LIDAR: Lightweight Adaptive Cue-Aware Fusion Vision Mamba for Multimodal Segmentation of Structural Cracks

Hui Liu
Tianjin University of Technology
Tianjin, China
liuhui1109@stud.tjut.edu.cn

Chen Jia*
Tianjin University of Technology
Tianjin, China
jiachen@email.tjut.edu.cn

Fan Shi
Tianjin University of Technology
Tianjin, China
shifan@email.tjut.edu.cn

Xu Cheng*
Tianjin University of Technology
Tianjin, China
xu.cheng@ieee.org

Mengfei Shi
Tianjin University of Technology
Tianjin, China
smf@stud.tjut.edu.cn

Xia Xie
Hainan University
Haikou, China
shelicy@hainanu.edu.cn

Shengyong Chen
Tianjin University of Technology
Tianjin, China
sy@ieee.org

Abstract

Achieving pixel-level segmentation with low computational cost using multimodal data remains a key challenge in crack segmentation tasks. Existing methods lack the capability for adaptive perception and efficient interactive fusion of cross-modal features. To address these challenges, we propose a Lightweight Adaptive Cue-Aware Vision Mamba network (LIDAR), which efficiently perceives and integrates morphological and textural cues from different modalities under multimodal crack scenarios, generating clear pixel-level crack segmentation maps. Specifically, LIDAR is composed of a Lightweight Adaptive Cue-Aware Visual State Space module (LacaVSS) and a Lightweight Dual Domain Dynamic Collaborative Fusion module (LD3CF). LacaVSS adaptively models crack cues through the proposed mask-guided Efficient Dynamic Guided Scanning Strategy (EDG-SS), while LD3CF leverages an Adaptive Frequency Domain Perceptron (AFDP) and a dual-pooling fusion strategy to effectively capture spatial and frequency-domain cues across modalities. Moreover, we design a Lightweight Dynamically Modulated Multi-Kernel convolution (LDMK) to perceive complex morphological structures with minimal computational overhead, replacing most convolutional operations in LIDAR. Experiments on three datasets demonstrate that our method outperforms other state-of-the-art (SOTA) methods. On the light-field depth dataset, our method achieves 0.8204 in F1 and 0.8465 in mIoU with only 5.35M parameters. Code and datasets are available at https://github.com/Karl1109/LIDAR-Mamba.

CCS Concepts

• **Computing methodologies** → **Image segmentation**.

*Chen Jia and Xu Cheng are the co-corresponding authors.

MM '25, Dublin, Ireland

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3755452

Keywords

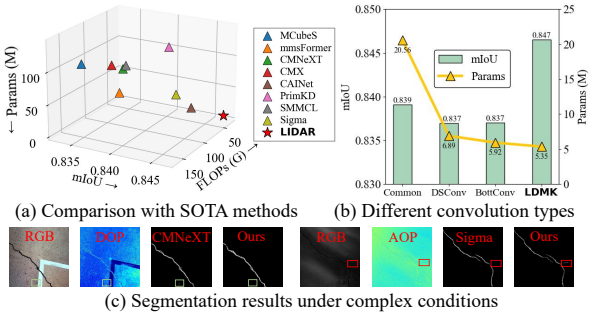
Structural Cracks; Multimodal Data; Crack Segmentation; Lightweight Network; Mamba Network

ACM Reference Format:
Hui Liu, Chen Jia, Fan Shi, Xu Cheng, Mengfei Shi, Xia Xie, and Shengyong Chen. 2025. LIDAR: Lightweight Adaptive Cue-Aware Fusion Vision Mamba for Multimodal Segmentation of Structural Cracks. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3746027.3755452

1 Introduction

Cracks with diverse morphologies frequently appear on the surfaces of real-world materials such as asphalt, concrete, plastic runways, and masonry, primarily due to shear stress. Thus, regular and automated structural health monitoring is essential to prevent losses in daily production and life [2, 14, 16, 19, 40, 49]. Recently, deep learning-based methods have demonstrated strong performance in automatic crack image segmentation [13, 23, 36, 38]. However, most methods rely solely on single-modality RGB data, making them vulnerable to lighting variations and background noise [20]. They fail to capture subsurface thermal anomalies in infrared images, model stress-induced polarization changes, or interpret spatial hierarchies in depth images [37], resulting in degraded performance under complex visual conditions such as uneven illumination, cluttered backgrounds, and ambiguous crack boundaries [21].

Recent multimodal semantic segmentation methods based on Convolutional Neural Networks (CNNs) and Transformers have achieved promising results [1, 35, 42, 43]. Representative methods such as PGDENet [52], CAINet [27], and ESANet [32] utilize CNNs to fuse high-level semantics with low-level spatial details through progressively guided strategies that reduce modality discrepancies. While CNN-based methods effectively capture morphological cues in key regions, their limited receptive fields and inductive biases hinder the modeling of continuous texture patterns. Additionally, lots of convolution operations lead to increased computational overhead. Transformer-based methods such as Omnivore [5], EMMA [51], and CMX [46] encode different modalities into interactive embeddings and employ self-attention to model long-range



(a) Comparison with SOTA methods (b) Different convolution types

(c) Segmentation results under complex conditions

Figure 1: Performance of LIDAR on light-field depth dataset. (a) Comparison with SOTA methods. (b) Impact of different convolution types on performance. (c) Segmentation results for dual-modality images under complex conditions.

dependencies. Although they capture both morphological and textural cues effectively, the quadratic scaling of attention mechanisms with input length makes training and inference on high-resolution images computationally expensive and unsuitable for edge deployment [44]. Despite their promising results, all of the above methods lack selective interaction and noise suppression across modalities and feature levels, leading to the loss of critical details.

Recently, the Selective State Space model (S6) Mamba [7, 9] has shown impressive performance in long-sequence modeling. Compared to traditional linear time-invariant state space models (S4) [6, 8], Mamba offers greater flexibility for complex data while reducing computational cost [10, 26, 29, 39]. Its success in vision tasks, such as Vision Mamba [54], confirms its effectiveness in modeling both local cues and long-range dependencies by scanning fixed-size pixel patches within each Visual State Space (VSS) block. Vision Mamba has since inspired various Mamba-based variants [17, 48], where the scanning strategy is key to capturing spatial dependencies across regions. PlainMamba [41], VMamba [24], and SCSegamba [22] adopt 2D parallel, bidirectional, or combined parallel-diagonal snake scanning to enhance the modeling of spatial continuity, multi-directional context, and complex textures. While such strategies improve irregular structure perception, they rely on uniform, rule-based scanning for all images, limiting adaptability to highly complex and image-specific texture cues and topologies. This often results in discontinuities or blurred segmentation outputs. Moreover, even with uniform rules, generating repeated scanning sequences for each image introduces latency and reduces efficiency. Sigma [34], the first to introduce Mamba into multimodal segmentation, adopts an inefficient parallel scanning strategy without selective interaction or noise suppression during fusion, resulting in high computational cost and missed detections in high-frequency critical regions, as shown in Figure 1(c). Additionally, existing Mamba-based methods stack many VSS blocks, increasing parameter count and computational overhead, which can lead to redundancy and feature degradation. Common convolutions with heavy parameters are used in feature processing and final segmentation, limiting the potential for deployment on resource-constrained devices.

To address the above challenges, we propose the Lightweight Adaptive Cue-Aware Vision Mamba network (LIDAR), which efficiently captures morphological and textural crack cues across various modalities—including RGB images, infrared thermography, polarization informations, and light-field depth cues, while handling arbitrary input sizes with low computational cost. This enables the generation of high-quality crack segmentation maps across diverse scenarios. As shown in Figure 1(a), LIDAR achieves the best performance while requiring the minimal computational resources. To make Mamba adaptable to complex modality-specific crack features, we design the Lightweight Adaptive Cue-Aware Visual State Space module (LacaVSS). It integrates the Efficient Dynamic Guided Scanning Strategy (EDG-SS) based on pre-scanned masks, which significantly accelerates the generation of scanning sequences. EDG-SS dynamically prioritizes crack regions based on image content, improving both the efficiency of texture modeling and the accuracy of crack-background separation. To reduce the computational cost of convolutions, we introduce the Lightweight Dynamically Modulated Multi-Kernel convolution (LDMK), which employs a dynamic intermediate channel selection mechanism and an adaptive selective kernel strategy. This design captures morphological cues through multiple receptive fields while maintaining low complexity. As shown in the Figure 1(b), LDMK enables the LIDAR to achieve the best performance while maintaining a low number of parameters. To enable effective cross-modal and hierarchical feature fusion, we propose the Lightweight Dual Domain Dynamic Collaborative Fusion module (LD3CF). It incorporates an Adaptive Frequency Domain Perceptron (AFDP) to enhance high-frequency crack features and suppress low-frequency background noise in both horizontal and vertical directions. Together with a dual-pooling strategy and dynamic gating, LD3CF enables efficient multi-level interaction with low computational cost.

In summary, our main contributions are as follows:

- We propose LIDAR for multimodal structural crack segmentation. Adaptively captures morphological and textural cues across multiple modalities at a low computational cost, generating high-quality segmentation maps.
- We design the LacaVSS based on the proposed EDG scanning strategy, which enables efficient and adaptive modeling of crack texture cues. The LDMK convolution significantly reduces computational cost while enhancing the perception of morphological information. The LD3CF module generates high-quality segmentation maps through efficient perception of frequency and spatial-domain cues, as well as a multi-level cross-modal interaction mechanism.
- We evaluate LIDAR on three datasets. Experimental results demonstrate that LIDAR outperforms existing SOTA methods while maintaining minimal computational requirements.

2 Related Works

2.1 Crack Segmentation Methods

RGB-based crack segmentation methods have achieved promising results [22, 23, 25]. For instance, SFIAN [3] selectively fuses high-resolution texture and low-resolution semantic information at multiple scales to capture crack geometries. MGCrackNet [45] adopts a learnable parallel CNN-Transformer hybrid module to repeatedly fuse global and local features. CIRL [2] introduces a clustering-inspired representation learning strategy that transforms supervised learning into an unsupervised clustering paradigm to extract more discriminative features. However, these methods rely solely on single-modality RGB data, making them unable to capture subsurface thermal anomalies, stress–polarization correlations, and

spatial or geometric information. Consequently, their performance deteriorates under complex lighting and environmental conditions.

Although no dedicated methods have been specifically developed for multimodal crack segmentation, CNN and Transformer-based methods have achieved strong performance in general semantic segmentation tasks [30, 35, 53]. For example, CAINet [27] improves accuracy through context-aware inference and detail aggregation, while PDCNet [42] utilizes pixel-difference convolution and cascaded large kernels for cross-modal feature fusion. However, CNNs suffer from limited receptive fields and strong inductive biases, making it difficult to model continuous texture patterns, while their dense convolutions result in high computational cost. Transformer-based methods such as CMNext [47] and MCubeSNet [18] fuse complementary information from arbitrary modalities and enhance segmentation via region-guided filter selection. Yet, their self-attention mechanisms introduce quadratic complexity with respect to sequence length, hindering deployment on resource-constrained devices. Moreover, both CNN and Transformer-based approaches lack mechanisms for selective semantic interaction and noise suppression across modalities and feature levels. Consequently, critical cues in fine-grained regions may be overwhelmed by redundant information. Dedicated multimodal crack segmentation methods are needed to effectively capture essential cues across modalities and enhance performance.

2.2 Selective State-Space Vision Model

The Selective State Space model Mamba [7] has attracted attention for its strong performance in sequence modeling tasks. Compared to traditional linear time-invariant models (S4) [8], Mamba provides greater flexibility and computational efficiency for handling complex data, leading to its adoption in vision tasks. At its core, the VSS block performs block-wise scanning over feature maps to capture both fine-grained local details and long-range dependencies. The scanning strategy is crucial for capturing diverse structural and textural patterns. PlainMamba [41] uses direction-aware 2D parallel scanning to preserve semantic continuity, VMamba [24] adopts bidirectional scanning to capture multi-directional dependencies, MaIR [15] applies S-shaped scans within strip regions to maintain locality, and SCSegamba [22] combines parallel and diagonal snake scanning to enhance perception of complex textures. While these methods improve continuity perception through multi-path scanning, they rely on fixed scanning rules and lack the ability to adaptively generate scan sequences per input image. This limits their effectiveness in modeling highly variable textures, especially in crack segmentation where fine details are critical, often resulting in blurred or fragmented outputs. Moreover, repeated static path generation for each image introduces unnecessary latency, reducing inference efficiency. These networks also stack many VSS blocks and use high-parameter convolutions for feature extraction and segmentation, leading to substantial computational cost.

3 Method

3.1 Preliminary

The overall architecture of our proposed LIDAR is illustrated in Figure 2. It comprises two key components: the LacaVSS, which hierarchically extracts morphological and textural cues from different modal inputs, and the LD3CF, which captures frequency and spatial-domain information and generates high-quality segmentation maps through multi-level cross-modal interaction. Given N input images from different modalities $\{X_1, X_2, ..., X_N\} \in \mathbb{R}^{B\times C\times 512\times 512}$, where B denotes the batch size and C denotes the number of channels, each is first processed by a multi-layer LacaVSS backbone. The image is divided into k patches, resulting in a sequence $\{P_1, P_2, ..., P_k\} \in \mathbb{R}^{B\times C\times 8\times 8}$. These patches are scanned and processed by four LacaVSS blocks to extract morphological and textural crack features, producing feature maps $\{F_1, F_2, F_3, F_4\} \in \mathbb{R}^{B\times 64\times 64\times 64}$ for each modality. Finally, LD3CF fuses the feature maps across modalities and levels, generating the final segmentation output $\in \mathbb{R}^{B\times 1\times 512\times 512}$.

3.2 Lightweight Dynamic Convolution

The structure of LDMK is given by Figure 3. To reduce the parameter count and computational cost of convolution operations, LDMK adopts a channel modulation mechanism to dynamically select the most important feature channels for processing. This avoids redundant computation, significantly lowers resource consumption, and enables the extraction of critical crack-related morphological cues through multi-scale dynamic kernel selection across multiple receptive fields. Specifically, given an input feature map $\alpha \in \mathbb{R}^{B\times C_{\text{in}}\times H\times W}$, LDMK first applies a pointwise convolution to project the input from C_{in} to an intermediate dimension C_m. It then models the importance of each channel. Concretely, the importance score s for each channel can be obtained from the following equation:

$$\alpha_m = Conv_{1\times 1}(\alpha) \in \mathbb{R}^{B\times C_m\times H\times W} \tag{1}$$

$$s = \sigma\left(W_2 \cdot ReLU(W_1 \cdot AvgPool(\alpha_m))\right) \in \mathbb{R}^{B\times C_m} \tag{2}$$

where $\sigma(\cdot)$ denotes the Sigmoid activation function, and W_1, W_2 are learnable linear layer parameters. Then, the top-k most influential channels are selected from s to generate a binary mask $M \in \{0,1\}^{B\times C_m}$, which is used for channel-wise pruning:

$$\tilde{\alpha} = \alpha_m \odot M \tag{3}$$

To prevent instability caused by the sharp fluctuation in the number of active channels k during training, LDMK adopts an Exponential Moving Average (EMA) strategy to smooth the channel activation ratio:

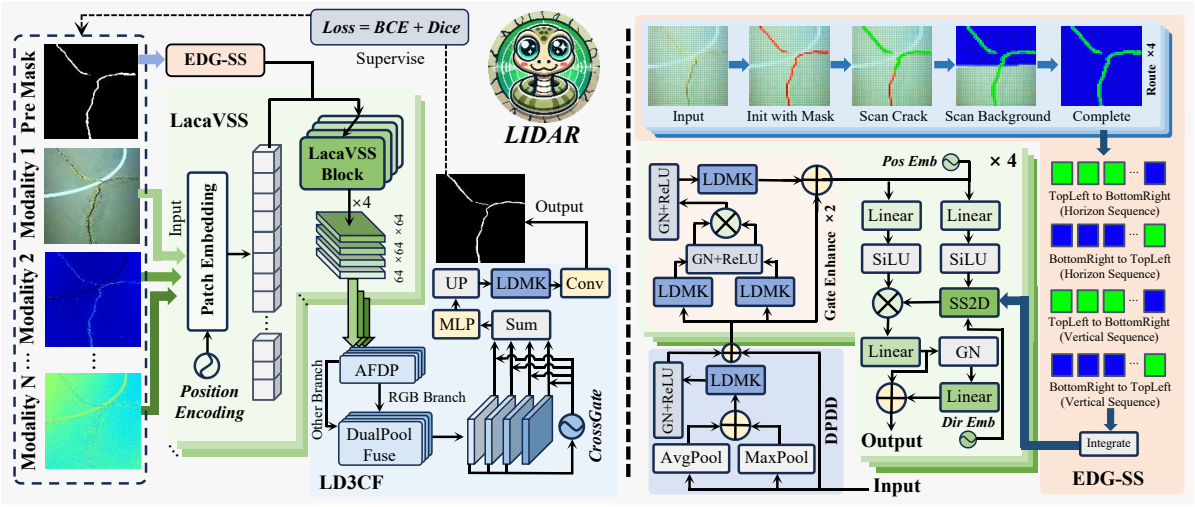
$$\hat{\rho}_t = \gamma \cdot \hat{\rho}_{t-1} + (1-\gamma)\cdot \rho_t, \quad \rho_t = \text{mean}(s) \tag{4}$$

where $\hat{\rho}_t$ denotes the smoothed activation ratio at iteration t, and $\gamma \in [0, 1)$ is the EMA decay factor. During training, the number of activated channels $k_t = \lfloor C_m \cdot \hat{\rho}_t \rfloor$ is adjusted according to $\hat{\rho}_t$ to dynamically control the channel width.

We construct multiple shared depthwise convolution kernels $W_i \in \mathbb{R}^{C_m\times 1\times k_i\times k_i}$, where kernel sizes $k_i \in \{3, 5, 7\}$ are used to capture texture features within different receptive fields. To enhance the adaptability of each receptive field, we introduce learnable scaling and shifting parameters α_i and β_i for each branch:

$$\hat{W}_i = (1+\alpha_i)\cdot W_i + \beta_i \tag{5}$$

where $\hat{W}_i$ denotes the dynamically reparameterized depthwise convolution kernel. Each branch incorporates the scaling factor α_i and bias term β_i to enhance adaptiveness. The output features of each convolution branch are concatenated along the channel dimension,



(a) Network Architecture of LIDAR (b) Architecture of LacaVSS Block

Figure 2: Overview of our LIDAR. Multimodal inputs are processed by LacaVSS to adaptively extract morphological and textural cues, and LD3CF fuses the multimodal feature maps to generate high-quality segmentation outputs. (a) illustrates the architecture of LIDAR and the processing flow for multimodal crack images. (b) illustrates the structure of LacaVSS.

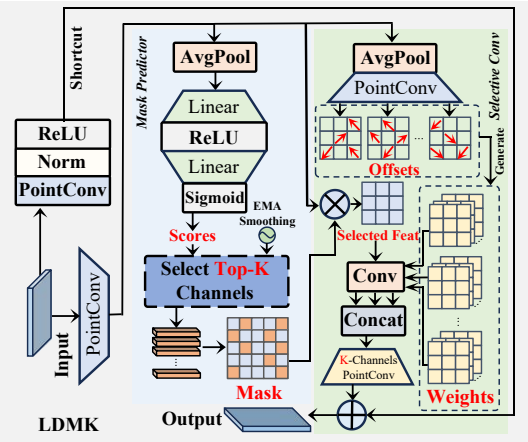


Figure 3: Architecture of LDMK. Adaptive multi-kernel feature extraction is applied to the Top-K most important channels selected from the input features.

and the fused feature is subsequently processed by a pointwise convolution to recover the output dimension C_{out}, followed by addition with a residual connection.

3.3 Lightweight Adaptive Cue Aware

The structure of LacaVSS is illustrated in Figure 2(b). When all feature maps from a single modality are fed into the LacaVSS block, they are first processed by a Dual-Pooling Dynamic Denoiser (DPDD), which we designed to suppress local noise while preserving prominent structural information, thereby enhancing the stability and representational capacity of downstream modeling. Given the input feature map $\omega \in \mathbb{R}^{B\times C\times H\times W}$, the output of the average and max pooling operations is computed as:

$$\omega_1 = \text{AvgPool}(\omega) + \text{MaxPool}(\omega) \tag{6}$$

Subsequently, ω_1 is passed through the LDMK module to adaptively extract local information and produce the denoised output:

$$\omega_{out} = \text{ReLU}(\text{GroupNorm}(\text{LDMK}(\omega_1))) + \omega \tag{7}$$

To further enhance the extraction of morphological crack cues, a gating enhancement unit is integrated into the LDMK module. After two layers of processing, a sequence of patch features enriched with crack morphology is obtained.

Notably, before formal training, LIDAR utilizes pre-generated masks and the EDG-SS to produce a personalized adaptive scanning sequence for each set of multimodal crack images. During pre-training, the EDG-SS in LacaVSS is replaced by the base parallel scanning strategy, and the model is trained for 10 epochs on the multimodal crack dataset to obtain a pre-trained weight file. This weight file is then used to traverse all RGB images in the dataset and generate initial masks that represent general crack contours. These masks are subsequently used by EDG-SS to generate personalized scanning sequences for each multimodal image group before formal training. This process guides the model to better focus on the morphological and textural cues of crack regions.

The EDG-SS scanning procedure is shown in Figure 2(b). Specifically, EDG-SS highlights salient regions based on the crack mask and incorporates an integral segmentation mechanism to rapidly assess the importance of each patch in various directions, thereby constructing adaptive scanning paths. This enables the model to prioritize regions with potential structural cues. Given a binary mask image $M \in \mathbb{R}^{H\times W}$ and patch size p, the integral image is defined as:

$$I(x,y) = \sum_{i=0}^{x}\sum_{j=0}^{y} M(i,j) \tag{8}$$

For a patch of size $p \times p$ starting from the top-left corner (i, j), the importance score $S_{i,j}$ can be computed as:

$$S_{i,j} = I(i+p, j+p) - I(i+p, j) - I(i, j+p) + I(i, j) \tag{9}$$

The EDG-SS separately traverses patches in horizontal and vertical directions to compute importance scores for each patch. Denote the importance score of the k-th patch in direction $d \in \{h, v\}$ as $S_k^{(d)}$, where h and v represent horizontal and vertical directions. Based on the importance values, all patches are divided into a crack region set C_d and a background region set B_d, defined as:

$$C_d = \{k \mid S_k^{(d)} > 0\}, \quad B_d = \{k \mid S_k^{(d)} = 0\} \tag{10}$$

In each direction d, two scanning orders are defined: from the top-left to the bottom-right (tb) and from the bottom-left to the top-right (bt). By combining the two directions and two orders, EDG-SS constructs four scanning sequences. These four sequences are unified and defined as:

$$O_s^{(d)} = \begin{cases} [C_d] + [B_d], & \text{if } s = \text{tb} \\ [C_d]^{\text{rev}} + [B_d]^{\text{rev}}, & \text{if } s = \text{bt} \end{cases}, \quad d \in \{h, v\} \tag{11}$$

where $[\cdot]$ denotes the operation that converts a set into an ordered sequence of patch indices, $[\cdot]^{\text{rev}}$ denotes a reverse operation. EDG-SS ultimately generates four scanning sequences: $O_{\text{tb}}^{(h)}$, $O_{\text{bt}}^{(h)}$, $O_{\text{tb}}^{(v)}$, and $O_{\text{bt}}^{(v)}$. Unlike traditional scanning strategies, EDG-SS only needs to generate scanning sequences once during the preprocessing phase. All scanning sequences of the training and testing images are saved in a JSON file. During model inference or training, LacaVSS only needs to read the file once to retrieve the sequence corresponding to a specific image, avoiding redundant scanning and greatly saving computational time.

After generating all scanning sequences, the patch sequence is fed into the core SS2D module of LacaVSS. The input patch sequence is reordered according to the corresponding sequence $O_s^{(d)}$, the position ϕ_{pos} and direction embedding $\psi_{dir}^{(d,s)}$ are added to each patch to form the input sequence:

$$\eta^{(d,s)} = \eta \left[O_s^{(d)} \right] + \psi_{\text{dir}}^{(d,s)} + \phi_{\text{pos}} \tag{12}$$

where η denotes the original input patch sequence extracted from the image. The sequence is subsequently passed into the core part of the LacaVSS block for state modeling, which is constructed based on discrete linear state space theory. At each scan position, the hidden state h_k is updated as follows:

$$h_k = e^{\Delta A} h_{k-1} + \left[(\Delta A)^{-1} \left(e^{\Delta A} - I \right) \cdot \Delta B \right] \eta_k \tag{13}$$

$$y_k = C h_k + D x_k \tag{14}$$

where $\eta_k \in \mathbb{R}^d$ denotes the input feature at the k-th scan position, h_k represents the current hidden state, and y_k is the model's response at this position. $\Delta A, \Delta B, C, D$ are learnable parameters used to dynamically adjust the update rate and strength of state propagation.

This modeling process is executed in parallel across all directions (d, s), resulting in four groups of state outputs, these outputs are aggregated and fused, and passed through a linear projection to obtain the output.

3.4 Lightweight Dual Domain Dynamic Fusion

As illustrated in Figure 2(a), the proposed LD3CF module aims to enhance and integrate multimodal features across levels by leveraging both frequency-domain perception and spatial-domain fusion.

The process begins with the AFDP, which receives the multimodal features extracted from LacaVSS. Given an input feature map $\zeta \in \mathbb{R}^{B \times C \times H \times W}$, a real-valued Fast Fourier Transform (rFFT) is first applied to project the features into the frequency domain. Direction-aware convolutions are then performed along horizontal and vertical axes to capture orientation-specific responses.

To isolate discriminative frequency components, we introduce a learnable soft masking mechanism. Let d_{center}^h and d_{center}^v represent the distance from each frequency bin to the spectral center in horizontal and vertical directions, respectively. The corresponding high-frequency and low-frequency masks are defined as:

$$\begin{aligned} \mathcal{M}_{\text{high}}^{h,v} &= \sigma \left((d_{\text{center}}^{h,v} - r) \cdot \tau \right), \\ \mathcal{M}_{\text{low}} &= 1 - \max \left(\mathcal{M}_{\text{high}}^{h}, \mathcal{M}_{\text{high}}^{v} \right) \end{aligned} \tag{15}$$

where r is a learnable frequency separation radius, τ is a temperature scaling factor, and $\sigma(\cdot)$ denotes the Sigmoid function. These masks are used to reconstruct directional frequency components, which are then fused in the spatial domain through a channel-wise gating strategy, producing frequency-enhanced features.

Based on these refined features, LD3CF proceeds to perform dual-branch fusion to integrate information from multiple modalities. Let $\delta^{(0)}$ be the RGB modality feature and $\{\delta^{(l)}\}_{l=1}^{M-1}$ the auxiliary modality features, where M denotes the number of modals. The RGB mode is first enhanced by the following operations:

$$\delta_{RGB} = \delta^{(0)} \cdot \sigma \left(Linear \left(AvgPool(\delta^{(0)}) \right) \right) \tag{16}$$

Then, each auxiliary modality interacts with δ_{RGB} through a dual pooling strategy:

$$\delta_{fuse}^{(l)} = w_1 \cdot AvgP(\delta_{RGB} + \delta^{(l)}) + w_{max} \cdot MaxP(\delta_{RGB} + \delta^{(l)}) \tag{17}$$

where w_1 and w_2 are learnable weights. The fused results are further transformed by LDMK convolutions to ensure compact and expressive representation. All modality-specific features are then aggregated to form a unified multimodal embedding:

$$\delta_{sum} = \delta_{RGB} + \sum_{l=1}^{M-1} \delta_{fuse}^{(l)} \tag{18}$$

To further ensure structural consistency and semantic complementarity across different feature levels, we introduce a cross-scale dynamic interaction mechanism. Let $\nu^{(n \in [0,3])}$ denote the output feature of LD3CF at level n, and $\nu^{(n-1)}$ the output from the previous level. These are adaptively fused via a learned gate:

$$\nu^{(n)} = \nu^{(n)} \cdot G^{(n)} + \nu^{(n-1)} \cdot (1 - G^{(n)}) \tag{19}$$

$$G^{(n)} = \sigma \left(Linear \left(AvgPool \left(\nu^{(n-1)} \right) \right) \right) \tag{20}$$

where $G^{(n)} \in [0, 1]^{B \times C \times 1 \times 1}$ serves as a gate to balance semantic reinforcement and structural preservation. For the lowest level ($n = 0$), no previous output exists, so $\nu^{(0)}$ is directly preserved without interaction.

All scale-level fused outputs are upsampled to the same spatial resolution and aggregated with learned weights. Subsequently, the result is processed by a linear layer, pointwise convolution, and further upsampling to generate the final segmentation map with rich awareness of crack shapes and texture structures.

Table 1: Performance comparison at dual-modal inputs. The best results are bolded and the second best results are underlined.

Method	IRTCrack (RGB+Infrared)				CrackDepth (RGB+Depth)				CrackPolar (RGB+AOP)				CrackDepth (RGB+DOP)			
	ODS	OIS	F1	mIoU	ODS	OIS	F1	mIoU	ODS	OIS	F1	mIoU	ODS	OIS	F1	mIoU
CMX[46]	0.8184	0.8244	0.8468	0.8463	0.8017	0.8032	0.8032	0.8336	0.7259	0.7376	0.7120	0.7847	0.7233	0.7261	0.7113	0.7843
CAINet[27]	0.8216	0.8294	0.8367	0.8453	0.8054	0.8080	0.8044	0.8358	0.7327	0.7438	0.7174	0.7886	0.7326	0.7364	0.7180	0.7886
PrimKD[11]	0.8271	0.8345	0.8457	0.8518	0.8010	0.8082	0.8013	0.8338	0.7385	0.7499	0.7258	0.7923	0.7378	0.7433	0.7279	0.7927
SMMCL[4]	0.8090	0.8105	0.8388	0.8384	0.8019	0.8044	0.7996	0.8330	0.7289	0.7322	0.7210	0.7861	0.7284	0.7389	0.7213	0.7855
SSRS[28]	0.8198	0.8274	0.8420	0.8441	0.8143	0.8158	0.8133	0.8425	0.7314	0.7335	0.7157	0.7888	0.7264	0.7295	0.7118	0.7857
Sigma[34]	0.8260	0.8322	0.8545	0.8513	0.8168	0.8176	0.8147	0.8437	0.7327	0.7456	0.7228	0.7895	0.7303	0.7320	0.7224	0.7875
MCubeS[18]	0.8284	**0.8350**	0.8430	0.8516	0.8167	0.8217	0.8114	0.8436	0.7432	0.7469	0.7299	0.7960	0.7449	0.7522	0.7347	0.7972
CMNeXT[47]	0.8256	0.8299	0.8406	0.8508	0.8031	0.8063	0.8026	0.8352	0.7359	0.7448	0.7256	0.7909	0.7405	**0.7572**	0.7321	0.7927
mmsFormer[50]	0.8186	0.8297	0.8421	0.8448	0.8138	0.8180	0.8124	0.8419	0.7430	0.7503	0.7307	0.7951	0.7362	0.7437	0.7258	0.7913
Ours	**0.8305**	0.8316	**0.8625**	**0.8548**	**0.8213**	**0.8237**	**0.8204**	**0.8465**	**0.7479**	**0.7512**	**0.7346**	**0.8000**	**0.7500**	0.7512	**0.7382**	**0.8015**

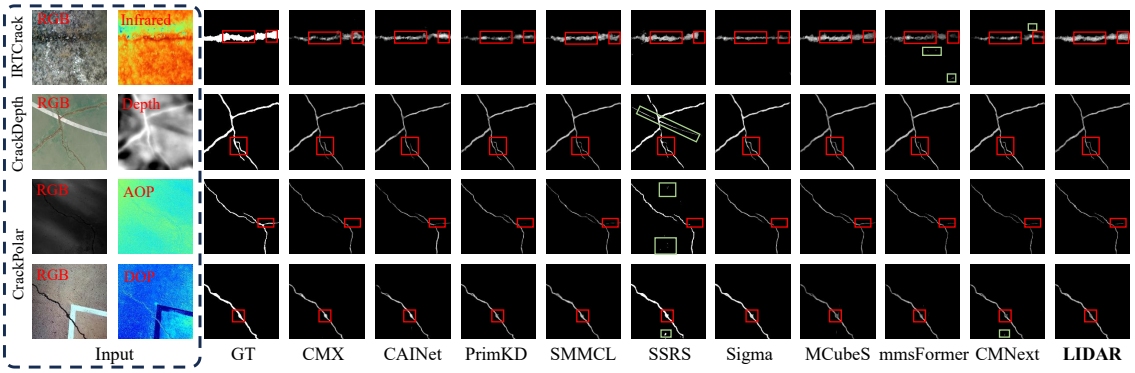


Figure 4: Visualization comparison for dual-modal inputs. Red boxes mark critical regions and green boxes mark noisy regions.

4 Experiments

4.1 Datesets

IRTCrack [20]: The IRTCrack dataset consists of 448 paired RGB and infrared thermal images captured using a thermal imager. It covers diverse conditions, including various crack types, background textures, and lighting. Infrared thermography captures surface and subsurface thermal anomalies, enabling effective crack detection.

CrackDepth: CrackDepth is a dataset we collected, including 655 paired RGB and light-field depth images acquired via a light-field camera. Depth images are generated through post-processing. The dataset covers four surface types under different lighting, and the spatial and geometric cues in the depth data help distinguish crack morphology in complex scenarios.

CrackPolar: CrackPolar is a dataset we collected, containing 986 groups of images, including RGB images, Angle of Polarization (AoP) images, Degree of Polarization (DoP) images, as well as polarization images captured at four typical angles: 0°, 45°, 90°, and 135°, acquired using a polarization camera. It spans four material types under varying lighting. Polarized images highlight stress–polarization variations, enhancing crack-background contrast and detail visibility.

4.2 Implementation Details

Experimental Settings. LIDAR is implemented using PyTorch v2.1.2 and trained on a server equipped with an Intel Xeon Platinum 8336C CPU and eight NVIDIA GeForce RTX 4090 GPUs running Ubuntu 20.04.6. During both pretraining and main training phases, we adopt the AdamW optimizer with an initial learning rate of 0.001. A polynomial learning rate decay strategy is used, and the weight decay is set to 0.01. LIDAR is pretrained for 10 epochs to generate initial structural masks, and then trained for 60 epochs in the main training phase. Input images are resized to a fixed resolution of 512×512 before being fed into the network. The loss function uses the sum of the BCE [12] and Dice [33] losses. All experiments were conducted under the same settings across all datasets.

Evaluation Metrics. We used four metrics to evaluate LIDAR's performance: F1 Score, Optimal Dataset Scale (ODS), Optimal Image Scale (OIS), and mean Intersection over Union (mIoU) [22].

4.3 Comparison with SOTA methods

Comparative experiments on two modality. As shown in Table 1 and Figure 4, we evaluate LIDAR on IRTCrack [20], CrackDepth and the CrackPolar using RGB, AoP, and DoP combinations. Our method consistently achieves superior performance across nearly all evaluation metrics. On the IRTCrack [20] dataset, LIDAR outperforms the second-best method by 0.94% in F1 score and 0.35% in mIoU, demonstrating its strong adaptability in modeling thermal crack responses. On CrackDepth, which includes light-field depth images, LIDAR surpasses Sigma [34] by 0.55%, 0.24%, 0.70%, and 0.33% in ODS, OIS, F1, and mIoU, respectively, indicating its effectiveness in capturing spatial hierarchies and geometric structures. On the CrackPolar dataset, LIDAR maintains SOTA performance. With RGB+AoP or RGB+DoP input, it exceeds the second-best method by an average of 0.56% in F1 and 0.52% in mIoU. This highlights its capability to integrate polarization cues reflecting surface stress and microstructural patterns, enabling precise segmentation in fine-detail regions and robust background suppression.

Table 2: Performance comparison at multi-modal inputs. The best results are bolded and the second best results are underlined.

Method	CrackPolar (RGB+[0°,45°,90°,135°])							CrackPolar (RGB+[0°,45°,90°,135°]+AOP)						
	ODS	OIS	F1	mIoU	FLOPs	Params	Size	ODS	OIS	F1	mIoU	FLOPs	Params	Size
MCubeS[18]	0.7486	0.7514	0.7331	0.7975	396.76G	293.35M	3594MB	0.7483	0.7509	0.7323	0.7975	473.62G	351.86M	4045MB
mmsFormer[50]	0.7403	0.7457	0.7275	0.7932	76.39G	59.99M	738MB	0.7356	0.7391	0.7214	0.7903	84.94G	73.34M	863MB
CMNeXT[47]	0.7473	0.7501	0.7302	0.7967	**48.46G**	57.66M	660MB	0.7515	**0.7654**	0.7420	0.7986	**49.69G**	57.67M	660MB
Ours	**0.7503**	**0.7543**	**0.7383**	**0.8015**	83.34G	**13.35M**	**193MB**	**0.7576**	0.7647	**0.7467**	**0.8023**	100.01G	**16.01M**	**231MB**
Method	CrackPolar (RGB+[0°,45°,90°,135°]+DOP)							CrackPolar (RGB+[0°,45°,90°,135°]+AOP+DOP)						
	ODS	OIS	F1	mIoU	FLOPs	Params	Size	ODS	OIS	F1	mIoU	FLOPs	Params	Size
MCubeS[18]	0.7530	0.7600	0.7417	0.8000	473.62G	351.86M	4045MB	0.7514	0.7555	0.7380	0.7993	550.48G	410.37M	4496MB
mmsFormer[50]	0.7523	0.7586	0.7384	0.8000	84.94G	73.34M	863MB	0.7445	0.7485	0.7296	0.7956	93.50G	86.68M	1034MB
CMNeXT[47]	0.7583	**0.7642**	0.7449	0.8035	**49.69G**	57.67M	660MB	0.7516	0.7583	0.7412	0.7991	**50.92G**	57.69M	660MB
Ours	**0.7608**	0.7633	**0.7476**	**0.8044**	100.01G	**16.01M**	**231MB**	**0.7588**	**0.7668**	**0.7479**	**0.8031**	116.68G	**18.68M**	**270MB**

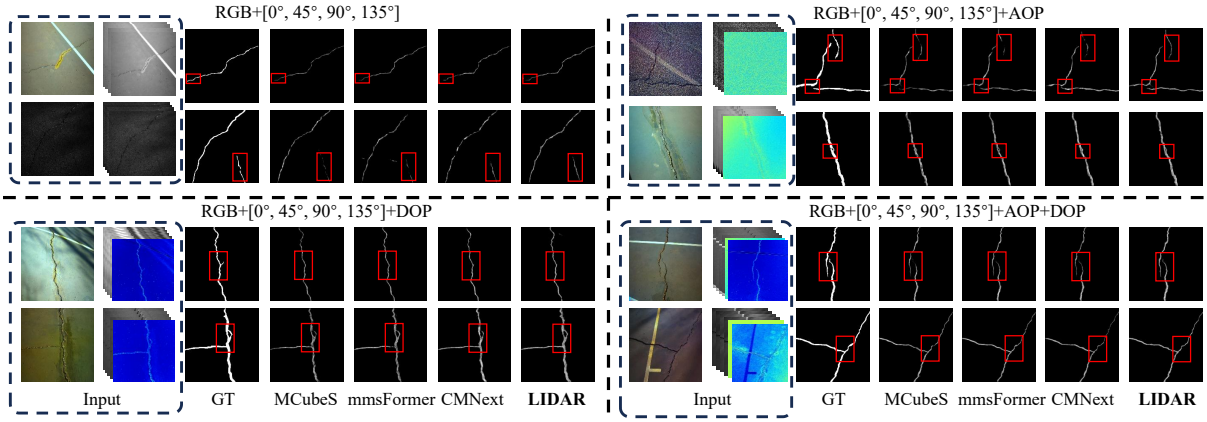


Figure 5: Visualization comparison for multi-modal inputs. Red boxes mark critical regions.

Table 3: Complexity comparison for dual-modal inputs.

Method	Year	FLOPs	Params	Size
MCubeS[18]	CVPR 2022	165.59G	117.76M	2253MB
mmsFormer[50]	MICCAI 2022	50.71G	19.96M	279MB
CMNeXT[47]	CVPR 2023	44.76G	57.63M	660MB
CMX[46]	TITS 2023	56.99G	66.56M	762MB
CAINet[27]	TMM 2024	37.28G	10.31M	138MB
PrimKD[11]	MM 2024	114.42G	139.85M	1597MB
SMMCL[4]	WACV 2024	61.95G	72.86M	826MB
SSRS[28]	TGRS 2024	679.09G	615.64M	7322MB
Sigma[34]	WACV 2025	63.43G	41.68M	553MB
Ours	MM 2025	**33.33G**	**5.35M**	**78MB**

As shown in Table 3, LIDAR also achieves the lowest complexity under dual-modal input at 512×512 resolution. Compared with CAINet [27], it reduces FLOPs, Params, and model size by 10.59%, 48.10%, and 43.48%, respectively. These gains stem from the LDMK convolution, the LacaVSS, and the LD3CF module, which contributes only 0.25 GFLOPs and 0.02M Params. These components enable LIDAR to efficiently extract geometric and spatial cues across modalities and produce high-quality segmentation maps with detail continuity and noise suppression.

Comparative experiments on multiple modalities. As shown in Table 2 and Figure 5, LIDAR achieves the best segmentation performance across all four polarization modality combinations compared to existing SOTA methods. When combining RGB with

Table 4: Comparison of different convolution types.

Conv Type	ODS	OIS	F1	mIoU	FLOPs	Params	Size
Common	0.8099	0.8126	0.8075	0.8391	156.14G	20.56M	241MB
DSConv	0.8074	0.8115	0.8082	0.8370	47.66G	6.89M	85MB
BottConv	0.8076	0.8137	0.8068	0.8370	39.33G	5.92M	**74MB**
LDMK	**0.8213**	**0.8237**	**0.8204**	**0.8465**	**33.33G**	**5.35M**	76MB

polarization angles (0°, 45°, 90°, and 135°), it surpasses the second-best method by an average of 0.71% in F1 score and 0.50% in mIoU. This highlights LIDAR's ability to leverage reflectance differences across polarization angles for extracting critical morphological cues from diverse surface materials. When using RGB in combination with all four angles and either AoP or DoP, LIDAR further outperforms CMNext [47], the second-best model, by 0.50% in F1 and 0.29% in mIoU on average. These results demonstrate LIDAR's strength in comprehensively capturing and fusing structural and textural cues encoded in polarization reflection behavior, microstructural directionality, and intensity variation.

In terms of computational efficiency, while LIDAR incurs slightly higher FLOPs under multimodal input, it maintains the lowest parameter count and model size among all methods. For example, under full-modal input, LIDAR reduces parameter count and model size by 67.63% and 59.09% compared to CMNext [47]. This efficiency is due to the lightweight LDMK design, LacaVSS's adaptive modeling of crack topology and textures, and the LD3CF module's capability to suppress irrelevant noise while enhancing key spatial and frequency-domain features. Together, these components enable

Table 5: Performance comparison with scanning strategies.

Scan Type	ODS	OIS	F1	mIoU	Delay Time
Para	0.7923	0.7935	0.7901	0.8269	1.63E-03s
Diag	0.8013	0.8030	0.7984	0.8332	2.17E-03s
ParaSnake	0.7993	0.8010	0.7971	0.8317	2.21E-03s
DiagSnake	0.8067	0.8093	0.8033	0.8366	2.40E-03s
bi_ParaSnake	0.7976	0.8010	0.7946	0.8303	1.55E-03s
bi_DiagSnake	0.7989	0.8002	0.7980	0.8319	6.25E-04s
SASS	0.8116	0.8158	0.8081	0.8406	2.16E-03s
w/o pre	0.8198	0.8232	0.8193	0.8458	2.50E-02s
w/o pre&integral	0.8189	0.8205	0.8186	0.8454	6.34E-02s
EDG-SS	**0.8213**	**0.8237**	**0.8204**	**0.8465**	**7.15E-07s**

Table 6: Performance comparison of different components combinations in LD3CF.

AFDP	DualPool	CrossGate	ODS	OIS	F1	mIoU
✓	✗	✗	0.7919	0.7932	0.7903	0.8265
✗	✓	✗	0.7894	0.7911	0.7891	0.8245
✗	✗	✓	0.7918	0.7929	0.7921	0.8273
✓	✓	✗	0.8073	0.8107	0.8050	0.8372
✓	✗	✓	0.8075	0.8094	0.8052	0.8371
✗	✓	✓	0.8119	0.8145	0.8102	0.8403
✓	✓	✓	**0.8213**	**0.8237**	**0.8204**	**0.8465**

LIDAR to produce high-quality segmentation maps with minimal computational overhead.

4.4 Ablation Studies

We performed ablation experiments on the CrackDepth dataset.

Performance comparison under different convolution types. Table 4 lists the performance of LIDAR with different convolution types, including common convolution, DSConv [31] and BottConv [22]. Our LDMK achieves the lowest FLOPs and parameter count while delivering the best performance across all metrics, including ODS, OIS, F1 score, and mIoU. Notably, compared to common convolution, LDMK reduces FLOPs, Params, and model size by 78.64%, 73.97%, and 68.05%, respectively. These results confirm that LDMK not only enhances the extraction of morphological cues across modalities but also significantly lowers computational overhead.

Ablation studies with different scanning strategies. Table 5 lists the performance comparison of LIDAR using the proposed EDG-SS and several classical scanning strategies, including parallel, diagonal, parallel snake, diagonal snake, bidirectional scanning, and SASS [22]. When EDG-SS is applied, LIDAR achieves the best results across all metrics, including ODS, OIS, F1 score, and mIoU, surpassing SASS by 1.19%, 0.97%, 1.52%, and 0.70%, respectively. It is noteworthy that models using conventional scanning strategies such as parallel, diagonal, and their snake variants suffer a clear performance drop. This indicates that fixed scanning paths lack adaptability and fail to effectively perceive the irregular texture cues present in different crack modalities, limiting the model's capacity to characterize crack variations and suppress noise. While bidirectional scanning shows moderate improvement, the disruption of contextual dependencies in the central region of the image reduces its ability to model continuous semantic structures.

In terms of sequence generation latency, EDG-SS significantly outperforms all other strategies. It is 874 times faster (7.15E-07s VS 6.25E-04s) than the next-fastest method, bidirectional diagonal snake scanning. Compared with variants that omit the integral image or pre-scanning mechanism, EDG-SS achieves substantial acceleration. These results demonstrate that EDG-SS, by leveraging an integral-image-based pre-scanning mechanism and mask-guided path selection, effectively reduces latency and improves the model's ability to capture continuous crack textures and suppress irrelevant background, thereby enhancing LIDAR's segmentation performance in irregular crack regions.

Ablation Study on Components of LD3CF. Table 6 lists the performance of LIDAR under different configurations of the LD3CF module components. Since LD3CF introduces only 0.25 GFLOPs and 0.02M parameters, the computational cost of each component is negligible. LIDAR achieved the best performance in all evaluation metrics when including all three components. Compared with the variant without AFDP, the complete model improves ODS, OIS, F1, and mIoU by 1.16%, 1.13%, 1.26%, and 0.74%, respectively. This demonstrates that AFDP enhances crack region representation by reinforcing high-frequency features and suppressing low-frequency background noise, leading to clearer textures for subsequent processing. Furthermore, removing both the dual pooling fusion and cross-level gating results in a significant performance drop, with F1 and mIoU decreasing by 3.81% and 2.42%, respectively. These findings indicate that the dual pooling fusion module effectively captures structural and texture cues in the spatial domain, while the cross-level gating mechanism enables adaptive and hierarchical interaction of multimodal features. Together, these components contribute to the generation of high-quality pixel-level crack segmentation maps.

5 Conclusion

In this paper, we propose LIDAR, a pioneering Lightweight Adaptive Cue-Aware Vision Mamba network for pixel-level multimodal structural crack segmentation. LIDAR integrates the LacaVSS and LD3CF modules and replaces most convolutional operations with the proposed LDMK convolution, enabling efficient extraction and fusion of geometric, morphological, and textural cues across multiple modalities at low computational cost. LacaVSS, guided by the EDG-SS, dynamically prioritizes crack regions for more effective modeling. LD3CF enhances segmentation quality by combining AFDP and a dual pooling fusion module, enabling robust spatial-frequency cue extraction and background suppression, ultimately producing high-quality segmentation maps. Extensive experiments on three multimodal crack datasets demonstrate that LIDAR consistently outperforms SOTA methods in both performance and efficiency. For example, LIDAR demonstrates optimal performance across a variety of datasets and combinations of multiple modalities, consistently achieving best results in diverse scenarios. In future work, we aim to further improve LIDAR's adaptability to modality-specific heterogeneity and explore more efficient learnable scanning strategies to enhance generalization on diverse crack datasets.

6 Acknowledgement

This work was supported by the National Natural Science Foundation of China (NSFC) under Grants 62272342, 62020106004, 62306212, and T2422015; the Tianjin Natural Science Foundation under Grants 23JCJQJC00070 and 24PTLYHZ00320; and the Marie Skłodowska-Curie Actions (MSCA) under Project No. 101111188.

References

[1] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. 2022. Multimae: Multi-modal multi-task masked autoencoders. In *European Conference on Computer Vision*. Springer, 348–367.

[2] Zhuangzhuang Chen, Zhuonan Lai, Jie Chen, and Jianqiang Li. 2024. Mind marginal non-crack regions: Clustering-inspired representation learning for crack segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12698–12708.

[3] Xu Cheng, Tian He, Fan Shi, Meng Zhao, Xiufeng Liu, and Shengyong Chen. 2023. Selective feature fusion and irregular-aware network for pavement crack detection. *IEEE Transactions on Intelligent Transportation Systems* 25, 5 (2023), 3445–3456.

[4] Xiaoyu Dong and Naoto Yokoya. 2024. Understanding dark scenes by contrasting multi-modal observations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 840–850.

[5] Rohit Girdhar, Mannat Singh, Nikhila Ravi, Laurens Van Der Maaten, Armand Joulin, and Ishan Misra. 2022. Omnivore: A single model for many visual modalities. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 16102–16112.

[6] Karan Goel, Albert Gu, Chris Donahue, and Christopher Ré. 2022. It's raw! audio generation with state-space models. In *International conference on machine learning*. PMLR, 7616–7633.

[7] Albert Gu and Tri Dao. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752* (2023).

[8] Albert Gu, Karan Goel, Ankit Gupta, and Christopher Ré. 2022. On the parameterization and initialization of diagonal state space models. *Advances in Neural Information Processing Systems* 35 (2022), 35971–35983.

[9] Albert Gu, Isys Johnson, Aman Timalsina, Atri Rudra, and Christopher Ré. 2022. How to train your hippo: State space models with generalized orthogonal basis projections. *arXiv preprint arXiv:2206.12037* (2022).

[10] Hang Guo, Jinmin Li, Tao Dai, Zhihao Ouyang, Xudong Ren, and Shu-Tao Xia. 2024. Mambair: A simple baseline for image restoration with state-space model. In *European Conference on Computer Vision*. Springer, 222–241.

[11] Zhiwei Hao, Zhongyu Xiao, Yong Luo, Jianyuan Guo, Jing Wang, Li Shen, and Han Hu. 2024. PrimKD: Primary Modality Guided Multimodal Fusion for RGB-D Semantic Segmentation. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 1943–1951.

[12] Shruti Jadon. 2020. A survey of loss functions for semantic segmentation. In *2020 IEEE conference on computational intelligence in bioinformatics and computational biology (CIBCB)*. IEEE, 1–7.

[13] Achref Jaziri, Martin Mundt, Andres Fernandez, and Visvanathan Ramesh. 2024. Designing a hybrid neural system to learn real-world crack segmentation from fractal-based simulation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 8636–8646.

[14] Hong Lang, Ye Yuan, Jiang Chen, Shuo Ding, Jian John Lu, and Yong Zhang. 2024. Augmented concrete crack segmentation: Learning complete representation to defend background interference in concrete pavements. *IEEE Transactions on Instrumentation and Measurement* (2024).

[15] Boyun Li, Haiyu Zhao, Wenxin Wang, Peng Hu, Yuanbiao Gou, and Xi Peng. 2025. MaIR: A Locality- and Continuity-Preserving Mamba for Image Restoration. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[16] Shuxin Li, Xu Cheng, Fan Shi, Hanwei Zhang, Hongning Dai, Houxiang Zhang, and Shengyong Chen. 2025. A Novel Robustness-Enhancing Adversarial Defense Approach to AI-Powered Sea State Estimation for Autonomous Marine Vessels. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 55, 1 (2025), 28–42.

[17] Yan Li, Yifei Xing, Xiangyuan Lan, Xin Li, Haifeng Chen, and Dongmei Jiang. 2025. AlignMamba: Enhancing Multimodal Mamba with Local and Global Cross-modal Alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[18] Yupeng Liang, Ryosuke Wakaki, Shohei Nobuhara, and Ko Nishino. 2022. Multimodal material segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19800–19808.

[19] Jianghai Liao, Yuanhao Yue, Dejin Zhang, Wei Tu, Rui Cao, Qin Zou, and Qingquan Li. 2022. Automatic tunnel crack inspection using an efficient mobile imaging module and a lightweight CNN. *IEEE Transactions on Intelligent Transportation Systems* 23, 9 (2022), 15190–15203.

[20] Fangyu Liu, Jian Liu, and Linbing Wang. 2022. Asphalt pavement crack detection based on convolutional neural network and infrared thermography. *IEEE Transactions on Intelligent Transportation Systems* 23, 11 (2022), 22145–22155.

[21] Fangyu Liu, Jian Liu, and Linbing Wang. 2022. Asphalt pavement fatigue crack severity classification by infrared thermography and deep learning. *Automation in Construction* 143 (2022), 104575.

[22] Hui Liu, Chen Jia, Fan Shi, Xu Cheng, and Shengyong Chen. 2025. SCSegamba: Lightweight Structure-Aware Vision Mamba for Crack Segmentation in Structures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[23] Huajun Liu, Xiangyu Miao, Christoph Mertz, Chengzhong Xu, and Hui Kong. 2021. Crackformer: Transformer network for fine-grained crack detection. In *Proceedings of the IEEE/CVF international conference on computer vision*. 3783–3792.

[24] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, and Yunfan Liu. 2024. VMamba: Visual State Space Model. *arXiv preprint arXiv:2401.10166* (2024).

[25] Yahui Liu, Jian Yao, Xiaohu Lu, Renping Xie, and Li Li. 2019. DeepCrack: A deep hierarchical feature learning architecture for crack segmentation. *Neurocomputing* 338 (2019), 139–153.

[26] Xiaoyong Lu and Songlin Du. 2025. JamMa: Ultra-lightweight Local Feature Matching with Joint Mamba. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[27] Ying Lv, Zhi Liu, and Gongyang Li. 2024. Context-aware interaction network for rgb-t semantic segmentation. *IEEE Transactions on Multimedia* 26 (2024), 6348–6360.

[28] Xianping Ma, Xiaokang Zhang, Man-On Pun, and Ming Liu. 2024. A multilevel multimodal fusion transformer for remote sensing semantic segmentation. *IEEE Transactions on Geoscience and Remote Sensing* (2024).

[29] Liu Mushui, Jun Dan, Ziqian Lu, Yunlong Yu, Yingming Li, and Xi Li. 2024. CM-UNet: Hybrid CNN-Mamba UNet for Remote Sensing Image Semantic Segmentation. *arXiv preprint arXiv:2405.10530* (2024).

[30] Y. Peng, M. Sonka, and D. Z. Chen. 2024. Group Vision Transformer. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 2623–2631.

[31] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. 4510–4520.

[32] Daniel Seichter, Söhnke Benedikt Fischedick, Mona Köhler, and Horst-Michael Groß. 2022. Efficient multi-task rgb-d scene analysis for indoor environments. In *2022 International joint conference on neural networks (IJCNN)*. IEEE, 1–10.

[33] Carole H Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M Jorge Cardoso. 2017. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. Springer, 240–248.

[34] Zifu Wan, Yuhao Wang, Silong Yong, Pingping Zhang, Simon Stepputtis, Katia Sycara, and Yaqi Xie. 2025. Sigma: Siamese Mamba Network for Multi-Modal Semantic Segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*.

[35] H. Wang, X. Liang, T. Zhang, et al. 2024. PSSD-Transformer: Powerful Sparse Spike-Driven Transformer for Image Semantic Segmentation. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 758–767.

[36] Jin Wang, Zhigao Zeng, Pradip Kumar Sharma, Osama Alfarraj, Amr Tolba, Jianming Zhang, and Lei Wang. 2024. Dual-path network combining CNN and transformer for pavement crack segmentation. *Automation in Construction* 158 (2024), 105217.

[37] Yingqian Wang, Longguang Wang, Zhengyu Liang, Jungang Yang, Wei An, and Yulan Guo. 2022. Occlusion-aware cost constructor for light field depth estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 19809–19818.

[38] Chao Xiang, Jingjing Guo, Ran Cao, and Lu Deng. 2023. A crack-segmentation algorithm fusing transformers and convolutional neural networks for complex detection scenarios. *Automation in Construction* 152 (2023), 104894.

[39] Chaodong Xiao, Minghan Li, Zhengqiang Zhang, Deyu Meng, and Lei Zhang. 2025. Spatial-Mamba: Effective Visual State Space Models via Structure-Aware State Fusion. In *The Thirteenth International Conference on Learning Representations*.

[40] X. Xue, D. Yu, L. Liu, et al. 2023. Transformer-based Open-world Instance Segmentation with Cross-task Consistency Regularization. In *Proceedings of the 31st ACM International Conference on Multimedia*. 2507–2515.

[41] Chenhongyi Yang, Zehui Chen, Miguel Espinosa, Linus Ericsson, Zhenyu Wang, Jiaming Liu, and Elliot J. Crowley. 2024. PlainMamba: Improving Non-Hierarchical Mamba in Visual Recognition. arXiv:2403.17695 [cs.CV]

[42] Jun Yang, Lizhi Bai, Yaoru Sun, Chunqi Tian, Maoyu Mao, and Guorun Wang. 2023. Pixel difference convolutional network for RGB-D semantic segmentation. *IEEE Transactions on Circuits and Systems for Video Technology* 34, 3 (2023), 1481–1492.

[43] Hanrong Ye and Dan Xu. 2022. Taskprompter: Spatial-channel multi-task prompting for dense scene understanding. In *The Eleventh International Conference on Learning Representations*.

[44] Weihao Yu and Xinchao Wang. 2025. MambaOut: Do We Really Need Mamba for Vision?. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[45] Hang Zhang, Allen A Zhang, Zishuo Dong, Anzheng He, Yang Liu, You Zhan, and Kelvin CP Wang. 2024. Robust semantic segmentation for automatic crack detection within pavement images using multi-mixing of global context and local image features. *IEEE Transactions on Intelligent Transportation Systems* 25, 9 (2024), 11282–11303.

[46] Jiaming Zhang, Huayao Liu, Kailun Yang, Xinxin Hu, Ruiping Liu, and Rainer Stiefelhagen. 2023. CMX: Cross-modal fusion for RGB-X semantic segmentation

with transformers. *IEEE Transactions on intelligent transportation systems* 24, 12 (2023), 14679–14694.

[47] Jiaming Zhang, Ruiping Liu, Hao Shi, Kailun Yang, Simon Reiß, Kunyu Peng, Haodong Fu, Kaiwei Wang, and Rainer Stiefelhagen. 2023. Delivering arbitrary-modal semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1136–1147.

[48] Jingwei Zhang, Anh Tien Nguyen, Xi Han, Vincent Quoc-Huy Trinh, Hong Qin, Dimitris Samaras, and Mahdi S Hosseini. 2025. 2DMamba: Efficient State Space Model for Image Representation with Applications on Giga-Pixel Whole Slide Image Classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[49] Tianjie Zhang, Donglei Wang, and Yang Lu. 2023. ECSNet: An accelerated real-time image segmentation CNN architecture for pavement crack detection. *IEEE Transactions on Intelligent Transportation Systems* 24, 12 (2023), 15105–15112.

[50] Yao Zhang, Nanjun He, Jiawei Yang, Yuexiang Li, Dong Wei, Yawen Huang, Yang Zhang, Zhiqiang He, and Yefeng Zheng. 2022. mmformer: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 107–117.

[51] Zixiang Zhao, Haowen Bai, Jiangshe Zhang, Yulun Zhang, Kai Zhang, Shuang Xu, Dongdong Chen, Radu Timofte, and Luc Van Gool. 2024. Equivariant multi-modality image fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 25912–25921.

[52] Wujie Zhou, Enquan Yang, Jingsheng Lei, Jian Wan, and Lu Yu. 2022. PGDENet: Progressive guided fusion and depth enhancement network for RGB-D indoor scene parsing. *IEEE Transactions on Multimedia* 25 (2022), 3483–3494.

[53] X. Zhou and T. Chen. 2024. Bsbp-RWKV: Background Suppression with Boundary Preservation for Efficient Medical Image Segmentation. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 4938–4946.

[54] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. 2024. Vision Mamba: Efficient Visual Representation Learning with Bidirectional State Space Model. In *International Conference on Machine Learning*.