

Article

GeoMamba: Geometric-Prior-Infused Multi-Scale Deformable Visual Mamba for Crack Semantic Segmentation

Sangning Li ¹, Bin Liu ², Haiyan Guan ^{1,*}, Lingfei Ma ², Yongtao Yu ³  and Yongming Xu ¹ 

¹ School of Remote Sensing and Geomatics Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China; lisangning30@nuist.edu.cn (S.L.); xym30@nuist.edu.cn (Y.X.)

² School of Geospatial Artificial Intelligence, East China Normal University, Shanghai 200241, China; lbin0110@stu.ecnu.edu.cn (B.L.); lfma@geoai.ecnu.edu.cn (L.M.)

³ Faculty of Computer and Software Engineering, Huaian University, Huaian 223003, China; allennessy@hyit.edu.cn

Highlights

What are the main findings?

- We propose a Multi-Scale Deformable Visual State Space (MDVSS) module, introducing a dynamic scanning strategy that adaptively captures irregular road crack morphologies.
- We develop a Geometric-Topology Prior Injection (GTPI) module, which leverages gated modulation skip connections to infuse edge priors, effectively resolving boundary ambiguity.

What are the implications of the main findings?

- This paper proposes GeoMamba for road crack semantic segmentation. It thus provides support for the structural health assessment of road infrastructure and is of vital importance to ensuring traffic safety and efficiency.

Abstract

Accurate pavement crack segmentation is critical for road infrastructure assessment, yet it remains challenging due to complex background noise and highly variable crack topologies. While emerging visual Mamba models excel in long-range contextual modeling, their inherent 1D sequence flattening process compromises local 2D spatial continuity. To address this limitation, we propose GeoMamba, a geometric-prior-infused multi-scale deformable visual Mamba network for road crack semantic segmentation. First, we design a Multi-Scale Deformable Visual State Space (MDVSS) module to extract multi-scale contextual features and dynamically adapt to tortuous crack paths through a novel deformable scanning mechanism. Second, a Geometric-Topology Prior Injection (GTPI) module is introduced to mitigate serialization artifacts. By leveraging deterministic, parameter-free analytical operators (i.e., Sobel and Laplace), the GTPI module explicitly extracts and adaptively infuses multi-scale structural priors into the Mamba decoder via gated skip connections, intrinsically reconstructing crack typologies while suppressing pseudo-structural noise. Comprehensive experiments on DeepCrack and Concrete3K datasets demonstrate that GeoMamba outperforms nine state-of-the-art methods. Specifically, it achieves peak performance on the DeepCrack dataset with an mIoU of 83.79% and an F1 score of 89.27%, demonstrating exceptional semantic segmentation performance, superior topological continuity, and robust generalization across diverse pavement materials.

Keywords: crack semantic segmentation; deep learning; Mamba



Academic Editor: Alessandro Mei

Received: 10 June 2026

Revised: 14 July 2026

Accepted: 20 July 2026

Published: 24 July 2026

Copyright: © 2026 by the author.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons](https://creativecommons.org/licenses/by/4.0/)

[Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

The structural health of road infrastructure is critical to ensuring transportation safety and efficiency. Road long-term service and environmental erosion inevitably generate various cracks on asphalt or concrete pavements. If these cracks are not detected and maintained in a timely manner, they will accelerate structural deterioration, cause safety hazards, and significantly increase maintenance costs [1,2]. To comprehensively evaluate infrastructure conditions, Structural Health Monitoring (SHM) systems (e.g., IoT, vibration) [3] and Non-Destructive Testing (NDT) techniques (e.g., GPR, infrared) [4] are indispensable for continuous, system-level, and subsurface diagnostics. As a vital complementary approach for rapid and large-scale surface screening, developing an efficient visual detection system is deemed equally essential. Driven by recent advancements in computer vision, deep learning-based semantic segmentation has been widely established to automatically extract crack features from RGB imagery, providing an accurate and scalable solution for infrastructure preservation.

Early efforts of pavement crack detection were primarily based on Convolutional Neural Networks (CNN). Zhang et al. [5] proposed an efficient CNN-based encoder–decoder network, named ECSNet, which, by integrating small-kernel convolutions and parallel down-sampling operations, significantly reduces model parameters while ensuring real-time pavement crack segmentation. Zaid Al-Huda et al. [6] improved the classic UNet with an asymmetric dual-decoder architecture. This innovative design simultaneously captures fine-grained crack details and coarse-grained spatial context, thereby significantly enhancing the overall segmentation accuracy. Liu et al. [7] proposed DeepCrack, an end-to-end deep hierarchical feature-learning architecture, for pixel-level crack segmentation. By integrating side-output layers and fusing multi-scale convolutional features from different network stages, this model effectively combined low-level details with high-level semantics to achieve high-precision detection. However, due to the inherent nature of local computation, the modeling of long-range dependencies in CNN architectures is severely hindered in capturing tortuous and widely extended crack topologies [8], frequently generating discontinuous segmentation results or false detections in complex backgrounds [9]. To overcome this bottleneck of long-range dependency modeling, Vision Transformer (ViT) [10] was introduced into dense prediction tasks. Following this advancement, self-attention mechanisms were incorporated into crack segmentation due to their robust global modeling capabilities [11–13]. Liu et al. [14] applied pure Transformers to crack segmentation in a seminal work, demonstrating the immense potential of global attention mechanisms in capturing subtle, meandering crack structures and significantly improving segmentation performance in complex backgrounds. Tao et al. [15] combined Transformers with a dedicated Boundary Awareness Module to effectively address the persistent issue of blurred boundaries in complex segmentation results. Xie et al. [16] introduced SegFormer, which, by unifying a hierarchical Transformer encoder with a lightweight Multi-Layer Perceptron (MLP) decoder, powerfully extracts multi-scale features for delineating complex pavement cracks. Furthermore, beyond crack detection, a recent study in remote sensing [17] demonstrates that multi-modal attention fusion effectively enhances the connectivity of fragmented road networks, confirming the power of attention mechanisms in recovering complex linear topologies. However, some inherent limitations exist within transformer architectures. For example, the computational complexity of self-attention mechanisms increases quadratically with image resolution, resulting in massive parameter counts and immense computational costs. Specifically in road crack segmentation tasks, high-resolution images are typically required as inputs to preserve minuscule crack branches and edge topologies. Therefore, a variety of efficient architectures that maintain

linear computational complexity when enabling global context modeling have emerged in recent years.

Recent advancements in State Space Models (SSMs) are drawing widespread attention [18]. SSMs, which are derived from the classical Kalman filter, excel at capturing long-range dependencies and enable efficient parallel training. These strengths quickly captured the attention of the computer vision community, spurring the development of dedicated visual models, such as Vision Mamba (Vim) [19] and VMamba [20]. In semantic segmentation, Liu et al. [21] proposed CM-UNet, a hybrid architecture that integrates CNNs with Mamba. CM-UNet enhances global-local feature modeling in remote sensing images by employing CSMamba blocks and multi-scale attention aggregation modules, which collectively contribute to its superior segmentation performance. Similarly, Li et al. [22] introduced UMFormer, a novel encoder–decoder network that integrates State Space Models with optimized self-attention mechanisms to significantly lower computational complexity while effectively capturing multi-scale global and local features. In contrast, aiming to achieve high-quality crack segmentation with low computational cost, Liu et al. [23] proposed SCSegamba, a lightweight structure-aware Vision Mamba network that effectively integrates morphological information and texture cues.

However, accurate crack segmentation is challenging due to irregular morphology, weak long-range continuity, and strong background interference. Existing visual Mamba networks predominantly employ predefined mapping strategies, including raster scanning [19,20], local scanning [24], and continuous scanning [25]. These approaches depend on fixed scanning paths, which fail to adapt to the inherent and often irregular spatial structure of images, particularly cracks. Moreover, the existing prevalent use of single-scale patch embedding struggles to capture fine cracks and lacks the flexibility to handle significant width variations.

Despite the advancements of CNNs, Transformers, and recent State Space Models (SSMs), pavement crack segmentation still faces critical challenges. First, the standard 1D sequence scanning in SSMs often disrupts the continuous 2D geometric topology of irregular cracks. Furthermore, rigid spatial modeling struggles to dynamically adapt to severe environmental noise, causing fragmented predictions. Finally, extracting fine-grained details from high-resolution inputs necessitates robust global context modeling, which incurs prohibitive computational costs. To address these limitations, we propose a dual-branch edge-guided multi-scale deformable visual Mamba model for road crack detection. Instead of using traditional scanning methods, the network employs dynamic scanning to adapt to the irregular patterns of road cracks. At the same time, it incorporates an auxiliary edge-detection path to explicitly extract boundary information, thereby refining the final segmentation output. The primary contributions can be summarized as follows:

- We design a Multi-Scale Deformable Visual State Space (MDVSS) module. It employs a deformable scanning strategy that dynamically adapts to crack morphologies. This design enhances long-range dependency modeling and optimizes attention allocation, improving computational efficiency and robust generalization.
- We develop a Geometric-Topology Prior Injection (GTPI) module to tackle boundary ambiguity and noise interference. It integrates multi-source edge priors into structured attention features. Facilitated by gated modulation skip connections, this module provides explicit guidance during feature fusion, which emphasizes authentic crack topologies and thereby achieves precise, coherent segmentation within complex backgrounds.

We conducted extensive experiments, and the results demonstrate that the proposed network significantly outperforms existing methods. Optimal performance is achieved on both the DeepCrack and Concrete3K datasets. The architectural design of GeoMamba is detailed in Section 2. Experimental datasets, evaluation metrics, parameter configurations,

and model performance comparisons are elaborated in Section 3. The analysis of visualization results and ablation experiments is shown in Section 4. The conclusion of this research is presented in Section 5.

2. Materials and Methods

In this section, we propose a Mamba-based architecture for road crack semantic segmentation. We break down our methodology by first introducing our two core innovative components before presenting the overall network framework. In Section 2.1, we detail the Multi-Scale Deformable Visual State Space (MDVSS) module, which uses deformable scanning to capture long-range dependencies of irregular cracks. Subsequently, Section 2.2 introduces the Geometric-Topology Prior Injection (GTPI) module, designed to inject multi-source edge priors into the attention features for precise boundary delineation. Finally, in Section 2.3, we illustrate the overall encoder–decoder architecture, demonstrating how the MDVSS and GTPI modules are integrated via gated modulation skip connections.

2.1. Multi-Scale Deformable Visual State Space Module

To address the limitations of rigid scanning methods and dynamically capture complex crack topologies under severe background interference, we design the Multi-Scale Deformable Visual State Space (MDVSS) module as our core backbone. The backbone network MDVSS first encodes the input image, which is then partitioned into patches via the newly proposed Multi-Scale Patch Embedding (MPE) (see Figure 1). The MPE module is designed to effectively capture thin cracks and adapt to cracks with varying widths. As shown in Figure 1, to extract multi-scale features, the MPE module adopts three parallel convolutional layers with distinct kernel sizes: a small-scale convolutional kernel G_1 that captures local details, a medium-scale convolutional kernel G_2 that maintains the standard resolution, and a large-scale convolutional kernel G_3 that extracts global contextual information. Subsequently, all feature maps are upsampled to a unified scale via bilinear interpolation and concatenation. Next, feature fusion and edge enhancement are performed to form a three-dimensional sequence B , which is flattened and conversed into a two-dimensional sequence $\{B_1, B_2, \dots, B_n\}$.

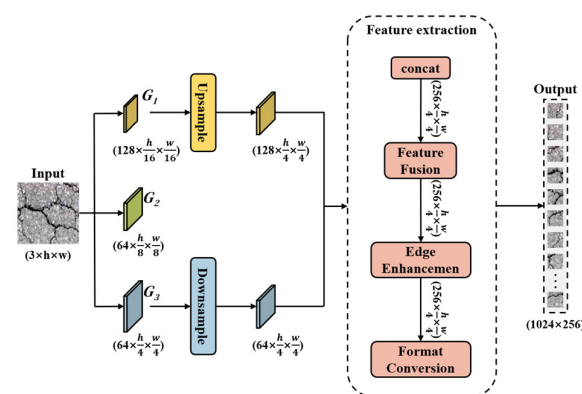


Figure 1. Illustration of the Multi-Scale Patch Embedding (MPE) module.

As shown on the left side of Figure 2, our designed MDVSS Block employs two-dimensional selective scanning (SS2D) tailored for visual tasks. It takes as input the sequence $\{B_1, B_2, \dots, B_n\}$ from the preceding MPE module. At its initial stage, the MDVSS Block integrates two layers of GBC [23]. The bottleneck convolution significantly reduces the number of parameters and computational cost, while the gating mechanism dynamically adjusts feature weights. This design enables accurate capture of crack morphology and effective noise suppression in complex backgrounds. Subsequently, PAF [26] is introduced

to fuse the initial feature sequence with those processed by the SS2D, enhancing the capture of crack shape and texture details. Residual connections are added after both the scan and fusion stages to preserve feature integrity and detail information. We modify the scan mode of the SS2D into a deformable scanning strategy [27], inspired by deformable convolution. This allows features to be extracted along crack textures, generating more refined and robust segmentation results than a fixed scan. Finally, the MDVSS module outputs multi-scale feature maps $\{F_1, F_2, F_3, F_4\}$, containing crack local details, middle-level contours, and deep global context.

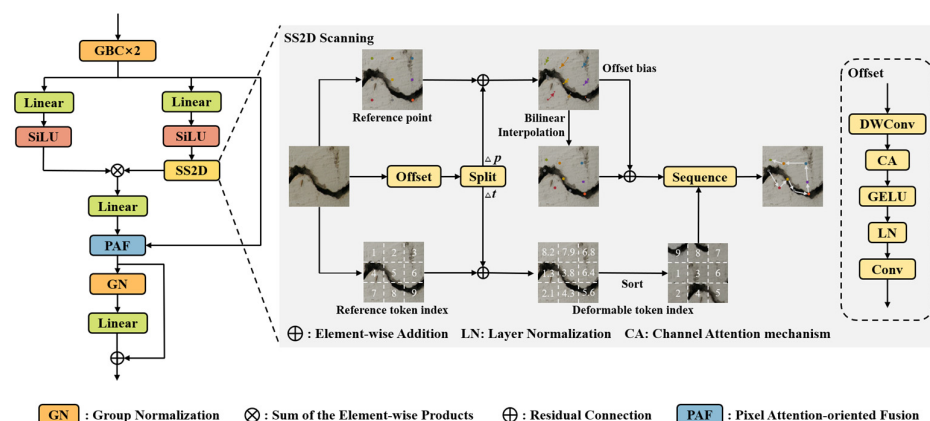


Figure 2. Illustration of the MDVSS module.

Figure 2 illustrates the newly introduced deformable scanning strategy. For a given input feature $x \in \mathbb{R}^{C \times h \times w}$, an offset subnetwork outputs the offset values $o \in \mathbb{R}^{3 \times h \times w}$, for the reference points and reference tokens. The architecture of the offset subnetwork is presented on the right side of Figure 2. As shown in the figure, the offset subnetwork extracts local features from the input features by depth-wise separable convolution. The features are then passed through a Channel Attention (CA) mechanism, a GELU activation function, and Layer Normalization (LN). Finally, a 1×1 convolution is applied to the normalized features to obtain offset values with a total dimension of 3. These offset values are subsequently split into two parts: the first two dimensions, which represent the offset of the reference points in two-dimensional space, termed as point offset Δp ; the third dimension, which denotes the index offset of reference tokens within the whole image patch, termed as token index offset Δt . The specific procedures are detailed as follows:

$$\Delta p, \Delta t = \text{Split}(\tanh(\text{Offset}(x)), \text{dims} = [2, 1]) \quad (1)$$

where $\text{Split}(\cdot)$ represents dividing the dimension into two parts, $\tanh(\cdot)$ suppresses the influence of extreme values, and $\text{Offset}(\cdot)$ denotes the subnetwork for the generation of offsets.

Subsequently, the point offset is implemented to dynamically adjust the spatial positions of feature points and obtain position-accurate offset features with offset bias. Then, based on the deformable index generated by the token index offset, we sort the offset features using the index values, thereby mapping two-dimensional features into a one-dimensional feature sequence. However, since the discrete sorting operation inherently truncates the gradient flow, we introduce a gradient compensation mechanism during backpropagation. Specifically, we approximate the gradient of the token index offset by computing the mean gradient of the sorted feature sequence and replicating it back to the offset tensor, which effectively reconnects the computational graph. Finally, we integrate the position adaptation from the point offset and the sequence order optimization from the token index offset to generate a feature sequence with dual adaptability to both content

and structure. The deformable scanning strategy incorporates both position compensation and gradient compensation (GC) strategies. This design directly addresses two key challenges—the loss of position information and the interruption of gradient propagation caused by deformation adjustment—improving feature representation capability and model training stability.

2.2. Geometric-Topology Prior Injection Module

To explicitly overcome the loss of geometric continuity caused by standard 1D sequence scanning and preserve the complex 2D topological structures of meandering cracks, we introduce the Geometric-Topology Prior Injection (GTPI) module. While the Mamba architecture exhibits exceptional capabilities in modeling long-range contextual dependencies, its inherent 1D sequence flattening operation inevitably disrupts the local 2D topological continuity of high-frequency regions, such as delicate crack branches and intersections. To fundamentally compensate for this spatial distortion, we propose the Geometric-Topology Prior Injection (GTPI) module (see Figure 3). Instead of employing task-driven auxiliary boundary branches that require additional training, the GTPI module leverages deterministic mathematical operators to explicitly extract multi-scale structural priors at the encoder front-end. These high-frequency topological representations are subsequently refined via an adaptive gating mechanism and directly infused into the Mamba decoder through skip connections, thereby intrinsically reconstructing the crack typologies without increasing the learning burden.

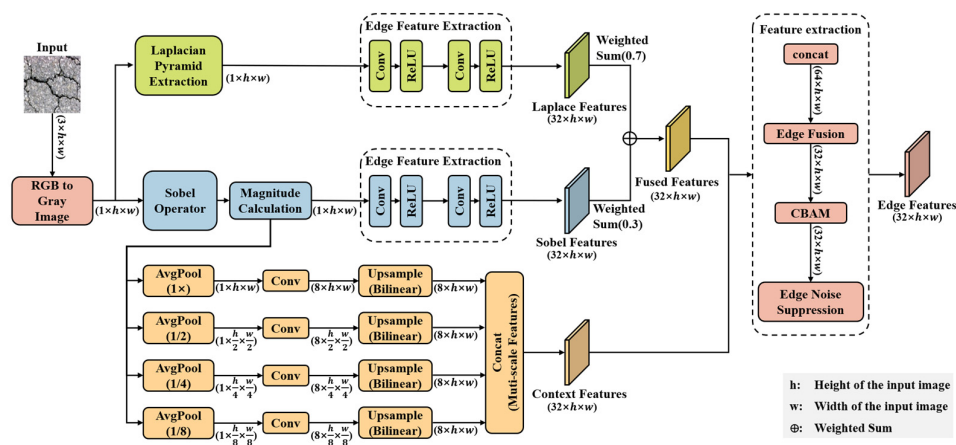


Figure 3. Illustration of the GTPI module.

To attain a balance between structural continuity and localization precision, we design a dual-branch deterministic prior extraction mechanism, which includes two first-order and second-order derivatives, i.e., a Sobel operator that captures gradient magnitude for structural continuity, and a Laplace operator that identifies zero-crossing points for precise localization. The dual-branch edge-guided mechanism is as follows:

$$F_f = W_1(L(I_{gray})) \oplus W_2(S(I_{gray})) \quad (2)$$

where $L(\cdot)$ denotes the Laplace operation, $S(\cdot)$ denotes the Sobel operation, and $W(\cdot)$ represents learnable projection layers. A static weighting strategy, empirically set to 0.7 and 0.3 [28], is introduced for integrating edge features from the Laplace and Sobel operators. This specific distribution is guided by the high-frequency recovery mechanism, which demonstrates that assigning a dominant weight to second-order derivatives optimally highlights subtle textural variations while utilizing first-order derivatives to preserve basic structural layouts. The Laplace operator yields a stronger response to step edge points in the image, thereby effectively capturing continuous and elongated structures typically

manifested by road cracks. whereby a higher weight is assigned to this operator. To address the variations in road cracks across different scales, we design a multi-scale edge feature extraction layer, which yields the gradient magnitude features through progressively increasing pooling ratios ($r \in \{1, 2, 4, 8\}$). Different pooling ratios enable the model to perceive features at different scales, thereby better extracting road cracks. Following the feature transformation and upsampling [29], multi-granularity edge representations are presented, which is detailed as follows:

$$F_{ms} = \text{Concat}\left(\text{Up}\left(\text{Conv}\left(\text{Pool}_r\left(F_{grad}\right)\right)\right)\right)_{r=1}^8 \quad (3)$$

Traditional edge operations are highly susceptible to background noise and interference, such as gaps in asphalt roads, rainwater, and mud. To tackle this issue, we design an attention gating mechanism in which the CBAM module recalibrates the fused features along the channel and spatial dimensions, and a suppression branch generates a pixel-level noise mask through a Sigmoid activation function. A residual scaling strategy is then adopted to eliminate gradient vanishing while preserving the details of faint cracks, which is detailed as follows:

$$F_{final} = \text{CBAM}\left(F_{fused}\right) \otimes (\alpha + (1 - \alpha) \cdot M) \quad (4)$$

where $\otimes$ denotes element-wise multiplication. Hyperparameter α (set to 0.7) [30] ensures that 70% of structural information is unconditionally retained, while the remaining 30% is adaptively filtered by the network, thereby balancing the recall and precision of crack segmentation. This specific ratio is adopted following the edge-guided mechanism, which has empirically demonstrated that retaining a dominant proportion of explicit boundary priors effectively prevents the loss of weak topological structures during deep feature abstraction.

2.3. Overall Architecture

Figure 4 illustrates the overall architecture of the proposed GeoMamba, which includes two core components: a Multi-Scale Deformable Visual State Space module for dynamic adaptation to the morphological characteristics of cracks, and a Dual-branch Edge-guided Geometric Structure Perception module for detecting thin edges and geometric topology of cracks.

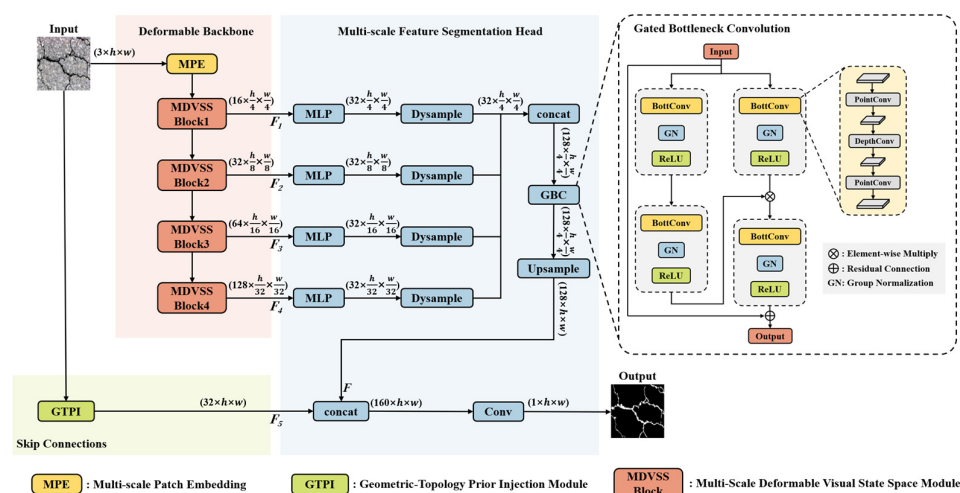


Figure 4. Overview of the GeoMamba.

Specifically, we process a given RGB input image $I \in \mathbb{R}^{3 \times h \times w}$ with height h and width w by dividing it into n patches via multi-scale patch embedding (MPE), thereby obtaining serialized features F . The obtained sequence is processed by four MDVSS modules. Through this process, key crack-pixel cues are embedded into multi-scale feature

maps $\{F_1, F_2, F_3, F_4\}$. Subsequently, these multi-scale features $\{F_1, F_2, F_3, F_4\}$ are fed into independent lightweight (MLP) layers separately. Each MLP layer adaptively transforms features to a unified channel dimension and learns the correlation between features and segmentation labels by nonlinear mapping. Then, dynamic upsampling is performed on the features of each scale processed by the MLP to preserve fine spatial details such as thin cracks and edges. Next, the upsampled multi-scale features are concatenated along the channel dimension to integrate complementary crack feature information, including shallow edge features and deep semantic features. After dynamic upsampling and channel-wise concatenation, the feature tensor is then fed into the Gated Bottleneck Convolution (GBC) module, where the gating mechanism and lightweight bottleneck convolution dynamically adjust feature weights. This process effectively suppresses background noise and enhances the representation of crack morphological details, thereby accomplishing the in-depth fusion and refinement of multi-scale features. Moreover, the input image I is also fed into the Geometric-Topology Prior Injection (GTPI) module to extract the high-frequency edge representation F_5 . With the Skip Connections mechanism, edge feature F_5 is efficiently injected into the MFS decoder, which provides explicit boundary constraints in the feature fusion stage and enables the model to intensively focus on the complex topological structure of cracks. Finally, within the MFS decoder, edge feature F_5 is concatenated with multi-scale feature F , and the resulting single prediction tensor is processed to output a refined pixel-level segmentation map $O \in \mathbb{R}^{1 \times h \times w}$.

The four feature maps $F_1, F_2, F_3, F_4 \in \mathbb{R}^{3 \times h \times w}$ generated by the MDVSS are input into the MFS. Processing is respectively conducted via efficient MLP operations and dynamic upsampling. This procedure restores the resolution of these maps to the original dimensions, finally yielding $F_1^{up}, F_2^{up}, F_3^{up}, F_4^{up} \in \mathbb{R}^{3 \times h \times w}$. The corresponding expression is detailed as follows:

$$F_i^{up} = DySample(MLP_i(F_i)) \quad (5)$$

where i denotes the layer index. The network integrates all multi-scale crack shape and texture features into a single tensor, thereby yielding a high-quality crack segmentation map $o \in \mathbb{R}^{1 \times h \times w}$, which is detailed as follows:

$$F = Upsample(GBC(Concat(F_1^{up}, F_2^{up}, F_3^{up}, F_4^{up}))) \quad (6)$$

$$o = Conv(Concat(F_5, F)) \quad (7)$$

Finally, we introduced a Binary Cross-Entropy loss (BCE) [31] with a Dice loss [32]. Through the hybrid loss function, the network effectively enhances its robustness to the imbalance of pixel data. The expression of the total loss function is detailed as follows:

$$L = \alpha \cdot L_{Dice} + \beta \cdot L_{BCE} \quad (8)$$

The hyperparameters α and β control the weights of the two loss components, with their ratio set at 1:5 [23]. This specific configuration is adopted because empirical evaluations demonstrate that it optimally balances pixel-level classification accuracy with structural overlap, effectively mitigating the severe class imbalance typically encountered in crack segmentation tasks.

3. Experimental Results

3.1. Experimental Details

- (1) The proposed GeoMamba and all nine comparative baseline algorithms were independently reproduced and evaluated under a fully unified experimental environment. Specifically, all networks were implemented using PyTorch v1.13.1 and executed on a

single NVIDIA GeForce RTX 4090 GPU. For all models, the input image resolution was uniformly set to 256×256 pixels, and the batch size was configured to 8. The AdamW optimizer was consistently employed across all models with an initial learning rate of 0.0005, a weight decay coefficient of 0.01, a globally fixed random seed of 42, and a PolyLR learning rate scheduling strategy. Each model was trained for up to 50 epochs, and the specific checkpoint exhibiting the optimal validation performance was selected for final testing to ensure that every comparative algorithm reached its best convergence state.

- (2) **Comparative Methods:** To comprehensively evaluate the performance of our GeoMamba network, we compared our GeoMamba with nine state-of-the-art methods. The compared networks included three parts: CNN- or Transformer-based models (i.e., ECSNet, UNet, DeepCrack, LETNet, CTCrack, SegFormer), SAM-based models (i.e., MFNet), and Mamba-based networks (i.e., UMFormer and SCSegamba).
- (3) **Evaluation Metrics:** The performance of our GeoMamba was evaluated by six metrics: Precision (P); Recall (R); F1 score; Optimal Dataset Scale (ODS); Optimal Image Scale (OIS); and mean Intersection over Union (mIoU). Within these metrics, ODS, which evaluates the overall performance using a single fixed optimal threshold across the entire dataset, was used to demonstrate the adaptability of the model to different dataset scales. Meanwhile, OIS, which captures the theoretical maximum performance by selecting the optimal threshold for each individual image, was used to illustrate the adaptability across varying image scales under perfectly tailored thresholds.

3.2. Datasets

For rigorous evaluation and absolute reproducibility, the DeepCrack and Concrete3K datasets were partitioned into training, validation, and testing sets at a strict 7:1:2 ratio. To mitigate potential class imbalance, a stratified random sampling strategy was implemented utilizing a globally fixed random seed. This ensures that the proportional distribution of crack versus background pixels remains perfectly balanced and consistent across all subsets. The validation set was exclusively used for hyperparameter tuning. Consequently, all quantitative comparisons between GeoMamba and the baseline models were evaluated on the identical, unseen fixed testing set.

In addition, the FCN dataset776 was utilized as an independent testing set to evaluate the cross-domain generalization capability of the models. All models were tested directly on this dataset using their pre-trained weights without any further fine-tuning. The detailed sample quantity and class distribution statistics for each subset are summarized in Table 1.

Table 1. Detailed sample quantities and class distribution statistics for each dataset.

Dataset	Subset	Images	Background Pixels (%)	Crack Pixels (%)
Deepcrack	train	368	96.30	3.70
	val	52	96.70	3.30
	test	106	96.91	3.09
Concrete3K	train	2100	93.80	6.20
	val	300	92.65	7.35
	test	600	93.52	6.48
FCN	all	776	93.72	6.28

- (1) **DeepCrack Dataset [7]:** It comprises 537 RGB images with a fixed resolution of 544×384 pixels. The images in this dataset are contained by a variety of noise interferences, such as shadows, oil stains, water spots, lane lines, and road surfaces with diverse textures. All images were employed as raw inputs, and manually annotated pixel-level binary masks are provided as the corresponding ground truth.

- (2) **Concrete3K Dataset [33]**: It includes a total of 3000 concrete pavement images with a resolution of 512×512 pixels. This dataset contains concrete pavements with gray and white shades, and all were captured from real-world construction sites. The images not only contain various cracks but are also filled with concrete-specific strong background noise, such as pitting, construction joints, exposed aggregate, and water stains. Mis-detection is easily induced by these complex texture interferences. This dataset is highly suitable for the evaluation of the robustness of algorithms under complex concrete environments.
- (3) **FCN dataset [34]**: It comprises 776 RGB and grayscale images with varying resolutions. The dataset contains a variety of complex crack morphologies, such as extremely thin, intersecting, and historical wide cracks on pavements and concrete walls. Each image is paired with a manually annotated pixel-level binary mask serving as the ground truth.

3.3. Model Performance Comparison

Table 2 lists the quantitative results by comparing our GeoMamba with the SOTA methods.

Table 2. Comparison with nine methods on two datasets.

Type	Method	DeepCrack					
		mIoU/%	F1/%	P/%	R/%	ODS/%	OIS/%
CNN	ECSNet [5]	76.35	69.99	61.88	81.3	60.60	63.61
	UNet [35]	81.13	77.11	72.25	85.23	65.88	68.77
	DeepCrack [7]	81.03	77.17	69.79	<u>86.53</u>	71.19	73.28
CNN+Transformer	LETNet [36]	77.12	70.81	71.18	72.88	66.23	70.12
	CTCrack [15]	82.91	79.87	75.30	85.30	66.15	70.22
	SegFormer [16]	78.86	73.82	69.49	79.62	67.53	70.86
SAM	MFNet [37]	76.69	70.21	64.24	77.70	58.46	60.93
Mamba	UMFormer [22]	80.89	77.05	72.51	82.54	70.30	72.80
	SCSegamba [23]	<u>83.02</u>	<u>84.91</u>	<u>85.12</u>	84.71	<u>78.63</u>	80.05
	GeoMamba (ours)	83.79	89.27	90.61	87.96	79.06	<u>79.78</u>
Type	Method	Concrete3K					
		mIoU/%	F1/%	P/%	R/%	ODS/%	OIS/%
CNN	ECSNet	84.93	84.34	87.61	81.3	84.49	85.44
	UNet	85.71	85.23	<u>89.77</u>	81.13	85.27	87.94
	DeepCrack	85.50	85.03	87.24	82.93	85.30	84.29
CNN+Transformer	LETNet	86.12	85.74	88.72	82.96	85.89	86.22
	CTCrack	86.16	85.76	90.27	81.68	85.76	85.9
	SegFormer	84.96	84.42	85.40	83.47	84.99	83.22
SAM	MFNet	85.35	84.94	83.51	86.41	85.28	84.31
Mamba	UMFormer	<u>86.59</u>	<u>86.28</u>	89.64	83.15	86.31	86.00
	SCSegamba	85.85	80.82	85.24	76.83	83.51	83.95
	GeoMamba (ours)	87.77	86.92	89.68	<u>84.34</u>	<u>86.04</u>	<u>87.59</u>

The best and second-best results in the table are marked in boldface and underlined, respectively.

3.3.1. DeepCrack Dataset

The DeepCrack dataset, characterized by complex road scenes and cracks with highly diverse topologies, poses significant challenges to segmentation algorithms. Specifically, it heavily tests a model's robustness against severe background interference (e.g., shadows, stains, and pavement markings) as well as its capacity to model long-range continuous dependencies inherent in crack structures. As quantitatively detailed in Table 2, our proposed

GeoMamba demonstrates clear superiority, achieving state-of-the-art performance across the core evaluation metrics when compared against other architectures.

Notably, GeoMamba obtains an F1 score of 89.27% and a Precision rate of 90.61%. These metrics surpass the second-best model (SCSegamba) by substantial margins of 4.36% and 5.49%, respectively. This significant improvement in Precision underscores the model's exceptional ability to distinguish true crack pixels from complex, noisy backgrounds—a common pitfall for baseline CNNs and standard Transformer networks. We attribute this success directly to the effectiveness of our proposed dual-branch edge-guided GTPI module. By systematically fusing multi-source edge information and transforming it into structured attention maps, the GTPI module empowers the network to focus strictly on the intrinsic geometric topology of the cracks. Consequently, it effectively mitigates false-positive detections that are frequently induced by background noise.

Furthermore, in terms of mIoU and Recall, which are metrics crucial for evaluating the structural integrity and continuity of the segmented cracks, GeoMamba continues to outperform all competitors. It achieves an mIoU of 83.79% and a Recall of 87.96%, yielding corresponding improvements of 0.77% over SCSegamba and 1.43% over the DeepCrack network. The high Recall specifically indicates that our method recovers more complete crack networks with minimal fragmentation. This success is primarily driven by the proposed MDVSS module. Within this module, the deformable scanning strategy breaks away from rigid grid constraints, enabling the network to dynamically adapt its spatial receptive field to the highly tortuous and irregular geometries of natural cracks. Concurrently, the multi-scale patch embedding ensures that the network can seamlessly extract multi-granularity features, allowing for the precise segmentation of cracks with widely varying widths, from prominent, wide fissures to subtle, hair-like fractures. Additionally, GeoMamba yields highly competitive boundary-based metrics, specifically an ODS of 79.06% and an OIS of 79.78%, further validating its comprehensive capability in precise crack delineation.

3.3.2. Concrete3K Dataset

To rigorously evaluate the cross-scenario generalization capabilities of GeoMamba across disparate materials and surface textures, further validation was conducted on the Concrete3K dataset. Unlike predominantly asphalt-based pavements, concrete surfaces present a highly heterogeneous, mottled appearance with significant intrinsic noise. This severe textural complexity drastically reduces the visual contrast between the foreground and the background, frequently blurring crack edges and posing a formidable challenge for accurate segmentation.

As shown in Table 2, while certain comparative methods excel in isolated metrics on this dataset, they fundamentally struggle to overcome the inherent Precision–recall trade-off. For instance, CTCrack achieves the highest Precision (90.27%) but suffers a drop in Recall (81.68%), whereas MFNet secures the top Recall (86.41%) but yields a comparatively lower Precision (83.51%). These baseline models fail to maintain an optimal balance, leaning either toward under-segmentation by omitting faint structural details or toward over-segmentation by misclassifying surface noise as cracks.

In contrast, GeoMamba successfully mitigates this trade-off, achieving an optimal equilibrium with a Precision of 89.68% and a Recall of 84.34%. Consequently, it secures the highest comprehensive F1 score of 86.92% and an mIoU of 87.77%, outperforming all other comparative methods. This robust performance indicates that the MDVSS module transcends mere superficial texture fitting; rather, it effectively captures the robust, intrinsic geometric topologies of the cracks. By focusing on these fundamental structural characteristics rather than dataset-specific background patterns, GeoMamba demonstrates

superior and stable cross-dataset generalization. Furthermore, the designed GTPI module proves indispensable in this specific low-contrast scenario. By systematically fusing multi-source edge information, the GTPI module explicitly guides the network to recognize the weak-contrast gradients typical of concrete surfaces. As a result, GeoMamba can sharply delineate even the most severely blurred crack boundaries, ensuring high-fidelity segmentation despite the challenging mottled background.

4. Discussion

4.1. Analysis of Visualization Results

Figure 5 presents the qualitative results of the proposed GeoMamba and other comparative methods on the DeepCrack dataset. As shown in the green squares of Figure 5, real road images contain abundant interfering elements, such as spots, water stains, and complex background textures. The comparative networks, lacking effective constraints on the intrinsic geometry of cracks, frequently misidentified these interfering elements as cracks, leading to obvious false-positive predictions. For instance, ECSNet, DeepCrack, and UMFormer produced numerous misdetections in such regions, lowering their crack segmentation performance. Conversely, compared with the competitors, our GeoMamba model filtered these background interferences out and obtained satisfactory crack segmentation results. This is because the GTPI module uses multi-source edge information as priors to ensure the network learns the distinction between authentic crack topology and disordered background noise in the feature fusion stage, contributing to salient feature representation of cracks under complex backgrounds.

Furthermore, as illustrated in the red squares of Figure 5, existing models generally extracted cracks with severe feature fragmentation or falsely detected extremely fine cracks with multiple branches and cracks with complex network topologies. However, our GeoMamba completely and continuously extracted these fine-grained structures. For example, in Figure 5(b), both UNet and GeoMamba predicted the slender crack in the lower left corner; in Figure 5(d), the GeoMamba captured ground residue within large cracks; and in Figure 5(e), our GeoMamba accurately captured the complex topological relationship in the lower left corner. This is because the MDVSS uses the deformable scanning strategy, rather than traditional fixed windows used in the comparative methods, for long-range dependency modeling, enabling the network to dynamically adapt to the irregular paths of cracks and finally precisely capture complex morphologies. Moreover, our multi-scale patch embedding strategy ensured the network with the preservation of minute local details in the deep feature space. In summary, the GeoMamba preserved the complete topological structure of cracks to the maximum extent using the model, while effectively suppressing background noise.

Figure 6 presents visual comparison results of the GeoMamba and other baseline networks on the Concrete3K dataset. We observed that the regions within the red bounding boxes contain complex topological branches and wide cracks accompanied by surface spalling. To be specific, there were fragmentations of fine terminal branches in Figure 6(a), inaccurate predictions in Figure 6(b), (c), as well as extremely blurry boundaries in Figure 6(d). Our GeoMamba accurately depicted the complex network boundaries of cracks and maintained internal continuity and topological integrity. This is because the dynamic scanning mechanism in the MDVSS module not only excels on slender cracks, but also flexibly adjusts its receptive field when encountering concrete cracks with drastic width variations, thereby preventing the loss of local features.

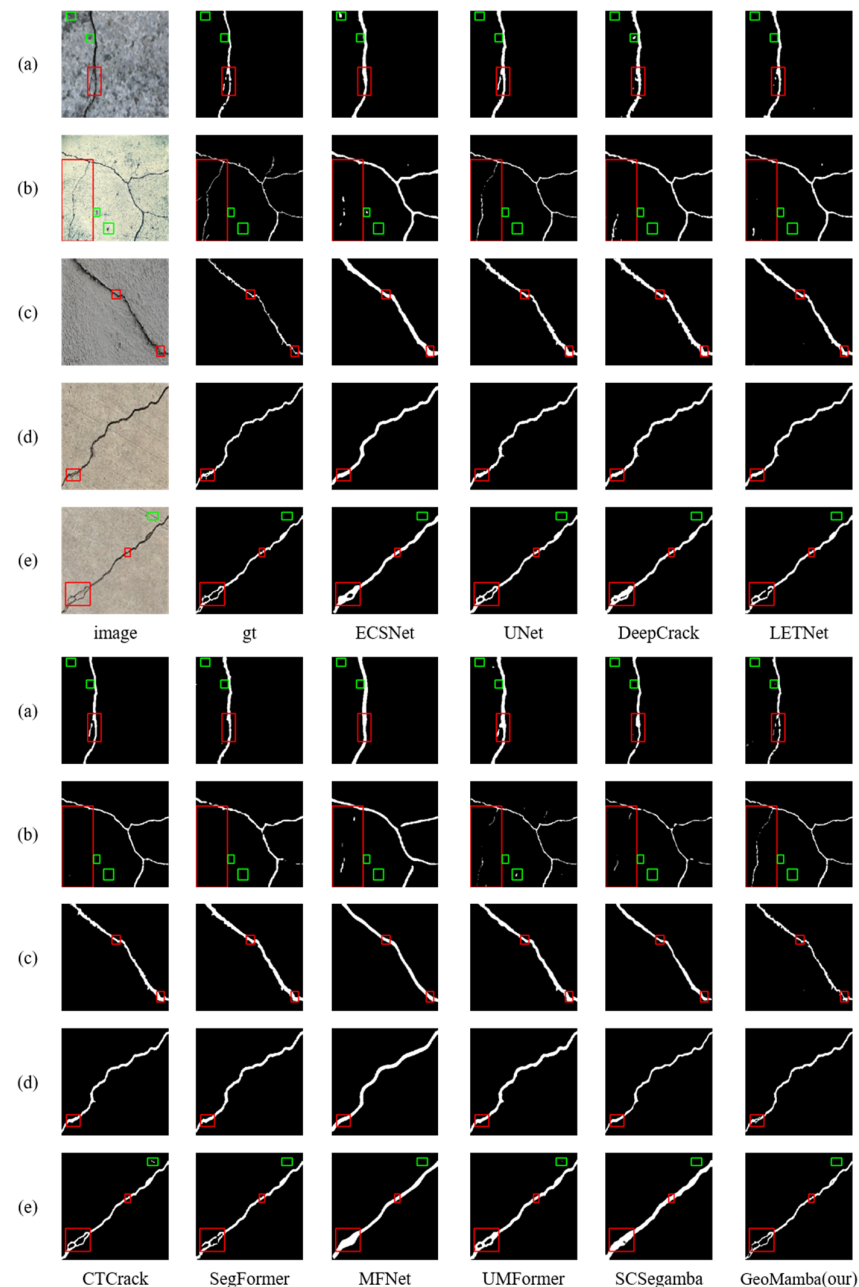


Figure 5. Visual analysis of crack detection results with DeepCrack dataset. (Within these boxes, background noise regions prone to misdetection are highlighted by green squares, while fine and topologically complex crack branches are emphasized by red squares).

Secondly, the GeoMamba also exhibited significant advantages in the suppression of background noise and in the trade-off of comprehensive performance. As illustrated in the green bounding boxes in Figure 6(e), concrete surface interferences, i.e., holes or shallow scratches that typically trigger false positives in common models (e.g., SegFormer), have minimal impact on the GeoMamba's performance in terms of recall. In contrast, the proposed GeoMamba uses the GTPI module to obtain multi-source edge priors, achieving strong structural constraints from blurred and low-contrast backgrounds. Consequently, our network strikes an optimal balance: it effectively filters out noise while preserving genuine crack structures.

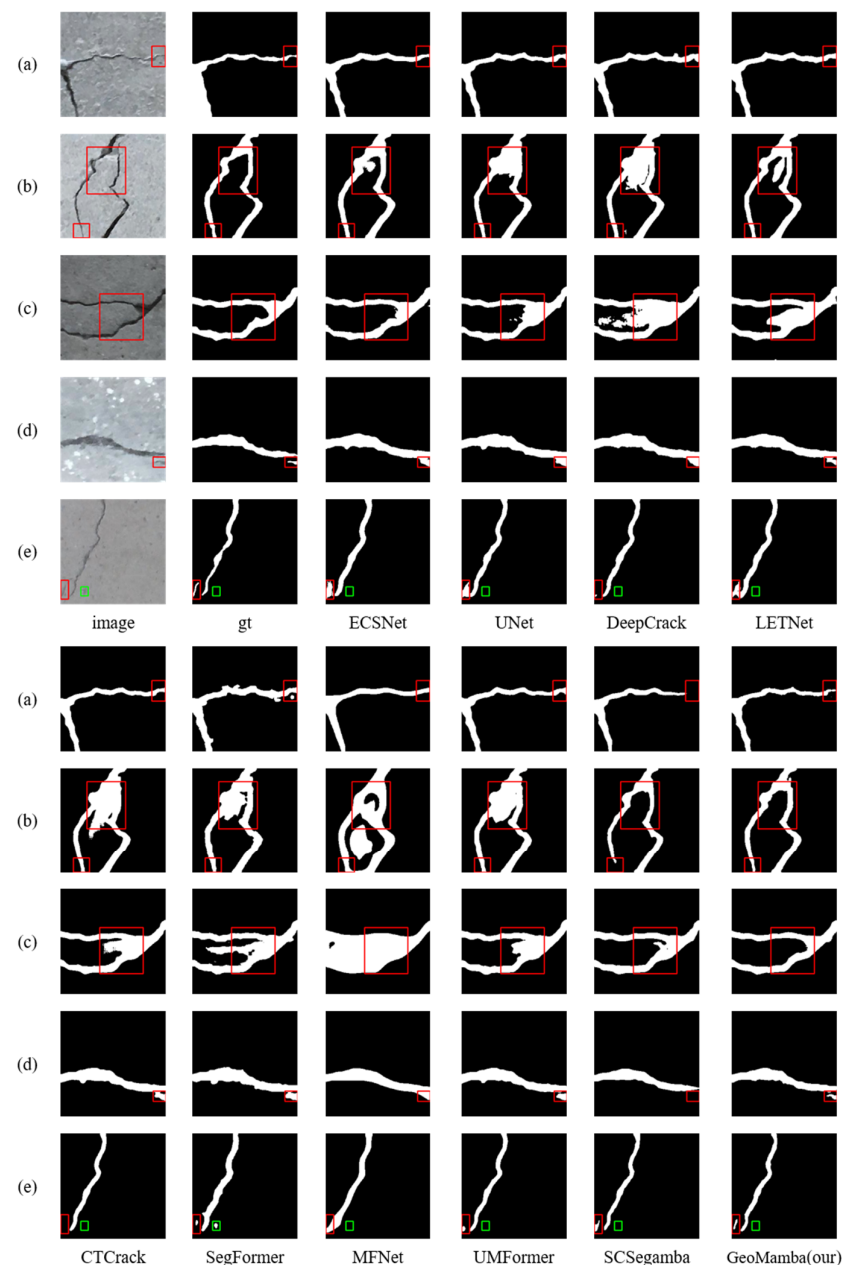


Figure 6. Visual analysis of crack detection results with Concrete3K dataset. (Within these boxes, background noise regions prone to misdetection are highlighted by green squares, while fine and topologically complex crack branches are emphasized by red squares).

The cases with poor model performance are visualized in Figure 7. For ultra-thin cracks with extremely low contrast against complex pavement textures, our GeoMamba model occasionally produces fragmented predictions. This limitation primarily occurs because the local pixel gradients of these microcracks are exceptionally weak. While our structural prior injection and deformable scanning modules effectively suppress background interference, they may inadvertently attenuate these exceptionally faint target features during the feature fusion process. Consequently, the network struggles to maintain the absolute topological continuity of these highly obscured structures.

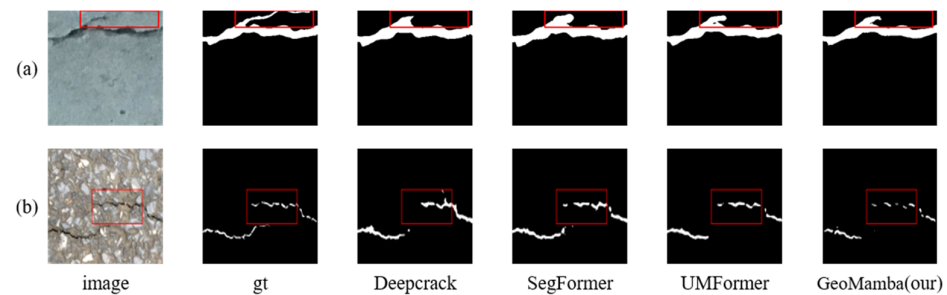


Figure 7. Visual analysis of crack detection results with Concrete3K dataset.

4.2. Complexity Analysis

Figure 8 illustrates the comprehensive performance comparison of various models on the Concrete3K dataset. The vertical axis shows the mIoU and the horizontal axis represents the model size in megabytes. The bubble area corresponds to the parameter count, and the color indicates the computational cost. Additionally, the border thickness of each bubble reflects the inference speed in frames per second (FPS). Compared to heavy architectures like MFNet with 146.02 G FLOPs and 210.21 MB size or DeepCrack with 136.80 G FLOPs and 117.89 MB size, GeoMamba requires significantly less computational overhead and storage space by utilizing only 5.81 G FLOPs and 17.65 MB. Furthermore, when evaluated against other recent models such as UMFormer at 86.59% and SCSEgamba at 85.85%, our method achieves a noticeably higher segmentation accuracy of 87.77%. By occupying the top left region of the chart, GeoMamba clearly demonstrates the optimal balance between high precision and lightweight design among all evaluated baselines.

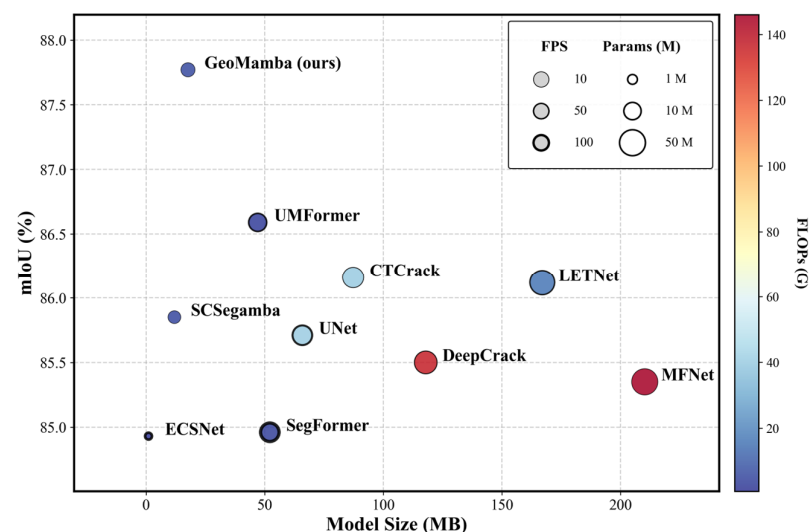


Figure 8. Comparison of complexity with other methods.

4.3. Generalization Analysis

To evaluate the cross-domain generalization capability of the proposed method, we conducted direct testing on the FCN dataset without any prior training. As shown in Table 3, GeoMamba achieves the best mIoU of 69.65% and OIS of 67.48%, demonstrating superior robustness compared to existing CNN, Transformer, and other SSM baselines. However, despite outperforming other competitive models, the absolute performance metrics including a Precision of 43.85% and an F1 score of 58.38% indicate a noticeable performance degradation compared to the in-domain evaluations. This degradation exposes the failure modes of the current architecture when encountering extreme domain shifts. Specifically, the model occasionally fails in extreme environments, such as extremely

low-contrast scenarios where the crack color closely blends with the asphalt background, or under extreme lighting conditions featuring severe shadows and overexposure. In these failure cases, although our deformable scanning strategy and GTPI module strive to maintain topological continuity, the severe environmental noise fundamentally disrupts the pixel-level gradients, occasionally leading to fragmented predictions or false positives.

Table 3. Generalization evaluation on the FCN dataset.

Methods	mIoU/%	F1/%	P/%	R/%	ODS/%	OIS/%
ECSNet	64.36	52.56	41.68	89.81	55.39	57.40
UNet	64.51	53.89	41.00	94.89	57.48	58.53
DeepCrack	64.33	52.93	42.76	90.67	57.86	59.04
LETNet	65.27	54.17	41.42	<u>93.07</u>	57.90	59.02
CTCrack	<u>65.65</u>	53.91	<u>44.62</u>	86.67	57.51	59.80
SegFormer	62.51	49.76	37.79	91.51	56.78	57.77
MFNet	62.20	50.84	40.80	90.59	56.16	57.89
UMFormer	62.50	51.13	40.21	92.11	57.26	59.54
SCSegamba	65.59	59.47	44.99	87.68	59.79	<u>64.58</u>
GeoMamba (ours)	69.65	<u>58.38</u>	43.85	87.29	<u>59.56</u>	67.48

The best and second-best results in the table are marked in boldface and underlined, respectively.

4.4. Ablation Experiments

To further validate the effectiveness of the proposed modules (i.e., MPE and GTPI), ablation experiments were conducted on the DeepCrack dataset. Table 4 lists the results. The baseline model devoid of the MPE, GTPI, GBC, PAF, and GC modules is denoted as Model I. Additionally, the network variants omitting the MPE, GTPI, GBC, PAF, and GC modules individually from the complete architecture are referred to as Model II, Model III, Model IV, Model V, and Model VI, respectively.

Table 4. Ablation experiment results for each module on DeepCrack dataset.

Model	MPE	GTPI	GBC	PAF	GC	mIoU/%	F1/%	P/%	R/%	ODS/%	OIS/%
I						73.66	44.82	29.21	96.22	61.76	64.21
II		✓	✓	✓	✓	<u>80.64</u>	<u>87.14</u>	87.57	86.73	<u>72.66</u>	<u>73.41</u>
III	✓		✓	✓	✓	75.23	82.89	94.34	73.61	63.75	63.84
IV	✓	✓		✓	✓	76.48	63.16	48.48	90.60	66.45	70.23
V	✓	✓	✓		✓	75.44	53.53	37.31	<u>94.72</u>	64.25	66.61
VI	✓	✓	✓	✓		74.43	62.04	46.84	91.85	63.11	66.19
Ours	✓	✓	✓	✓	✓	83.79	89.27	<u>90.61</u>	87.96	79.06	79.78

The best and second-best results in the table are marked in boldface and underlined, respectively.

As listed in Table 4, the baseline architecture Model I demonstrates an inherent bias toward severe over-segmentation. While it achieves an exceptionally high Recall of 96.22%, this comes at the severe expense of Precision 29.21%, indicating a strong propensity to misclassify complex background noise as crack pixels, which consequently degrades the overall F1 score to 44.82%.

The incorporation of the Multi-Scale Patch Embedding MPE module, as observed in Model III, which omits the GTPI module, directly addresses this over-segmentation. It drastically boosts Precision to an impressive 94.34% by effectively filtering out false positives. However, this aggressive suppression inadvertently sacrifices structural continuity, causing the Recall rate to drop to 73.61%. Conversely, Model II, which omits the MPE module but retains the Geometric Topological Prior Injection GTPI module, yields a vastly superior trade-off between Precision 87.57% and Recall 86.73%, significantly elevating the overall segmentation integrity and pushing the mIoU to 80.64%.

Furthermore, the drastic performance degradation observed in Models IV, V, and VI warrants a deeper discussion. Specifically, the individual omission of the GBC, PAF, or GC modules leads to a substantial decline in Precision and a simultaneous rebound in Recall. This phenomenon occurs because the MPE module and the GTPI module extract fundamentally distinct representations: MPE focuses on local high-frequency textures, while GTPI provides global topological priors. The GBC, PAF, and GC modules do not merely act as auxiliary components; rather, they serve as indispensable feature-alignment mechanisms and gating buffers. When these synergistic fusion modules are removed, the raw and potentially conflicting multi-modal features are directly fed into the classification head. This significant semantic gap overwhelms the classifier, causing it to numerically degenerate and default to a conservative overprediction strategy to minimize the training loss. Consequently, the network instinctively predicts almost all suspected pixels as cracks, explaining the sharp return to a high Recall and low Precision state. This behavior strongly proves that our proposed modules must operate together as an inseparable collaborative framework to effectively bridge feature conflicts and achieve optimal geometric and textural balance.

Ultimately, the final proposed architecture demonstrates that the concurrent integration of all modules acts as a highly synergistic framework, achieving an optimal precision–recall equilibrium. In this unified model, the MPE module reliably suppresses background false positives to guarantee a high Precision of 90.61%, while the GTPI module serves as a critical structural compensator. Through its dual-branch geometric topological prior injection, GTPI explicitly restores the continuous 2D spatial topology that might be overly suppressed by the MPE module, thereby preserving a high Recall of 87.96%. This seamless, complementary collaboration empowers the complete model to achieve peak performance across all comprehensive evaluation metrics, culminating in the highest mIoU of 83.79%, F1 score of 89.27%, ODS of 79.06%, and OIS of 79.78%.

To meticulously isolate and evaluate the specific contribution of the deformable scanning strategy to precise crack segmentation, we conducted a comparative ablation study against several established sequence scanning mechanisms, including parallel, snake, bidirectional, and Structure Aware Scanning Strategy (SASS) [23]. Specifically, SASS employs a comprehensive four-route scanning strategy consisting of two parallel snake paths and two diagonal snake paths to effectively capture continuous semantic information across multiple directions. As demonstrated in Table 5, when evaluated under identical experimental settings on the DeepCrack dataset, the network equipped with our deformable scanning strategy consistently delivers the most superior and balanced performance.

Table 5. Ablation study results for different scanning methods on the DeepCrack dataset.

Scan	mIoU/%	F1/%	P/%	R/%	ODS/%	OIS/%
Parallel	80.28	80.80	85.01	76.99	74.64	76.39
Snake	80.00	83.50	82.37	84.67	73.02	74.10
Bidirectional	<u>82.11</u>	<u>87.91</u>	<u>86.74</u>	89.12	<u>76.30</u>	<u>78.18</u>
SASS	78.12	74.02	67.18	82.40	69.63	71.97
Deformable	83.79	89.27	90.61	<u>87.96</u>	79.06	79.78

The best and second-best results in the table are marked in boldface and underlined, respectively.

Notably, traditional rigid scanning methods inherently struggle to align with irregular crack structures, as evidenced by the severe drop in Precision (67.18%) observed with SASS, or the merely moderate baseline performance of parallel and snake scanning. The bidirectional scanning strategy emerges as a strong sub-optimal baseline due to its expanded spatial context, achieving a high Recall of 89.12%. However, it still falls short in accurately filtering out background noise, securing a Precision of only 86.74%.

The deformable scanning strategy decisively overcomes these limitations. Compared to the bidirectional approach, it yields absolute improvements of 1.68% in mIoU and 1.36% in F1 score, reaching 83.79% and 89.27%, respectively. More critically, the deformable approach significantly elevates the Precision to a peak of 90.61%, representing a substantial 3.87% margin over the bidirectional method, without severely compromising Recall. This quantitative leap physically translates to the fundamental structural advantage of our method: whereas traditional fixed scanning trajectories artificially fragment the continuous, tortuous morphologies of natural cracks due to rigid grid constraints, the deformable scanning strategy dynamically adapts its sampling locations. By actively adjusting feature extraction strictly along the intrinsic, irregular textures and semantic flow of the cracks, the network efficiently avoids assimilating irrelevant background noise, thereby achieving unprecedented segmentation precision and boundary fidelity, as evidenced by an ODS of 79.06% and an OIS of 79.78%.

5. Conclusions

This paper proposed GeoMamba, a geometric-prior-infused and multi-scale deformable visual Mamba model for road crack detection. Aiming to extract cracks with highly variable morphologies and severe background interference, the GeoMamba achieved superior performance by integrating the MDVSS and GTPI modules. Specifically, the MDVSS module used adaptive deformable scanning for capturing irregular extension trends and long-range dependencies. Crucially, GTPI compensates for the spatial topology disruption caused by Mamba-based 1D sequence flattening by extracting deterministic high-frequency structural priors, which impose explicit geometric constraints during feature fusion. Synergistically, the network is guided to focus on geometric topologies of cracks, ensuring that feature extraction is strictly aligned with continuous crack structures in complex backgrounds. In summary, the GeoMamba achieved precise and coherent crack segmentation. The experiments were conducted on two public datasets, and the results demonstrated that the GeoMamba obtained mIoU and F1-score values reaching up to 83.79% and 89.27%, respectively. We demonstrated the stability and generalization ability of the proposed method in complex scenarios.

Nevertheless, to address extreme cross-domain shifts, further improving model generalization remains a key challenge. In future research, we plan to investigate vision-language foundation models. By integrating text-prompt-driven semantic priors with our geometric structure perception module, we believe that it will be a promising direction for breaking the generalization bottlenecks under complex scenarios.

Author Contributions: Conceptualization, S.L. and B.L.; methodology, S.L. and B.L.; software, S.L. and B.L.; validation, L.M. and Y.X.; formal analysis, S.L.; investigation, L.M.; resources, B.L. and H.G.; data curation, S.L.; writing—original draft preparation, S.L. and B.L.; writing—review and editing, H.G. and Y.Y.; visualization, S.L.; supervision, H.G. and L.M.; project administration, H.G.; funding acquisition, H.G., L.M. and Y.X. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Key Laboratory of Land Satellite Remote Sensing Application, Ministry of Natural Resources of the People's Republic of China under Grant KLSMNR-G202506 and by the National Natural Science Foundation of China Grants (42371447).

Data Availability Statement: The datasets used are available on <https://github.com/yhlleo/DeepCrack> (accessed on 22 January 2019), <https://github.com/CHDyshi/HrSegNet4CrackSegmentation> (accessed on 12 August 2023) and https://drive.google.com/open?id=1cplcUBmgHfD82YQTWnn1dssK2Z_xRpjx (accessed on 29 August 2018).

Acknowledgments: During the preparation of this manuscript, the authors used Gemini 3.1 (Google) for the purposes of English-language polishing and grammatical correction. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Kheradmandi, N.; Mehranfar, V. A critical review and comparative study on image segmentation-based techniques for pavement crack detection. *Constr. Build. Mater.* **2022**, *321*, 126162. [\[CrossRef\]](#)
2. Hsieh, Y.; Tsai, Y. Machine Learning for Crack Detection: Review and Model Performance Comparison. *J. Comput. Civ. Eng.* **2020**, *34*, 04020038. [\[CrossRef\]](#)
3. Frankowski, P.K.; Matysik, S. Integrating Electromagnetic NDT and IoT for Enhanced Structural Health Monitoring of Corrosion in Reinforced Concrete as a Key to Sustainable Smart Cities. *Sustainability* **2025**, *17*, 10307. [\[CrossRef\]](#)
4. Frankowski, P.K.; Chady, T.; Zieliński, A. Magnetic force induced vibration evaluation (M5) method for frequency analysis of rebar-debonding in reinforced concrete. *Measurement* **2021**, *182*, 109655. [\[CrossRef\]](#)
5. Zhang, T.; Wang, D.; Lu, Y. ECSNet: An Accelerated Real-Time Image Segmentation CNN Architecture for Pavement Crack Detection. *IEEE Trans. Intell. Transp. Syst.* **2023**, *24*, 15105–15112. [\[CrossRef\]](#)
6. Al-Huda, Z.; Peng, B.; Algburi, R.; Al-antari, M.; AL-Jarazi, R.; Al-maqdari, O.; Zhai, D. Asymmetric dual-decoder-U-Net for pavement crack semantic segmentation. *Autom. Constr.* **2023**, *156*, 105138. [\[CrossRef\]](#)
7. Liu, Y.; Yao, J.; Lu, X.; Xie, R.; Li, L. DeepCrack: A deep hierarchical feature learning architecture for crack segmentation. *Neurocomputing* **2019**, *338*, 139–153. [\[CrossRef\]](#)
8. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*; NIPS: La Jolla, CA, USA, 2017.
9. Cheng, X.; He, T.; Shi, F.; Zhao, M.; Liu, X.; Chen, S. Selective feature fusion and irregular-aware network for pavement crack detection. *IEEE Trans. Intell. Transp. Syst.* **2023**, *25*, 3445–3456. [\[CrossRef\]](#)
10. Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, Munich, Germany, 8–14 September 2018; pp. 801–818.
11. Yu, Y.; Guan, H.; Li, D.; Zhang, Y.; Jin, S.; Yu, C. CCapFPN: A Context-Augmented Capsule Feature Pyramid Network for Pavement Crack Detection. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 3324–3335. [\[CrossRef\]](#)
12. Feng, H.; Li, W.; Ma, L.; Chen, Y.; Guan, H.; Yu, Y.; Marcato Junior, J.; Li, J. Crack-U2Net: Multiscale Feature Learning Network for Pavement Crack Detection from Large-Scale MLS Point Clouds. *IEEE Trans. Intell. Transp. Syst.* **2024**, *25*, 17952–17964. [\[CrossRef\]](#)
13. Liu, B.; Kang, J.; Guan, H.; Ma, L.; Yu, Y.; Li, S.; Li, J. Boundary-Guided Real-Time Semantic Segmentation and Pixel-Level Quantification of Pavement Cracks. *IEEE Trans. Intell. Transp. Syst.* **2026**, *27*, 3459–3475. [\[CrossRef\]](#)
14. Liu, H.; Miao, X.; Mertz, C.; Xu, C.; Kong, H. Crackformer: Transformer network for fine-grained crack detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Montreal, BC, Canada, 11–17 October 2021; pp. 3783–3792.
15. Tao, H.; Liu, B.; Cui, J.; Zhang, H. A convolutional-transformer network for crack segmentation with boundary awareness. In *Proceedings of the 2023 IEEE International Conference on Image Processing (ICIP)*, Kuala Lumpur, Malaysia, 8–11 October 2023; pp. 86–90.
16. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 12077–12090.
17. Yuan, Y.; Cheng, Y.; Pan, B.; Jin, G.; Yu, D.; Ye, M.; Zhang, Q. A Multi-Modal Attention Fusion Framework for Road Connectivity Enhancement in Remote Sensing Imagery. *Mathematics* **2025**, *13*, 3266. [\[CrossRef\]](#)
18. Gu, A.; Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. In *Proceedings of the First Conference on Language Modeling*, Philadelphia, PA, USA, 7–9 October 2024.
19. Zhu, L.; Liao, B.; Zhang, Q.; Wang, X.; Liu, W.; Wang, X. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv* **2024**, arXiv:2401.09417.
20. Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; Jiao, J.; Liu, Y. Vmamba: Visual state space model. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 103031–103063. [\[CrossRef\]](#)
21. Liu, M.; Dan, J.; Lu, Z.; Yu, Y.; Li, Y.; Li, X. CM-UNet: Hybrid CNN-Mamba UNet for remote sensing image semantic segmentation. *arXiv* **2024**, arXiv:2405.10530.
22. Li, L.; Yi, J.; Fan, H.; Lin, H. A lightweight semantic segmentation network based on self-attention mechanism and state space model for efficient urban scene segmentation. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 4703215. [\[CrossRef\]](#)

23. Liu, H.; Jia, C.; Shi, F.; Cheng, X.; Chen, S. SCSegamba: Lightweight structure-aware vision mamba for crack segmentation in structures. In Proceedings of the Computer Vision and Pattern Recognition Conference, Nashville, TN, USA, 11–15 June 2025; pp. 29406–29416.
24. Huang, T.; Pei, X.; You, S.; Wang, F.; Qian, C.; Xu, C. Localmamba: Visual state space model with windowed selective scan. In Proceedings of the European Conference on Computer Vision, Milan, Italy, 29 September–4 October 2024; pp. 12–22.
25. Yang, C.; Chen, Z.; Espinosa, M.; Ericsson, L.; Wang, Z.; Liu, J.; Crowley, E.J. Plainmamba: Improving non-hierarchical mamba in visual recognition. *arXiv* **2024**, arXiv:2403.17695.
26. Liu, H.; Jia, C.; Shi, F.; Cheng, X.; Wang, M.; Chen, S. Staircase cascaded fusion of lightweight local pattern recognition and long-range dependencies for structural crack segmentation. *arXiv* **2024**, arXiv:2408.12815.
27. Liu, L.; Zhang, M.; Yin, J.; Liu, T.; Ji, W.; Piao, Y.; Lu, H. Defmamba: Deformable visual state space model. In Proceedings of the Computer Vision and Pattern Recognition Conference, Nashville, TN, USA, 11–15 June 2025; pp. 8838–8847.
28. Song, Y.; Liu, Y. Lightweight Image Super-Resolution Reconstruction Based on Edge Detection. In Proceedings of the 2024 5th International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE), Shenzhen, China, 25–27 October 2024; pp. 292–296.
29. Liu, W.; Lu, H.; Fu, H.; Cao, Z. Learning to upsample by learning to sample. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 6027–6037.
30. Bui, N.-T.; Hoang, D.-H.; Nguyen, Q.-T.; Tran, M.-T.; Le, N. Meganet: Multi-scale edge-guided attention network for weak boundary polyp segmentation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2024; pp. 7985–7994.
31. Li, Q.; Jia, X.; Zhou, J.; Shen, L.; Duan, J. Rediscovering bce loss for uniform classification. *arXiv* **2024**, arXiv:2403.07289.
32. Sudre, C.H.; Li, W.; Vercauteren, T.; Ourselin, S.; Jorge Cardoso, M. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In Proceedings of the International Workshop on Deep Learning in Medical Image Analysis, Québec City, QC, Canada, 14 September 2017; pp. 240–248.
33. Li, Y.; Ma, R.; Liu, H.; Cheng, G. Real-time high-resolution neural network with semantic guidance for crack segmentation. *Autom. Constr.* **2023**, *156*, 105112. [[CrossRef](#)]
34. Yang, X.; Li, H.; Yu, Y.; Luo, X.; Huang, T.; Yang, X. Automatic Pixel-Level Crack Detection and Measurement Using Fully Convolutional Network. *Comput.-Aided Civ. Infrastruct. Eng.* **2018**, *33*, 1090–1109. [[CrossRef](#)]
35. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; pp. 234–241.
36. Xu, Z.; Guan, H.; Kang, J.; Lei, X.; Ma, L.; Yu, Y.; Chen, Y.; Li, J. Pavement crack detection from CCD images with a locally enhanced transformer network. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *110*, 102825. [[CrossRef](#)]
37. Ma, X.; Zhang, X.; Pun, M.-O.; Huang, B. A unified framework with multimodal fine-tuning for remote sensing semantic segmentation. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5405015. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.