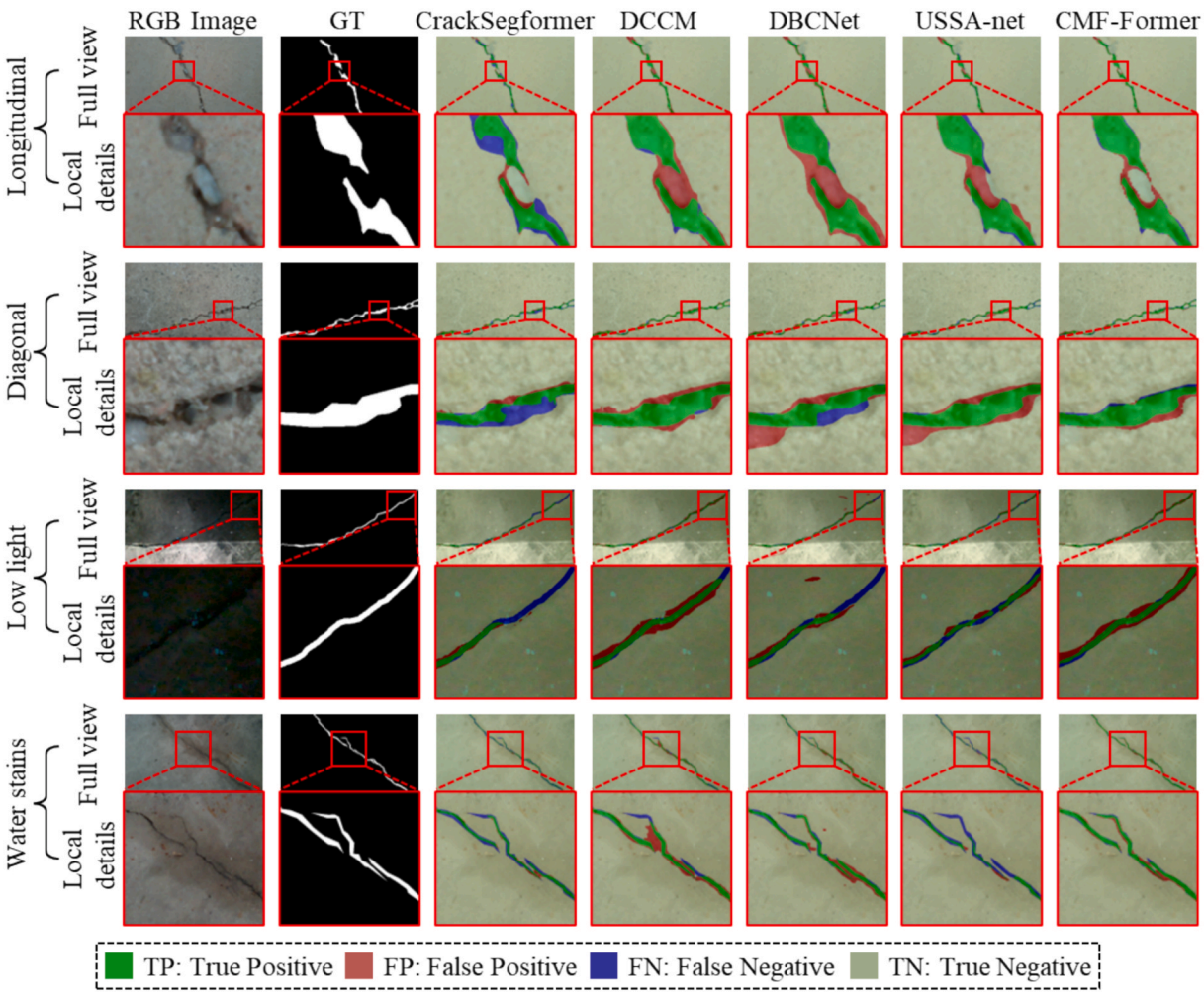

Fig. 14. Qualitative comparison of crack-specific networks and CMF-Former under different conditions.

6.3. Crack geometry quantification and evaluation

This section presents the results of crack quantification, focusing on the measurement of crack width, length, and depth. To highlight the significance of depth feature information for accurate quantification, the results from CMF-Former are compared with those from DeepLabV3+ (DeLab), Swin Transformer (Swin), and the ground truth (GT). The ground truth geometric information of the cracks was obtained through multiple manual measurements and averaged. Fig. 15 illustrates the four cracks analyzed in this section, along with the detection results of various networks and the extracted skeleton lines.

6.3.1. Crack width

Among various crack morphological features, crack width plays a particularly important role, as it provides critical indicators for structural assessment in fracture mechanics. To evaluate the accuracy of width prediction, 100 crack widths were manually measured from four representative test images. Measurements were taken along the direction of the crack skeleton to ensure consistency and objectivity. Two quantitative metrics are adopted to assess prediction performance: Absolute Error Rate and RMSE. The Absolute Error Rate reflects the average absolute difference between the predicted and ground truth values, while RMSE quantifies the standard deviation of the prediction errors, as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i}^{n}(y_i - \hat{y}_i)^2} \tag{14}$$

where y_i denotes the actual width of a crack, $\hat{y}_i$ denotes the predicted value, and n is the total number of cracks. A lower RMSE indicates higher accuracy of the implemented method.

Fig. 15 presents the segmentation results of different networks, and to better visualize the quantitative performance, Fig. 16 provides spline plots of predicted crack widths with shaded regions, alongside absolute error curves. In the plots, the shaded areas of different colors represent the prediction width ranges of various networks, while the red lines indicate the corresponding absolute error rates.

For Crack A, DeepLabV3 + significantly overestimates the width due to FP regions, resulting in an average absolute error rate of 27.83 %, which is 9.96 % and 12.54 % higher than that of Swin and CMF-Former, respectively. In Crack B, both DeepLabV3 + and Swin suffer from discontinuous detection, leading to maximum absolute error rates of 179.29 % and 122.27 %. In contrast, CMF-Former leverages depth-aware feature alignment to reconstruct a coherent crack path and produces smoother and more realistic predictions. For geometrically complex cases such as Cracks C and D, CMF-Former consistently achieves sub-millimeter absolute errors of 0.36 mm and 0.53 mm. It also produces more consistent results, reflecting a better ability to detect small changes in crack width. This is particularly important in engineering practice, where even minor width differences may indicate underlying structural issues.

Quantitative comparisons in Table 5 further support these observations. For Crack B and Crack D, the actual average widths are 4.3 mm and 9.5 mm. DeepLabV3 + predicts these as 2.68 mm and 11.5 mm,

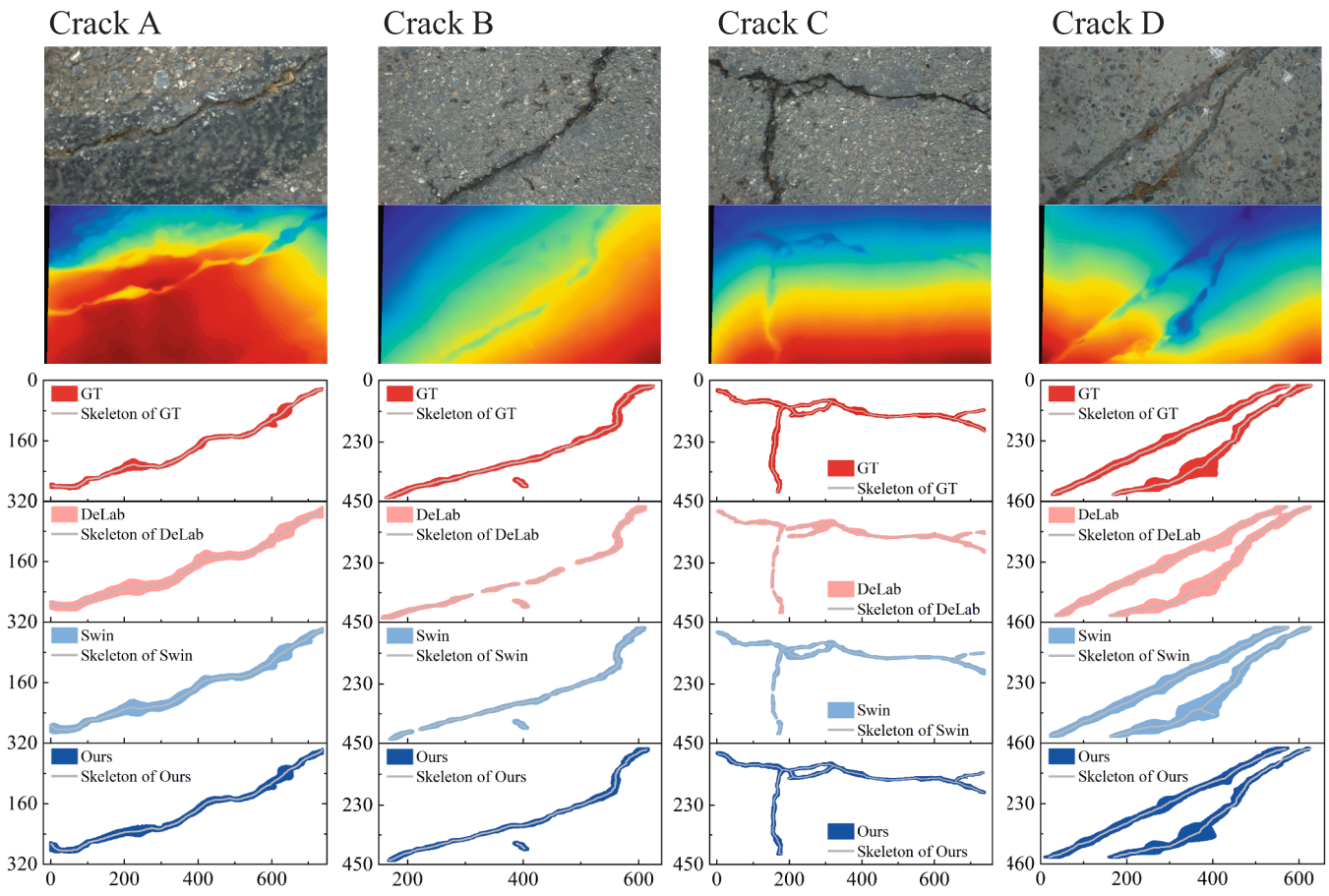

Fig. 15. Comparison of crack detection results between various networks.

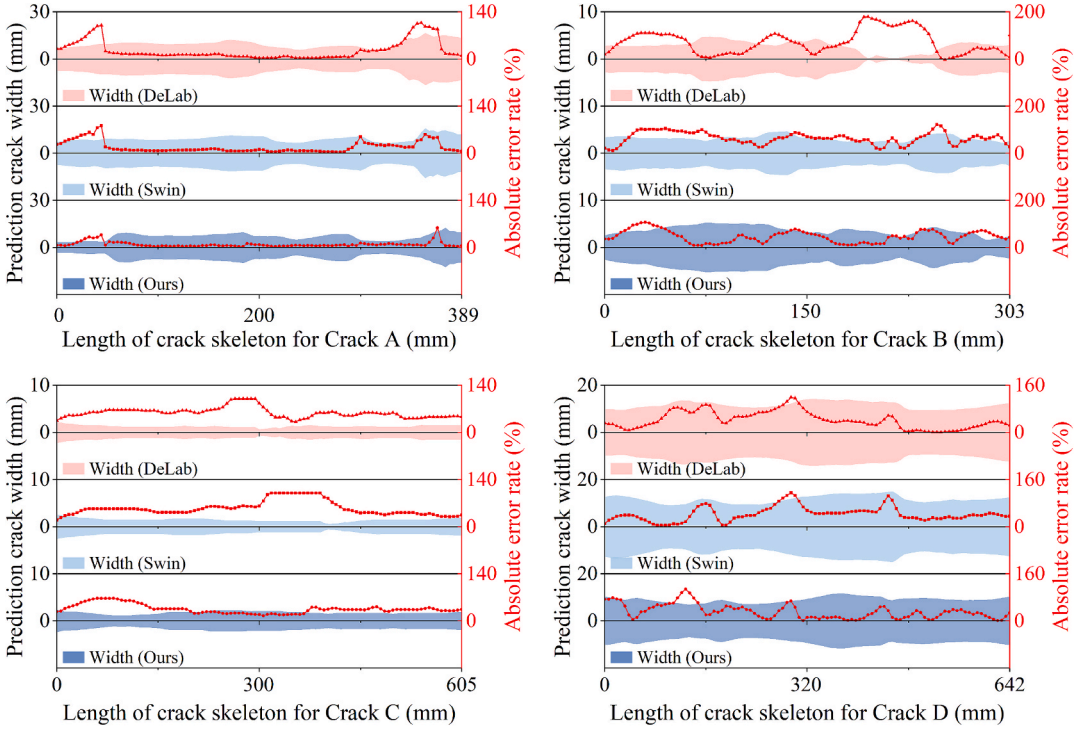

Fig. 16. Spline plots of predicted crack width and Absolute error rate.

Table 5
Quantification results of crack width.

Crack No.		Mean crack width (mm)			RMSE		
	Manual Measurement	DeLab	Swin	Ours	DeLab	Swin	Ours
A	6.54	9.52	9.38	6.9	7.038	4.133	1.578
B	4.3	2.68	3.19	3.68	1.948	1.604	1.322
C	2.3	1.28	1.39	1.97	1.414	1.306	0.889
D	9.5	11.5	11.68	9.04	2.722	2.457	0.878
		Mean RMSE			3.281	2.375	1.167

Table 6
Quantification results of crack length.

Crack No.		Crack length (mm)			Absolute error (mm)			Relative error rate (%)		
	Manual Measurement	DeLab	Swin	Ours	DeLab	Swin	Ours	DeLab	Swin	Ours
A	389	407	419	390	18	30	1	4.63	7.71	0.26
B	303	259	287	308	44	16	5	14.52	5.28	1.65
C	605	542	562	589	63	43	16	10.41	7.11	2.64
D	642	521	543	615	121	99	27	18.85	15.42	4.21

while Swin gives 3.19 mm and 11.68 mm, both showing significant deviations. In contrast, CMF-Former, which incorporates depth features, achieves more accurate predictions of 3.68 mm and 9.04 mm, closely matching the ground truth. For Crack A and Crack C, although CMF-Former shows slight deviations, the errors between its predicted and actual mean widths are only 0.36 mm and 0.33 mm, respectively. On average, CMF-Former has the smallest RMSE across all cracks, reducing the prediction error by 1.174 compared to DeepLabV3 + and by 1.201 compared to Swin.

Overall, CMF-Former achieves more reliable and accurate crack width estimation by effectively capturing both surface texture and geometric depth cues. Through the integration of RGB and depth information using cross-modal attention, the model can better distinguish true crack regions from background noise, restore incomplete or blurry boundaries, and track subtle width variations along irregular crack paths.

6.3.2. Crack length

The crack lengths are computed based on the proposed skeleton extraction method, and the predicted values from different networks are summarized in Table 6. Evaluation is conducted using both absolute and relative error metrics. For Crack A, CMF-Former achieves the highest accuracy, with an absolute error of only 1 mm and a relative error rate of 0.26 %, significantly outperforming DeepLabV3+ (18 mm, 4.63 %) and Swin (30 mm, 7.71 %).

As shown in Fig. 16, the advantage becomes more evident in challenging cases like Crack B. Both DeepLabV3 + and Swin fail to maintain continuous detection along the crack path, resulting in broken skeletons and large underestimations of crack length. In comparison, CMF-Former produces a more continuous and complete skeleton, resulting in a length prediction that more closely matches the ground truth. Similar trends are observed in Crack D, where DeepLabV3 + suffers from extensive missed regions, leading to a large absolute error of 121 mm, which is 94 mm higher than that of CMF-Former. These results confirm that CMF-Former is more reliable for crack length estimation, particularly in irregular or noisy conditions. By integrating RGB and depth features through cross-modal attention, the model improves its ability to trace the full extent of cracks, including weak or disconnected regions, which is critical for accurate damage assessment in engineering applications.

6.3.3. Crack depth

To evaluate the effectiveness of the proposed method in crack depth estimation, its results were compared with manual measurements conducted using a depth vernier caliper. Given that each crack skeleton typically consists of hundreds of pixels, performing manual depth measurement at every location is infeasible, and such measurements cannot be perfectly aligned with the discrete pixel grid. To ensure representative coverage, 25 points were selected at approximately uniform intervals along each crack skeleton, so that the measurements span the entire crack in a consistent manner. The average of these measurements was used as the ground truth reference for each crack.

Table 7
Quantification results of crack depth.

Crack No.	Mean depth (mm)		Absolute error (mm)	RMSE (mm)	Relative error rate (%)
	Manual Measurement	Ours			
A	6.574	6.876	0.302	0.452	4.59
B	5.428	5.152	0.276	0.368	5.08
C	7.399	6.536	0.863	0.917	11.66
D	13.289	13.744	0.455	0.503	3.42
	Mean RMSE			0.56	
		Mean Relative error rate			6.188

The quantitative results presented in Table 7 show that the average depths measured manually for the four test cracks range from 5.428 mm to 13.289 mm. The proposed method, which utilizes RGB-D information, achieves an average RMSE of 0.56 mm and an average relative error rate of 6.188 %. In contrast, Crack C shows a relatively large discrepancy, with a relative error of 11.66 %. One of the main contributing factors to this deviation is environmental interference, such as lighting variation and surface reflectance, which introduce noise into the depth sensing process and compromise the accuracy of the captured depth map.

Since previous crack detection networks are incapable of measuring crack depth, this section focuses on comparing the depth measurements obtained by the proposed method with those acquired through manual measurement. The crack depth was manually measured using a depth vernier caliper. Owing to the difficulty of measuring the entire crack manually, 25 points along the crack were selected, and the depth at each point was recorded. The average depth across all measured points was then calculated to represent the overall depth of the crack. To further assess the consistency of the proposed method, the Pearson correlation coefficient between the predicted and manually measured depths was calculated. The resulting coefficient of 0.987 indicates a strong positive correlation, supporting the reliability of the proposed depth estimation approach.

In summary, the proposed method demonstrates reliable performance in estimating crack depth, with sub-millimeter average RMSE and strong correlation to manual measurements. By leveraging RGB-D sensing, it provides a practical alternative to traditional manual

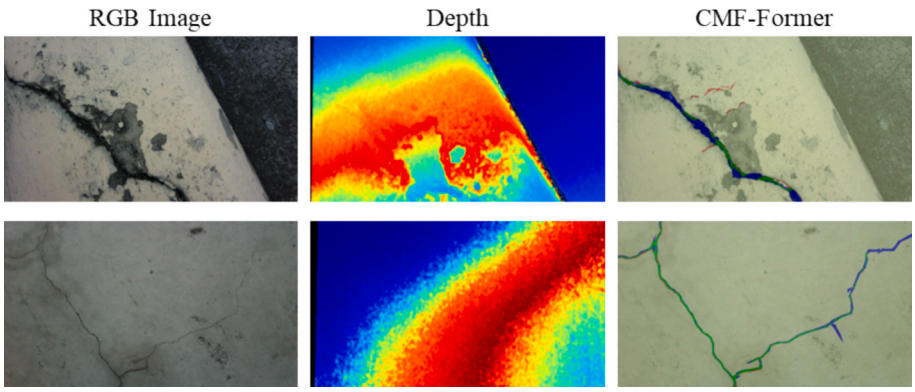

Fig. 17. Failure cases of CMF-Former.

measurement, particularly when manual depth acquisition is difficult or inefficient. Despite some limitations under challenging environmental conditions, such as in Crack C, the method maintains acceptable accuracy and consistency. These results highlight its potential for efficient and accurate crack depth quantification in real-world engineering inspections.

7. Limitations and future work

Although CMF-Former achieves promising performance compared with various CNN– and Transformer-based models as well as recent crack-specific algorithms, several limitations should be acknowledged. As illustrated in Fig. 17, the first example shows that low contrast between cracks and background, combined with unreliable depth estimation, leads to imprecise boundary delineation. In the second example, extremely fine cracks are not adequately represented in the depth map, limiting the effectiveness of cross-modal fusion and resulting in incomplete segmentation. In addition, the incorporation of depth information substantially increases computational complexity, leading to high FLOPs and a relatively low inference speed, which constrains its suitability for real-time inspection tasks. Furthermore, reliance on RGB-D sensors imposes practical constraints, as depth measurements can be degraded under strong illumination, reflective surfaces, or long-range scenarios, thereby affecting robustness in outdoor environments.

Potential directions for future work include the development of lightweight fusion mechanisms and model compression techniques, such as knowledge distillation, to reduce computational overhead and improve deployment efficiency while preserving accuracy. Another promising avenue is the integration of advanced depth enhancement strategies, for instance self-supervised reconstruction or multi-view consistency, which could mitigate the impact of noisy or incomplete depth maps and thereby strengthen the reliability of RGB-D fusion in diverse real-world environments.

In summary, while CMF-Former represents a significant step forward in RGB-D crack analysis, addressing these limitations through more efficient fusion designs and enhanced depth processing will be essential for improving its robustness and practicality in large-scale engineering applications.

8. Conclusion

This paper proposes CMF-Former, an RGB-D fusion network that enhances crack segmentation and enables accurate quantification of geometric features. The main conclusions are summarized as follows:

1) CMF-Former introduces progressive cross-modal interaction across multiple semantic levels, enhancing structural sensitivity and boundary precision. The architecture integrates several key components, including a modality-specific Patch Embedding for preserving crack-relevant cues in each stream, a GLSA block for capturing both long-range context and fine local details, a CAF module that selectively aligns RGB and depth features to enhance complementary representation, and a DPFR module in the decoder to facilitate accurate recovery of thin and fragmented crack contours through spatially aligned feature fusion.
2) A depth-assisted quantification method is developed to directly estimate crack width, length, and depth by leveraging spatial cues from the depth map. This method incorporates automatic spatial ratio estimation by integrating the RGB-D camera with the YOLOv5 detection framework, thereby facilitating geometric measurement under diverse conditions.
3) A dedicated RGB-D crack dataset was collected to validate the effectiveness of the approach. The proposed fusion framework demonstrates strong generalizability, achieving consistent performance gains of 2.14 % to 3.25 % when applied to both CNN-based and Transformer-based architectures.
4) CMF-Former achieves the highest segmentation performance on the dataset with an mIoU of 86.51 %, and yields more accurate geometric measurements compared to RGB-only baselines. In particular, the method achieves a mean RMSE of 1.174 in width estimation, a relative error of 2.19 % in length prediction, and 6.188 % in depth inference, demonstrating reliable and comprehensive crack quantification suitable for practical engineering assessments.

Overall, the proposed RGB-D feature fusion approach presents a significant advancement in crack detection and quantification, offering both enhanced segmentation accuracy and reliable geometric measurements.

CRediT authorship contribution statement

Yingjie Wu: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis. **Shaoqi Li:** Writing – review & editing, Supervision, Project administration, Methodology. **Yancheng Li:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition, Conceptualization.

Declaration of competing interest

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

Acknowledgements

The authors would like to acknowledge the generous supports from Australian Government Research Training Program, SmartCrete CRC (Project ID: 21.PP.0112) and Nanjing Municipal Commission of Urban

and Rural Construction (Project ID: Ks2514) to conduct this research.

Appendix A. . Process of coordinate conversion

In this appendix, the process of coordinate conversion is described. Firstly, a series of checkerboard images were captured at various orientations and distances to form a calibration dataset for camera calibration. Then, the captured data were processed to obtain the intrinsic and extrinsic matrices of camera using the Zhang Zhengyou calibration method. The calibration process involves three coordinate systems, including 3D world coordinate system (X_W, Y_W, Z_W), camera coordinate system (X_C, Y_C, Z_C), and image coordinate system (o, u, v), as shown in Fig. A.1.

The extrinsic matrix can be expressed as Eq. (5), which is represented by a combination of a rotation matrix R and a translation vector, and is used to define the position and orientation of the camera in the real-world.

$$k_1 = [R \quad t] \tag{A.1}$$

The role of extrinsic matrix is to transform points from the 3D world coordinate system to the camera coordinate system, defined as Eq. (A.2):

$$\begin{bmatrix} X_C \\ Y_C \\ Z_C \end{bmatrix} = R \begin{bmatrix} X_W \\ Y_W \\ Z_W \end{bmatrix} + t \tag{A.2}$$

where (X_C, Y_C, Z_C) are the points in the camera coordinate system, and (X_W, Y_W, Z_W) are the points in the 3D world coordinate system.

The intrinsic matric of the camera is expressed as Eq. (A.3):

$$k = \begin{pmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{pmatrix} \tag{A.3}$$

where f_x and f_y denote the focal length in terms of pixels in the x and y direction, respectively, c_x and c_y denote the coordinates of the principal point of the image. The intrinsic matric serves to map the 3D points in the camera coordinate system to 2D points in the image coordinate system, defined as Eq. (A.4):

$$\begin{aligned} x' &= f_x \frac{X_C}{Z_C} \\ Y' &= f_y \frac{Y_C}{Z_C} \end{aligned} \tag{A.4}$$

where (X', Y') are the points in the image coordinate system.

According to the extrinsic and intrinsic matrices of the RGB-D camera and the transformation relationships of each coordinate system, the crack images will be calibrated to eliminate the distortion effect.

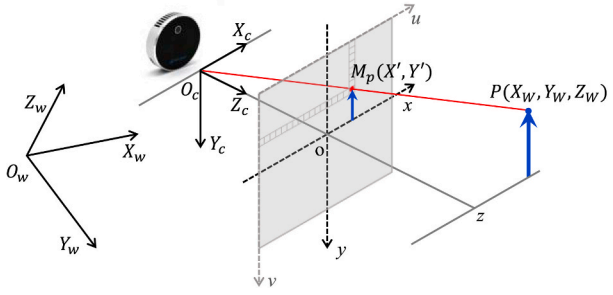

Fig. A.1. . Coordinate system illustration of 3D world, camera and image.

Appendix B. . Algorithm for distance acquisition

.

Table B1
. Pseudocode for distance acquisition.

Algorithm 1: Distance acquisition algorithm flow

Input: — bounding box coordinates $(x_{min}, y_{min}, x_{max}, y_{max})$

— depth map D of size $H \times W$
— window size = 3 (i.e., offsets dx, dy $\in \{-1, 0, 1\}$)
Output: distance Z

(*continued on next page*)

Table B1 (*continued*)

Algorithm 1: Distance acquisition algorithm flow **Input:** — bounding box coordinates $(x_{min}, y_{min}, x_{max}, y_{max})$	
1	Compute the center point $\left(x_p, y_p\right)$
2	$x_p \leftarrow (x_{min} + x_{max})/2$
3	$y_p \leftarrow (y_{min} + y_{max})/2$
4	Initialize depth sum S ← 0
5	Initialize valid count N ← 0
6	**for** each dx in $\{-1, 0, 1\}$ **do**
7	**for** each dy in $\{-1, 0, 1\}$ **do**
8	$x_i \leftarrow x_p + dx$
9	$y_i \leftarrow y_p + dy$
10	**if** $0 \le x_i < W$ and $0 \le y_i < H$ **then**
11	$d \leftarrow D[y_i][x_i]$
12	**if** $d > 0$ **then**
13	S ← S + d
14	N ← N + 1
15	**end if**
16	**end if**
17	**end for**
18	**end forSS**
N	**if** $N > 0$ **then**
20	Z ← S / N
21	**else**
22	Z ← −1 ⟶ invalid or missing depth
23	**Return** Z

Data availability

Data will be made available on request.

References

[1] H. Zhuang, Y. Cheng, M. Zhou, Z. Yang, Deep learning for surface crack detection in civil engineering: a comprehensive review, Measurement 248 (2025) 116908, https://doi.org/10.1016/j.measurement.2025.116908.

[2] X. Xu, B. Ran, N. Jiang, L. Xu, P. Huan, X. Zhang, Z. Li, A systematic review of ultrasonic techniques for defects detection in construction and building materials, Measurement 226 (2024) 114181, https://doi.org/10.1016/j.measurement.2024.114181.

[3] X. Lu, Q. Li, J. Li, L. Zhang, Deep learning-based method for detection and feature quantification of microscopic cracks on the surface of concrete dams, Measurement 240 (2025) 115587, https://doi.org/10.1016/j.measurement.2024.115587.

[4] L. Cui, L. Li, W. Zhang, F. Sun, D. Fan, H. Zhang, F. Jiao, J. Xin, T. Ling, Advances of deep learning application in qualitative and quantitative detection of road subsurface distress using ground penetrating radar: a review, Measurement 247 (2025) 116760, https://doi.org/10.1016/j.measurement.2025.116760.

[5] Aci-224r-01, ACI 224R-01, 2001. https://www.concrete.org/Portals/0/Files/PDF/224R_01Ch3.pdf.

[6] En-1992-1-1, EN 1992-1-1, 2005. https://www.phd.eng.br/wp-content/uploads/2015/12/en.1992.1.1.2004.pdf.

[7] Iso-13822:2010, ISO 13822:2010, 2010. https://cdn.standards.iteh.ai/samples/46556/7d0859948a6848c3bdd5c6dfdb298b71/ISO-13822-2010.pdf.

[8] As-5100.7:2017, AS 5100.7:2017, 2017. https://www.scribd.com/document/776033568/AS5100-7-2017.

[9] R. Ali, J.H. Chuah, M.S.A. Talip, N. Mokhtar, M.A. Shoaib, Structural crack detection using deep convolutional neural networks, Autom. Constr. 133 (2022) 103989, https://doi.org/10.1016/j.autcon.2021.103989.

[10] Y. Wu, S. Li, J. Li, Y. Yu, J. Li, Y. Li, Deep learning in crack detection: a comprehensive scientometric review, J. Infrastruct Intell. Resil. 4 (2025) 100144, https://doi.org/10.1016/j.iintel.2025.100144.

[11] X. Ma, Y. Li, Z. Yang, S. Li, Y. Li, Lightweight network for millimeter-level concrete crack detection with dense feature connection and dual attention, J. Build. Eng. 94 (2024) 109821, https://doi.org/10.1016/j.jobe.2024.109821.

[12] Y. Wu, S. Li, J. Zhang, Y. Li, Y. Li, Y. Zhang, Dual attention transformer network for pixel-level concrete crack segmentation considering camera placement, Autom. Constr. 157 (2024) 105166, https://doi.org/10.1016/j.autcon.2023.105166.

[13] J. Hang, Y. Wu, Y. Li, T. Lai, J. Zhang, Y. Li, A deep learning semantic segmentation network with attention mechanism for concrete crack detection, Struct. Health Monit. 22 (2023) 3006–3026, https://doi.org/10.1177/14759217221126617.

[14] M. Ren, Y. Li, T. Hussain, Y. Wu, J. Li, Pixel-level concrete crack quantification through super resolution reconstruction and multi-modality fusion, Advanced Engineering Informatics 69 (2026) 103807, https://doi.org/Doi: 10.1016/j.aei.2025.103807.

[15] T. Hussain, Y. Li, M. Ren, J. Li, Pixel-level crack segmentation and quantification enabled by multi-modality cross-fusion of RGB and depth images, Construction and Building Materials 487 (2025) 141961, https://doi.org/Doi: 10.1016/j.conbuildmat.2025.141961.

[16] X. Yang, H. Li, Y. Yu, X. Luo, T. Huang, X. Yang, Automatic Pixel-Level Crack Detection and Measurement using fully Convolutional Network, Comput. Aided Civ. Inf. Eng. 33 (2018) 1090–1109, https://doi.org/10.1111/mice.12412.

[17] H. Ahmed, C.P. Le, H.M. La, Pixel-level classification for bridge deck rebar detection and localization using multi-stage deep encoder-decoder network, Dev. Built Environ. 14 (2023) 100132, https://doi.org/10.1016/j.dibe.2023.100132.

[18] Z. Liu, Y. Cao, Y. Wang, W. Wang, Computer vision-based concrete crack detection using U-net fully convolutional networks, Autom. Constr. 104 (2019) 129–139, https://doi.org/10.1016/j.autcon.2019.04.005.

[19] F. Song, B. Liu, G. Yuan, Pixel-Level Crack Identification for Bridge Concrete Structures using Unmanned Aerial Vehicle Photography and Deep Learning, Struct. Control Health Monit. 2024 (2024) 1299095, https://doi.org/10.1155/2024/1299095.

[20] Z. Yu, Y. Shen, Z. Sun, J. Chen, W. Gang, Cracklab: a high-precision and efficient concrete crack segmentation and quantification network, Dev. Built Environ. 12 (2022) 100088, https://doi.org/10.1016/j.dibe.2022.100088.

[21] K. Li, J. Yang, S. Ma, B. Wang, S. Wang, Y. Tian, Z. Qi, Rethinking Lightweight Convolutional Neural Networks for Efficient and High-Quality Pavement Crack Detection, IEEE Trans. Intell. Transp. Syst. 25 (2024) 237–250, https://doi.org/10.1109/TITS.2023.3307286.

[22] Y. Xu, Y. Xia, Q. Zhao, K. Yang, Q. Li, A Road Crack Segmentation Method Based on Transformer and Multi-Scale Feature Fusion (2024), https://doi.org/10.3390/electronics13122257.

[23] E. Asadi Shamsabadi, C. Xu, A. S. Rao, T. Nguyen, T. Ngo, D. Dias-Da-Costa, Vision transformer-based autonomous crack detection on asphalt and concrete surfaces, Autom. Constr. 140 (2022) 104316, Doi: 10.1016/j.autcon.2022.104316.

[24] H. Liu, J. Yang, X. Miao, C. Mertz, H. Kong, CrackFormer Network for Pavement Crack Segmentation, IEEE Trans. Intell. Transp. Syst. 24 (2023) 9240–9252, https://doi.org/10.1109/TITS.2023.3266776.

[25] C. Wang, H. Liu, X. An, Z. Gong, F. Deng, SwinCrack: Pavement crack detection using convolutional swin-transformer networkImage 1, Digit. Signal Process. 145 (2024) 104297, https://doi.org/10.1016/j.dsp.2023.104297.

[26] D.A. Beyene, D.Q. Tran, M.B. Maru, T. Kim, S. Park, S. Park, Unsupervised domain adaptation-based crack segmentation using transformer network, J. Build. Eng. 80 (2023) 107889, https://doi.org/10.1016/j.jobe.2023.107889.

[27] W. Zhao, Y. Liu, J. Zhang, Y. Shao, J. Shu, Automatic pixel-level crack detection and evaluation of concrete structures using deep learning, Struct. Control Health Monit. 29 (2022) e2981.

[28] Y. Wang, S. Li, Y. Zhang, Y. Li, Lightweight concrete crack segmentation network for drone image with complex backgrounds using multi-scale feature fusion and optimized architecture, Construction and Building Materials 495 (2025) 143667, https://doi.org/Doi: 10.1016/j.conbuildmat.2025.143667.

[29] X. Yang, R. Guo, H. Li, Comparison of multimodal RGB-thermal fusion techniques for exterior wall multi-defect detection, J. Infrastruct Intell. Resil. 2 (2023) 100029, https://doi.org/10.1016/j.iintel.2023.100029.

[30] H. Huang, Y. Cai, C. Zhang, Y. Lu, A. Hammad, L. Fan, Crack detection of masonry structure based on thermal and visible image fusion and semantic segmentation, Autom. Constr. 158 (2024) 105213, https://doi.org/10.1016/j.autcon.2023.105213.

[31] M. Shi, H. Li, Q. Yao, J. Zeng, J. Wang, Vision based nighttime pavement cracks pixel level detection by integrating infrared visible fusion and deep learning, Constr. Build. Mater. 442 (2024) 137662, https://doi.org/10.1016/j.conbuildmat.2024.137662.
[32] D. Zhang, Q. Zou, H. Lin, X. Xu, L. He, R. Gui, Q. Li, Automatic pavement defect detection using 3D laser profiling technology, Autom. Constr. 96 (2018) 350–365, https://doi.org/10.1016/j.autcon.2018.09.019.
[33] Z. Pan, J. Guan, X. Yang, K. Fan, J.C. Ong, N. Guo, X. Wang, One-stage 3D profile-based pavement crack detection and quantification, Autom. Constr. 153 (2023) 104946, https://doi.org/10.1016/j.autcon.2023.104946.
[34] H. Kim, S. Lee, E. Ahn, M. Shin, S.-H. Sim, Crack identification method for concrete structures considering angle of view using RGB-D camera-based sensor fusion, Struct. Health Monit. 20 (2020) 500–512, https://doi.org/10.1177/1475921720934758.
[35] W. Lin, X. Li, H. Han, Q. Yu, Y.-H. Cho, A novel approach for pavement distress detection and quantification using RGB-D camera and deep learning algorithm, Constr. Build. Mater. 407 (2023) 133593, https://doi.org/10.1016/j.conbuildmat.2023.133593.
[36] W. Wang, U. Neumann, Year. Depth-aware cnn for rgb-d segmentation, Eur. Conf. Comput. vis. ECCV (2018) 135–150, https://doi.org/10.48550/arXiv.1803.06791.
[37] S.-J. Park, K.-S. Hong, S. Lee, Year. Rdfnet: Rgb-d multi-level residual feature fusion for indoor semantic segmentation, IEEECVF Int, Conf. Comput. vis. ICCV (2017) 4980–4989, https://doi.org/10.1109/ICCV.2017.533.
[38] H. Zhang, V.S. Sheng, X. Xi, Z. Cui, H. Rong, Overview of RGBD semantic segmentation based on deep learning, J. Ambient. Intell. Humaniz. Comput. 14 (2023) 13627–13645, https://doi.org/10.1007/s12652-022-03829-6.
[39] J. Wang, K. Song, D. Zhang, M. Niu, Y. Yan, Collaborative Learning attention Network based on RGB image and Depth image for Surface defect Inspection of No-Service Rail, IEEE/ASME Trans. Mechatron. 27 (2022) 4874–4884, https://doi.org/10.1109/TMECH.2022.3167412.
[40] Y. Zhang, B. Chen, Y. Li, H. Wang, L. Tan, C. Wang, H. Zhang, RGBD-based method for segmenting apparent pores within bridge towers, Meas. Sci. Technol. 35 (2024) 115407, https://doi.org/10.1088/1361-6501/ad6897.
[41] D. Hendrycks, K. Gimpel, Gaussian Error Linear Units (gelus) (2016), https://doi.org/10.48550/arXiv.1606.08415.
[42] A.F. Agarap, Deep Learning Using Rectified Linear Units (relu) (2018), https://doi.org/10.48550/arXiv.1803.08375.
[43] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, Y. Wei, Year. Deformable convolutional networks, IEEECVF Int, Conf. Comput. vis. ICCV (2017) 764–773, https://doi.org/10.48550/arXiv.1703.06211.
[44] S. Santurkar, D. Tsipras, A. Ilyas, A. Madry, How does batch normalization help optimization? Adv. Neural Inf. Process Syst. 31 (2018) https://doi.org/10.48550/arXiv.1805.11604.
[45] A. Gu, C. Gulcehre, T. Paine, M. Hoffman, R. Pascanu. Year. Improving the gating mechanism of recurrent neural networks, Int. Conf. Mach. Learn. ICML (2020) 3800-3809, Doi: 10.48550/arXiv.1910.09890.
[46] E. Ranjan, S. Sanyal, P. Talukdar. Year. Asap: Adaptive structure aware pooling for learning hierarchical graph representations, AAAI Conf. Artif. Intell. AAAI 34 (04) (2020) 5470-5477, Doi: 10.48550/arXiv.1911.07979.
[47] J. Yao, J. Qi, J. Zhang, H. Shao, J. Yang, X. Li, A Real-Time Detection Algorithm for Kiwifruit Defects Based on YOLOv5, (2021). Doi: 10.3390/electronics10141711.
[48] S. Soni, R. Chadha, S. Kaur, A Review Paper on Thinning of image using Zhang and Suen Algorithm and Neural Network, IOSR J. Comput. Eng. 18 (2016) 48–51, https://doi.org/10.9790/0661-1802054851.
[49] W. Wang, A. Zhang, K.C. Wang, A.F. Braham, S. Qiu, Pavement crack width measurement based on Laplace's equation for continuity and unambiguity, Comput. Aided Civ. Inf. Eng. 33 (2018) 110–123, https://doi.org/10.1111/mice.12319.
[50] Z. Zhou, J. Zhang, C. Gong, W. Wu, Automatic tunnel lining crack detection via deep learning with generative adversarial network-based data augmentation, Undergr. Space 9 (2023) 140–154, https://doi.org/10.1016/j.undsp.2022.07.003.
[51] F. Lourenço, H. Araujo, Year. Intel RealSense SR305, D415 and L515: Experimental Evaluation and Comparison of Depth Estimation, Int. Joint Conf. Comput. vis. Imaging Comput. Graph. Theory Appl. VISIGRAPP (2021) 362–369, https://doi.org/10.5220/0010254203620369.
[52] H. Li, H. Zhang, H. Zhu, K. Gao, H. Liang, J. Yang, Automatic crack detection on concrete and asphalt surfaces using semantic segmentation network with hierarchical Transformer, Eng. Struct. 307 (2024) 117903, https://doi.org/10.1016/j.engstruct.2024.117903.
[53] J. Zhang, S. Zhang, D. Li, J. Wang, J. Wang, Crack segmentation network via difference convolution-based encoder and hybrid CNN-Mamba multi-scale attention, Pattern Recogn. 167 (2025) 111723, https://doi.org/10.1016/j.patcog.2025.111723.
[54] J. Zhang, D. Li, Z. Zeng, R. Zhang, J. Wang, Dual-branch crack segmentation network with multi-shape kernel based on convolutional neural network and Mamba, Eng. Appl. Artif. Intel. 150 (2025) 110536, https://doi.org/10.1016/j.engappai.2025.110536.
[55] K.-W. Tse, R. Pi, W. Yang, X. Yu, C.-Y. Wen, Advancing UAV-based inspection system: the USSA-net segmentation approach to crack quantification, IEEE Trans. Instrum. Meas. 73 (2024) 1–14, https://doi.org/10.1109/TIM.2024.3418073.
[56] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Year. Encoder-decoder with atrous separable convolution for semantic image segmentation, Eur. Conf. Comput. vis. ECCV (2018) 801–818, https://doi.org/10.1007/978-3-030-01234-2_49.
[57] O. Ronneberger, P. Fischer, T. Brox, Year. U-net: Convolutional networks for biomedical image segmentation, Int. Conf. Med. Image Comput. Comput.-Assist. Interv. MICCA I (2015) 234–241, https://doi.org/10.1007/978-3-319-24574-4_28.
[58] H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, Year. Pyramid scene parsing network, IEEECVF Conf, Comput. vis. Pattern Recognit. CVPR (2017) 2881–2890, https://doi.org/10.1109/CVPR.2017.660.
[59] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, J. Sun, Year. Unified perceptual parsing for scene understanding, Eur. Conf. Comput. vis. ECCV (2018) 418–434, https://doi.org/10.1007/978-3-030-01228-1_26.
[60] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, L. Shao, Year. Pyramid vision transformer: a versatile backbone for dense prediction without convolutions, IEEECVF Int, Conf. Comput. vis. ICCV (2021) 568–578, https://doi.org/10.1109/ICCV48922.2021.00061.
[61] A. Dosovitskiy, An image is worth 16x16 words: Transformers for image recognition at scale, (2020). Doi: 10.48550/arXiv.2010.11929.
[62] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Year. Swin transformer: Hierarchical vision transformer using shifted windows, IEEECVF Int. Conf. Comput. vis. ICCV (2021) 10012–10022, https://doi.org/10.48550/arXiv.2103.14030.