MambaLiteUNet: Cross-Gated Adaptive Feature Fusion for Robust Skin Lesion Segmentation

Supplementary Material

6. Evaluation Metrics

We evaluate our segmentation performance using overlap-based, boundary-based, and classification-based metrics. The Intersection over Union (IoU), also known as the Jaccard index, calculates the ratio of the intersection between the predicted and ground truth masks relative to their union. The Dice similarity coefficient (DSC), which is equivalent to the F1 score, emphasizes correct overlaps by giving twice the weight to true positives. For boundary quality, we use the 95th percentile Hausdorff Distance (HD95), which measures the alignment of lesion contours while reducing the influence of outliers. Additionally, we report Accuracy (AC), Sensitivity (SE), and Specificity (SP) to reflect pixel-level classification. The metrics are defined as follows:

$$\text{IoU} = \frac{TP}{TP + FP + FN}, \tag{11}$$

$$\text{DSC} = \frac{2\,TP}{2\,TP + FP + FN}, \tag{12}$$

$$\text{AC} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{13}$$

$$\text{SE} = \frac{TP}{TP + FN}, \tag{14}$$

$$\text{SP} = \frac{TN}{TN + FP}, \tag{15}$$

$$\text{HD}(G, P) = \max(\max_{g \in G} \min_{\hat{p} \in P} d(g, \hat{p}), \max_{\hat{p} \in P} \min_{g \in G} d(\hat{p}, g)) \tag{16}$$

where TP, FP, FN, and TN represent true positives, false positives, false negatives, and true negatives, respectively. G and P denote the ground-truth and predicted masks, and $d(\cdot, \cdot)$ is the Euclidean distance between boundary points. HD95 corresponds to the 95th percentile of these distances, providing a robust measure of boundary alignment.

7. Additional Experiments and Results

To complement the main paper, we present additional experiments that further broaden the evaluation and justification of our framework. These cover boundary-focused metrics such as HD95, cross-dataset generalization, and tests on non-dermoscopic datasets.

7.1. HD95 Evaluation across Four Datasets

To evaluate boundary accuracy, we present HD95 results on ISIC2017 [6], ISIC2018 [5], HAM10000 [32], and PH2 [22], as shown in Table 11. The HD95 measures the boundary accuracy, which is crucial in medical image segmentation, where precise lesion contours are as important as region overlap.

Model	HD95 (↓)			
	ISIC2017	ISIC2018	HAM10000	PH2
U-Net [26]	$16.48^{1.50}$	$19.67^{1.65}$	$15.35^{1.08}$	$18.40^{1.60}$
SCR-Net [35]	$17.06^{1.52}$	$15.82^{1.32}$	$13.90^{0.98}$	$22.80^{1.80}$
TransFuse [42]	$15.04^{1.32}$	$16.77^{1.42}$	$14.76^{1.00}$	$16.10^{1.28}$
UTNetV2 [11]	$17.22^{1.55}$	$17.23^{1.50}$	$18.50^{1.22}$	$22.00^{1.75}$
ASwin U-Net [1]	$15.84^{1.48}$	$19.79^{1.68}$	$16.17^{1.15}$	$18.10^{1.55}$
C^2SDG [15]	$14.30^{1.25}$	$15.21^{1.28}$	$16.00^{1.08}$	$17.30^{1.42}$
UNeXt-S [33]	$14.30^{1.18}$	$15.03^{1.20}$	$14.20^{0.95}$	$16.85^{1.28}$
MALUNet [28]	$14.66^{1.12}$	$14.72^{1.12}$	$13.70^{0.92}$	$15.10^{1.10}$
EGE-UNet [29]	$12.49^{1.02}$	$15.40^{1.35}$	$12.97^{0.86}$	$14.90^{1.02}$
VM-UNet [27]	$14.43^{1.08}$	$14.31^{1.18}$	$13.40^{0.90}$	$15.90^{1.20}$
VM-UNet2 [41]	$14.06^{1.08}$	$14.77^{1.20}$	$12.84^{0.86}$	$15.30^{1.15}$
LightM-UNet [17]	$13.80^{1.12}$	$15.10^{1.24}$	$12.52^{0.80}$	$15.40^{1.12}$
LB-UNet [39]	$12.05^{1.02}$	$14.61^{1.10}$	$9.72^{0.74}$	$14.70^{1.00}$
ULVM-UNet [36]	$12.93^{0.98}$	$15.06^{1.20}$	$12.23^{0.88}$	$12.40^{1.05}$
Ours	$\mathbf{10.73}^{0.92}$	$\mathbf{12.94}^{1.02}$	$\mathbf{8.65}^{0.62}$	$\mathbf{9.88}^{0.90}$

Table 11. HD95 (in pixels) results on ISIC2017, ISIC2018, HAM10000, and PH2. All results are averaged over five runs. Values are reported as mean$^{\text{SD}}$ (equivalent to mean±SD). All SOTA baselines are reproduced using their publicly available implementations with identical train–val–test splits for fair comparison. (↓) indicates Lower is better. Best results are in bold.

Traditional CNN-based models such as U-Net and SCR-Net exhibit significant errors, indicating poor boundary localization. Transformer-based methods, including TransFuse, UTNetV2, and ASwin U-Net, reduce errors in some cases but still face limitations in accuracy. More compact CNN–MLP hybrid models, such as MALUNet, EGE-UNet, LB-UNet, and ULVM-UNet, demonstrate better performance, with LB-UNet achieving 9.72 on HAM10000 and ULVM-UNet achieving 12.40 on PH2. Recent Mamba-based architectures (VM-UNet, VM-UNet2, LightM-UNet, ULVM-UNet) show clear improvements in boundary detection over CNN and Transformer models, reducing HD95 to the 12–15-pixel range, yet they still struggle to achieve consistent accuracy across datasets.

Our model achieves the lowest HD95 scores on all four datasets: 10.73 on ISIC2017, 12.94 on ISIC2018, 8.65 on HAM10000, and 9.88 on PH2. These results reduce boundary error by approximately 1.1–2.5 pixels compared to the best-performing baselines, including recent Mamba-based models. The consistent gains demonstrate that our framework produces sharper lesion boundaries across all four

Model	P(M)↓	F(G)↓	ISIC2017			ISIC2018			HAM10000			PH2			Ours−Model(Avg.)	Cost vs Ours
			IoU↑	DSC↑	HD95↓	IoU↑	DSC↑	HD95↓	IoU↑	DSC↑	HD95↓	IoU↑	DSC↑	HD95↓	IoU/DSC/HD95	Params×/GFLOPs×
H-vmunet [37]	8.97	0.742	84.22	91.43	12.81	81.78	89.98	14.67	89.54	94.48	9.47	87.30	93.22	10.06	+1.41 / +0.81 / -1.20	18.2× / 2.3×
WTCM-UNet [10]	28.74	3.12	80.21	89.02	15.67	80.90	89.44	15.24	86.31	92.65	12.75	86.72	92.89	14.76	+3.58 / +2.09 / -4.06	58.2× / 9.6×

Table 12. Comparison with recent Mamba-based segmentation models. H-vmunet and WTCM-UNet are re-implemented and evaluated under our unified training and evaluation protocol for fair comparison. IoU/DSC are reported in %, and HD95 is reported in pixels. Params (M) and GFLOPs are denoted by P(M) and F(G), respectively.

benchmarks, ensuring reliable segmentation performance.

7.2. Additional Comparison with Recent Mamba-based Segmentation Models

Table 12 provides an additional comparison with two recent Mamba-Based segmentation models, H-vmunet [37], and WTCM-UNet [10], evaluated under the same protocol. Although both methods use substantially larger model capacity, MambaLiteUNet remains more effective and efficient. On average, our model surpasses H-vmunet by 1.41 points in IoU and 0.81 points in DSC while reducing HD95 by 1.20 pixels. Compared with WTCM-UNet, our model improves IoU and DSC by 3.58 and 2.09 points, respectively, and reduces HD95 by 4.06 pixels, while using much fewer parameters and GFLOPs.

7.3. Cross-Dataset Generalization Analysis

We evaluate cross-dataset generalization by training on ISIC2018 (train split) and directly testing on PH2 (whole dataset) without fine-tuning. This setting assesses whether models can reliably segment images of the same modality collected at different centers under varying acquisition conditions, providing a rigorous measure of domain robustness relevant to real-world applications.

As shown in Table 13, U-Net obtains 77.02% IoU, 87.02% DSC, and 22.95 HD95, showing limited transferability. TransFuse (80.56% IoU, 89.23% DSC, 18.70 HD95) and UTNetV2 (79.94% IoU, 88.85% DSC, and 18.82 HD95) improve overlap but remain inconsistent. Recent methods EGE-UNet (81.11% IoU, 89.57% DSC, 17.36 HD95) and ULVM-UNet (81.35% IoU, 89.72% DSC, 17.07 HD95) narrow the gap but still suffer from boundary-precision errors.

Our model achieves the best overall results with 81.71% IoU, 89.93% DSC, 93.19% accuracy, and 15.58 HD95, outperforming CNN-, Transformer-, and Mamba-based baselines. Compared with ULVM-UNet, we gain +0.36 IoU, +0.21 DSC, and a reduction of 1.49 HD95. Compared with EGE-UNet, the improvements are +0.60 IoU, +0.36 DSC, and a reduction of 1.78 in HD95. Compared with LightM-UNet, we reduce HD95 by 1.05 while maintaining IoU and DSC. These improvements highlight stronger overlap accuracy and sharper boundaries under domain shift.

Model	Train on ISIC2018 → Test on PH2					
	IoU↑	DSC↑	AC↑	SP↑	SE↑	HD95↓
U-Net [26]	77.02	87.02	90.89	89.11	94.63	22.95
SCR-Net [35]	78.93	88.23	92.17	92.76	90.94	19.54
ASwin U-Net [1]	75.01	85.72	90.53	91.71	88.07	21.76
TransFuse [42]	80.56	89.23	92.58	91.29	**95.29**	18.70
UTNetV2 [11]	79.94	88.85	92.69	**93.86**	90.23	18.82
C^2SDG [15]	79.83	88.79	92.29	91.22	94.54	21.53
UNeXt-S [33]	80.70	89.32	92.71	91.85	94.52	18.42
MALUNet [28]	79.87	88.81	92.46	92.32	92.74	19.62
EGE-UNet [29]	81.11	89.57	93.08	93.56	92.07	17.36
VM-UNet [27]	80.75	89.35	92.79	92.34	93.74	17.12
VM-UNet2 [41]	80.94	89.47	92.76	91.58	95.25	17.76
LightM-UNet [17]	81.10	89.56	92.97	92.71	93.52	16.63
LB-UNet [39]	81.17	89.61	92.91	92.03	94.76	17.38
ULVM-UNet [36]	81.35	89.72	92.96	91.90	95.18	17.07
Ours	**81.71**	**89.93**	**93.19**	92.68	94.26	**15.58**

Table 13. Cross-dataset generalization performance when training on ISIC2018 and testing on the full PH2 dataset without fine-tuning. (↑) indicates higher is better. (↓) indicates lower is better. Best results are in bold.

7.4. Generalization to Non-Dermoscopic Datasets

To evaluate generalization beyond dermoscopic images, we extend our analysis to the BUS [2] and GlaS [31] datasets. BUS contains breast ultrasound scans with heavy speckle noise, low contrast, and irregular lesion boundaries, while GlaS comprises colorectal histopathology images characterized by structural complexity and staining variability. Both datasets present substantially different challenges compared to dermoscopic benchmarks. Table 14 summarizes the results. On BUS, U-Net achieves 67.03% IoU, 80.26% DSC, and 22.72 HD95, while EGE-UNet records 65.81% IoU, 79.38% DSC, and 19.29 HD95. Transformer-based methods perform better but remain inconsistent. TransFuse obtains 70.16% IoU, 82.46% DSC, and 18.46 HD95 with sensitivity at 79.68, while UTNetV2 achieves 68.63% IoU, 81.40% DSC, and 25.15 HD95 with sensitivity at 86.81. Mamba-based architectures improve boundary accuracy, with VM-UNet reaching 72.02% IoU, 83.74% DSC, and 15.29 HD95, and LightM-UNet 71.44% IoU, 83.34% DSC, and 15.37 HD95, but their overlap scores remain limited.

Our model achieves 77.68% IoU, 87.44% DSC, and 11.55 HD95 on BUS, improving over the strongest baseline (C^2SDG: 73.11 IoU, 84.47 DSC, 13.32 HD95) by +4.57 IoU, +2.97 DSC, and a reduction of 1.77 in HD95. This

Model	BUS (Ultrasound) [2]						GlaS (Histopathology) [31]					
	IoU↑	DSC↑	AC↑	SP↑	SE↑	HD95↓	IoU↑	DSC↑	AC↑	SP↑	SE↑	HD95↓
U-Net [26]	67.03	80.26	97.95	98.31	90.46	22.72	72.69	84.19	83.86	83.75	83.96	25.30
TransFuse [42]	70.16	82.46	98.44	99.34	79.68	18.46	73.49	84.72	83.47	77.10	**89.55**	25.49
UTNetV2 [11]	68.63	81.40	98.17	98.72	86.81	25.15	67.67	80.72	81.52	87.74	75.59	25.12
C^2SDG [15]	73.11	84.47	98.59	99.33	83.30	13.32	75.49	86.04	85.49	83.56	87.34	24.28
UNeXt-S [33]	72.11	83.80	98.51	99.23	83.67	15.88	74.41	85.33	85.00	84.71	85.27	25.17
MALUNet [28]	67.19	80.37	98.29	99.35	76.19	22.75	74.64	85.48	85.70	89.29	82.27	24.17
EGE-UNet [29]	65.81	79.38	98.35	**99.77**	68.95	19.29	71.25	83.21	83.70	88.69	78.93	25.06
VM-UNet [27]	72.02	83.74	98.37	98.72	**91.11**	15.29	72.64	84.15	84.67	90.05	79.54	24.94
LightM-UNet [17]	71.44	83.34	98.59	99.67	76.32	15.37	69.40	81.94	81.72	82.42	81.04	28.06
LB-UNet [39]	63.75	77.86	98.14	99.45	71.07	14.49	71.30	83.24	84.23	**92.26**	76.56	24.66
ULVM-UNet [36]	70.19	82.49	98.49	99.53	76.97	15.23	73.33	84.61	84.24	83.80	84.67	26.66
Ours	**77.68**	**87.44**	**98.88**	99.51	85.53	**11.55**	**78.63**	**88.04**	**87.91**	88.90	86.96	**21.62**

Table 14. Performance comparison of SOTA models on BUS and GlaS datasets. For BUS, 163 samples are split into 114 for training, 18 for validation, and 31 reserved for testing. For GlaS, 165 samples are split into 70 for training and 15 for validation, with 80 reserved for testing. All models are trained and evaluated on the same partitions. (↑) indicates higher is better. (↓) indicates Lower is better. Best results are in bold.

performance underlines the model's ability to retain both overlap accuracy and precise boundary localization under heavy noise and low contrast. On GlaS, MALUNet delivers 74.64% IoU, 85.48% DSC, and 24.17 HD95, while our model achieves 78.63% IoU, 88.04% DSC, and 21.62 HD95, improving by +3.99 IoU, +2.56 DSC, and a reduction of 2.55 in HD95. Here, the gains show that our approach adapts to structural irregularities and staining variations that cause other baselines, including Mamba-based ones, to degrade. Therefore, our model demonstrates robustness across imaging modalities by maintaining consistent improvements on BUS and GlaS.

7.5. Robustness to Reduced Training Data

To assess robustness under limited supervision, we train the model with only {50, 70, 100}% of the original training split on ISIC2017 and ISIC2018, while keeping the test sets unchanged. As shown in Table 15, performance degrades steadily as the amount of training data decreases. On ISIC2017, reducing the training data from 100% to 50% lowers mIoU from 85.55 to 83.30 and DSC from 92.21 to 90.89, while HD95 increases from 10.73 to 13.24. A similar trend is observed on ISIC2018, where mIoU and DSC decrease from 83.60/91.07 to 81.49/89.80, while HD95 rises from 12.94 to 14.99. The performance drop remains modest on both datasets, which suggests that the model can still learn stable and discriminative representations even when annotation is substantially reduced. This behavior is especially important in medical image segmentation, where collecting dense pixel-level labels is costly and often limited.

8. Additional Ablation Study

This section presents additional ablation studies to evaluate the impact of our design decisions further.

Table 15. Robustness to reduced training data on ISIC2017 and ISIC2018. mIoU and DSC are reported in %, and HD95 is reported in pixels. Results are obtained by randomly subsampling {50, 70, 100}% of the training split, while keeping the test set unchanged. All metrics are computed on the full test set.

Training Data Size	ISIC2017			ISIC2018		
	mIoU↑	DSC↑	HD95↓	mIoU↑	DSC↑	HD95↓
50%	83.30	90.89	13.24	81.49	89.80	14.99
70%	84.14	91.39	12.06	82.23	90.25	13.61
100%	**85.55**	**92.21**	**10.73**	**83.60**	**91.07**	**12.94**

Loss		Performance Metrics				
BCE	Dice	IoU↑	DSC↑	AC↑	SP↑	SE↑
✓		86.95	93.02	93.28	98.16	88.51
	✓	86.59	92.81	93.14	**98.84**	87.57
✓	✓	**88.54**	**93.92**	**94.08**	97.79	**90.45**

Table 16. Results for different Loss functions on PH2. (↑) indicates higher is better. ✓ indicates loss selection. Our results are averaged over five independent runs, and the best are in bold.

8.1. Loss Function Analysis on PH2

Table 16 compares three loss variants on the PH2 dataset [22]. Using binary cross-entropy (BCE) [3] alone, we achieve 86.95% IoU and 93.02% DSC, while Dice loss (Dice) [3] alone gives 86.59% IoU and 92.81% DSC. Combining BCE and Dice loss, MambaLiteUNet produces the best results (88.54% IoU, 93.92% DSC) and increases sensitivity to 90.45%, representing an absolute sensitivity gain of 1.94 points over BCE loss alone and 2.88 points over Dice loss alone. Therefore, our findings suggest that the hybrid loss stabilizes training and improves both overlap and boundary alignment.

Modules w/o Mamba			Complexity		ISIC2017					ISIC2018				
AMF	LGFM	CGA	Params (M)↓	GFLOPs↓	IoU↑	DSC↑	AC↑	SP↑	SE↑	IoU↑	DSC↑	AC↑	SP↑	SE↑
✓			0.321	**0.237**	83.80	91.18	96.57	98.02	90.59	81.96	90.09	95.81	**98.39**	86.66
	✓		**0.194**	0.332	82.82	90.60	96.27	97.38	**91.72**	81.20	89.62	95.50	97.52	88.33
		✓	0.664	0.383	83.70	91.13	96.53	97.91	90.87	82.17	90.21	95.79	97.87	88.38
✓	✓		0.420	0.352	83.15	90.80	96.44	98.07	89.75	82.02	90.12	95.72	97.62	88.95
✓		✓	0.559	0.359	84.03	91.32	**96.71**	**98.77**	88.28	82.48	90.40	95.81	97.55	**89.66**
	✓	✓	0.420	0.344	83.94	91.27	96.60	98.03	90.74	82.50	90.41	95.89	98.07	88.16
✓	✓	✓	0.658	0.374	**84.26**	**91.46**	96.70	98.34	90.01	**82.66**	**90.51**	**95.91**	97.90	88.83

Table 17. Ablation study of AMF, LGFM, and CGA modules under the Mamba-off/Mamba-free control settings on ISIC2017 and ISIC2018 datasets. Complexity is measured by Params (M) and GFLOPs. (↑) indicates higher is better, while (↓) indicates lower is better.

Module	Design Goal	Our Mechanism	Closest Prior	Key Distinction → Expected Benefit
AMF	Scale–adaptive multi-branch feature fusion under tight compute.	Channels are split into parallel Mamba SSM branches, then merged through a two-stage DW→PW gating pipeline that adapts routing to the input. Residual reweighting ensures stability.	ResNeXt (grouped conv with fixed cardinality).	Fixed partitions in ResNeXt vs. dynamic gating with Mamba branches + dual gates → Content-aware allocation of capacity, sharper interiors, and reduced under-/over-emphasis structures at similar cost.
LGFM	Fuse local detail with long-range context inside a single block.	A dual-path block: DWConv(3×3) for textures and edges, + MHA for global dependencies. Features are concatenated and projected back in a single residual unit (no external fusion head).	TransFuse (CNN and Transformer encoders fused by external BiFusion).	Separate encoders + late fusion vs. In-block local–global mixing. → Less redundancy, balanced paths, and stronger boundary retention (lower HD95).
CGA	Denoise or reduce background information and regulate skip connections before decoder fusion.	Cross-gated skip aggregation: encoder–decoder pairs refined with Mamba, projected through DWConv+sigmoid, and gated bidirectionally before fusion.	Attention U-Net (decoder-driven, one-way gating of encoder skips).	One-way decoder gating vs. Bidirectional pre-fusion gating. → Cleaner skips, suppressed background noises, and sharpens edges with minimal overhead.

Table 18. Comparative analysis of our AMF, LGFM, and CGA modules against their closest priors (ResNeXt [38], TransFuse [42], and Attention U-Net [23]). The table highlights how architectural differences in design goals and mechanisms translate into measurable segmentation gains. Abbreviations: DW: depthwise convolution; PW: pointwise convolution; SSM: state space model [12]; MHA: multi-head self-attention [7].

8.2. Effect of Core Architectural Modules without (w/o) Mamba Integration

To isolate the contributions of our Adaptive Multi-branch Feature Fusion (AMF), Local-Global Feature Mixing (LGFM), and Cross-Gated Attention (CGA) from SSM-based long-range modeling, we design Mamba-off control experiments. In this configuration, all Mamba layers are replaced with token-MLPs (two-layer feed-forward networks with LayerNorm, GELU nonlinearity, and residual connection). In this substitution, we aim to preserve a similar parameter count and the same channel dimensionality while removing Mamba's structured recurrent dynamics.

Our originally proposed AMF, LGFM, and CGA modules are inspired by selective gating principles and incorporate Mamba layers in the full model. In the Mamba-off control, however, these modules adopt the same selective gating principles but are implemented using convolutional layers, multi-head self-attention, and token-MLPs, without invoking Mamba kernels or SSM recurrence. Consequently, we can disentangle the contribution of the Mamba-integrated and Mamba-off approaches.

Table 17 shows the impact of AMF, LGFM, and CGA under the Mamba-off setting. On ISIC2017, AMF alone achieves the best single-module improvement (83.80% IoU, 91.18% DSC) with minimal cost, while LGFM improves sensitivity (91.72%). CGA provides balanced gains but requires higher complexity. Pairwise combinations further enhance performance, with AMF+CGA achieving the highest accuracy and specificity, and the full configuration reaching the best overall results (84.26% IoU, 91.46% DSC). On ISIC2018, results trends are consistent: AMF improves overlap, LGFM boosts sensitivity, and CGA strengthens boundary quality. The full setup (AMF+LGFM+CGA without Mamba) achieves 82.66% IoU and 90.51% DSC, with 0.658M parameters and 0.374 GFLOPs.

However, when comparing with the Mamba-integrated

configuration (see Table 7 in the main manuscript), we observe clear performance boosts. With Mamba, individual modules improve their performance. The full design (AMF+LGFM+CGA with Mamba) achieves 85.55% IoU and 92.21% DSC on ISIC2017 and 83.60% IoU and 91.07% DSC on ISIC2018, outperforming all Mamba-off results while remaining lightweight (0.494M parameters, 0.326 GFLOPs). Therefore, Mamba integration with AMF, LGFM, and CGA is critical, which demonstrates consistent improvements across datasets with minimal overhead.

9. Comparative Analysis of Module Designs

To further clarify the novelty of our proposed AMF, LGFM, and CGA, we provide a detailed comparison with their closest prior designs. Table 18 summarizes each module's design goal, mechanism, nearest prior, and the architectural differences that lead to the expected improvements in lesion segmentation.

10. Module-wise Feature Map Visualization

Figure 5 provides a qualitative comparison of representative feature maps with and without the key modules in MambaLiteUNet. The top row presents the feature responses from the full model with AMF, LGFM, and CGA, while the bottom row shows the corresponding feature responses after removing each module. This comparison highlights how each component shapes the internal spatial representation.

With AMF, the feature response is more structured and lesion-aware, which reflects its role in adaptive feature refinement. LGFM produces the clearest and most coherent lesion-focused activation; when it is removed, the feature map becomes weaker and less discriminative, indicating the importance of local-global feature integration. CGA mainly strengthens boundary-sensitive structure. With CGA, the lesion contour is more clearly emphasized, whereas removing it yields a smoother and less selective response around the lesion region. Therefore, these visualizations show that the three modules contribute in complementary ways. AMF improves adaptive refinement, LGFM strengthens lesion-focused representation, and CGA enhances boundary-aware filtering. Together, they produce more informative and spatially coherent intermediate features.

11. Stage-wise Feature Map Visualization

This section provides stage-wise qualitative evidence of how MambaLiteUNet processes lesion images throughout its encoder–decoder pipeline, complementing the quantitative results presented in the main manuscript. Figure 6 illustrates how our proposed MambaLiteUNet progressively transforms feature representations. We visualize intermediate feature maps using a model pre-trained on ISIC2018 [5] and tested on a held-out image. The top row shows the input image, followed by activation maps from each encoder stage (Encoder1–Encoder5) and the bottleneck. As depth increases, the model learns progressively more abstract and localized features that emphasize lesion boundaries and suppress background noise.

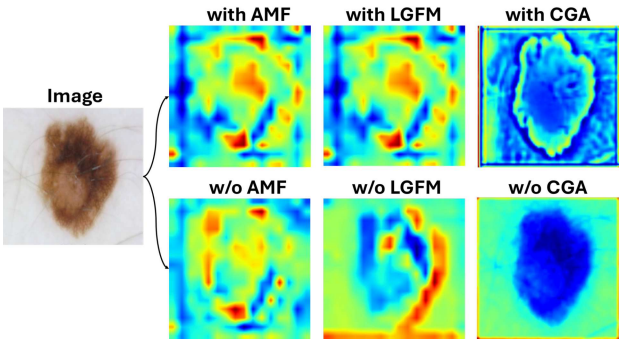

Figure 5. Qualitative visualization of module contributions. The top row shows representative feature maps when each module is enabled, while the bottom row shows the corresponding maps when the module is removed. "with" denotes the full model containing the module, and "w/o" denotes the variant where the module is disabled.

The bottom row (left→right) shows the ground-truth mask, the model's final output, and then decoder activations from Decoder5 through Decoder1. Early decoder blocks (Decoder1 and Decoder2) recover coarse structure, while later blocks (Decoder3–Decoder5) refine contours and sharpen lesion boundaries. This validates our model's hierarchical encoding, skip-guided decoding, and reconstruction. Arrows indicate the forward flow of information through the network.

Model	Sec/Image ↓	Memory (MB) ↓
VM-UNet [27]	0.1718	582.5
VM-UNet2 [41]	0.1836	613.7
LightM-UNet [17]	0.0194	63.6
ULVM-UNet [36]	**0.0058**	**17.4**
Ours	0.0167	54.5

Table 19. Inference efficiency comparison among Mamba-based models. Latency (Sec/Image) and peak GPU memory (MB) at 256×256 with batch size 1 evaluated on an NVIDIA RTX 3090 Ti (24 GB) using CUDA timing and PyTorch memory profiling. (↓) indicates lower is better. Best results in bold.

12. Inference Time and Memory Usage

Table 19 presents a comparative analysis of inference time and memory for Mamba-based models. VM-UNet (0.1718 Sec/Image, 582.5 MB) and VM-UNet2 (0.1836 Sec/Image, 613.7 MB) are the most computationally expensive. LightM-UNet is considerably lighter (0.0194

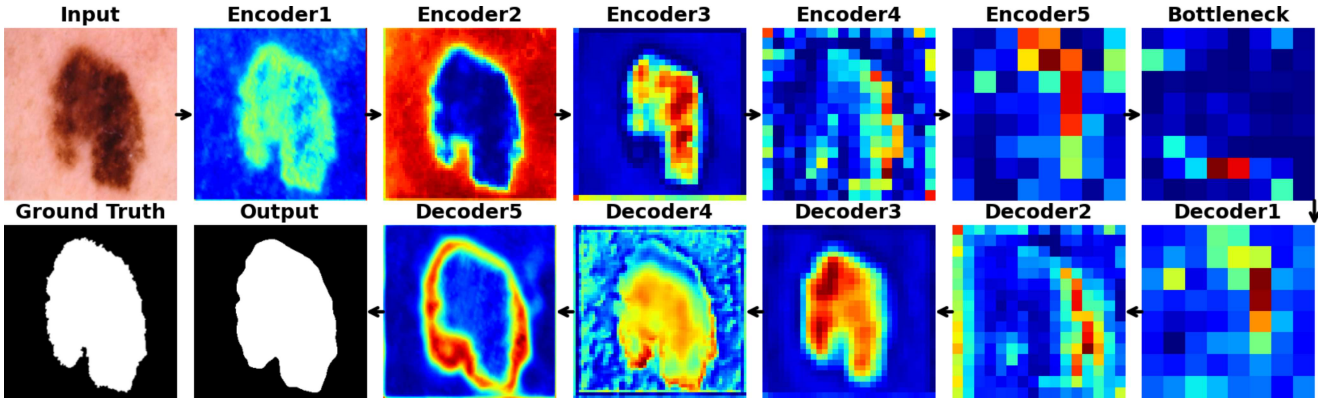

Figure 6. Feature map visualization in MambaLiteUNet. Top row (left→right): input image, Encoder1–Encoder5, Bottleneck—features grow more abstract and emphasize lesion boundaries. Bottom row (left→right): ground-truth mask, model output, and Decoder5–Decoder1 activations. Decoding flows from Decoder1 through Decoder5 to the output (arrows): early blocks (Decoder1/Decoder2) recover coarse structure, while later blocks (Decoder4/Decoder5) refine and sharpen lesion contours. Arrows indicate the forward flow of information.

Sec/Image, 63.6 MB), and ULVM-UNet achieves the best efficiency at 0.0058 Sec/Image and 17.4 MB. In comparison, our MambaLiteUNet operates at 0.0167 Sec/Image with 54.5 MB, slightly above ULVM-UNet in cost but offering stronger representational power and higher segmentation accuracy (see Sec. 7), making it a balanced choice for accuracy and deployment efficiency.