DS4Net: Pseudo-Depth Injection and Spectral-Spatial State Space Modeling for Label-Efficient Crack Segmentation

Xiaohu Chen[1], Ruiqiang Xiao[2], Hongyuan Liao[2], Wei Li[1] and Mingzhu Wang[2*]

[1]School of Transportation, Southeast University, China
[2]Department of Architecture and Civil Engineering, City University of Hong Kong, Hong Kong SAR, China

[*]Corresponding Author: mwang489@cityu.edu.hk,

Abstract -

Crack segmentation is a critical yet challenging task in structural health monitoring, as cracks typically appear as thin, low-contrast structures that are easily confounded by complex background textures. Although recent deep learning methods have substantially improved crack segmentation accuracy in conventional RGB imagery, they often overfit dataset-specific textures, resulting in poor cross-domain generalization. As an alternative, RGB-D methods provide complementary geometric cues but are limited by costly and imperfect depth sensors. To address these limitations, we propose the Dual-Stream Spectral-Spatial State Space Network (DS4Net), a hardware-free framework that injects pseudo-depth priors estimated by a frozen depth foundation model into crack segmentation. First, instead of physical depth, we introduce a Pseudo-Depth Injection strategy driven by the zero-shot capability of Depth Anything V2, effectively transferring universal geometric priors to the crack segmentation task. Second, to resolve the conflict between global topology and local edge sharpness, we design a Dual-Path Frequency-Mamba (DFM) Block. It synergizes the linear complexity of State Space Models (Mamba) for long-range dependency modeling with a Fast Fourier Transform (FFT) branch that explicitly enhances high-frequency edge details in the spectral domain. Third, to handle cross-modal distribution discrepancy and varying modality reliability, we propose an Information-Theoretic KL-Gating mechanism, which uses Kullback-Leibler divergence to adaptively weight RGB and pseudo-depth features, thereby suppressing unreliable signals and promoting more informative cues. Experimental results show that DS4Net outperforms state-of-the-art RGB and RGB-D methods in both in-domain and cross-domain settings.

Keywords -

Crack segmentation; Pseudo-depth injection; Multimodal fusion; Visual state space model; Domain generalization

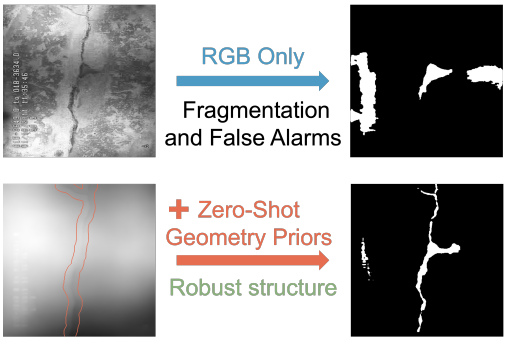

Figure 1. **Motivation**. RGB-only crack segmentation is vulnerable to environmental noise and domain shifts, causing fragmented masks and false alarms. In contrast, pseudo-depth estimated by a depth foundation model reveals more stable geometric crack structures. By injecting such zero-shot geometric priors, DS4Net achieves robust segmentation without auxiliary sensors.

1 Introduction

Critical infrastructure, such as bridges, tunnels, and pavement systems, faces an accelerating aging crisis globally. As the earliest and most prevalent symptom of structural degradation, surface cracks serve as vital indicators for Structural Health Monitoring (SHM)[1]. Traditionally, crack inspection relied on manual visual assessments, which are labor-intensive, subjective, and often hazardous [2]. In recent years, the rapid advancement of deep learning has revolutionized this field, enabling automated pixel-level crack segmentation. State-of-the-art methods, ranging from Convolutional Neural Networks (CNNs) [3, 4] to recent Vision Transformers (ViTs) [5] and State Space Models (SSMs/Mamba) [6, 7], have achieved impressive performance on standard benchmarks.

Despite these advancements, a fundamental bottleneck remains: most existing approaches rely exclusively on the RGB modality. While effective in controlled environments, purely RGB-based models suffer from inherent visual ambiguity in complex, real-world scenarios [8]. Cracks, characterized by low pixel intensity, are often visually indistinguishable from shadows, oil stains, water spots, and dark aggregate textures. This ambiguity leads to two critical issues: (1) In-domain precision loss: Even within standard datasets, models frequently generate false positives in cluttered regions (e.g., mistaking a tire mark for a crack) or fail to detect fine cracks in low-light areas. (2) Cross-domain generalization failure: RGB models tend to overfit to the color distributions and textures of the training domain (e.g., asphalt), leading to severe performance degradation when transferred to unseen materials (e.g., concrete) or varying lighting conditions. This suggests

that relying solely on appearance cues is insufficient for robust and precise segmentation. As shown in Fig 1, under cross-domain settings and challenging illumination, RGB-only crack segmenters often yield fragmented masks and false alarms.

In principle, integrating depth information can resolve this ambiguity, as cracks represent physical geometric discontinuities that are invariant to illumination and surface texture. However, the adoption of RGB-D fusion has been hindered by a practical hardware paradox. High-precision sensors such as LiDAR are prohibitively expensive for large-scale deployment, while consumer-grade depth cameras often suffer from limited resolution and sensor-camera misalignment, introducing artifacts that impair boundary segmentation [9]. Fortunately, the emergence of Vision Foundation Models, such as Depth Anything V2[10], offers a paradigm shift. Trained on massive-scale data, these models possess strong zero-shot capability to estimate fine-grained geometric structures from single RGB images. This presents a unique opportunity: **can we distill these universal, hardware-free geometric priors to improve both the accuracy and robustness of crack segmentation?** Indeed, Fig 1 suggests that pseudo-depth predicted by Depth Anything V2 highlights a more stable crack structure than RGB, making it a promising source of geometric priors for robust segmentation.

To harness this potential, we propose the Dual-Stream Spectral-Spatial State Space Network (DS4Net), a novel framework that synergizes geometric priors with efficient structural modeling. Our approach addresses two key technical challenges. First, simply concatenating RGB and pseudo-depth features is suboptimal due to their distinct characteristics: RGB contains rich texture details but high noise, while pseudo-depth offers clean structure but may lack high-frequency edge sharpness. To resolve this, we design the Dual-Path Frequency-Mamba (DFM) Block. This module combines the linear complexity of State Space Models (Mamba) for modeling long-range crack topology with a Fast Fourier Transform (FFT) branch that enhances high-frequency signals in the spectral domain, preserving fine crack edges. Second, to handle the dynamic reliability of the two modalities (e.g., RGB is unreliable in shadows; pseudo-depth is unreliable on flat markings), we introduce an Information-Theoretic KL-Gating mechanism. By dynamically measuring the Kullback-Leibler divergence between feature distributions, the network adaptively re-weights fusion, suppressing unreliable modality cues and prioritizing the more informative modality under challenging conditions.

The main contributions of this work are summarized as follows:

- We propose a hardware-free RGB-D segmentation framework that leverages zero-shot geometric priors from Foundation Models (Depth Anything V2), significantly enhancing robustness without the cost of physical sensors.
- We design the DFM Block, a hybrid unit that integrates the global modeling capability of Mamba with spectral-domain learning (FFT). This design effectively balances long-range crack connectivity with local boundary precision.
- We introduce an Information-Theoretic KL-Gating mechanism that dynamically regulates multimodal fusion based on distributional divergence.
- Extensive experiments demonstrate that DS4Net achieves State-of-the-Art (SOTA) results on standard benchmarks and exhibits superior robustness in cross-domain scenarios compared to existing CNN, Transformer, and RGB-D-based methods.

2 Related Work

2.1 Crack Segmentation using deep learning methods

Deep learning has shifted crack segmentation from handcrafted pipelines to pixel-level prediction using CNN and Transformer models. CNN approaches such as FCN [11], U-Net [3], DeepLabV3+ [4], and DeepCrack [12] learn strong local cues but may produce fragmented masks for long curvilinear cracks under occlusion or low contrast. Transformer designs such as CrackFormer [13] and SCDeepLab [2] improve connectivity via self-attention, yet their computational and memory costs hinder large-scale training and edge deployment. Both paradigms rely mainly on RGB appearance, making them sensitive to illumination changes and background clutter such as shadows, oil stains, water marks, and patch repairs, which limits cross-domain generalization and motivates the use of complementary cues beyond photometric appearance.

2.2 Visual State Space Models

Visual State Space Models, abbreviated as VSSMs, offer a linear-time alternative to self-attention, capturing global context through selective scanning [14, 15]. Vision adaptations such as VMamba [6] extend selective scanning to multidirectional two-dimensional passes and have been integrated into encoder-decoder segmenters such as U-Mamba [16] and VM-UNet [17] to improve the accuracy-efficiency balance in dense prediction. For crack segmentation, CrackMamba [7] and SCSegamba [18] adapt serialization and scanning to curvilinear topology, improving long-range continuity while remaining more deployment-friendly than Transformers. Nonetheless, existing VSSM-based pipelines operate mainly in the spatial domain and may under-represent high-frequency boundaries of extremely thin cracks, motivating explicit frequency cues.

2.3 Foundation Models for Geometry Estimation

Monocular depth estimation has evolved from supervised learning on scarce RGB-depth pairs to foundation models that learn transferable geometric priors from diverse data. MiDaS [19] improves cross-dataset generalization via mixed-domain training, while foundation depth models further strengthen zero-shot transfer. Depth Anything V2 [10] improves fine-detail recovery through synthetic-to-real distillation, helping resolve thin structural discontinuities. These priors enable hardware-free pseudo-depth for crack inspection without costly RGB-D sensors such as LiDAR [20] or active depth cameras [9], improving robustness and cross-domain generalization.

2.4 Multi-modal Fusion and Frequency Domain Learning

Frequency-domain learning uses the Fast Fourier Transform (FFT) to emphasize high-frequency edges while decoupling low-frequency illumination, benefiting crack segmentation dominated by thin boundaries [21]. Yet frequency cues are rarely integrated with VSSM or Mamba encoders, leaving the combination of linear-complexity topology modeling and explicit edge enhancement underexplored. Meanwhile, RGB and depth priors are often fused by concatenation, summation, or cross-attention [22, 8], but such designs assume both modalities are reliable, even though RGB suffers from shadows and stains and pseudo-depth fails on textureless or reflective surfaces. Measuring cross-modal distributional discrepancy can adaptively weight modalities for robust inspection.

Recent works introduce multimodal cues or frequency analysis for crack segmentation to improve robustness and boundary quality [22, 8, 21]. However, multimodal systems often require extra sensors, while VSSM or frequency designs still rely mainly on RGB. Moreover, spatial token mixing alone fails to explicitly preserve high-frequency edges. To bridge this gap, we transfer foundation depth priors, combine VSSM global modeling with spectral enhancement, and use discrepancy-aware gating to weight modalities, enabling robust crack segmentation.

3 The Proposed DS4Net

3.1 Overview of DS4Net

We propose the **Dual-Stream Spectral-Spatial State Space Network (DS4Net)** to achieve robust crack segmentation without physical depth sensors. The overall architecture of DS4Net is depicted in Fig 2. The architecture follows a hierarchical encoder-decoder design driven by two parallel streams: the RGB Stream (processing texture) and the Pseudo-Depth Stream (processing geometry injected from the frozen Depth Anything V2 foundation model). To effectively fuse these heterogeneous modalities, we introduce two key components at each encoder stage: the **Dual-Path Frequency-Mamba (DFM) Block** for intra-modal feature refinement, and the **Information-Theoretic KL-Gating** for inter-modal adaptive fusion.

3.2 Dual-Path Frequency-Mamba (DFM) Block

Cracks exhibit a unique dual characteristic: they possess global topological continuity but define their presence through high-frequency local edges. To capture these complementary cues, the DFM Block processes the pre-fused feature through two parallel branches: a spatial Mamba branch for modeling long-range topological dependencies and a spectral attention branch for enhancing high-frequency boundary cues. The outputs of both branches are then concatenated and fused by a 1×1 convolution.

3.2.1 Spatial Mamba Branch

We employ the Visual State Space (VSS) module based on the Mamba [14] architecture to capture long-range dependencies with linear complexity $O(N)$. This branch ensures that the connectivity of cracks is preserved even when partially occluded. Since the implementation follows standard SS2D scanning mechanisms [6], we omit the details here for brevity.

3.2.2 Learnable Spectral Attention Branch

Standard convolutions often struggle to decouple high-frequency crack details from low-frequency background variations, such as illumination gradients. We propose a **Learnable Spectral Attention** mechanism to explicitly filter spectral components.

Let $X \in \mathbb{R}^{H\times W\times C}$ be the input feature map. We first transform it to the frequency domain via 2D Fast Fourier Transform (FFT):

$$F(X) = \mathrm{FFT}(X) \in \mathbb{C}^{H\times W\times C} \tag{1}$$

We decouple the complex spectrum into Amplitude $\mathcal{A}(X)$ and Phase $\mathcal{P}(X)$ components:

$$\mathcal{A}(X) = |F(X)|, \quad \mathcal{P}(X) = \angle F(X) \tag{2}$$

To adaptively highlight crack-sensitive frequencies, we employ a channel-wise attention mechanism in the spectral domain. We first aggregate spectral statistics using both Average Pooling and Max Pooling along the channel dimension, followed by a 7×7 convolution layer to generate the attention masks. Formally, the Amplitude Attention Map $M_{\mathcal{A}}$ is computed as:

$$M_{\mathcal{A}} = \sigma\left(\mathrm{Conv}_{7\times 7}\left([\mathrm{AvgPool}_c(\mathcal{A}); \mathrm{MaxPool}_c(\mathcal{A})]\right)\right) \tag{3}$$

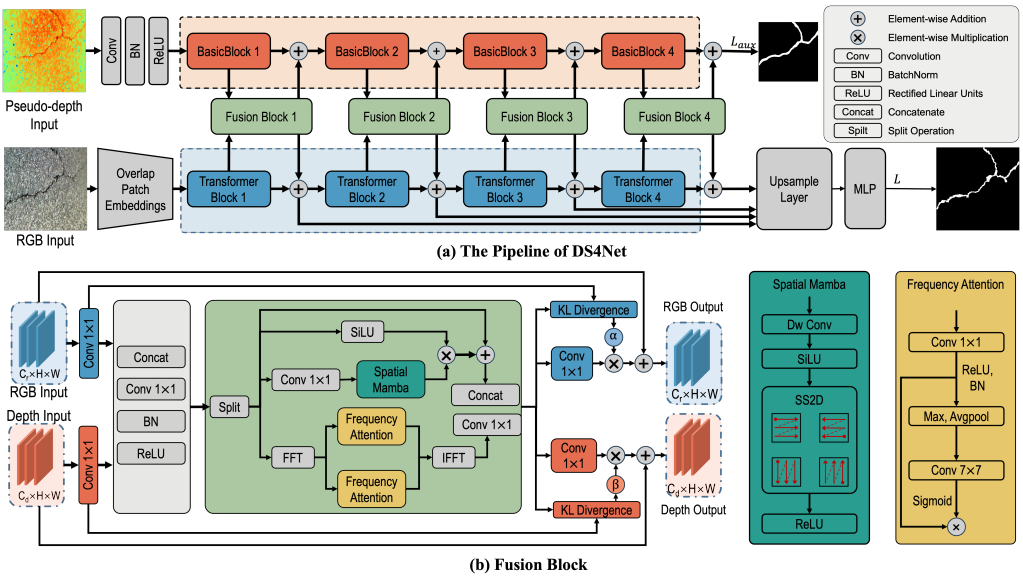

Figure 2. (a) Architecture of the proposed DS4Net, including a dual-stream encoder–decoder with an RGB stream and a pseudo-depth stream generated by a frozen depth foundation model. (b) The fusion block integrates the Dual-Path Frequency-Mamba (DFM) module for spectral–spatial refinement and employs KL-Gating to adaptively reweight cross-modal fusion.

Similarly, the Phase Attention Map $M_{\mathcal{P}}$ is computed as:

$$M_{\mathcal{P}} = \sigma\left(\text{Conv}_{7\times7}\left([\text{AvgPool}_c(\mathcal{P}); \text{MaxPool}_c(\mathcal{P})]\right)\right) \tag{4}$$

where $[\cdot;\cdot]$ denotes concatenation along the channel dimension, and σ is the Sigmoid activation function. These masks effectively learn to suppress background noise in the amplitude spectrum while preserving structural cues in the phase spectrum. The modulated features are reconstructed via Inverse FFT:

$$X_{freq} = \text{IFFT}\left((\mathcal{A}(X) \odot M_{\mathcal{A}}) \cdot e^{j(\mathcal{P}(X)\odot M_{\mathcal{P}})}\right) \tag{5}$$

This branch acts as a frequency-domain high-pass filter, explicitly sharpening crack boundaries before fusion with the parallel Mamba branch.

3.3 Information-Theoretic KL-Gating

Simply adding or concatenating RGB and Pseudo-Depth features assumes both modalities are equally reliable. However, in real-world scenarios, one modality may fail (e.g., RGB in shadows, Depth on flat stains). To address this, we propose an **Information-Theoretic Gating** mechanism that dynamically assesses the "Information Gain" of each modality at the image level.

3.3.1 Global Distribution Modeling

We interpret the feature channels as a semantic probability distribution. Let $F_{rgb}, F_{depth} \in \mathbb{R}^{B\times C\times H\times W}$ be the features from the two branches, while F_{mamba} represents the intermediate feature processed by the Mamba fusion block that serves as a context-aware consensus.

To capture the **holistic reliability**, we first apply Global Average Pooling (GAP) to compress the spatial dimensions, followed by a Softmax normalization along the channel dimension:

$$p_k = \text{Softmax}(\text{GAP}(F_k)), \quad k \in \{rgb, depth, mamba\} \tag{6}$$

where $p_k \in \mathbb{R}^{B\times C}$ represents the global channel activation distribution of modality k.

3.3.2 Divergence-Based Weight Generation

We employ the **Kullback-Leibler (KL) Divergence** [23] to measure how much unique information each modality contributes relative to the fusion consensus (p_{mamba}). A higher divergence implies that the modality contains distinct, potentially complementary information that the current consensus lacks.

$$D_{KL}(p_k||p_{mamba}) = \sum_{c=1}^{C} p_k^c \log \frac{p_k^c}{p_{mamba}^c + \epsilon} \tag{7}$$

This yields a scalar reliability score for each modality. The final fusion weights α_{rgb} and β_{depth} are generated by normalizing these divergence scores:

$$[\alpha_{rgb}, \beta_{depth}] = \text{Softmax}([D_{KL}(rgb), D_{KL}(depth)]) \tag{8}$$

3.3.3 Residual Injection

Finally, the fused features are dynamically re-weighted and injected back into the main streams via residual connections:

$$F_{out}^{rgb} = F_{in}^{rgb} + \alpha_{rgb} \cdot \Psi(F_{mamba}) \quad (9)$$

$$F_{out}^{depth} = F_{in}^{depth} + \beta_{depth} \cdot \Psi(F_{mamba}) \quad (10)$$

where Ψ denotes a projection layer (implemented as 1×1 convolution). This mechanism allows the network to prioritize the modality with higher information value for the current scene, achieving **Sample-Adaptive Fusion**.

4 Experiments and Results

4.1 Dataset

We evaluate DS4Net on a public RGB-T asphalt pavement crack segmentation benchmark [22], which provides RGB images and pixel-wise crack annotations. The benchmark contains 448 images with a resolution of 640×480, and we follow the standard split of 358 images for training and 90 images for testing using a predefined random seed to guarantee experimental consistency and reproducibility. To construct RGB-D inputs without additional sensing hardware, we employ Depth Anything V2 [10] to generate a pseudo-depth map for each RGB image, producing aligned RGB-D pairs for both training and evaluation.

To further assess cross-domain robustness, we additionally introduce the TUT dataset [24] as an unseen test set. Unlike the relatively structured benchmark, TUT features dense and cluttered backgrounds and cracks with more elaborate, intricate geometries, which better reflects challenging real-world conditions. It contains 1,408 RGB images covering diverse materials, backgrounds, and surface conditions; in our experiments, it is used exclusively for cross-domain testing without any fine-tuning.

4.2 Implementation Details

To enhance the diversity of the training data and improve the model's robustness, spatial, color, and numerical transformations were applied to the training set, such as random horizontal or vertical flips in spatial, randomly altering brightness or contrast in color. For images in the validation set, only resizing to 480×480 pixels and normalization processes were applied. The proposed DS4Net was implemented using the PyTorch framework and optimized with AdamW, incorporating a weight decay of 1e-4. A batch size of 8 and 150 training epochs were designated for training. All experiments were performed on an NVIDIA GeForce RTX 4090 for training and testing.

4.3 Evaluation Metrics

To accurately and objectively evaluate the performance of various models, we employed four widely used evaluation metrics in image segmentation: Dice, IoU (Intersection over Union), Accuracy and Precision [25]. Higher values for these metrics indicate superior segmentation performance. Simultaneously, these metrics are also used as the evaluation criteria for the RGB-D asphalt pavement crack segmentation benchmark.

4.4 Comparison with State-of-the-Art Methods

To validate the effectiveness of DS4Net, we compared it with nine representative methods on the Infrared dataset. Specifically, U-Net[3], DeepLabV3[4], DeepCrack[12], and CrackFormer[13] rely only on RGB inputs. In contrast, CRM_RGBT_Seg[26], PotCrackSeg, Evi-RoadSeg, CMNeXt[27], and IRFusionFormer take both RGB and depth images as inputs. Quantitative results are summarized in Table 1 and Table 2. Table 1 reports in-domain performance on the benchmark test set, whereas Table 2 evaluates cross-domain generalization on TUT. Qualitative visualizations are provided in Fig 3.

Table 1. In-domain quantitative comparison on the Infrared dataset. The best results are highlighted in **bold**, and the second-best results are underlined.

Model	Type	Dice ↑	IoU ↑	Acc. ↑	Prec. ↑
Unet	RGB	0.7890	0.6515	0.9794	0.8160
DeepLabV3	RGB	0.8337	0.7149	0.9828	0.8133
DeepCrack	RGB	0.8211	0.6964	0.9831	0.8806
CrackFormer	RGB	0.8487	0.7372	0.9847	0.8458
CRM_RGBTSeg	RGB-D	0.8391	0.7286	0.9809	0.8311
PotCrackSeg	RGB-D	0.8220	0.6978	0.8222	0.8109
Evi-RoadSeg	RGB-D	0.8380	0.7212	0.9775	0.8172
CMNeXt	RGB-D	0.8589	0.7527	0.9808	0.8565
IRFusionFormer	RGB-D	0.8795	0.7850	0.9879	0.8808
Ours	**RGB-D**	**0.8982**	**0.8152**	**0.9898**	**0.9053**

Results on in-domain setting. Table 1 shows that DS4Net delivers the best performance across all four metrics, achieving a Dice score of 0.8982 and an IoU of 0.8152. It substantially outperforms the strongest RGB-only baseline, CrackFormer, with Dice and IoU of 0.8487 and 0.7372, as well as the best RGB-D competitor, IRFusionFormer, with corresponding values of 0.8795 and 0.7850. DS4Net also improves accuracy and precision, suggesting more reliable crack delineation under in-domain setting.

Results on cross-domain setting. As shown in Table 2, DS4Net achieves the strongest cross-domain performance on TUT, attaining the highest Dice score of 0.6806, IoU of 0.5158, and accuracy of 0.9790. Although some methods yield higher precision under domain shift, such as PotCrackSeg, DS4Net maintains strong precision at

Table 2. Cross-domain generalization on the TUT dataset (trained on the Infrared dataset, tested on TUT without fine-tuning). The best results are highlighted in **bold**, and the second-best results are underlined.

Model	Type	Dice ↑	IoU ↑	Acc. ↑	Prec. ↑
Unet	RGB	0.3012	0.1773	0.9507	0.2729
DeepLabV3	RGB	0.3594	0.2191	0.9534	0.3180
DeepCrack	RGB	0.5231	0.3542	0.9687	0.5044
CrackFormer	RGB	0.4694	0.3067	0.9512	0.3576
CRM_RGBTSeg	RGB-D	0.5471	0.4095	0.9466	0.4743
PotCrackSeg	RGB-D	0.3118	0.1847	0.4721	**0.7472**
Evi-RoadSeg	RGB-D	0.5558	0.3848	0.9706	0.5317
CMNeXt	RGB-D	0.5437	0.3733	0.9537	0.3951
IRFusionFormer	RGB-D	0.5340	0.3643	0.9654	0.4648
Ours	**RGB-D**	**0.6806**	**0.5158**	**0.9790**	0.6545

0.6545 while substantially improving overlap-based metrics, which indicates more complete and consistent crack segmentation in cluttered scenes. A category-wise analysis on TUT indicates that DS4Net performs best on the road category, with a Dice score of 0.7158, while alternator, weld, and pipeline remain more challenging, with Dice scores of 0.4895, 0.5398, and 0.5488, respectively.

4.5 Ablation Study

Our DS4Net comprises two core components, the DFM block and the information theoretic KL-Gating mechanism. Accordingly, we conduct ablation studies to assess the contribution of each component under both in-domain and cross-domain settings.

Ablation protocol. We construct three variants by progressively enabling the proposed components while maintaining an identical SegFormer backbone, training schedule, and evaluation protocol. DS4Net (Backbone) denotes an RGB-only baseline that excludes the depth branch and RGB-D fusion, and is used as the reference model. DS4Net (Fixed) introduces the pseudo-depth branch and DFM fusion, but replaces adaptive gating with fixed late fusion, in which RGB and depth features are combined with constant weights of 0.5 and 0.5 at each fusion site. DS4Net (without KL-G) further employs the proposed gating network, but removes KL-based discrepancy supervision, such that the gating is learned solely from the segmentation objective. DS4Net (Full) uses the complete design, incorporating pseudo-depth priors, DFM fusion, and KL-Gating together with KL-based regularization.

Ablation analysis across settings. Quantitative comparisons in Tables 3 and 4, reported using Dice, IoU, Accuracy and Precision, show that incorporating depth cues and fusion improves upon the RGB-only backbone. However, fixed-weight fusion becomes suboptimal when modality reliability varies across spatial locations. Moreover, removing the KL loss consistently degrades performance relative to the full model, which indicates that KL-based discrepancy supervision is critical for learning robust, sample-adaptive fusion and for maintaining stable segmentation quality under domain shift.

Table 3. Ablation study of the proposed components on the Infrared dataset (in-domain). The best results are highlighted in **bold**, and the second-best results are underlined.

Model	Dice ↑	IoU ↑	Acc. ↑	Prec. ↑
DS4Net (Backbone)	0.8518	0.7419	0.9839	0.7939
DS4Net (Fixed)	0.8834	0.7911	0.9885	0.9003
DS4Net (without KL-G)	0.8858	0.7950	0.9886	0.8916
DS4Net (Full)	**0.8982**	**0.8152**	**0.9898**	**0.9053**

Table 4. Cross-domain ablation study on the TUT dataset (trained on the Infrared dataset). All models are trained on the Infrared dataset. The best results are highlighted in **bold**, and the second-best results are underlined.

Model	Dice ↑	IoU ↑	Acc. ↑	Prec. ↑
DS4Net (Backbone)	0.5126	0.3446	0.9580	0.4048
DS4Net (Fixed)	0.6536	0.4855	0.9778	0.6452
DS4Net (w/o KL-G)	0.6669	0.5003	0.9765	0.6046
DS4Net (Full)	**0.6806**	**0.5158**	**0.9790**	**0.6545**

5 Conclusion

This paper presented DS4Net, a hardware-free dual-stream framework for robust crack segmentation that integrates foundation-model geometric priors with efficient spectral–spatial modeling. By injecting pseudo-depth generated from a frozen depth foundation model, DS4Net augments RGB appearance with geometry-aware cues without requiring additional sensing devices. To reconcile the inherent conflict between global crack continuity and local boundary sharpness, we introduced the Dual-Path Frequency-Mamba (DFM) block, which combines linear-complexity state space modeling for long-range topology with an FFT-based spectral attention branch that explicitly strengthens high-frequency edge details. Furthermore, we proposed an information-theoretic KL-Gating mechanism to dynamically regulate cross-modal fusion, enabling the network to suppress unreliable modality signals and emphasize the more informative modality under challenging conditions.

Extensive evaluations on the in-domain benchmark and cross-domain TUT dataset demonstrated that DS4Net achieves state-of-the-art performance and substantially improves generalization under domain shift. Ablation studies further verified the effectiveness of both the DFM block

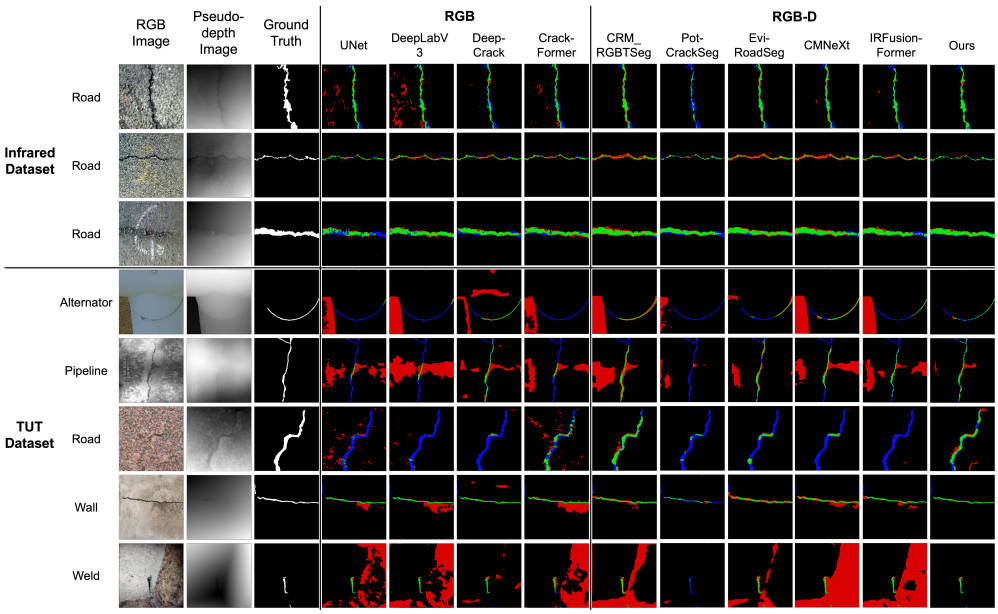

Figure 3. Qualitative results on in-domain and cross-domain scenes. Columns show RGB input, pseudo-depth prior, ground truth, and predictions. Error maps denote TP (green), FP (red), and FN (blue). DS4Net yields more continuous masks with fewer false alarms under noise and domain shift.

and KL-guided adaptive fusion, showing consistent gains when each component is enabled. In future work, we plan to further improve pseudo-depth robustness in textureless, reflective, and low-contrast scenarios, where modality reliability may degrade, explore lightweight designs for real-time edge deployment, and extend the proposed sensor-free paradigm to broader surface-defect segmentation tasks in large-scale infrastructure inspection.

References

[1] Ruiqiang Xiao and Xiaohu Chen. Irfusionformer: Enhancing pavement crack segmentation with rgb-t fusion and topological-based loss. *arXiv preprint arXiv:2409.20474*, 2024.

[2] Zhong Zhou, Junjie Zhang, and Chenjie Gong. Hybrid semantic segmentation for tunnel lining cracks based on swin transformer and convolutional neural network. *Computer-Aided Civil and Infrastructure Engineering*, 38(17):2491–2510, 2023.

[3] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[4] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[5] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021.

[6] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. Vmamba: Visual state space model. *Advances in neural information processing systems*, 37:103031–103063, 2024.

[7] Wanqiang Cai, Xudong Wang, Yifan Xue, Yingyao Ma, Jiasong Wu, Zongyuan Ge, and Bin Wang. Crackmamba with normalized soft-frangi-filter enhancement towards accurate crack segmentation. In *Proceedings of the 2025 International Conference on Multimedia Retrieval*, pages 43–51, 2025.

[8] Yingjie Wu, Shaoqi Li, and Yancheng Li. Depth-aware rgb-d concrete crack segmentation and quantification using progressive cross-modal attention. *Measurement*, page 119453, 2025.

[9] Yu-Ting Huang, Mohammad Reza Jahanshahi, Nikkhil Vijaya Sankar, and Fangjia Shen. Affordable, autonomous, and comprehensive road condition assessment using rgb-d sensors: Enhancing pavement condition evaluation. *IEEE Transactions on Intelligent Transportation Systems*, 2025.

[10] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024.

[11] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[12] Qin Zou, Zheng Zhang, Qingquan Li, Xianbiao Qi, Qian Wang, and Song Wang. Deepcrack: Learning hierarchical convolutional features for crack detection. *IEEE Transactions on Image Processing*, 28 (3):1498–1512, 2019.

[13] Huajun Liu, Jing Yang, Xiangyu Miao, Christoph Mertz, and Hui Kong. Crackformer network for pavement crack segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 24(9):9240–9252, 2023.

[14] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First conference on language modeling*, 2024.

[15] Badri Narayana Patro and Vijay Srinivas Agneeswaran. Mamba-360: Survey of state space models as transformer alternative for long sequence modelling: Methods, applications, and challenges. *Engineering Applications of Artificial Intelligence*, 159: 111279, 2025.

[16] Jun Ma, Feifei Li, and Bo Wang. U-mamba: Enhancing long-range dependency for biomedical image segmentation. *arXiv preprint arXiv:2401.04722*, 2024.

[17] Jiacheng Ruan, Jincheng Li, and Suncheng Xiang. Vm-unet: Vision mamba unet for medical image segmentation. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2024.

[18] Hui Liu, Chen Jia, Fan Shi, Xu Cheng, and Shengyong Chen. Scsegamba: lightweight structure-aware vision mamba for crack segmentation in structures. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29406–29416, 2025.

[19] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3): 1623–1637, 2020.

[20] Huifang Feng, Wen Li, Zhipeng Luo, Yiping Chen, Sarah Narges Fatholahi, Ming Cheng, Cheng Wang, José Marcato Junior, and Jonathan Li. Gcn-based pavement crack detection using mobile lidar point clouds. *IEEE Transactions on Intelligent Transportation Systems*, 23(8):11052–11061, 2021.

[21] Zhenguang Zhang, Bo Peng, and Tingyu Zhao. An ultra-lightweight network combining mamba and frequency-domain feature extraction for pavement tiny-crack segmentation. *Expert Systems with Applications*, 264:125941, 2025.

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

[23] Jiequan Cui, Beier Zhu, Qingshan Xu, Zhuotao Tian, Xiaojuan Qi, Bei Yu, Hanwang Zhang, and Richang Hong. Generalized kullback-leibler divergence loss. *arXiv preprint arXiv:2503.08038*, 2025.

[24] Hui Liu, Chen Jia, Fan Shi, Xu Cheng, Mianzhao Wang, and Shengyong Chen. Staircase cascaded fusion of lightweight local pattern recognition and long-range dependencies for structural crack segmentation. *arXiv preprint arXiv:2408.12815*, 2024.

[25] Esraa Elhariri, Nashwa El-Bendary, and Shereen A Taie. Automated pixel-level deep crack segmentation on historical surfaces using u-net models. *Algorithms*, 15(8):281, 2022.

[26] Ukcheol Shin, Kyunghyun Lee, In So Kweon, and Jean Oh. Complementary random masking for rgb-thermal semantic segmentation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11110–11117. IEEE, 2024.

[27] Jiaming Zhang, Huayao Liu, Kailun Yang, Xinxin Hu, Ruiping Liu, and Rainer Stiefelhagen. Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers. *IEEE Transactions on Intelligent Transportation Systems*, 2023.