MixerCSeg: An Efficient Mixer Architecture for Crack Segmentation via Decoupled Mamba Attention

Supplementary Material

7. Additional Ablation Experiments

7.1. Ablation studies of the TransMixer Module

To thoroughly investigate the effectiveness of the TransMixer module, we conducted comprehensive ablation experiments focusing on this component. Specifically, we evaluated the model performance under three conditions: removing the TransMixer module, removing global tokens, and removing local tokens. As shown in Table 7, when the TransMixer module is removed, the model performance decreases by 0.76% and 0.49% on the DeepCrack [17] and CamCrack789 [31] datasets, respectively. Similarly, models with either the local tokens or the global tokens removed also exhibit varying degrees of performance decline. These experimental results validate the effectiveness of the TransMixer module, demonstrating that each subcomponent plays a specific and meaningful role.

Furthermore, we replaced the maxpooling with avgpooling in the Local Refinement module. As shown in Table 8, compared to avgpooling, maxpooling is more effective in enhancing the most critical and salient crack features within local regions.

Table 7. Experimental results of decoding global tokens and local tokens in the TransMixer module on the DeepCrack and CamCrack789 datasets.

Local	Global	DeepCrack				CamCrack789			
		mIoU	ODS	OIS	F1	mIoU	ODS	OIS	F1
		0.9082	0.9009	0.9054	0.9155	0.8368	0.8057	0.8158	0.8252
✓		0.9120	0.9057	0.9138	**0.9206**	0.8382	0.8081	0.8179	0.8247
	✓	0.9136	0.9079	0.9169	0.9204	0.8384	0.8076	0.8200	**0.8275**
✓	✓	**0.9151**	**0.9094**	**0.9197**	0.9205	**0.8409**	**0.8115**	**0.8202**	0.8244

Table 8. Experimental results of different pooling operations in the Local Refinement Module on the DeepCrack and CamCrack789 Datasets.

Methods	DeepCrack				CamCrack789			
	mIoU	ODS	OIS	F1	mIoU	ODS	OIS	F1
Avgpooling	0.9147	0.9089	0.9192	0.9140	0.8399	0.8100	**0.8227**	**0.8273**
Maxpooling	**0.9151**	**0.9094**	**0.9197**	**0.9205**	**0.8409**	**0.8115**	0.8202	0.8244

7.2. Ablation studies of the DEGConv Module

In Section 3.4, the DEGConv module consists of four components: Rearrange, DEG (Directional Embedding Generation), Edge Convolution, and Gating. To systematically evaluate the necessity and contribution of each component, we conducted the experiments shown in Table 9. The results demonstrate that each component contributes to the improvement of segmentation performance. In particular, the DEG operation leads to notable performance gains, achieving an mIoU increase of 0.36% and 0.46% on the DeepCrack [17] and CamCrack789 [31] datasets, respectively. Although the Rearrange operation alone yields relatively modest improvements, with mIoU gains of 0.24% and 0.05% on the two datasets, respectively, it plays a critical role by optimizing the spatial arrangement of features through a spatial block approach. This results in a more structured feature input, which serves as a solid foundation for the subsequent DEG and Edge Convolution steps, thereby contributing to the overall performance enhancement of the module.

Table 9. The ablation experiment results of DEGConv on the DeepCrack and CamCrack789 datasets, including the Rearrange, DEG, Edge Convolution, and Gating operations.

Rearrange	DEG	EdgeConv	Gating	DeepCrack				CamCrack789			
				mIoU	ODS	OIS	F1	mIoU	ODS	OIS	F1
				0.9052	0.8980	0.9042	0.9052	0.8334	0.8008	0.8122	0.8114
✓				0.9074	0.9004	0.9127	0.9130	0.8338	0.8009	0.8079	0.8162
✓	✓			0.9107	0.9044	0.9143	0.9174	0.8376	0.8065	**0.8237**	0.8252
✓	✓	✓		0.9132	0.9072	0.9125	**0.9205**	0.8398	0.8092	0.8193	**0.8254**
✓	✓	✓	✓	**0.9151**	**0.9094**	**0.9197**	**0.9205**	**0.8409**	**0.8115**	0.8202	0.8244

7.3. Impact of Layer Depth

In the design of the MixerCSeg network architecture, the depth of each TransMixer block is set to 1 to balance computational resources and performance. We quantitatively analyzed the impact of block depth on model performance through controlled experiments in Table 10. While keeping other parameters constant, we tested configuration schemes with TransMixer block depths of 2, 4, and 6 layers, respectively. The experimental results show that when the depth of each layer is set to 1, the model demonstrates optimal segmentation performance with lower computational resource consumption. In comparison, when the network depth is increased to 2 layers, the model's computational load is 3.51 GFLOPs, the number of parameters is 4.76 M, and the memory usage is 1550 MiB. These values represent an increase of 71.2% in computation, 87.4% in parameters, and 30.2% in memory usage. However, since crack segmentation tasks are highly dependent on local fine-grained features such as pixel-level crack morphology and topological continuity, the finer details in deeper networks may gradually become blurred through multiple layers of propagation. This can lead to issues such as over-smoothing at edges.

Table 10. Experiments on different layer depths on the DeepCrack and CamCrack789 Datasets.

Depth	FLOPs (G)	Params (M)	Memory (MiB)	DeepCrack				CamCrack789			
				mIoU	ODS	OIS	F1	mIoU	ODS	OIS	F1
1	**2.05**	**2.54**	**1190**	**0.9151**	**0.9094**	0.9197	**0.9205**	**0.8409**	**0.8115**	0.8202	0.8244
2	3.51	4.76	1550	0.9141	0.9084	**0.9213**	0.9071	0.8381	0.8069	0.8222	0.8208
4	6.42	9.20	2080	0.9126	0.9065	0.9166	0.9173	0.8375	0.8065	**0.8237**	**0.8252**
6	9.33	13.63	3818	0.9073	0.9004	0.9127	0.9130	0.8336	0.8005	0.8134	0.8203

In addition, the difficulty of parameter optimization significantly increases, resulting in degraded performance. Therefore, setting the depth of each TransMixer block to 1 not only ensures the accuracy of crack segmentation but also effectively controls model complexity, providing feasibility for deployment on edge devices.

7.4. Impact of the number of bins on the Crack500

In Section 4.2 of the experimental part, we mentioned that the number of bins n was set to 36 rather than 180 on the Crack500 dataset. To systematically investigate the influence of this parameter, we conducted detailed ablation experiments on the Crack500 [29] dataset, as shown in Table 11, evaluating the model performance with nset to 9, 18, 36, 90, and 180, respectively. The experimental results indicate that the model achieves the best performance when $n = 36$.

To further explore the intrinsic causes of this phenomenon, we conducted a detailed observation of the crack morphologies in the Crack500 dataset. As shown in Figure 5, most cracks are characterized by large width and small curvature with significant background noise interference, while a few cracks (the last one) face the challenge of complex topological structures, and the ratio of the two types is close to 15:1. This observation is consistent with the phenomenon that the optimal performance is achieved when $n = 36$. A moderate n can not only effectively capture the features of cracks with gentle curvature but also maintain good representation ability for wide cracks, and is also capable of handling a small proportion of morphologically complex cracks.

Table 11. Ablation studies of the number of bins n on the Crack500 dataset.

n	mIoU	ODS	OIS	F1
9	0.7805	0.7269	0.7467	0.7575
18	0.7814	0.7273	0.7441	0.7631
36	**0.7824**	**0.7281**	**0.7483**	**0.7755**
90	0.7813	0.7264	0.7455	0.7692
180	0.7774	0.7212	0.7409	0.7592

Figure 5. Representative crack images from the Crack500 dataset.

8. Qualitative Analysis of SRF Module

In the component ablation experiments presented in Table 3, after replacing the decoder of SegFormer [27] with the SRF module, the model's computational cost decreased by 89.3%, the number of parameters was reduced by 33.8%, and GPU memory usage dropped by 67.2%. Meanwhile, on two crack segmentation benchmark datasets, DeepCrack [17] and CamCrack789 [31], the mIoU achieved improvements of 0.59% and 0.32%, respectively.

To gain a deeper insight into the underlying mechanism of this performance gain, we conducted visual analyses on both the original and SRF-processed multi-scale feature maps. As illustrated in Figure 6, after being processed by the SRF module, the semantic discriminability between crack regions and background in feature maps across various scales was significantly enhanced. The raw F_4' only exhibited weak texture responses; however, after refinement by F_1', the activation values in crack regions were substantially strengthened, forming clearer semantic boundaries. This improvement provides more discriminative cues for the segmentation head. Additionally, the alignment capability of the optimized features with high-resolution details was enhanced, effectively alleviating the issue of insufficient fusion between high-level semantics and low-level details in the decoder. Ultimately, these enhancements contributed to the improvement in segmentation accuracy.

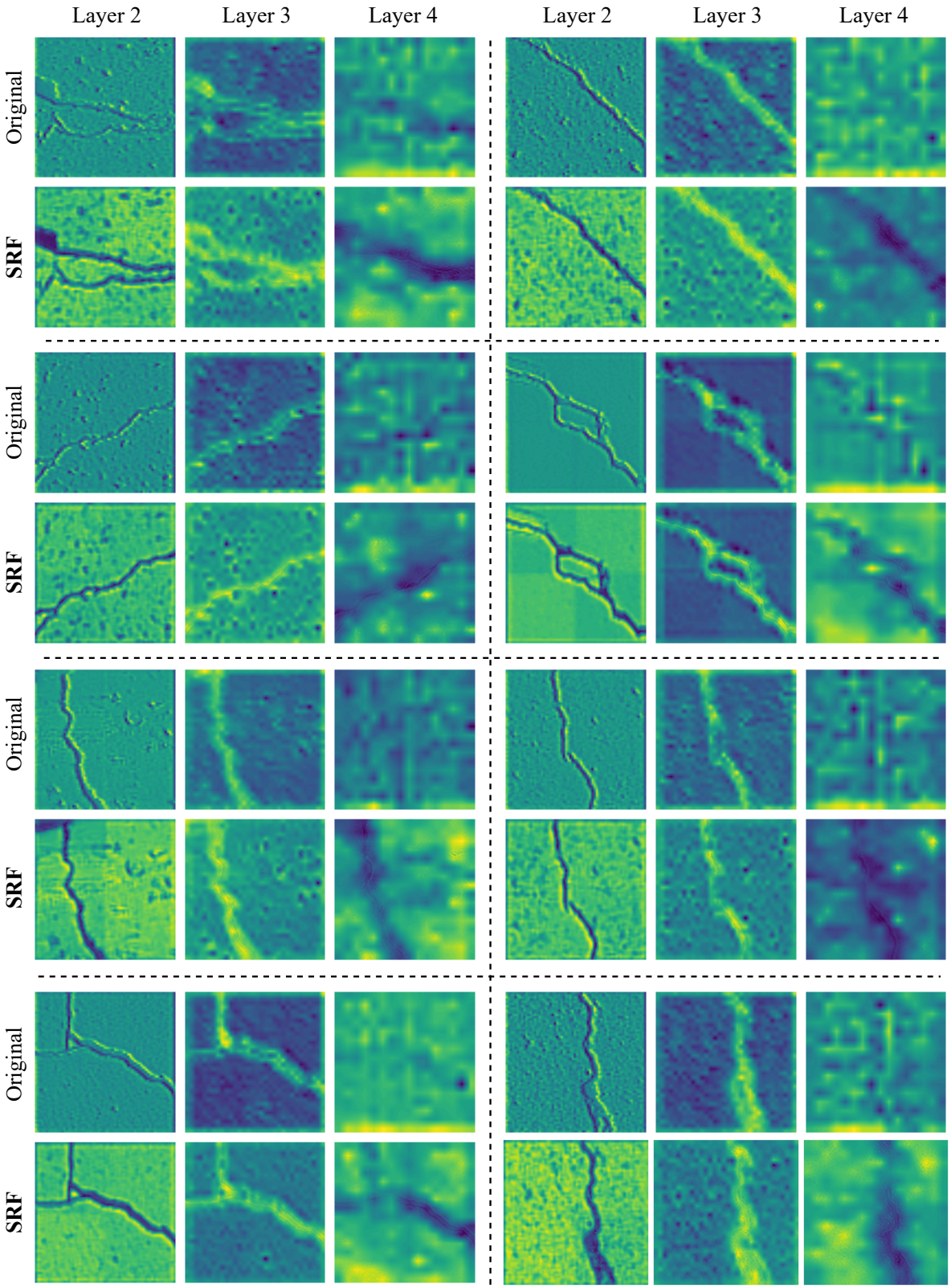

Figure 6. Visual comparison results of multi-scale features before and after SRF module processing. The distinction between cracks and the background is more prominent, and more notably, the crack features in the layer 4 have been significantly enhanced.