技術論文

DiffMamba: Leveraging Mamba for Effective Fusion of Noise and Conditional Features in Diffusion Models for Skin Lesion Segmentation

Amit Shakya　Shruti Phutke　Chetan Gupta　Rupesh Kumar　Lalit Sharma
Chetan Arora

This paper, titled "DiffMamba: Leveraging Mamba for Effective Fusion of Noise and Conditional Features in Diffusion Models for Skin Lesion Segmentation," was presented at CVIP-2024 (9th International Conference on Computer Vision & Image Processing), held at IIITDM Kancheepuram, Chennai, India, from December 19–21, 2024. It received the Best Industry Paper Award.

Abstract

Effective Skin Lesion Segmentation is crucial for dermato- logical care, it enables the early identification and accurate diagnosis of skin cancer. Denoising Diffusion Probabilistic Models (DDPMs) have recently become a major focus in computer vision. Its applications in image generation, such as Stable Diffusion, Latent Diffusion Models and Imagen, have showcased remarkable abilities in creating high-quality generative outputs. Recent research highlights that DDPMs also perform exceptionally well in medical image analysis, specifically in medical image segmentation tasks. Even though a U-Net backbone served as the foundation for these models initially, there is a promising opportunity to boost their performance by incorporating other mechanisms. Recent research includes a transformer-based framework for diffusion models, but the advancement comes with the challenge of inherent quadratic complexity. Research has shown that state space models (SSMs), like Mamba efficiently capture long-range dependencies while maintaining linear computational complexity. Due to these benefits, it outperforms many of the mainstream foundational architectures. However, we found that simply merging Mamba with diffusion results in suboptimal performance. To truly harness the power of these two advanced technologies for medical image segmentation, a more effective integration is required, we formulate a novel Mamba-Based Diffusion framework, called Diff- Mamba for skin lesion segmentation. We access its performance on the ISIC 2018 dataset for skin lesion segmentation, and our method out- performs existing state-of-the-art techniques. The code is available at: https://github.com/amit-shakya-28/DiffMamba

1 INTRODUCTION

Background Melanoma, the most fatal type of skin cancer, results from the abnormal growth of melanocyte cells due to the activation of mutation by unusual Deoxyribonucleic Acid (DNA) damages. Melanocyte cells create melanin, which is the substance responsible for skin colour[1]. Melanoma cases have increased sharply over the last 30 years, with around 10,000 deaths annually in the USA[35]. It is highly curable, if detected in early stage[19]. In dermatology, early melanoma diagnosis is achieved by visual examination through methods like the ABCD criteria[27] (asymmetry, border irregularity, colour patterns, and diameter) and the seven-point checklist[3]. However, these methods are time consuming and face subjectivity issue. Automated medical image segmentation methods gained considerable interest recently for their potential to diagnose the diseases accurately. The effectiveness of these models in medical image segmentation stems from progress in deep learning technologies, ranging from widely used convolutional neural networks (CNN)[4][31][17] to the newer

vision transformer architectures (ViT)[5][39].

Denoising Diffusion Probabilistic Models Recently, the DDPM has gained significant recognition as a potent class of generative models. The rising acknowledgment of DDPMs has sparked significant interest and research, fueled by their remarkable ability to produce high-quality and diverse samples, as demonstrated by models like DALLE2[29], Imagen[34], and Stable Diffusion[30]. Building on these advancements, researchers have introduced innovative techniques for medical image segmentation utilizing diffusion models. By leveraging DDPMs, many approaches have achieved cutting-edge results across various benchmarks. The outstanding results of these models arise from their built-in stochastic sampling process[28]. DDPMs can produce varied segmentation predictions through repeated runs, with the diversity of these outputs effectively reflecting the inherent uncertainty in medical images. This is particularly valuable for segments where organs or lesions often have uncleared or indistinct boundaries.

Mamba Based Diffusion Framework However, it is noteworthy that all these approaches are built upon traditional U-Net backbone. Compared to the growing trend of using state space models (Mamba), traditional U-Net models compromised the segmentation quality, due to which it generates a diverse but incorrect mask during ensemble. Ultimately, this can introduce persistent noise that permanently degrades performance. Building on this momentum, we want to combine the Mamba-based U-Net model such as U-Mamba[26] with the diffusion model. However, we observed that a straightforward implementation yielded inadequate performance. One of the main reasons behind it is that the features produced by the Mamba are not compatible with those from the diffusion backbone. The Mamba extracts detailed semantic information directly from the original image, while the diffusion backbone handles features from a noisy and corrupted mask, which complicates the process of combining these features. To mitigate these shortcomings, we formulate a novel Mamba-driven Diffusion Framework for skin lesion segmentation, called DiffMamba.

The main concept is to combine conditional embedding and diffusion embedding. To effectively link these two embeddings, we proposed a novel Mamba-based Fusion module for their integration. The feature fusion module merges noise and semantic features using a cross-mamba block (CroMB) and the merged features further in succeeding Mamba block (see Section 3.3). To align the noise and conditions at each step, CroMB utilizes cross-input features to the mamba block to enrich features, while integrated output of CroMB processed in further Mamba block are refined features through selective scanning and concatenation to produce the final fused output. The proposed work provides the following key aspects.

- We formulate a novel method that combine Mamba and diffusion model for skin lesion segmentation.
- We introduce an innovative Mamba-based fusion module that seamlessly merges conditional semantic features with diffusion noise. As far as the authors are aware, this is the first successful method to combine diffusion and condition embeddings in skin lesion segmentation.
- The proposed method is evaluated using the ISIC 2018 dataset for comparison.

The enhancement in Intersection over Union (IoU) and Dice score relative to current leading medical image segmentation methods demonstrates the efficacy of the proposed approach.

2 RELATED WORKS

2-1. Skin Lesion Image Segmentation

Traditional Approaches Before the advent of deep learning, image segmentation was predominantly driven by classical methods and machine learning techniques. These included approaches like adaptive thresholding[7], support vector machines[48], region growing[18], unsupervised clustering[48], and active contours[13]. These methods were heavily dependent on manually designed features, which were difficult to create and often lacked the adaptability and effectiveness required for handling

more complex datasets. As a result, they struggled with larger and more intricate data.

Deep Learning Approaches This section offers a summary of deep learning methods used for skin lesion segmentation, with a focus on the U-Net model introduced by Ronneberger et al.[32], which efficiently leverages data augmentation to make the most of limited labeled samples in biomedical imaging. Bi et al.[6] created a multi-stage fully convolutional network for skin lesion segmentation, which includes stages for both coarse and fine boundary learning, along with a parallel integration method to enhance detection. Yuan et al.[45] proposed a fully automated skin lesion segmentation approach that employs a 19-layer DCNN and uses Jaccard Distance as the loss function. Despite extensive parameter optimization and testing on the ISBI 2016 and PH2 datasets, their approach did not perform as good as state-of-the-art methods, notably when handling low-contrast images. The 2020s saw a shift in computer vision with the rise of vision transformers, disrupting the dominance of CNNs and leading to innovations like Swin-Unet[15], which integrates Swin Transformer[24] blocks into U-Net models. Hybrid architectures like UCTransNet[38] and MCTrans[46] combine CNNs and transformers, while all-transformer models such as SMESwin-Unet[47] utilize transformers throughout the entire U-Net structure. Despite these advancements, challenges remain due to the limited availability of annotated segmentation data compared to classification data, which affects the precision and reliability of segmentation algorithms.

2-2. Diffusion Models for Medical Image Segmentation

Building on recent advancements, researchers have introduced innovative medical image segmentation techniques that utilize diffusion models to address this complex problem. For example, EnsDiff[41] uses ground truths for training and treats input images as priors to create segmentation distributions, which aids in generating uncertainty maps and an implicit ensemble of segmentation maps. Kim et al.[20] introduce an innovative method for self-supervised vessel segmentation. On the other hand, MedSegDiff[42] utilizes diffusion probabilistic models (DPM) for medical image segmentation, incorporating dynamic conditional encoding alongside the FF-Parser helps to mitigate the impact of high-frequency noise. The subsequent MedSegDiff-V2[43] improves upon this method by incorporating a conditional U-Net, which strengthens the interaction between semantic features and noise.

2-3. Mamba in Computer Vision

Previous methods in semantic segmentation either use CNNs[21], which are scalable but constrained by small receptive fields and weight-sharing limitations, or Vision Transformers (ViTs)[36], which provide better global context but suffer from quadratic complexity and efficiency issues. To overcome these challenges, Selective Structured State Space Models (Mamba)[9] have become increasingly popular for their ability to cover global receptive fields and use dynamic weights while maintaining linear complexity. Mamba has proven highly effective in long sequence modeling tasks, particularly in natural language processing[9]. Its capabilities are also being investigated in vision tasks like image classification[23], medical image segmentation[26][33], and 3D scene comprehension[22] introduced a residual state space block that combines channel attention with Mamba and a 2-D selective scanning technique for better image restoration. Inspired from its success, we introduced a fusion module in our proposal.

3 METHODOLOGY

This part provides a thorough explanation of our proposed Mamba-based Diffusion framework for semantic segmentation. We commence with an explanation of the foundational ideas of DDPMs and State Space Models. Then, we provide a summary of the architecture we propose, followed by a detailed examination of the fusion module.

3-1. Denoising Diffusion Probabilistic Models

We offer a concise summary of the diffusion models as described in[14]. These generative models are defined by a Markov chain and consist of a forward process, in

which the data undergoes gradual degradation by adding Gaussian noise, and a backward process, in which the degradation is progressively reversed by removing the noise. The forward process, denoted as $\boldsymbol{q}$, is given by the following formulation:

$$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}) \tag{1}$$

Here, T are number of steps x_1, x_2, ..., x_T are the latent variables at each step, and x_0 is the initial data sample. For every step in the forward pass, the Gaussian noise is given as:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t \mathbf{I}_{n\times n}) \tag{2}$$

where β_t specifies the noise schedule with a constant and $\mathbf{I}_{n\times n}$ is the identity matrix having size $n \times n$, The forward pass allows for sampling at any arbitrary timestamp t, is described in [14], which can be reparametrised to

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\,\epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}_{n\times n}) \tag{3}$$

where,

$$\begin{aligned} \alpha_t &= 1-\beta_t \\ \bar{\alpha}_t &= \prod_{s=0}^{t} \alpha_s \end{aligned} \tag{4}$$

The reverse pass, parameterized with θ, is defined as:

$$p_\theta(x_{0:T-1}|x_T) = \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t). \tag{5}$$

Starting with $p_\theta(x_T) = \mathcal{N}(x_T; 0, \mathbf{I}_{n\times n})$, this process transforms the distribution from $p_\theta(x_T)$ to $p_\theta(x_0)$. The reverse pass is carried out by applying Gaussian steps described with:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t,t), \Sigma_\theta(x_t,t)). \tag{6}$$

As shown in [14], we can then predict x_{t-1} from x_t with

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\alpha_t}}\epsilon_\theta(x_t,t)\right) + \sigma_t z, \tag{7}$$

where $z \sim \mathcal{N}(0, \mathbf{I})$, and σ_t represents the variance scheme to be learned, as proposed in [14]. As seen in Equation 7, the sampling process includes a random component z, resulting in stochastic behavior. Note that ϵ_θ refers to the U-Net model containing a Mamba-based fusion module we train, with input $x_t = \sqrt{\alpha_t}x_0 + \sqrt{1-\alpha_t}\,\epsilon$. The model $\epsilon_\theta(x_t, t)$ that is subtracted from x_t during sampling, as described in Equation 7, needs to be learned by the model. This U-Net with Mamba is trained using the loss objectives specified in [14].

3-2. State Space Models

State Space Models (SSMs) [10][11] are a framework used for sequence-to-sequence tasks, characterized by their time-invariant dynamics, also referred to as linear time-invariant (LTI) properties. Because of their linear complexity, SSMs are ideal for capturing the dynamics of systems by mapping them to latent states, described as follows:

$$y(t) = Ch(t) + Dx(t), \dot{h}(t) = Ah(t) + Bx(t) \tag{8}$$

Here, $x(t) \in \mathbb{R}$ represents the input, $h(t) \in \mathbb{R}^N$ is hidden state, and $y(t) \in \mathbb{R}$ is the output. N denotes the state size, and $\dot{h}(t)$ refers to the time derivative of $h(t)$. Additionally, $A \in \mathbb{R}^{N\times N}$, $B \in \mathbb{R}^{N\times 1}$, $C \in \mathbb{R}^{1\times N}$, and $D \in \mathbb{R}$ are the system matrices. Since the matrix D is considered a residual connection between the input and output, State Space Models (SSMs) are often represented by omitting D, leading to the following simplified forms:

$$\begin{aligned} h'(t) &= Ah(t) + Bx(t); \quad \textit{state equation} \\ y(t) &= Ch(t); \qquad\quad \textit{output equation} \end{aligned} \tag{9}$$

This representation captures the global feature dependency of SSMs, as the current output depends on all preceding states and the input. While the above equations assume continuous-time inputs, deep learning approaches treat the input as discrete in time. To handle discrete sequences such as text and images, SSMs use zero-order hold (ZOH) discretization[10]. This approach maps the continuous $\{x_1, x_2, ..., x_K\}$ (input sequence) to the discrete $\{y_1, y_2, ..., y_K\}$ (output sequence). To achieve this, a time scale parameter $\Delta \in \mathbb{R}^D$ is introduced, transforming the matrices A to $\bar{A}$ and B to $\bar{B}$ as follows:

$$\begin{aligned} \bar{A} &= \exp(\Delta A), \\ \bar{B} &= (\Delta A)^{-1}\left(\exp(A) - I\right) \cdot \Delta B, \\ \bar{C} &= C, \\ y_k &= \bar{C}h_k + \bar{D}x_k, \\ h_k &= \bar{A}h_{k-1} + \bar{B}x_k. \end{aligned} \tag{10}$$

In this context, the dimensions of all matrices remain consistent throughout each iteration of the process. In addition, as detailed in Mamba[9], a first-order Taylor expansion is used to approximate the matrix B:

$$\bar{B} = (\exp(A) - I)A^{-1}B \approx \Delta B$$

where we simplify $(\Delta A)(\Delta A)^{-1}\Delta B$ to ΔB. This approximation facilitates the analysis and implementation of the model by reducing the computational complexity associated with matrix exponentiation. By simplifying $(\exp(A) - I)A^{-1}B$ to ΔB, we make the model more tractable and easier to work with in practical applications.

3-3. Proposed Architecture

Our approach utilizes the diffusion model[42], comprising a forward diffusion process adding the Gaussian noise and the reverse process with neural network restoring the original data. This process is explained mathematically in Section 3.1. To ensure consistency with forward process, the noisy image is iteratively refined with the reverse process, progressively restoring it to achieve the final clear segmentation (see Figure 1). Similar to the standard diffusion probabilistic model (DPM) setup, we use an encoder-decoder network for training. For segmentation purpose, we enhance the model by conditioning the noise prediction function ϵ_θ on information from the original image.

$$\epsilon_\theta(x_t, I, t) = DC\left(\text{MambaF}\left(F_{I_t}, F_{x_t}\right), t\right) \tag{11}$$

Here, F_{I_t} and F_{x_t} (see Figure 1), signifies the feature embeddings of the original image (condition) and diffusion input at the current stage t respectively. These two embeddings are merge together using Mamba based fusion module (MambaF) and then fed into a UNet decoder (DC) for the reconstruction process. The complete process of DiffMamba is shown in Figure 1. To explain this, we consider the single step t in diffusion process. Initially, the noisy mask x_T is passed through a UNet, which is called the Diffusion Model. The model is directed by semantic features of the raw images using a separate UNet, referred to as the condition UNet. Following this, the semantic condition is integrated into the encoder features of the Diffusion UNet, combining the semantic segmentation embeddings from the condition UNet. This integration is controlled by the Fusion Module, which refines the representation by connecting noise and semantic embeddings, utilizing the global and adaptive features of Mamba. The proposed approach used a standard noise prediction loss L_n for training, analogous to diffusion probabilistic models (DPMs)[14].

Features Fusion with Mamba [37] proposed a feature fusion module that integrate multi-modal features using two blocks, Concat-Mamba Block (ConMB), and Cross-Mamba Block (CroMB). The CroMB enhances features through cross-multiplication and selective scanning, while ConMB combines outputs from CroMB using concatenation and selective scanning. Inspired from this approach[37], we propose the Mamba Fusion block to merge the semantic (conditional) and noise features. The combined features are then forwarded to UNet decoder in the diffusion model for further processing. Let us suppose the conditional features and noisy features are represented by $F^o_{I_t}$ and $F^o_{x_t}$ respectively, the complete fusion process is formulated as:

$$\begin{aligned} F^o_{I_t}, F^o_{x_t} &= CroMB\left(F_{I_t}, F_{x_t}\right) \\ MambaF_{Output} &= Mamba(Conv < F^o_{I_t}, F^o_{x_t} >). \end{aligned} \tag{12}$$

where, $<\cdot>$ is concatenation, $Conv$ is 1×1 Convolution, $Mamba$ is a mamba block (see Mamba in Figure 1 for more details). Here, all the features remain in original dimension. The Mamba block basically works as a gating mechanism with the State space model. In Cross mamba block we utilize the cross features for generating the gating mechanism (see the inputs F_{I_t} and F_{x_t} provided differently to the two parallel mamba blocks in Figure 1). The Mamba block uses the 2D-selective scanning mechanism (SS2D)[12] as shown in Figure 1. Furthermore, refer to equation 11 to see how $MambaF_{output}$ features are used to condition the noise prediction function.

4 EXPERIMENTS AND RESULT DISCUSSION

This section provides details about the dataset used in the

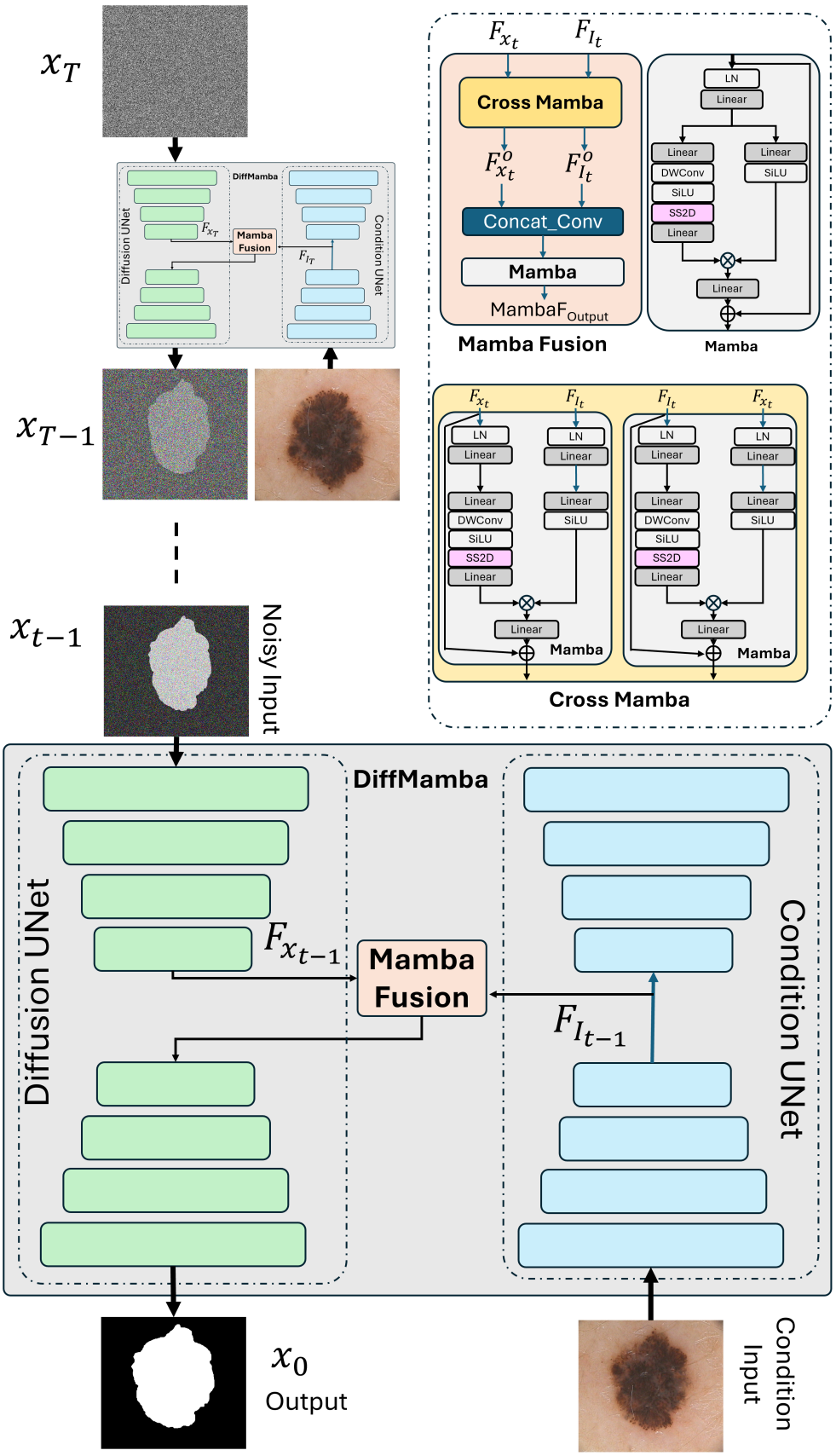

Fig. 1 Details of the Proposed Architecture for Skin Lesion Segmentation: We present a Mamba-based fusion method that integrates noise from the diffusion encoder with conditional features to enhance the effectiveness of image segmentation

experiments, outlines the implementation process, explains the evaluation metrics, and discusses the results.

4-1. Dataset

The ISIC 2018 dataset[8], from International Skin Imaging Collaboration (ISIC), offers a diverse and extensive collection of dermoscopy images. In which, Task 1 is dedicated to lesion segmentation having a total of 3,694 images, of these, 2,594 images are designated for training, consisting of 72% nevi, 20% melanomas, and 8% seborrheic keratoses. Additionally, 100 images are allocated for validation, and 1,000 for testing. The image resolutions range from 0.5 to 29 megapixels, with dimensions spanning from 540 × 576 to 4,499 × 6,748 pixels. Figure 2 illustrates sample images from the dataset.

4-2. Training and Implementation Details

The proposed segmentation network was trained on images which are resized to a resolution of 256 × 256. The model was trained for 100k iterations, employing the AdamW[25] optimizer with a *batch size* = 8 and a *learning rate* = 0.0001. For inference, 100 diffusion steps were used, and the ensemble of 5 times model execution is considered, which is less than 25 runs performed in MedSegDiff[44]. The STAPLE algorithm[40] was applied to merge the samples. The experiments are implemented on PyTorch framework and executed on an NVIDIA A100 GPU.

4-3. Evaluation Metrics

The proposed network's performance was measured using the IoU and Dice score to compare it with leading medical image segmentation methods.

Dice Score The Dice coefficient is measured using the recall and precision from prediction, evaluating the similarity of the prediction and the ground truth. It also accounts for false positives, which is generally useful in datasets with significant class imbalance, such as those found in medical image segmentation. Mathematically it is defined as:

$$\text{Dice} = \frac{2 \times \text{True Positive}}{2 \times \text{True Positive} + \text{False Positive} + \text{False Negative}} \tag{13}$$

Intersection Over Union (IoU) It evaluates the overlap of the prediction and the ground truth by evaluating the ratio of their common area to the total area covered by both. In mathematical terms, it is defined as:

$$\text{IoU} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive} + \text{False Negative}} \tag{14}$$

The difference between the two metrics is that the IoU penalizes under-and over-segmentation more than Dice score.

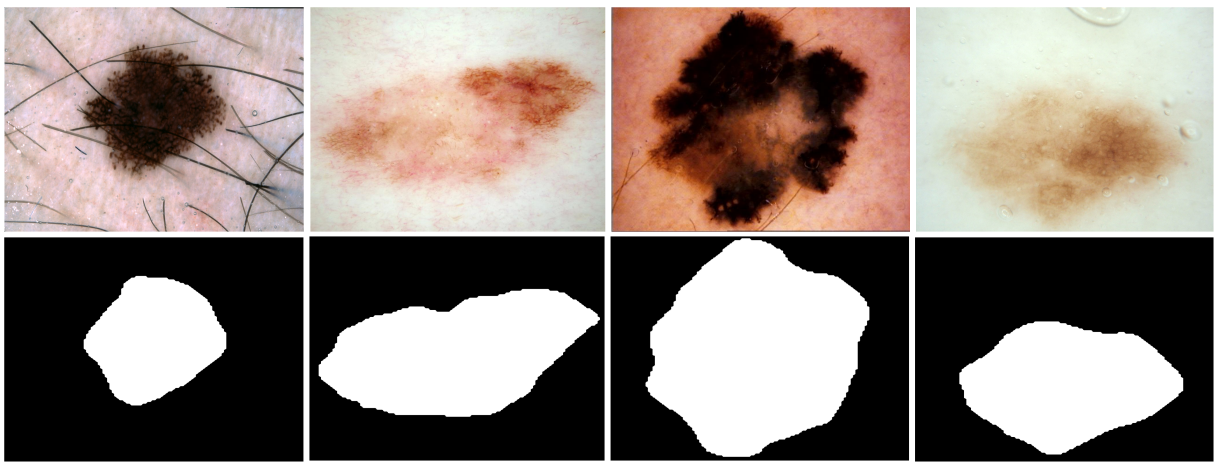

Fig. 2 Sample Images from ISIC 2018 dataset. First and second row represent the image and Ground Truth respectively

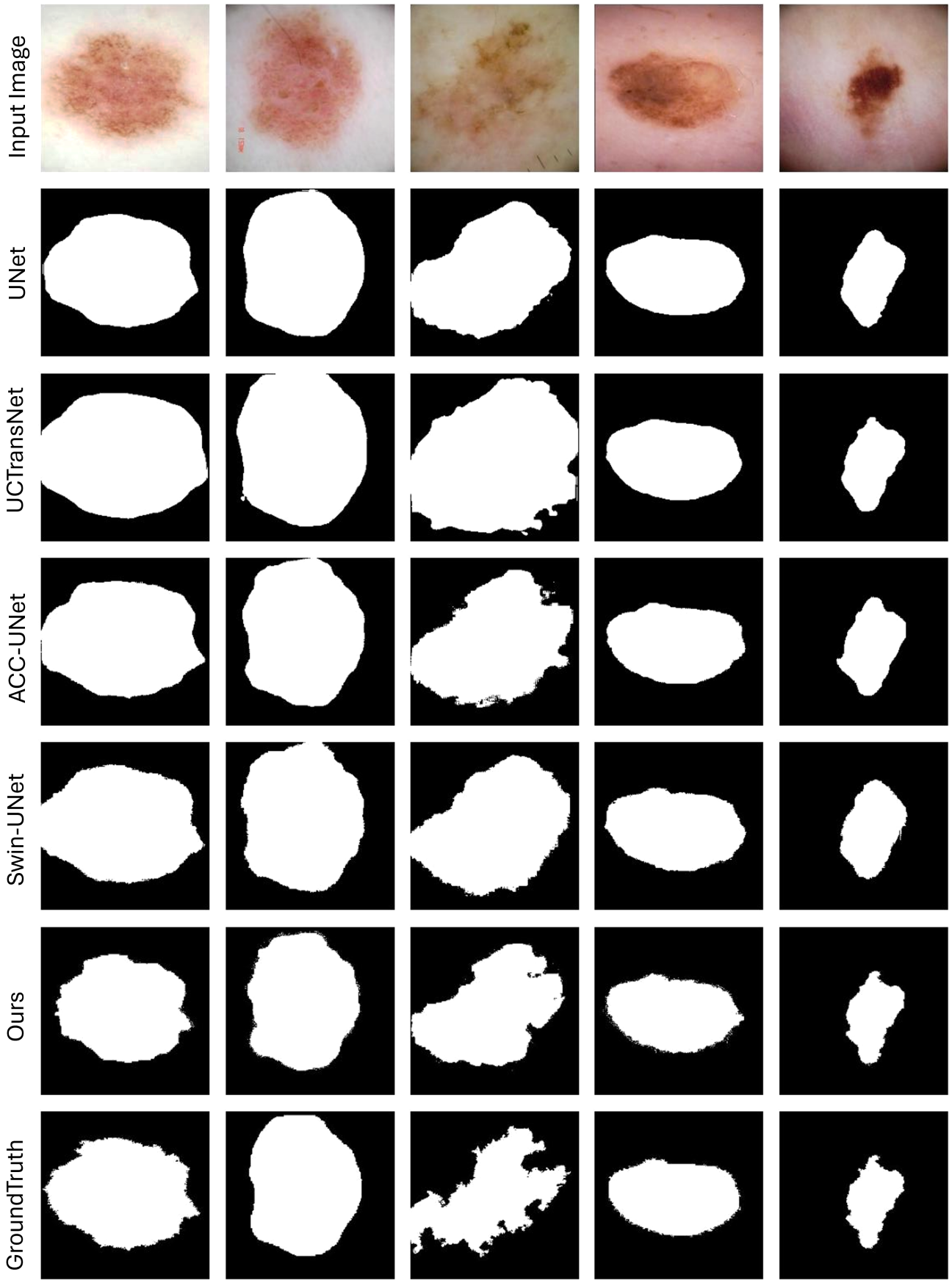

Fig. 3 Methods comparison based on qualitative results from the ISIC 2018 dataset

Table 1 A comparison of the proposed method with existing state-of-the-art approaches for skin lesion segmentation. Note: The **best** is indicated with **bold**, and second best is indicated with underline

Dataset	ISIC 2018	
Metric	Dice score	IoU
UNet[32]	85.41	76.85
UCTransNet[38]	86.69	78.35
ACC-UNet[16]	86.57	78.81
Swin-UNet[15]	88.03	80.02
SegDiff[2]	87.30	79.42
EnsemDiff[41]	88.21	80.72
MedSegDiff[42]	91.30	84.14
Ours	**92.72**	**86.73**

4-4. Result and Discussion

Table 1 provides an analysis of our proposed DiffMamba on the ISIC 2018 dataset. We consider IoU and Dice score (DSC) for evaluation. According to the results, our method outshines both CNN and Transformer based strategies, emphasizing its ability to accurately capture boundaries on the ISIC 2018 dataset. In particular, our proposed approach demonstrates better performance compared to methods that rely on transformers such as Swin-Unet, CNNs such as UNet, and hybrid models such as UCTransNet, ACC-UNet, and diffusion-based methods[38][16][15][42]. Additionally, DiffMamba outperforms the baseline model (Medsegdiff) with improvements of +1.42% in the DSC score and +2.59% in IoU on the ISIC 2018 dataset. Moreover, Figure 3 shows that our method excels in visually capturing detailed structures and defining boundaries more accurately than other models. This visual analysis highlights the superior performance of the Mamba-based fusion module, which effectively captures long-range dependencies while keeping computational complexity linear during the learning process.

5 CONCLUSION

This study proposed the DiffMamba diffusion network for segmenting skin lesions. We improve the diffusion-based framework for medical image segmentation by integrating Mamba mechanism with the original UNet backbone. We introduced a fusion module to align noise and semantic features. The comparative analysis is carried out on ISIC 2018 dataset which shows that our approach outperforms the existing SOTA methods. Considering its effectiveness on ISIC 2018 dataset, in future, the approach will be evaluated on the other medical image segmentation datasets to verify its generalizability. A Mamba-based diffusion model for skin lesion segmentation is being introduced for the first time as per our knowledge, we anticipate that DiffMamba will set a new standard for future research in this field.

REFERENCES

[1] American Cancer Society: Cancer Facts Figures 2016. American Cancer Society, Atlanta, GA, USA (2016)

[2] Amit, T., Shaharbany, T., Nachmani, E., Wolf, L.: Segdiff: Image segmentation with diffusion probabilistic models. arXiv preprint arXiv:2112.00390 (2021)

[3] Argenziano, G., Fabbrocini, G., Carli, P., Giorgi, V. D., Sammarco, E., Delfino, M.: Epiluminescence microscopy for the diagnosis of doubtful melanocytic skin lesions: Comparison of the abcd rule of dermatoscopy and a new 7-point checklist based on pattern analysis. Archives of Dermatology 134(12), 1563–1570 (1998)

[4] Azad, R., Aghdam, E. K., Rauland, A., Jia, Y., Avval, A. H., Bozorgpour, A., Karim-ijafarbigloo, S., Cohen, J. P., Adeli, E., Merhof, D.: Medical image segmentation review: The success of u-net. arXiv preprint arXiv:2211.14830 (2022)

[5] Azad, R., Kazerouni, A., Heidari, M., Aghdam, E. K., Molaei, A., Jia, Y., Jose, A., Roy, R., Merhof, D.: Advances in medical image analysis with vision transformers: A comprehensive review. arXiv preprint arXiv:2301.03505 (2023)

[6] Bi, L., Kim, J., Ahn, E., Kumar, A., Fulham, M., Feng, D.: Dermoscopic image segmentation via multistage fully convolutional networks. IEEE Transactions on Biomedical Engineering 64(9), 2065–2074 (2017)

[7] Celebi, M. E., Wen, Q., Iyatomi, H., Shimizu, K., Zhou, H., Schaefer, G.: A state-of-the-art survey on lesion border detection in dermoscopy images. Dermoscopy image analysis 10(1), 97–129 (2015)

[8] Codella, N., Rotemberg, V., Tschandl, P., Celebi, M. E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M., et al.: Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). arXiv preprint arXiv:1902.03368 (2019)

[9] Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312. 00752 (2023)

[10] Gu, A., Goel, K., Ré, C.: Efficiently modeling long sequences with structured state spaces. arXiv preprint arXiv:2111. 00396 (2021)

[11] Gu, A., Johnson, I., Goel, K., Saab, K., Dao, T., Rudra, A., Ré, C.: Combining recurrent, convolutional, and continuous-time models with linear state space layers. Advances in Neural Information Processing Systems 34, 572–585 (2021)

[12] Guo, H., Li, J., Dai, T., Ouyang, Z., Ren, X., Xia, S. T.: Mambair: A simple baseline for image restoration with state-space model. In: European Conference on Computer Vision. pp. 222–241. Springer (2025)

[13] Hemalatha, R., Thamizhvani, T., Dhivya, A. J. A., Joseph, J. E., Babu, B., Chan-drasekaran, R.: Active contour based segmentation techniques for medical image analysis. Medical and biological image analysis 4(17), 2 (2018)

[14] Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems 33, 6840–6851 (2020)

[15] Hu, C., Wang, Y., Joy, C., Dongsheng, J., Xiaopeng, Z., Qi, T., Manning, W.: Swin-Unet: Unet-like pure transformer for medical image segmentation. In: Computer Vision–ECCV 2022 Workshops. pp. 205–218. Springer Nature Switzerland, Cham (2023)

[16] Ibtehaz, N., Kihara, D.: Acc-unet: A completely convolutional unet model for the 2020s. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 692–702. Springer (2023)

[17] Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J., Maier-Hein, K. H.: nnu-net: A self-configuring method for deep learning-based biomedical image segmentation. Nature Methods 18(2), 203–211 (2021)

[18] Iyatomi, H.: Computer-based diagnosis of pigmented skin lesions. book: Campolo D, ed. New developments in biomedical engineering. pp. 183–200 (2010)

[19] Jerant, A. F., Johnson, J. T., Sheridan, C., Caffrey, T. J.: Early detection and treatment of skin cancer. American Family Physician 61(2), 357–386 (2000)

[20] Kim, B., Oh, Y., Ye, J.: Diffusion adversarial representation learning for self-supervised vessel segmentation. In: The Eleventh International Conference on Learning Representations (2023)

[21] LeCun, Y., Bengio, Y.: Convolutional Networks for Images, Speech, and Time Series. MIT Press, Cambridge, MA, USA (1998)

[22] Liang, D., Zhou, X., Wang, X., Zhu, X., Xu, W., Zou, Z., Ye, X., Bai, X.: Point-mamba: A simple state space model for point cloud analysis. arXiv preprint (2024)

[23] Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., Liu, Y.: Vmamba: Visual state space model. arXiv preprint arXiv:2401.10166 (2024)

[24] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9992–10002. IEEE (October 2021)

[25] Loshchilov, I., Hutter, F.: Fixing weight decay regularization in adam. arXiv preprint arXiv:1711.05101(5) (2017)

[26] Ma, J., Li, F., Wang, B.: U-mamba: Enhancing long-range dependency for biomedical image segmentation. arXiv preprint arXiv:2401.04722 (2024)

[27] Nachbar, F., Stolz, W., Merkle, T., Cognetta, A. B., Vogt, T., Landthaler, M., Bilek, P., B. -Falco, O., Plewig, G.: The abcd rule of dermatoscopy: high prospective value in the diagnosis of doubtful melanocytic skin lesions. Journal of the American Academy of Dermatology 30(4), 551–559 (1994)

[28] Rahman, A., Valanarasu, J. M. J., Hacihaliloglu, I., Patel, V. M.: Ambiguous medical image segmentation using diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11536–11546 (2023)

[29] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 (2022)

[30] Rombach, R., Blattmann, A., Lorenz, D., Esser, P.,

Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10684–10695 (2022)

[31] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 234–241. Springer (2015)

[32] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. pp. 234–241. Springer (2015)

[33] Ruan, J., Xiang, S.: Vm-unet: Vision mamba unet for medical image segmentation. arXiv preprint arXiv:2402.02491 (2024)

[34] Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S., Ayan, B., Mahdavi, S., Lopes, R., et al.: Photorealistic text-to-image diffusion models with deep language understanding. arXiv preprint arXiv:2205.11487 (2022)

[35] Skincancer. org: Melanoma - skincancer. org (2016), http://www.skincancer.org/skin-cancer-information/

[36] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems 30. Curran Associates, Inc. (2017)

[37] Wan, Z., Wang, Y., Yong, S., Zhang, P., Stepputtis, S., Sycara, K., Xie, Y.: Sigma: Siamese mamba network for multi-modal semantic segmentation. arXiv preprint arXiv:2404.04256 (2024)

[38] Wang, H., Cao, P., Wang, J., Zaiane, O.: Uctransnet: Rethinking the skip connections in u-net from a channel-wise perspective with transformer. Proceedings of the AAAI Conference on Artificial Intelligence 36(3), 2441–2449 (June 2022)

[39] Wang, J., Wei, L., Wang, L., Zhou, Q., Zhu, L., Qin, J.: Boundary-aware transformers for skin lesion segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference. pp. 206–216. Springer (2021)

[40] Warfield, S. K., Zou, K. H., Wells, W. M.: Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation. IEEE Transactions on Medical Imaging 23(7), 903–921 (2004)

[41] Wolleb, J., Sandkühler, R., Bieder, F., Valmaggia, P., Cattin, P.: Diffusion models for implicit image segmentation ensembles. In: International Conference on Medical Imaging with Deep Learning. pp. 1336–1348. PMLR (2022)

[42] Wu, J., Fang, H., Zhang, Y., Yang, Y., Xu, Y.: Medsegdiff: Medical image segmentation with diffusion probabilistic model. arXiv preprint arXiv:2211.00611 (2022)

[43] Wu, J., Fu, R., Fang, H., Zhang, Y., Xu, Y.: Medsegdiff-v2: Diffusion based medical image segmentation with transformer. arXiv preprint arXiv:2301.11798 (2023)

[44] Wu, J., Fu, R., Fang, H., Zhang, Y., Yang, Y., Xiong, H., Liu, H., Xu, Y.: Medsegdiff: Medical image segmentation with diffusion probabilistic model. In: Medical Imaging with Deep Learning. pp. 1623–1639. PMLR (2024)

[45] Yuan, Y., Chao, M., Lo, Y. C.: Automatic skin lesion segmentation using deep fully convolutional networks with jaccard distance. IEEE transactions on medical imaging 36(9), 1876–1886 (2017)

[46] Yuanfeng, J., Zhang, R., Huijie, W., Zhen, L., Lingyun, W., Shaoting, Z., Ping, L.: Multi-compound transformer for accurate biomedical image segmentation. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. pp. 326–336. Springer International Publishing, Cham (2021)

[47] Ziheng, W., Min, X., Fangyu, S., Ruinian, J. S. N., Ichen, Y., Ryoichi, N.: Smeswin unet: Merging cnn and transformer for medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention (MICCAI) 2021 (2021)

[48] Zortea, M., Skrøvseth, S. O., Schopf, T. R., Kirchesch, H. M., Godtliebsen, F.: Automatic segmentation of dermoscopic images by iterative classification. International journal of biomedical imaging 2011(1), 972648 (2011)

DiffMamba: Leveraging Mamba for Effective Fusion of Noise and Conditional Features in Diffusion Models for Skin Lesion Segmentation

■著者

Amit Shakya
Emerging Technology and Innovation Lab,
Yamaha Motor Solutions India

Shruti Phutke
Emerging Technology and Innovation Lab,
Yamaha Motor Solutions India

Chetan Gupta
Emerging Technology and Innovation Lab,
Yamaha Motor Solutions India

Rupesh Kumar
Emerging Technology and Innovation Lab,
Yamaha Motor Solutions India

Lalit Sharma
Emerging Technology and Innovation Lab,
Yamaha Motor Solutions India

Chetan Arora
Indian Institute of Technology Delhi