InceptionMamba: Efficient Multi-Stage Feature Enhancement with Selective State Space Model for Microscopic Medical Image Segmentation

Daniya Najiha Abdul Kareem[1] DANIYA.KAREEM@MBZUAI.AC.AE
Abdul Hannan[2] ABDUL.HANNAN@UNITN.IT
Mubashir Noman[1] MUBASHIR.NOMAN@MBZUAI.AC.AE
Jean Lahoud[1] JEAN.LAHOUD@MBZUAI.AC.AE
Mustansar Fiaz[3] MUSTANSAR.FIAZ@GMAIL.COM
Hisham Cholakkal[1] HISHAM.CHOLAKKAL@MBZUAI.AC.AE
[1] *Mohamed bin Zayed University of AI, UAE*
[2] *University of Trento, Italy*
[3] *IBM Research, UAE*

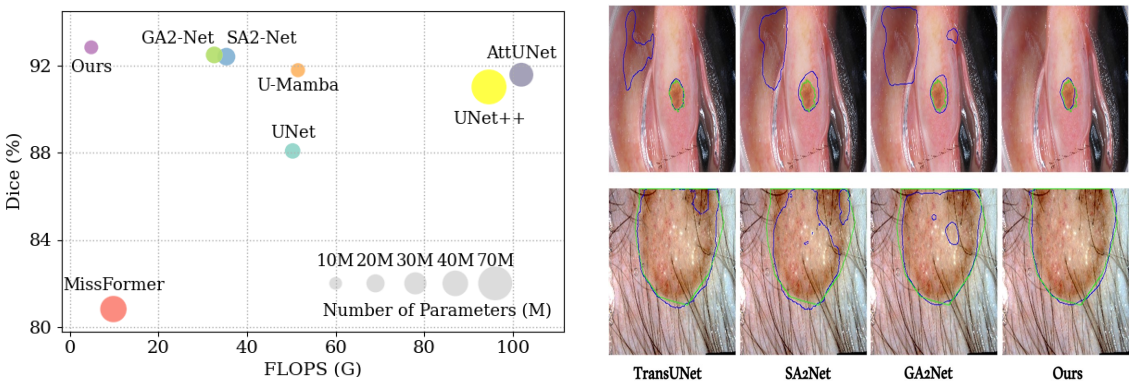
Editors: Under Review for MIDL 2026

Abstract

Accurate microscopic medical image segmentation plays a crucial role in diagnosing various cancerous cells and identifying tumors. Driven by advancements in deep learning, convolutional neural networks (CNNs) and transformer-based models have been extensively studied to enhance receptive fields and improve medical image segmentation task. However, they often struggle to capture complex cellular and tissue structures in challenging scenarios such as background clutter and object overlap. Moreover, their reliance on the availability of large datasets for improved performance, along with the high computational cost, limit their practicality. To address these issues, we propose an efficient framework for the segmentation task, named InceptionMamba, which encodes multi-stage rich features and offers both performance and computational efficiency. Specifically, we exploit semantic cues to capture both low-frequency and high-frequency regions to enrich the multi-stage features to handle the blurred region boundaries (e.g., cell boundaries). These enriched features are input to a hybrid model that combines an Inception depth-wise convolution (Szegedy et al., 2015) with a Mamba block (Gu and Dao, 2023), to maintain high efficiency and capture inherent variations in the scales and shapes of the regions of interest. These enriched features, along with low-resolution features, are fused to get the final segmentation mask. Our model achieves state-of-the-art performance on two challenging microscopic segmentation datasets (SegPC21 and GlaS) and two skin lesion segmentation datasets (ISIC2017 and ISIC2018), while reducing computational cost by about 5 times compared to the previous best performing method.

1. Introduction

Microscopic image segmentation is an essential task in the medical domain as it assists specialists in diagnosing various diseases accurately and thoroughly. Additionally, it provides valuable information in other medical applications such as drug discovery and cellular analysis (Leygeber et al., 2019). However, medical image segmentation is a challenging task due to the presence of irregularly shaped objects consisting of indistinct boundaries having



(a) (b)

Figure 1: (a) Comparison of our method with existing models based on Dice score, parameters, and GFLOPs on the SegPC21 dataset. Our method performs favorably compared to the state-of-the-art GA2-Net while requiring 4.85× fewer GFLOPs and 1.45× fewer parameters. (b) Qualitative comparisons of our method on the ISIC2018 dataset. Our method performs better in cluttered backgrounds and correctly detects the affected region. (Ground truth - green, Prediction - blue)

various sizes, especially for the microscopic images (Fiaz et al., 2023). The problem becomes further challenging in the presence of complex backgrounds and clutter. For instance, the texture, color, and shape of cells and tissues in microscopic images from different organs present significant difficulties for accurate segmentation.

Earlier segmentation methods in the medical domain were composed of convolutional neural networks (CNNs) in an encoder-decoder fashion. UNet (Ronneberger et al., 2015), introduced a decade ago, gained popularity for the medical segmentation task. Subsequently, many researchers adapted this architecture due to its promising performance, leveraging dense skip connections and multi-scale feature extraction (Xu and Duan, 2021; Wang et al., 2020; Bala and Kant, 2020). Although CNN-based methods exhibit satisfactory performance in segmentation tasks, they still struggle to capture long-range dependencies and lack the ability to accurately segment targets of varying sizes. Later, transformer-based methods were introduced due to their potential to capture global contextual relationships (Cao et al., 2023; Lin et al., 2022; Valanarasu et al., 2021; Yuan et al., 2023; Tragakis et al., 2023). However, the presence of complex backgrounds and small structures in the microscopic images constrains the performance of transformer-based methods. Moreover, the limited data in the medical domain degrade the performance of the transformer-based methods due to their data-hungry properties. Recently, Mamba-based models (Ma et al., 2024b; Xing et al., 2024; Liu et al., 2024) have gained attention with their ability to dynamically focus on relevant information based on the image characteristics. Numerous studies have been conducted based on the improvement of long-range modeling capabilities in CNNs when paired with Mamba blocks (Ma et al., 2024b; Wang et al., 2024b). However, such approaches are computationally expensive and still struggle to perform better in the presence of overlapped object structures having a wide range of shape and size variations. Therefore, we propose a novel CNN and Mamba architecture, named InceptionMamba, which offers a simple yet effective framework for medical image segmentation tasks. As shown in Fig.

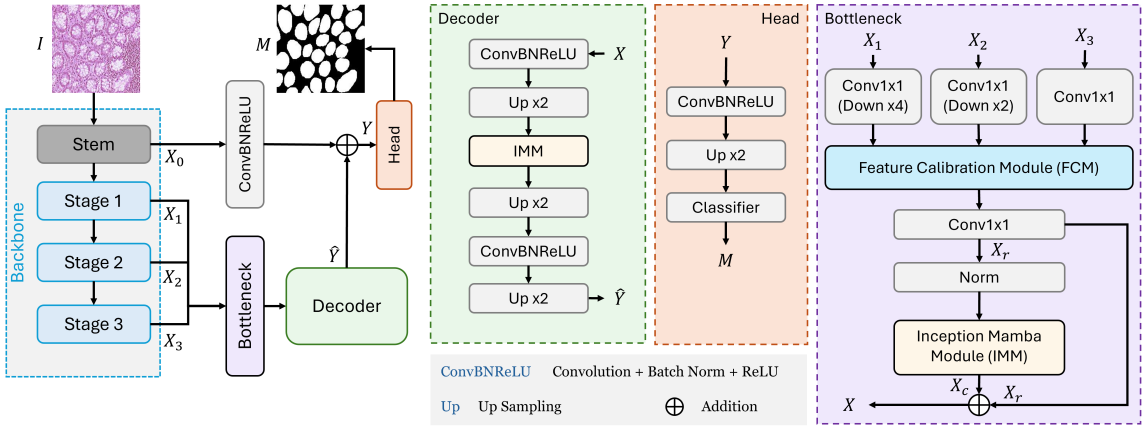


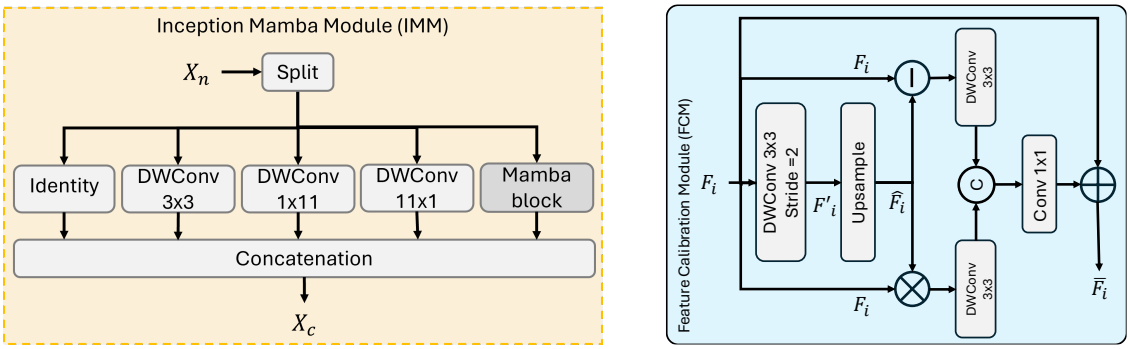
Figure 2: Illustration of the overall architecture of our proposed framework. **(left)** We feed the image to the backbone and extract features X_i where $i \in 0, 1, 2, 3$. Multi-stage features from the backbone are effectively combined and enhanced by the bottleneck block. Afterwards, the decoder upsamples and further refines multi-contextual features. Finally, the output of the decoder is combined with features of the stem, and the segmentation head is utilized to obtain mask M. **(middle)** We also illustrate the structure of the decoder (light green box) and segmentation head (light orange box). **(right)** The bottleneck block, which includes the Feature Calibration Module, is shown in light purple.

1(a) and Fig. 1(b), this framework enhances segmentation performance while significantly reducing computational complexity. Our contributions are as follows:

- We propose an efficient module, named Inception Mamba module (IMM), to capture multi-contextual representations by utilizing convolutions and state space models.
- We introduce a bottleneck block that effectively combines the multi-scale information from the backbone network while highlighting the fine details to improve segmentation performance. In addition, we utilize a simple decoder without dense connections that reduces computational complexity.
- We perform extensive experimentation on four medical datasets, showing that our method achieves state-of-the-art performance while reducing the computational cost by about 5 times compared to the previous top-performing method (Fiaz et al., 2024).

2. Method

This section demonstrates the proposed approach in detail. First, we introduce the overall architecture of the proposed approach, followed by an elaboration on our Inception Mamba module (IMM) that efficiently captures the multi-contextual features and provides rich semantic information necessary for the segmentation task. Finally, we discuss the proposed bottleneck block that effectively combines the multi-scale information from backbone stages while preserving the boundary information.



(a) (b)

Figure 3: (a) Inception Mamba module (IMM) takes input features X_n, splits them channel-wise, and utilizes various convolution kernels and state space module to extract multi-contextual features X_c for better segmentation performance as depicted above. (b) Illustration of the feature calibration module. Feature maps F_i are first down-sampled using depth-wise convolution followed by an upsampling operation. Then, we multiply features F_i with $\hat{F}_i$ to highlight blob regions. Similarly, subtraction of F_i and $\hat{F}_i$ is performed to focus on fine details. Finally, highlighted features are combined and projected to get rich semantic features.

2.1. Overall Architecture

We illustrate the overall architecture of the proposed framework in Fig. 2. We adopt a U-Net (Ronneberger et al., 2015) based architecture with ResNet (He et al., 2016) backbone as our baseline model. As depicted, the model takes an image $I \in \mathbb{R}^{C\times H\times W}$ as input and feeds it to the ResNet backbone to extract multi-scale features X_i where $i = 0, 1, 2, 3$. Here, $i = 0$ refers to the features of the stem layer. To reduce the computational complexity of the model, we utilize the features of the first three stages of the ResNet backbone. We then pass the multi-scale features to the bottleneck block for enrichment. The bottleneck block takes the features of three stages and passes them to a feature calibration module (FCM) that is responsible for highlighting the boundary information by utilizing the low and high-frequency information in a parallel manner. Afterward, IMM is utilized to capture the multi-contextual features from the refined feature maps. The proposed bottleneck block provides the rich multi-contextual information necessary for better segmentation performance. The output of the bottleneck block is input to a decoder that uses several convolution and upsampling layers in a cascade manner. We carefully utilize the IMM in the decoder to further enhance the multi-contextual information as shown in Fig. 2 (light green box). Finally, we add the low-level semantics of the stem layer of the backbone network with the semantically rich feature maps of the decoder and feed it to a segmentation head module (as shown in Fig. 2 (light orange box)) which provides the output prediction mask M.

2.2. Inception Mamba Module

The Inception Mamba module (IMM) is illustrated in Fig. 3(a). As mentioned earlier, medical microscopic images contain complex object structures having varying shapes and sizes. Therefore, an explicit module is required to effectively detect medical objects in microscopic

images. To this end, we introduce a simple and effective module, the Inception Mamba module (IMM), that simultaneously captures different contextual representations while aiming to minimize the computational complexity. During the depth-wise convolutions, square and rectangular convolutional kernels are utilized to obtain feature representations having smaller and larger receptive fields. We choose depth-wise convolutions to build a computationally efficient module. In parallel, the state-space Mamba block is employed to capture larger contextual representations while reducing the computational cost. In contrast to the Inception block (Szegedy et al., 2015, 2016) that uses the max-pooling operation to add another context branch, we utilize the identity operation. We empirically observe that the identity branch when utilized in IMM provides better segmentation performance. Finally, we concatenate the different representations in the channel dimension to obtain the rich multi-contextual feature maps (X_c).

Mamba Block: Mamba (Gu and Dao, 2023) effectively encodes global representations while benefiting from low computational costs, striking a balanced trade-off between performance and efficiency. In addition, the selective scan mechanism of Mamba has the potential to mine the core semantics for long sequences, which reduces the semantic redundancy limitation. Motivated by the selective scan mechanism and high efficiency of the Mamba block, we integrate it into our inception Mamba module to capture the various contextual representations. This fusion of the Mamba module within our IMM provides a mutually beneficial outcome by capturing the global representations while maintaining high efficiency.

2.3. Bottleneck Block

The critical component of the proposed framework is the bottleneck block as shown in Fig. 2 (right). The bottleneck block takes the feature maps, X_i where $i \in 1, 2, 3$, of three stages from the backbone network and enhances them by employing the feature calibration module (FCM). The FCM is responsible for highlighting the fine details that are necessary to separate the overlapped cells and tissues as well as enhancing the feature contrast. The feature representations of each stage are separately enhanced by the FCM blocks. For the first stage, i.e. $i = 1$, feature maps are first downsampled by a factor of four before inputting them into the FCM block. Similarly, we downsample the feature representations of the second stage by a factor of two and then feed them to the FCM block. After feature enhancement, we concatenate the feature maps of three stages and utilize the 1×1 convolution operation to intermix the different responses and obtain rich multi-scale representations X_r. Later, we utilize the IMM to further boost the contextual representations and obtain features X_c. Finally, we add the features X_r and X_c to obtain semantically rich features. Both FCM and IMM complement each other to provide better representations for medical segmentation tasks. Our approach effectively combines the low-level semantics of the first and second stages with the high-level semantics of the third stage of the backbone network to achieve better segmentation performance.

Feature Calibration Module: As shown in Fig. 3(*b*), the feature calibration module takes input features F_i and utilizes a 3×3 depth-wise convolution having a stride value of 2 to reduce the spatial resolution of the features producing F'_i. We then use interpolation operation to upsample the spatial resolution of the features to obtain feature maps $\hat{F}_i$. The objective of the convolutional down sampling followed by upsampling is to produce

smoothness in the feature maps. Subsequently, we subtract the features $\hat{F}_i$ from the F_i to highlight the fine details. Simultaneously, we multiply the features $\hat{F}_i$ and F_i to focus on the blob regions. Afterwards, we utilize the convolution operation on the subtracted and multiplied features and concatenate them. Finally, we utilize the 1×1 convolution to project the feature representations and add the input F_i to obtain enhanced feature maps $\bar{F}_i$. The feature calibration module enhances boundary information addressing the complex nature of medical image segmentation.

2.4. Decoder

To reduce the computational cost of the framework, we introduce an effective decoder as shown in Fig. 2. We empirically verify that the dense skip connections in the baseline increase the computational cost while negligibly improving the segmentation performance. Additionally, we observe that the high-level semantics of the fourth stage of the backbone are not profitable for microscopic medical images. We therefore utilize a decoder that does not require skip connections while still achieving better segmentation performance.

3. Experimental Study

3.1. Datasets & Experimental details

We benchmark our results on the SegPC21 (Gupta et al., 2021), GlaS (Sirinukunwattana et al., 2017), ISIC2017 (Codella et al., 2018), and ISIC2018 (Codella et al., 2019) datasets. In this study, we perform all the experiments using 32GB Tesla V100 GPU based on PyTorch 2.1.1+cu118. We utilize ResNet50 (He et al., 2016), pre-trained on ImageNet (Deng et al., 2009), as our backbone network to extract the features from the stem layer and the first three-layer features. We set the input resolution to 224×224. During the training, we apply rotation and random flipping augmentation techniques. We train the model using the combined cross-entropy and DICE loss functions. For SegPC21, ISIC2018, and ISIC2017 datasets, we follow (Fiaz et al., 2023; Azad et al., 2024) and set batch size of 16, learning rate 0.0001, and use the Adam optimizer to train the model for 100 epochs. For the GlaS dataset, we follow UCTransNet (Wang et al., 2022a) and set the batch size to 4, the initial learning rate to 0.001, and employ the Adam optimizer to train our model. To ensure the results are robust for a smaller dataset, similar to (Fiaz et al., 2023; Wang et al., 2022a), we perform three times 5-fold cross-validation. During inference, we employ an ensemble technique to obtain the final prediction masks by taking the mean for all five models. We evaluate our method using the Dice and IoU metrics.

3.2. Quantitative Comparison

We compare our method with existing CNN-based, transformer-based, and Mamba-based methods on the SegPC21 dataset in Table 1. We report Dice score, IoU, number of parameters (in millions), and FLOPs (GFLOPs). From Table 1, compared to CNN methods, we notice that AttUNet (Oktay et al., 2018) attains the best Dice and IoU scores of 91.58% and 91.44%, respectively. However, among hybrid approaches, GA2Net (Fiaz et al., 2024) achieves 92.49% Dice and 92.32% IoU scores. In contrast, our approach achieves state-of-the-art Dice and IoU scores of 92.56% and 92.37%, respectively. In addition, we notice that

Table 1: Comparison of our method on SegPC21 dataset with state-of-the-art methods. The best two results are in red and blue, respectively. ⋆ means the model backbone is PVT-V2-B2 ((Wang et al., 2022b)).

Method	Params (M)	GFLOPs	SegPC21	
			Dice (%)	IoU (%)
U-Net (Ronneberger et al., 2015)	14.8	50.3	88.08	88.2
UNet++ (Zhou et al., 2018)	74.5	94.6	91.02	90.92
AttUNet (Oktay et al., 2018)	34.9	101.9	91.58	91.44
MultiResUNet (Ibtehaz and Rahman, 2020)	57.2	78.4	86.49	86.76
TransUNet (Chen et al., 2021)	105.0	56.7	82.33	83.38
MissFormer(Huang et al., 2021)	42.46	9.86	80.82	82.09
UCTransNet (Wang et al., 2022a)	65.6	63.2	91.74	91.59
SA2-Net (Fiaz et al., 2023)	19.3	35.36	92.41	92.23
DwinFormer (Kareem et al., 2024)	198.0	113.13	91.10	90.99
UDTransNet (Wang et al., 2024a)	33.8	63.2	89.91	-
UN-SAM (Chen et al., 2025)	17.36	32.61	89.01	82.14
U-Mamba (Ma et al., 2024b)	12.36	51.5	91.79	91.83
LKMUNet (Wang et al., 2024b)	123.8	251.5	92.20	92.02
GA2-Net (Fiaz et al., 2024)	17.36	32.61	92.49	92.32
(Ours)	11.92	6.72	92.56	92.37
(Ours*)	27.27	4.86	92.84	92.63

Table 2: Comparison of our method on the GlaS dataset with state-of-the-art methods. The best two results are in red and blue, respectively.

Method	GlaS	
	Dice (%)	IoU (%)
U-Net (Ronneberger et al., 2015)	85.45 ± 1.3	74.78 ± 1.7
UNet++ (Zhou et al., 2018)	87.56 ± 1.2	79.13 ± 1.7
AttUNet (Oktay et al., 2018)	88.80 ± 1.1	80.69 ± 1.7
MultiResUNet (Ibtehaz and Rahman, 2020)	88.73 ± 1.2	80.89 ± 1.7
TransUNet (Chen et al., 2021)	88.40 ± 0.7	80.40± 1.0
MedT (Valanarasu et al., 2021)	85.93 ± 2.9	75.47 ± 3.5
Swin-Unet (Cao et al., 2023)	89.58 ± 0.6	82.07 ± 0.7
UCTransNet (Wang et al., 2022a)	90.18 ± 0.7	82.96 ± 1.1
SA2-Net (Fiaz et al., 2023)	91.38 ± 0.4	84.90 ± 0.6
UDTransNet (Wang et al., 2024a)	91.03 ± 0.6	-
TriConvUNeXt (Ma et al., 2024a)	91.41 ± 0.4	84.87 ± 0.3
Ours	91.88 ± 0.3	85.65 ± 0.4

our approach obtains a significant amount of reduced parameters and GFLOPs. Compared to GA2Net, our method has around 1.45 times fewer parameters and 4.85 times fewer GFLOPs. Moreover, we also demonstrate that while employing the PVT-V2-B2 (Wang et al., 2022b) backbone, our method further improves the Dice and IoU scores to 92.84% and 92.63%, respectively. We notice that this further reduces the computational cost to 4.86 GFLOPs. Furthermore, our method achieves an impressive HD95 score of 1.92 ± 0.32 compared to the HD95 score of 3.87 ± 3.56 for the top-performing GA2Net model, which indicates better boundary segmentation capability.

In Table 2, we compare our method on the GlaS dataset. Among the compared methods, we observe that SA2Net (Fiaz et al., 2023), TriConvUNeXt (Ma et al., 2024a), and UDTransNet (Wang et al., 2024a) attain around 91.00% Dice score. In contrast, we achieve 91.88% Dice and 85.65% IoU score, which demonstrates the state-of-the-art performance of our approach. We also compare our method on ISIC2017 and ISIC2018 datasets in Table 3. Among CNN-based methods, MultiResUNet (Ibtehaz and Rahman, 2020) ranks first with a Dice score of 86.94 % and an IoU score of 85.37 on ISIC2018 dataset. However, the re-

Table 3: Comparison with state-of-the-art methods on ISIC2017 and ISIC2018 datasets using the Dice and IoU metrics. The best two results are in red and blue, respectively.

Method	ISIC2017		ISIC2018	
	Dice (%)	IoU (%)	Dice (%)	IoU (%)
U-Net (Ronneberger et al., 2015)	81.59	79.32	86.71	84.91
UNet++ (Zhou et al., 2018)	82.32	80.13	88.22	86.51
AttUNet (Oktay et al., 2018)	86.45	85.97	88.20	86.49
MultiResUNet (Ibtehaz and Rahman, 2020)	86.83	85.91	86.94	85.37
TransUNet (Chen et al., 2021)	86.11	84.98	84.99	83.65
MissFormer(Huang et al., 2021)	87.82	85.79	86.57	84.84
UCTransNet (Wang et al., 2022a)	88.81	87.22	88.98	87.29
SA2-Net (Fiaz et al., 2023)	89.59	88.14	88.88	87.21
UDTransNet (Wang et al., 2024a)	90.11	88.49	89.91	88.02
U-Mamba (Ma et al., 2024b)	89.60	88.45	89.37	88.12
VMKLA-UNet (Su et al., 2025)	89.16	87.72	89.82	86.90
LKMUNet (Wang et al., 2024b)	90.25	88.89	89.33	87.60
GA2-Net (Fiaz et al., 2024)	89.99	88.58	89.29	87.69
Ours	91.16	89.82	90.56	88.98

cently introduced UDTransNet (Wang et al., 2024a), GA2Net (Fiaz et al., 2024), U-Mamba (Ma et al., 2024b), and LKMUNet (Wang et al., 2024b) achieve a Dice score of more than 89.00%. Our approach demonstrates better performance with a Dice score of 90.56% and an IoU score of 88.98% on ISIC2018 dataset. We have computed the HD95 score for our proposed method. On the ISIC2018 dataset, our method achieves an HD95 score of 4.15 ± 0.20 (where lower is better), significantly outperforming the state-of-the-art UDTransNet, which has an HD95 score of 10.85 ± 0.26. Similarly, our method shows consistent improvement on ISIC2017.

3.3. Qualitative Comparison

To further demonstrate the effectiveness of our method, we perform a qualitative comparison in Fig. 4. We compare our method with UNet (Ronneberger et al., 2015), TransUNet (Chen et al., 2021), SA2Net (Fiaz et al., 2023) and GA2Net (Fiaz et al., 2024). We compare the methods on SegPC21 dataset samples and observe that our method exhibits better capabilities to learn the multi-scale features and preserve the boundary regions of the cytoplasm in the cluttered and unclear backgrounds.

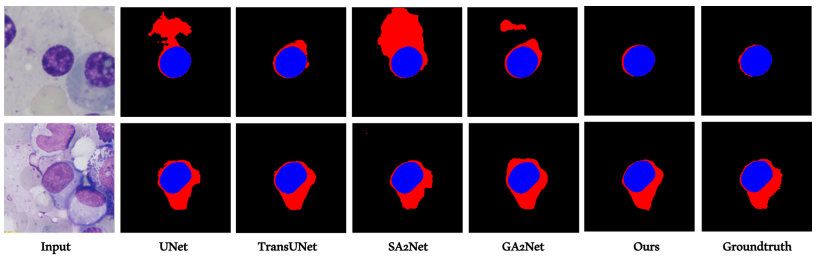


Figure 4: Qualitative comparison of InceptionMamba on SegPC21 dataset. For a fair comparison, we follow (Fiaz et al., 2023; Azad et al., 2024) and perform the cytoplasm segmentation (red mask) for a given input nucleus mask (blue). Our method provides improved segmentation performance compared to other methods.

Table 4: Ablation study comparing the Dice score and computational requirements of our method over SegPC21 dataset. Proposed FCM and IMM modules positively impact the model performance with lesser computational cost.

Exp.	Methods	Dice	Params(M)	GFLOPs
1	Baseline	89.05	11.42	5.27
2	Baseline + FCM	90.4	11.53	5.29
3	Baseline +IMM (with only Inception DWC)	89.6	11.42	5.27
4	Baseline +IMM (with only Mamba)	90.2	11.42	5.70
5	Baseline +IMM (with Inception DWC and Mamba)	90.9	11.43	5.28
6	Baseline+ Inception Module (Szegedy et al., 2015)	89.4	11.43	5.27
7	Baseline+ Inception Module (Szegedy et al., 2016)	89.3	11.43	5.27
8	Baseline+ Inception Module (Szegedy et al., 2016)	89.5	11.44	5.29
9	Baseline+ Self-Attention (SA)	89.9	12.07	6.7
10	Baseline+IMM (SA replacing Mamba)	90.3	**12.46**	8.3
11	Baseline + FCM + IMM	91.5	11.54	5.30
12	Baseline + FCM + IMM + Decoder with IMM	92.05	11.67	6.09
13	Ours	**92.56**	11.92	6.72

3.4. Ablation Study

In Table 4, we present an ablation study on SegPC21 dataset to validate the effectiveness of our contributions. We adopt modified UNet (Ronneberger et al., 2015) without skip connections as the baseline, pre-trained on ImageNet (Deng et al., 2009). The baseline utilizes the stem features and the first three stages feature as multi-stage features, which are fused and passed to a decoder with upsampling and convolution layers. We experiment with combinations of different modules with baseline including FCM (Exp 2), IMM with only Inception depth-wise convolutions (Exp 3), IMM with only Mamba block (Exp 4), the proposed IMM (Exp 5), variants of Inception Module (Exp 6 - 8), Self-Attention (Exp 9) and IMM utilizing Self-Attention instead of Mamba block (Exp 10). It is evident that baseline combined with FCM (Exp 2) and with IMM (Exp 5) outperform other combinations with comparable parameters and GFLOPs. Although Self-Attention and Inception Modules are capable of encoding global details, they fail to explicitly encode complex semantic details for the identification of targets with varying shapes and blurred boundaries. Hence, we integrate the IMM and FCM modules into to baseline, which is set as our design choice. Moreover, directly using the convolutions and upsampling in a decoder might lose complementary information. We modified the decoder and introduced our IMM module within the decoder which resulted in a significant gain in performance (Exp 12). Finally, we utilize a skip connection from stem features using a convolution layer, which serves a complementary function, which further improves the performance (Exp 13). In addition, we carefully analyze the impact of the IMM design by evaluating the individual contributions of the Inception DWC and Mamba blocks within the IMM architecture (Exp 3, Exp 4). It can be noticed that integration of IMM (with only Inception DWC) into the baseline improves the performance while maintaining the computational cost, whereas the integration of IMM (with only Mamba block) increases both the performance and computational cost due to the large number of channels for the Mamba block. When both the Inception DWC and Mamba modules of the IMM block are included (Exp 11), the model achieves further Dice score improvement (+1.85%) over the baseline, while maintaining a computational cost (GFLOPs) nearly identical to the baseline. Note that splitting the feature channels between

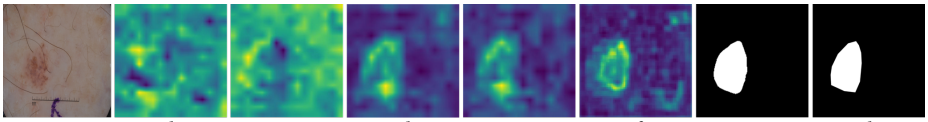


Figure 5: Illustration of the impact of our contributions. (a) is the input, (b) and (c) presents stage 2 and 3 output feature maps. The (d) and (e) are output features by FCM and IMM, respectively. The (f) is the decoder feature map. Finally (g) and (h) are the prediction and ground truth. These feature visualizations show that our model preserve the precise contours for the unclear tissue boundaries.

Table 5: Ablation studies on the kernel sizes for DWC1, DWC2 and DWC3 in IMM. Arrangement in Exp 4 showed improved performance with optimal computation cost.

Exp	Kernel Size	GFLOPs	Dice
1	(3,3)(1,5)(5,1)	6.32	91.73
2	(7,7)(1,5)(5,1)	7.15	92.28
3	(3,3)(1,7)(7,1)	6.57	91.98
4	(3,3)(1,11)(11,1)	6.72	**92.56**

Table 6: Ablation studies on the position of IMM in the decoder. We notice that placing IMM in the middle of decoder results in the optimal solution.

Method	Dice
First place of IMM in decoder	92.21
Second place of IMM in decoder	**92.56**
Third place of IMM in decoder	92.32

the Identity, Inception DWC, and Mamba blocks within the IMM enhances performance while limiting the computational cost compared to using Mamba blocks alone. From these experiments, it is clear that our contributions exhibit better capabilities to learn refined multi-context information for accurate segmentation tasks.

We extend our ablation studies by analyzing the impact of depthwise convolution kernal sizes in IMM. From Table 5, it is evident that increasing the kernel sizes simultaneously in horizontal and vertical directions increases the computational cost (Exp 2). So we set DWC1 with kernel size (3,3) to capture fine-grained local textures followed by DWC2 and DWC3 with kernel sizes (1×11) and (11×1) to efficiently maximize the receptive field for capturing elongated structural patterns with lesser computational cost (Exp 4). Furthermore, we perform an ablation study to find the optimal position of our IMM in the decoder. In Table 6, it can be seen that IMM in the middle of the decoder provides optimal performance gain. The mid-resolution stage balances high-level semantic information with fine-grained spatial details, compared to higher and lower resolution stages. Finally, we show the feature map visualizations of our contribution in Fig. 5, which highlights that our model can learn the complex structures of tissues in the presence of cluttered and noisy backgrounds.

4. Conclusion

In this paper, we propose a segmentation framework that exploits multi-stage features from a backbone architecture and passes them to the bottleneck which is responsible for generating enriched multi-contextual feature representations using feature calibration and Inception Mamba modules, while remaining computationally efficient. Furthermore, our simple yet effective decoder incorporating Inception Mamba module provides better segmentation results. Experimental results on two microscopic and two skin lesion segmentation datasets reveal the significance of our approach.

References

Reza Azad, Ehsan Khodapanah Aghdam, and et al. Medical image segmentation review: The success of u-net. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Surayya Ado Bala and Shri Kant. Dense dilated inception network for medical image segmentation. *international Journal of Advanced Computer Science and Applications*, 11 (11), 2020.

Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part III*, pages 205–218. Springer, 2023.

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.

Zhen Chen, Qing Xu, Xinyu Liu, and Yixuan Yuan. Un-sam: Domain-adaptive self-prompt segmentation for universal nuclei images. *Medical Image Analysis*, page 103607, 2025.

Noel Codella 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.

Noel CF Codella et al. 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)*, pages 168–172. IEEE, 2018.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Mustansar Fiaz, Moein Heidari, Rao Muhammad Anwar, and Hisham Cholakkal. Sa2-net: Scale-aware attention network for microscopic image segmentation. *arXiv preprint arXiv:2309.16661*, 2023.

Mustansar Fiaz, Mubashir Noman, and et al. Guided-attention and gated-aggregation network for medical image segmentation. *Pattern Recognition*, page 110812, 2024.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Anubha Gupta, Ritu Gupta, Shiv Gehlot, and Shubham Goswami. Segpc-2021: Segmentation of multiple myeloma plasma cells in microscopic images. *IEEE Dataport*, 1(1):1, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Xiaohong Huang, Zhifang Deng, Dandan Li, and Xueguang Yuan. Missformer: An effective medical image segmentation transformer. *arXiv preprint arXiv:2109.07162*, 2021.

Nabil Ibtehaz and M Sohel Rahman. Multiresunet: Rethinking the u-net architecture for multimodal biomedical image segmentation. *Neural networks*, 121:74–87, 2020.

Daniya Najiha Abdul Kareem, Mustansar Fiaz, Noa Novershtern, and Hisham Cholakkal. Medical image segmentation using directional window attention. *arXiv preprint arXiv:2406.17471*, 2024.

Markus Leygeber, Dorina Lindemann, Christian Carsten Sachs, Eugen Kaganovitch, Wolfgang Wiechert, Katharina Nöh, and Dietrich Kohlheyer. Analyzing microbial population heterogeneity—expanding the toolbox of microfluidic single-cell cultivations. *Journal of molecular biology*, 431(23):4569–4588, 2019.

Ailiang Lin, Bingzhi Chen, Jiayu Xu, Zheng Zhang, Guangming Lu, and David Zhang. Ds-transunet: Dual swin transformer u-net for medical image segmentation. *IEEE Transactions on Instrumentation and Measurement*, 71:1–15, 2022.

Jiarun Liu, Hao Yang, Hong-Yu Zhou, Yan Xi, Lequan Yu, Yizhou Yu, Yong Liang, Guangming Shi, Shaoting Zhang, Hairong Zheng, and Shanshan Wang. Swin-umamba: Mamba-based unet with imagenet-based pretraining. *arXiv preprint arXiv:2402.03302*, 2024.

C. Ma, Y. Gu, and Z. Wang. Triconvunext: A pure cnn-based lightweight symmetrical network for biomedical image segmentation. *Journal of Digital Imaging*, 37:2311–2323, 2024a. doi: 10.1007/s10278-024-01116-8. URL https://doi.org/10.1007/s10278-024-01116-8.

Jun Ma, Feifei Li, and Bo Wang. U-mamba: Enhancing long-range dependency for biomedical image segmentation. *arXiv preprint arXiv:2401.04722*, 2024b.

Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

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: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

Korsuk Sirinukunwattana, Josien PW Pluim, Hao Chen, Xiaojuan Qi, Pheng-Ann Heng, Yun Bo Guo, Li Yang Wang, Bogdan J Matuszewski, Elia Bruni, Urko Sanchez, et al. Gland segmentation in colon histology images: The glas challenge contest. *Medical image analysis*, 35:489–502, 2017.

C. Su, X. Luo, S. Li, et al. Vmkla-unet: vision mamba with kan linear attention u-net. *Scientific Reports*, 15(13258), 2025. doi: 10.1038/s41598-025-97397-2. URL https://doi.org/10.1038/s41598-025-97397-2.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

Athanasios Tragakis, Chaitanya Kaul, Roderick Murray-Smith, and Dirk Husmeier. The fully convolutional transformer for medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3660–3669, 2023.

Jeya Maria Jose Valanarasu, Poojan Oza, Ilker Hacihaliloglu, and Vishal M Patel. Medical transformer: Gated axial-attention for medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*, pages 36–46. Springer, 2021.

Haonan Wang, Peng Cao, Jiaqi Wang, and Osmar R Zaiane. Uctransnet: rethinking the skip connections in u-net from a channel-wise perspective with transformer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 2441–2449, 2022a.

Haonan Wang, Peng Cao, Jinzhu Yang, and Osmar Zaiane. Narrowing the semantic gaps in u-net with learnable skip connections: The case of medical image segmentation. *Neural Networks*, page 106546, 2024a.

Jinhong Wang, Jintai Chen, Danny Chen, and Jian Wu. Large kernel vision mamba unet for medical image segmentation. *arXiv preprint arXiv:2403.07332*, 2024b.

Shuhang Wang, Szu-Yeu Hu, Eugene Cheah, Xiaohong Wang, Jingchao Wang, Lei Chen, Masoud Baikpour, Arinc Ozturk, Qian Li, Shinn-Huey Chou, et al. U-net using stacked dilated convolutions for medical image segmentation. *arXiv preprint arXiv:2004.03466*, 2020.

Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3):415–424, 2022b.

Zhaohu Xing, Tian Ye, Yijun Yang, Guang Liu, and Lei Zhu. SegMamba: Long-range Sequential Modeling Mamba For 3D Medical Image Segmentation . In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume LNCS 15008. Springer Nature Switzerland, October 2024.

Qing Xu and Wenting Duan. An automatic nuclei image segmentation based on multi-scale split-attention u-net. In *MICCAI Workshop on Computational Pathology*, pages 236–245. PMLR, 2021.

Feiniu Yuan, Zhengxiao Zhang, and Zhijun Fang. An effective cnn and transformer complementary network for medical image segmentation. *Pattern Recognition*, 136:109228, 2023.

Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, pages 3–11. Springer, 2018.