SFCI-MAMBA: FACTORIZED AMPLITUDE–PHASE MODELING WITH SPATIAL–FREQUENCY INTERACTION FOR SKIN LESION SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Skin lesion segmentation requires both boundary detail and global structure. We study two questions in spatial–frequency state-space modeling: how to model Fourier amplitude and phase, and how spatial and spectral features should interact before fusion. We present SFCI-Mamba, a compact encoder–decoder built around spatial-frequency collaborative state-space blocks (SFCBs). Each SFCB independently maps amplitude- and phase-domain representations with Mamba, while dual-context interaction (DCI) lets each branch retrieve from a shared spatial-frequency context using its own query before fusion. SFCI-Mamba achieves micro-averaged Dice scores of 0.9121, 0.9001, and 0.9228 on ISIC2017, ISIC2018, and PH^2, respectively. The PH^2 result is obtained using the ISIC2018-trained model without fine-tuning. Among the tested variants, frequency modeling outperforms spatial replacements, joint amplitude–phase modeling gives the strongest overlap, and DCI improves over the tested interaction alternatives. These results support factorized spectral state-space modeling with pre-fusion spatial–frequency interaction for dermoscopic lesion segmentation.

1 INTRODUCTION

Skin lesion segmentation delineates lesion regions in dermoscopic images and supports quantitative image analysis (Codella et al., 2018). Accurate delineation remains difficult when lesion borders fade into surrounding skin, lesion appearance varies substantially, or hair occludes lesion regions (Xia et al., 2025). These challenges call for representations that preserve boundary detail while capturing the global lesion structure.

Deep segmentation networks have made substantial progress in addressing these challenges. U-Net (Ronneberger et al., 2015) recovers spatial detail through encoder–decoder skip connections, while Transformer-based designs capture broader context (Chen et al., 2021; Zhang et al., 2021). Visual state-space networks further introduce selective scanning for medical segmentation (Ruan et al., 2025), including compact variants (Wu et al., 2025b). Frequency representations provide complementary structural information: DD-Mamba (Liu et al., 2025) uses spatial and frequency Mamba paths with low/high-frequency separation, while HFSTI-Net (He et al., 2026) develops hierarchical spatial-frequency interaction. Neither explicitly factorizes Fourier amplitude and phase inside the state-space transformation.

Fourier amplitude measures component strength, whereas phase contributes to spatial arrangement (Oppenheim & Lim, 1981). Spatial-Frequency Enhanced Mamba (Sun et al., 2025) decomposes amplitude and phase for image fusion, but maps them with convolutions. This motivates two questions for segmentation: *(i)* can amplitude- and phase-domain features be modeled separately with state-space mappings, and *(ii)* can spectral and spatial features exchange information before fusion?

We address these questions with SFCI-Mamba, built around spatial-frequency collaborative state-space blocks (SFCBs). Each SFCB contains a spatial branch and a frequency branch that factorizes the spectral input into amplitude and phase and maps the two domains independently with Mamba. Dual-context interaction (DCI) then uses branch-specific queries to retrieve from a shared spatial-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

frequency context before fusion. SFCBs are used throughout a U-shaped encoder–decoder to predict the lesion mask.

Our contributions are summarized as follows:

- We introduce factorized spectral state-space modeling that decomposes the Fourier representation into amplitude and phase, maps the two domains independently with Mamba, and reconstructs a learned complex representation.
- We propose DCI, which uses branch-specific queries to retrieve from a shared spatial-frequency context and update both branches before fusion.
- We analyze spatial versus frequency modeling, amplitude/phase variants, and cross-branch interaction, with cross-dataset evaluation from ISIC2018 to PH^2.

2 Related Work

2.1 Skin Lesion Segmentation

Skin lesion segmentation requires delineating irregular boundaries while retaining contextual information. U-Net (Ronneberger et al., 2015) combines an encoder–decoder architecture with skip connections to recover spatial detail. Transformer (Vaswani et al., 2017) and ViT (Dosovitskiy et al., 2021) use attention mechanisms for contextual modeling. TransFuse (Zhang et al., 2021) combines CNN and Transformer branches in parallel for medical image segmentation. For lightweight skin lesion segmentation, MALUNet (Ruan et al., 2022) integrates multiple attention mechanisms, and EGE-UNet (Ruan et al., 2023) uses grouped attention and multiscale feature aggregation.

Mamba (Gu & Dao, 2024) introduces selective state-space modeling, while VMamba (Liu et al., 2024) adapts it to two-dimensional visual data. Building on these foundations, VM-UNet (Ruan et al., 2025) uses visual state-space blocks within an encoder–decoder architecture, and H-vmunet (Wu et al., 2025a) integrates high-order interactions and local feature processing with two-dimensional selective scanning. UltraLight VM-UNet (Wu et al., 2025b) uses parallel Mamba processing to reduce parameter cost. These advances motivate incorporating complementary feature representations into state-space segmentation networks.

2.2 Spatial-Frequency Representation Learning

Spatial and frequency representations have been jointly studied in medical image segmentation and related vision tasks. Y-Net (Farshad et al., 2022) uses dual encoders to extract spatial and spectral features. DD-Mamba (Liu et al., 2025) uses Mamba paths in both domains and processes low- and high-frequency components separately. HFSTI-Net (He et al., 2026) integrates spatial and frequency features through hierarchical interactions while modeling temporal information for video polyp segmentation. Spatial-Frequency Enhanced Mamba (Sun et al., 2025) decomposes Fourier features into amplitude and phase spectra and applies separate convolutional mappings before reconstruction for multimodal image fusion.

These studies differ from our method in either spectral representation or spatial–frequency interaction. DD-Mamba and HFSTI-Net use spatial–frequency modeling without explicit amplitude/phase factorization, whereas Spatial-Frequency Enhanced Mamba factorizes amplitude and phase but maps them with convolutions for image fusion. SFCB instead combines amplitude/phase factorization with independent state-space mappings, while DCI lets each branch query a shared spatial-frequency context before fusion.

3 Method

3.1 Overview of SFCI-Mamba

We propose SFCI-Mamba for skin lesion segmentation, integrating spatial and frequency modeling within a U-shaped encoder–decoder network. Figure 1 presents the overall architecture. Its core unit, the spatial-frequency collaborative state-space block (SFCB), contains a spatial branch and a

frequency branch that models amplitude and phase separately. A dual-context interaction (DCI) module then enables information exchange between the two branches before fusion.

Given an RGB image $I \in \mathbb{R}^{H\times W\times 3}$, shallow convolutions first produce features at one-quarter resolution. Four encoder stages progressively extract multiscale representations using SFCBs, followed by a convolutional bottleneck. The decoder progressively restores spatial resolution and incorporates encoder features through additive skip connections. Each encoder and decoder stage contains two SFCBs. Finally, convolutional projection and bilinear upsampling produce a lesion probability map $P \in [0,1]^{H\times W}$ through a sigmoid activation.

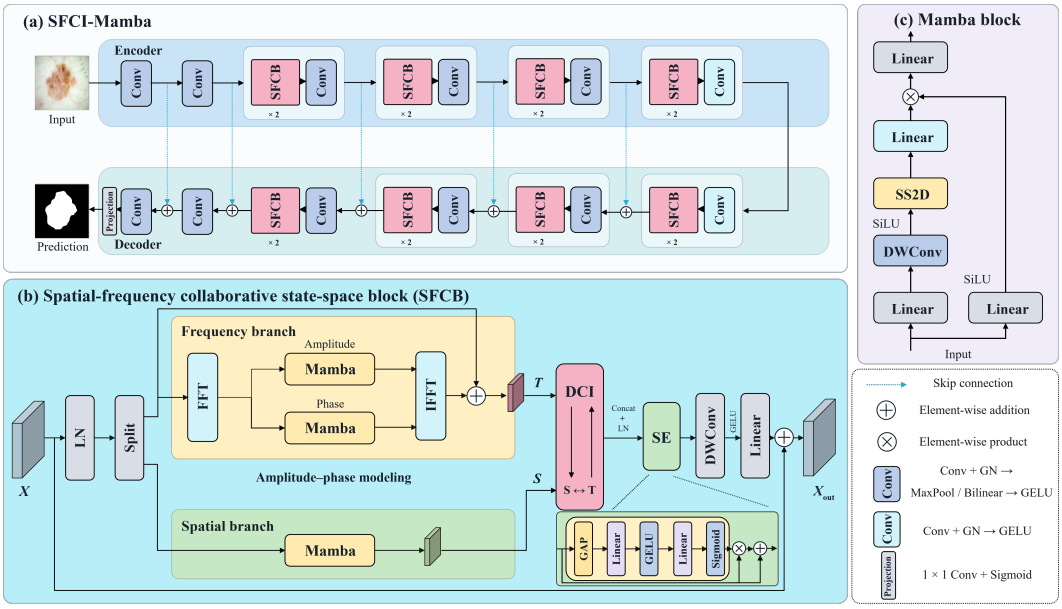

Figure 1: Overview of SFCI-Mamba and its core modeling unit. The encoder and decoder use SFCBs to extract multiscale features. Within each SFCB, the frequency branch models amplitude- and phase-domain representations with separate Mamba mappings, while DCI enables bidirectional interaction with the spatial branch. SE recalibrates the fused features.

3.2 Spatial-Frequency Collaborative State-Space Block

Skin lesions exhibit irregular contours and heterogeneous appearance. To represent these patterns in both spatial and spectral domains, SFCB processes features through two parallel branches. For an input $X \in \mathbb{R}^{h\times w\times C}$, we apply layer normalization and split the channels equally:

$$[Z_s, Z_f] = \mathrm{Split}(\mathrm{LN}(X)), \qquad S = \mathcal{M}_s(Z_s), \tag{1}$$

where $Z_s, Z_f \in \mathbb{R}^{h\times w\times C/2}$. The mapping $\mathcal{M}_s$ uses two-dimensional selective scanning (SS2D) (Liu et al., 2024) along rows, columns, and their reverse directions to capture long-range spatial dependencies.

The frequency branch separates the strength of spectral components from their phase relationships. For each channel of Z_f, we apply a two-dimensional real FFT (RFFT2) over the spatial dimensions and obtain the amplitude and phase of the resulting complex spectrum:

$$F = \mathrm{RFFT2}(Z_f), \qquad A = |F|, \qquad \Phi = \arg(F). \tag{2}$$

The amplitude and phase representations are fed into two independent Mamba mappings, respectively:

$$\widetilde{A} = \mathcal{M}_a(A), \qquad \widetilde{\Phi} = \mathcal{M}_p(\Phi). \tag{3}$$

The processed representations are combined into a complex spectrum by forming its real and imaginary parts:

$$\widetilde{F} = \widetilde{A} \odot \cos(\widetilde{\Phi}) + \mathrm{i}\, \widetilde{A} \odot \sin(\widetilde{\Phi}), \tag{4}$$

where i is the imaginary unit. We restore the original spatial size with $T_f = \mathrm{IRFFT2}(\widetilde{F}, s = (h, w))$ and use a residual connection, $T = T_f + Z_f$. The three mappings use independent parameters. Since $\mathcal{M}_a$ and $\mathcal{M}_p$ are unconstrained, $\widetilde{A}$ and $\widetilde{\Phi}$ are learned amplitude- and phase-domain representations rather than physically constrained spectral variables.

DCI refines the two branch outputs using their joint context, as detailed in Section 3.3. The updated features are concatenated and normalized:

$$(S', T') = \mathrm{DCI}(S, T), \qquad U = \mathrm{LN}(\mathrm{Cat}(S', T')). \tag{5}$$

We then apply squeeze-and-excitation (SE) channel gating (Hu et al., 2018), using GELU between its two linear mappings and a residual formulation $U_{\mathrm{SE}} = U + U \odot g(U)$, where $g(U)$ is obtained by global average pooling, the two mappings, and sigmoid activation. A 3×3 depthwise convolution followed by GELU and a linear projection completes the block:

$$X_{\mathrm{out}} = X + \mathrm{Linear}\big(\mathrm{GELU}(\mathrm{DWConv}_{3\times3}(U_{\mathrm{SE}}))\big). \tag{6}$$

3.3 Dual-Context Interaction

The spatial and frequency branches provide different views of the same feature map. We design DCI to allow each branch to retrieve information from both views before fusion. As shown in Figure 2, DCI forms branch-specific queries and a joint context for bidirectional updates.

Let $\bar{S} = \mathrm{LN}_s(S)$ and $\bar{T} = \mathrm{LN}_f(T)$. We construct the joint context $J = \mathrm{Cat}(\bar{S}, \bar{T})$. For each branch $b \in \{s, f\}$, with $B_s = \bar{S}$ and $B_f = \bar{T}$, queries, keys, and values are computed as

$$Q_b = \mathcal{P}(L_b^q(B_b)), \qquad K_b = \mathcal{P}(L_b^k(J)), \qquad V_b = \mathcal{P}(L_b^v(J)), \tag{7}$$

where L_b^q, L_b^k, and L_b^v are pointwise projections to C channels. The operator $\mathcal{P}$ applies adaptive average pooling when needed to cap the interaction grid at 32×32, then flattens the grid into tokens. The two paths use independently learned projections: each branch forms its query from its own representation, while keys and values are computed from the same joint spatial-frequency context.

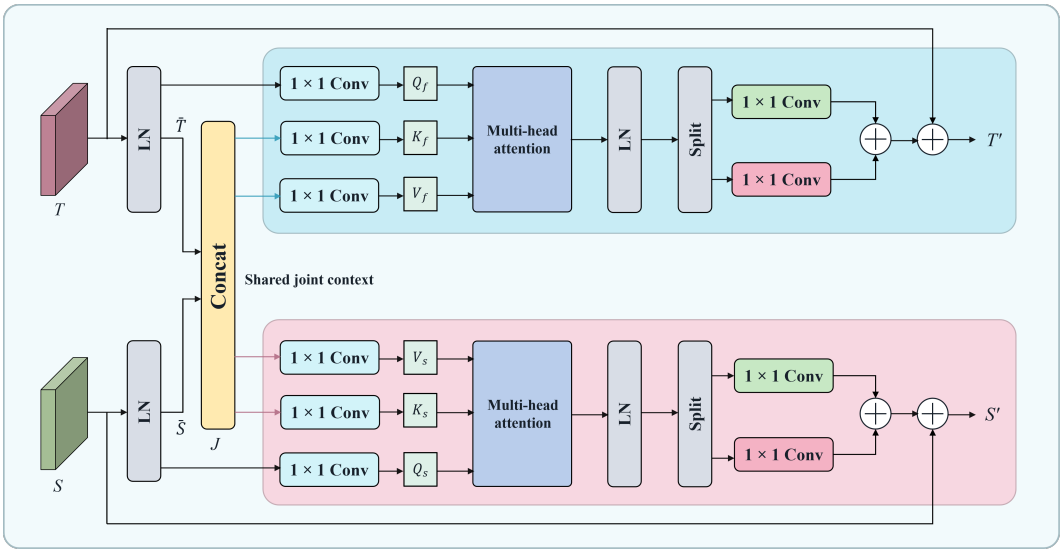

Figure 2: Dual-context interaction. Each branch forms its own query and attends to the joint spatial-frequency context. The attention response is projected into a residual update for the corresponding branch. The implementation uses two channel-group projections whose outputs are summed. The updated features are concatenated for subsequent fusion.

For attention head j in branch b, the response is

$$H_b^{(j)} = \mathrm{Softmax}\left(\frac{Q_b^{(j)}(K_b^{(j)})^{\mathsf{T}}}{\sqrt{d}}\right) V_b^{(j)}, \tag{8}$$

where j indexes the four heads and $d = C/4$. Concatenating the heads, restoring the spatial grid when needed, and applying layer normalization gives O_b. The branch updates can be written compactly as

$$S' = S + P_s(O_s), \qquad T' = T + P_f(O_f), \tag{9}$$

where P_s and P_f are pointwise output projections. In the implementation shown in Figure 2, each O_b is split into two channel groups and $P_b(O_b)$ is parameterized as the sum of two independent pointwise projections. Both groups are derived from the same joint-context response. The "self" and "cross" labels denote projection paths rather than source-pure information flows.

3.4 Loss Function

We use an equally weighted combination of binary cross-entropy (BCE) and Dice loss in our experiments. For a predicted probability map P and binary mask Y, the objective is

$$\begin{aligned} \mathcal{L} &= \mathcal{L}_{\mathrm{BCE}} + \mathcal{L}_{\mathrm{Dice}}, \\ \mathcal{L}_{\mathrm{BCE}} &= -\frac{1}{N}\sum_{n=1}^{N}\big[y_n \log p_n + (1-y_n)\log(1-p_n)\big], \\ \mathcal{L}_{\mathrm{Dice}} &= 1 - \frac{2\sum_{n=1}^{N} p_n y_n + \epsilon}{\sum_{n=1}^{N} p_n + \sum_{n=1}^{N} y_n + \epsilon}, \end{aligned} \tag{10}$$

where N is the number of pixels, p_n and y_n are the prediction and label at pixel n, and $\epsilon = 1$ is a smoothing constant. The loss is averaged over the training batch. BCE provides pixelwise supervision, while Dice loss optimizes region overlap.

4 Experiments

4.1 Datasets and Implementation Details

Datasets. We evaluate SFCI-Mamba on three public skin lesion segmentation datasets: ISIC2017 (Codella et al., 2018), ISIC2018 (Codella et al., 2019; Tschandl et al., 2018), and PH^2 (Mendonca et al., 2013). Following prior work (Wu et al., 2025b), we split 2,000 ISIC2017 images and 2,594 ISIC2018 images as shown in Table 1. SFCI-Mamba and all retrained comparison methods use the same splits. A separate model is trained on each ISIC dataset and evaluated on its corresponding test set. For the common comparison protocol, the 200 PH^2 images are evaluated using each method's corresponding ISIC2018-trained model, without fine-tuning.

Table 1: Dataset splits. PH^2 is evaluated without fine-tuning using the corresponding ISIC2018-trained model.

Dataset	Train	Validation	Test
ISIC2017	1250	150	600
ISIC2018	1815	259	520
PH^2	–	–	200

Evaluation metrics. For model comparison, we report seven metrics: intersection over union (IoU), Dice similarity coefficient (DSC), sensitivity (SEN), specificity (SPE), accuracy (ACC), 95th-percentile Hausdorff distance (HD95), and average symmetric surface distance (ASSD). Predicted probabilities and normalized ground-truth masks are binarized at 0.5, with the lesion as foreground. The five pixel-count metrics are computed as

$$\begin{aligned} \mathrm{IoU} &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}}, \qquad \mathrm{DSC} = \frac{2\mathrm{TP}}{2\mathrm{TP}+\mathrm{FP}+\mathrm{FN}}, \\ \mathrm{SEN} &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \qquad \mathrm{SPE} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}, \\ \mathrm{ACC} &= \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}. \end{aligned} \tag{11}$$

where TP, TN, FP, and FN are accumulated over the test set. The resulting five metrics are therefore micro-averaged over test-set pixels. Higher values indicate better performance.

HD95 and ASSD assess boundary quality (px; lower is better). For each 256×256 test image i, let P_i and G_i be the binarized prediction and ground-truth masks, with four-connected boundaries ∂P_i and ∂G_i. For nonempty masks, the directed boundary-distance multiset is

$$\mathcal{D}(P, G) = \left\{ \min_{g \in \partial G} \|p - g\|_2 : p \in \partial P \right\}. \tag{12}$$

For each image, we combine the distances in $\mathcal{D}(P_i, G_i)$ and $\mathcal{D}(G_i, P_i)$ into one list $\mathcal{D}_i$, retaining repeated values. The per-image metrics are

$$\begin{aligned} \mathrm{HD95}_i &= Q_{0.95}(\mathcal{D}_i), \\ \mathrm{ASSD}_i &= \mathrm{mean}(\mathcal{D}_i), \end{aligned} \tag{13}$$

where $Q_{0.95}$ is the empirical 95th percentile and mean is the arithmetic mean. For a test set of M images, the reported scores average the per-image values:

$$\mathrm{HD95} = \frac{1}{M} \sum_{i=1}^{M} \mathrm{HD95}_i, \qquad \mathrm{ASSD} = \frac{1}{M} \sum_{i=1}^{M} \mathrm{ASSD}_i. \tag{14}$$

We use MedPy 0.5.2 with unit pixel spacing. When both masks are empty, both per-image values are set to 0. When only one mask is empty, both are set to $\sqrt{255^2 + 255^2}$ px.

Implementation details. We conducted all experiments using Python 3.9 and PyTorch 2.0.1 on a single NVIDIA RTX 4090 GPU. Following prior work (Wu et al., 2025b), we resize images to 256×256 and normalize them. Random rotation, horizontal flipping, and vertical flipping are applied for data augmentation. All models are trained for 250 epochs with a batch size of 8 using AdamW (Loshchilov & Hutter, 2019). The initial learning rate is set to 10^{-3}, with a weight decay of 10^{-2}, and cosine annealing (Loshchilov & Hutter, 2017) is adopted as the learning-rate scheduler with a minimum learning rate of 10^{-5}. The training objective combines BCE and Dice losses with equal weights. The checkpoint with the lowest validation loss is used for testing.

Model complexity. Table 2 reports parameter counts and computational cost at 256×256 resolution. SFCI-Mamba integrates spatial and frequency modeling within a compact architecture, while requiring fewer parameters and FLOPs than most of the evaluated networks.

Table 2: Model complexity at an input resolution of 256×256.

Method	Params (M) ↓	GFLOPs ↓
U-Net (MICCAI, 2015)	31.038	54.738
MALUNet (BIBM, 2022)	0.175	0.083
EGE-UNet (MICCAI, 2023)	0.053	0.072
TransUNet (arXiv, 2021)	105.322	32.229
TransFuse (MICCAI, 2021)	26.273	11.528
VM-UNet (TOMM, 2025)	27.427	4.112
UltraLight VM-UNet (Patterns, 2025)	0.049	0.060
H-vmunet (Neurocomputing, 2025)	8.972	0.741
ours	2.538	0.354

4.2 Comparison with Existing Methods

We compare SFCI-Mamba with eight segmentation networks: U-Net (Ronneberger et al., 2015), MALUNet (Ruan et al., 2022), EGE-UNet (Ruan et al., 2023), TransUNet (Chen et al., 2021), TransFuse (Zhang et al., 2021), VM-UNet (Ruan et al., 2025), UltraLight VM-UNet (Wu et al., 2025b), and H-vmunet (Wu et al., 2025a). These methods cover convolutional, Transformer-based, and visual state-space architectures, including lightweight designs.

Quantitative comparison. Table 3 shows the highest IoU, DSC, sensitivity, and accuracy for SFCI-Mamba on all three datasets under the common protocol. It also achieves the lowest HD95 and

Table 3: Three-dataset comparison. HD95 and ASSD are reported in px. Best values for each metric are shown in bold.

Method	IoU ↑	DSC ↑	SEN ↑	SPE ↑	ACC ↑	HD95 ↓	ASSD ↓
				ISIC2017			
U-Net	0.8254	0.9043	0.8743	0.9858	0.9643	14.6339	5.9533
MALUNet	0.8233	0.9030	0.8728	0.9856	0.9639	13.0943	5.2555
EGE-UNet	0.8201	0.9011	0.8599	**0.9884**	0.9636	13.4943	5.4640
TransUNet	0.8220	0.9023	0.8666	0.9870	0.9638	13.9552	5.9434
TransFuse	0.8319	0.9082	0.8826	0.9854	0.9656	12.9227	5.8638
VM-UNet	0.8200	0.9011	0.8807	0.9823	0.9627	15.2376	6.0027
UltraLight VM-UNet	0.8174	0.8995	0.8732	0.9837	0.9624	13.4948	5.4710
H-vmunet	0.8320	0.9083	0.8792	0.9864	0.9658	12.3157	5.0641
ours	**0.8385**	**0.9121**	**0.8932**	0.9844	**0.9668**	**11.8265**	**4.9305**
				ISIC2018			
U-Net	0.7959	0.8863	0.8583	0.9786	0.9528	18.4583	7.5515
MALUNet	0.8058	0.8924	0.8877	0.9723	0.9542	**14.2970**	5.9847
EGE-UNet	0.8060	0.8925	0.8913	0.9711	0.9540	15.8431	6.4484
TransUNet	0.8015	0.8898	0.8717	0.9761	0.9538	16.1897	6.2748
TransFuse	0.7949	0.8857	0.8535	**0.9799**	0.9528	16.2027	6.9293
VM-UNet	0.7909	0.8832	0.8866	0.9670	0.9498	18.0767	7.3693
UltraLight VM-UNet	0.7885	0.8817	0.8857	0.9664	0.9491	17.6780	7.0493
H-vmunet	0.8074	0.8934	0.8942	0.9707	0.9543	14.9930	6.2868
ours	**0.8183**	**0.9001**	**0.9100**	0.9695	**0.9567**	14.3276	**5.8219**
				PH^2			
U-Net	0.8270	0.9053	0.9066	**0.9542**	0.9388	17.1301	6.2939
MALUNet	0.8394	0.9127	0.9352	0.9456	0.9423	14.1203	6.0046
EGE-UNet	0.8404	0.9132	0.9332	0.9473	0.9428	14.3751	6.0132
TransUNet	0.8326	0.9087	0.9323	0.9430	0.9395	18.0226	6.3806
TransFuse	0.8387	0.9123	0.9250	0.9510	0.9426	14.6301	5.9979
VM-UNet	0.8236	0.9032	0.9562	0.9233	0.9339	18.4093	7.4832
UltraLight VM-UNet	0.8438	0.9153	0.9596	0.9346	0.9427	15.0154	6.0295
H-vmunet	0.8520	0.9201	0.9564	0.9416	0.9464	13.2868	5.6105
ours	**0.8567**	**0.9228**	**0.9638**	0.9405	**0.9480**	**13.2284**	**5.5968**

ASSD on ISIC2017 and PH^2, as well as the lowest ASSD on ISIC2018. MALUNet obtains a slightly lower HD95 on ISIC2018. Higher sensitivity but lower specificity than some baselines indicates a coverage trade-off. For every method, PH^2 uses the corresponding ISIC2018-trained checkpoint without fine-tuning.

Qualitative comparison. Figure 3 compares segmentation results under indistinct boundaries, varied lesion appearance, and hair occlusion. With indistinct boundaries, SFCI-Mamba reduces background inclusion and follows the lesion extent more closely, as shown in the first ISIC2017 example. For variations in shape and appearance, it preserves the narrow lesion in the first ISIC2018 example and produces a more complete region with fewer holes in the second PH^2 example. Under dense hair occlusion, several methods miss large lesion areas, while SFCI-Mamba recovers more of the obscured lesion in the second ISIC2017 example.

4.3 Ablation Studies

We examine the effects of DCI and SE, frequency-domain modeling, and separate modeling of amplitude and phase.

Effectiveness of DCI and SE. Table 4 evaluates DCI and SE while retaining both branches. Either component improves lesion overlap on both datasets, and their combination gives a further improvement. DCI enables information exchange before fusion, while SE recalibrates the fused channels.



Figure 3: Qualitative comparison on ISIC2017, ISIC2018, and PH^2, with two examples from each dataset. White denotes the lesion and black denotes the background in the ground truth and predicted masks.

Table 4: Ablation of DCI and SE channel gating. Both branches are retained. A checkmark indicates an enabled component. Bold marks the highest value within each dataset.

DCI	SE	IoU ↑	DSC ↑	SEN ↑	SPE ↑	ACC ↑
			ISIC2017			
		0.8204	0.9013	0.8861	0.9809	0.9626
✓		0.8296	0.9068	0.8915	0.9821	0.9647
	✓	0.8321	0.9084	0.8914	0.9830	0.9653
✓	✓	**0.8385**	**0.9121**	**0.8932**	**0.9844**	**0.9668**
			ISIC2018			
		0.7985	0.8880	0.8946	0.9672	0.9517
✓		0.8161	0.8987	**0.9313**	0.9615	0.9550
	✓	0.8083	0.8940	0.9063	0.9669	0.9540
✓	✓	**0.8183**	**0.9001**	0.9100	**0.9695**	**0.9567**

On ISIC2018, adding SE to DCI improves specificity with lower sensitivity, yielding better overlap and pixel accuracy. This pattern supports combining branch interaction with channel selection.

Comparison with alternative interaction methods. Table 5 compares DCI with no explicit interaction and bidirectional Cross-Attention on ISIC2018. All variants retain both branches and SE under the same data split and training settings. Cross-Attention uses each branch as query and the other as keys/values, with four heads on a grid capped at 32×32, as in DCI. DCI gives higher IoU and DSC, while also using more parameters and FLOPs. Since the variants are not parameter-matched, this comparison supports the complete interaction design rather than an efficiency advantage at equal capacity.

Frequency modeling versus spatial replacements. We replace the frequency branch with either one spatial Mamba path or two independent spatial Mamba paths whose outputs are averaged. Table 6 shows that both replacements reduce lesion overlap and sensitivity, indicating less complete recovery of lesion pixels. Using two spatial paths does not improve overlap compared with the sin-

Table 5: Comparison of cross-branch interaction methods on ISIC2018.

Interaction method	IoU ↑	DSC ↑	Params (M) ↓	GFLOPs ↓
No DCI + SE	0.8083	0.8940	2.002	0.247
Bidirectional Cross-Attention + SE	0.8080	0.8938	2.182	0.283
DCI + SE	0.8183	0.9001	2.538	0.354

Table 6: Frequency-branch replacements on ISIC2018. Bold indicates the highest value in each column.

Configuration	IoU ↑	DSC ↑	SEN ↑	SPE ↑	ACC ↑
Single spatial replacement	0.8119	0.8962	0.8780	**0.9778**	0.9564
Dual spatial replacement	0.8098	0.8949	0.8879	0.9737	0.9553
Spatial + frequency (full)	**0.8183**	**0.9001**	**0.9100**	0.9695	**0.9567**

Table 7: Amplitude and phase modeling on ISIC2018. The spatial branch, DCI, and SE are retained. Bold indicates the highest value in each column.

Configuration	IoU ↑	DSC ↑	SEN ↑	SPE ↑	ACC ↑
Identity	0.8087	0.8942	0.9078	0.9666	0.9540
Amplitude only	0.8149	0.8980	0.9017	**0.9710**	0.9561
Phase only	0.8159	0.8986	**0.9138**	0.9673	0.9558
Amplitude + phase	**0.8183**	**0.9001**	0.9100	0.9695	**0.9567**

gle spatial replacement. These results support the value of frequency features beyond adding spatial processing paths.

Effectiveness of amplitude- and phase-domain modeling. Table 7 shows that either learned component improves IoU and DSC over identity. Amplitude-only gives the highest specificity among these variants, whereas phase-only gives the highest sensitivity. Joint modeling gives the highest IoU, DSC, and ACC among the tested spectral variants, with modest margins over the single-component variants.

4.4 Discussion and Limitations

Representative failure cases under weak visual cues and extreme lesion scales are shown in Figure A.2 in the appendix. They include a displaced prediction for a tiny low-contrast lesion, extension into normal skin under atypical framing, and missed peripheral regions for a large border-touching lesion. Our evidence is limited to dermoscopic segmentation: PH^2 tests cross-dataset, not cross-modality, transfer. The spectral branch also models wrapped phase directly and uses an unconstrained amplitude-domain latent. Periodic phase representations and constrained magnitude mappings are left for future work.

5 Conclusion

We presented SFCI-Mamba for dermoscopic lesion segmentation. SFCB combines spatial selective scanning with a frequency branch factorized at the amplitude/phase input and mapping level, while DCI uses branch-specific queries to retrieve from shared spatial-frequency context before fusion. Under a common protocol, SFCI-Mamba improves IoU and DSC over the evaluated baselines on ISIC2017, ISIC2018, and cross-dataset PH^2. The ablations favor retaining the frequency branch, joint amplitude–phase modeling, and DCI within the evaluated configurations. The conclusions are limited to the dermoscopic setting studied here.

AI USE STATEMENT

We used ChatGPT for language polishing, manuscript organization, reference-related editing, and feedback on the evaluation methodology and presentation of experimental results. It was not used to originate the research idea, implement the proposed method, conduct experiments, or generate experimental results. The authors reviewed all AI-assisted material against the original sources and experimental records and take responsibility for the final content.

REPRODUCIBILITY STATEMENT

To support reproducibility, we provide the dataset splits, preprocessing and data augmentation procedures, training objective, optimizer, learning-rate schedule, batch size, number of training epochs, checkpoint-selection criterion, software environment, and hardware configuration in Section 4.1. All comparison methods are retrained using the same dataset splits, and PH^2 is evaluated using the corresponding ISIC2018-trained checkpoints without fine-tuning. The evaluation metrics and their computation, including the handling of empty masks for boundary-based metrics, are explicitly defined. Architectural details of SFCI-Mamba, including the SFCB and DCI modules, are described in Section 3, with additional ablation studies and qualitative analyses provided in the appendix.

REFERENCES

Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L. Yuille, and Yuyin Zhou. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. *arXiv preprint arXiv:2102.04306*, 2021. URL https://arxiv.org/abs/2102.04306.

Noel Codella, Veronica Rotemberg, Philipp Tschandl, M. Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, Harald Kittler, and Allan Halpern. Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC). *arXiv preprint arXiv:1902.03368*, 2019. URL https://arxiv.org/abs/1902.03368.

Noel C. F. Codella, David Gutman, M. Emre Celebi, Brian Helba, Michael A. Marchetti, Stephen W. Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, and Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge at the 2017 International symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC). In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pp. 168–172, 2018. doi: 10.1109/isbi.2018.8363547.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Azade Farshad, Yousef Yeganeh, Peter Gehlbach, and Nassir Navab. Y-Net: A spatiospectral dual-encoder network for medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pp. 582–592, 2022. doi: 10.1007/978-3-031-16434-7_56.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *Conference on Language Modeling*, 2024.

Yuanqin He, Guilian Chen, Yuhua Zhang, Huisi Wu, and Jing Qin. HFSTI-Net: Hierarchical frequency-spatial-temporal interactions for video polyp segmentation. In *International Conference on Learning Representations*, 2026.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141, 2018. doi: 10.1109/CVPR.2018.00745.

Shangwang Liu, Mengjiao Zhao, and Yinghai Lin. Dual-domain guided vision Mamba network for medical image segmentation. *Expert Systems with Applications*, 291:128512, 2025. doi: 10.1016/j.eswa.2025.128512.

Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. VMamba: Visual state space model. In *Advances in Neural Information Processing Systems*, volume 37, pp. 103031–103063, 2024. doi: 10.52202/079017-3273.

Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Teresa Mendonca, Pedro M. Ferreira, Jorge S. Marques, Andre R. S. Marcal, and Jorge Rozeira. PH^2 - A dermoscopic image database for research and benchmarking. In *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 5437–5440, 2013. doi: 10.1109/embc.2013.6610779.

Alan V. Oppenheim and Jae S. Lim. The importance of phase in signals. *Proceedings of the IEEE*, 69(5):529–541, 1981. doi: 10.1109/PROC.1981.12022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pp. 234–241, 2015. doi: 10.1007/978-3-319-24574-4_28.

Jiacheng Ruan, Suncheng Xiang, Mingye Xie, Ting Liu, and Yuzhuo Fu. MALUNet: A Multi-Attention and Light-weight UNet for Skin Lesion Segmentation. In *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 1150–1156, 2022. doi: 10.1109/BIBM55620.2022.9995040.

Jiacheng Ruan, Mingye Xie, Jingsheng Gao, Ting Liu, and Yuzhuo Fu. EGE-UNet: An Efficient Group Enhanced UNet for Skin Lesion Segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pp. 481–490, 2023. doi: 10.1007/978-3-031-43901-8_46.

Jiacheng Ruan, Jincheng Li, and Suncheng Xiang. VM-UNet: Vision Mamba UNet for medical image segmentation. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2025. doi: 10.1145/3767748.

Hui Sun, Long Lv, Pingping Zhang, Tongdan Tang, Feng Tian, Weibing Sun, and Huchuan Lu. Spatial-frequency enhanced Mamba for multi-modal image fusion. *IEEE Transactions on Image Processing*, 34:7684–7696, 2025. doi: 10.1109/TIP.2025.3632221.

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5: 180161, 2018. doi: 10.1038/sdata.2018.161.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5998–6008, 2017.

Renkai Wu, Yinghao Liu, Pengchen Liang, and Qing Chang. H-vmunet: High-order Vision Mamba UNet for medical image segmentation. *Neurocomputing*, 624:129447, 2025a. doi: 10.1016/j.neucom.2025.129447.

Renkai Wu, Yinghao Liu, Guochen Ning, Pengchen Liang, and Qing Chang. UltraLight VM-UNet: Parallel Vision Mamba significantly reduces parameters for skin lesion segmentation. *Patterns*, 6 (11):101298, 2025b. doi: 10.1016/j.patter.2025.101298.

Qingling Xia, Hong Zheng, Haonan Zou, Dinghao Luo, Honggan Tang, Lingxiao Li, and Bin Jiang. A comprehensive review of deep learning for medical image segmentation. *Neurocomputing*, 613: 128740, 2025. doi: 10.1016/j.neucom.2024.128740.

Yundong Zhang, Huiye Liu, and Qiang Hu. TransFuse: Fusing Transformers and CNNs for Medical Image Segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pp. 14–24, 2021. doi: 10.1007/978-3-030-87193-2_2.

A Appendix

A.1 Feature Visualization and Analysis

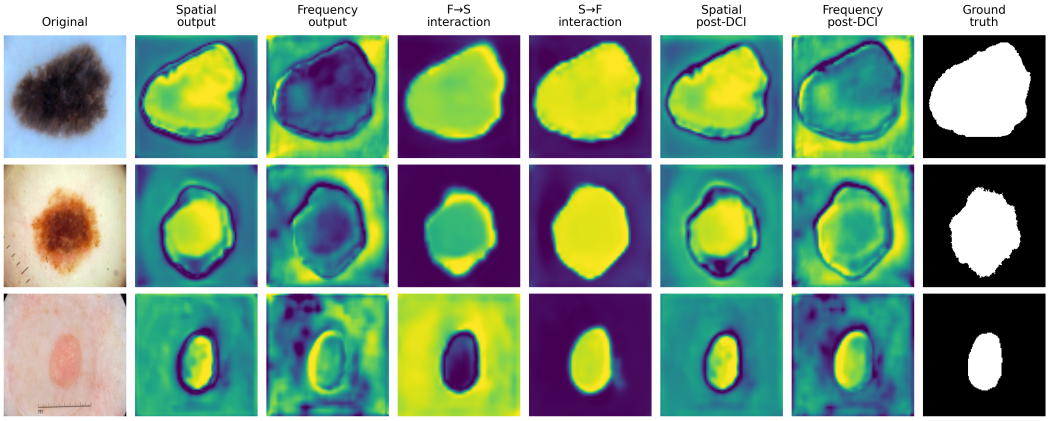

Figure A.1: Visualization of spatial and frequency branch outputs and their interaction through DCI. F→S and S→F are shorthand labels for the designated residual projection paths of the two branch updates. Both are derived from joint-context responses and do not represent source-pure information flow. Post-DCI outputs include the full residual updates. Response maps show channel-wise root-mean-square (RMS) magnitudes with independent normalization.

Figure A.1 qualitatively shows different spatial patterns across the two branches. Spatial outputs emphasize lesion interiors, while frequency outputs show contour-like and surrounding responses. The two designated DCI projection paths also exhibit different spatial patterns across examples. After interaction, spatial outputs retain localization and frequency outputs combine interior and contour structure. Because maps are normalized independently, the figure compares spatial patterns rather than absolute response magnitudes.

A.2 Failure Cases

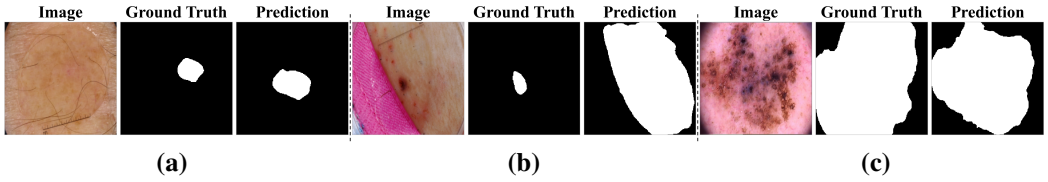

Figure A.2: Failure cases of SFCI-Mamba. Each example shows the input image, ground truth, and prediction from left to right.