Expert Systems With Applications 261 (2025) 125518

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Selective and multi-scale fusion Mamba for medical image segmentation

Guangju Li [a,b], Qinghua Huang [b,*], Wei Wang [c], Longzhong Liu [d]

[a] School of Computer Science, Northwestern Polytechnical University, Xian, 710129, China
[b] School of Artificial Intelligence, OPtics and ElectroNics (iOPEN), Northwestern Polytechnical University, Xian, 710072, China
[c] Department of Medical Ultrasonics, Institute of Diagnostic and Interventional Ultrasound, Ultrasomics Artificial Intelligence X-Lab, The First Affiliated Hospital of Sun Yat-Sen University, Guangzhou 510060, China
[d] Department of Ultrasound, State Key Laboratory of Oncology in South China, Guangdong Provincial Clinical Research Center for Cancer, Sun Yat-sen University Cancer Center, Guangzhou 510060, China

ARTICLE INFO

Keywords:
Medical image segmentation
Mamba
U-shape network
Multi-scale fusion

ABSTRACT

Given the high variability in the morphology and size of lesion areas in medical images, accurate medical image segmentation requires both precise positioning of global contours and careful processing of local boundaries. This emphasizes the importance of fusing multi-scale local and global features. However, existing CNN and Transformer-based models are often limited by high parameter counts and complex calculations, making it difficult to efficiently integrate these features. To address this challenge, we proposed two innovative optimization architectures: Selective Fusion Mamba (SF-Mamba) and Multi-Scale Fusion Mamba (MF-Mamba). SF-Mamba can flexibly and dynamically adjust the fusion strategy of local and global features according to the characteristics of the lesions, effectively handling segmentation tasks with variable morphology. MF-Mamba enhances the model's segmentation ability for lesions of different sizes by capturing global information across scales. Based on these two structures, we constructed a lightweight SMM-UNet model, which not only significantly reduces the computational burden (with only 0.038M parameters) but also demonstrates excellent generalization ability and can efficiently adapt to various types of medical images. Extensive tests on the ISIC2017, ISIC2018, and BUSI public datasets show that SMM-UNet achieves excellent segmentation performance with an extremely low parameter cost.

1. Introduction

In the traditional process of medical image segmentation, doctors usually need to spend a lot of time on tedious manual outlining and measurement. However, with the development of segmentation models, doctors' workloads have begun to decrease, saving medical resources (Hatamizadeh et al., 2022; Ibtehaz & Rahman, 2020; Srivastava et al., 2023). Convolutional neural network (CNN) models, such as UNet (Ronneberger, Fischer, & Brox, 2015), have played a central role in the early stages of medical image segmentation. However, due to the inherent locality of convolution operations, CNN-based models struggle to capture global features, which limits their potential to achieve higher segmentation accuracy. To address this issue, researchers have made several improvements to UNet, resulting in more complex model structures like UNet++ (Zhou, Rahman Siddiquee, Tajbakhsh, & Liang, 2018), Attention-UNet (Oktay et al., 2018), LM-Net (Lu, She, Wang, & Huang, 2024), and NAG-Net (Huang et al., 2023). These enhancements primarily focus on strategies such as expanding the convolution kernel's receptive field, integrating the strengths of multiple models, and introducing attention mechanisms, all aimed at optimizing the model's segmentation performance. Despite these advancements, these methods still fall short in effectively capturing global feature information (Liao et al., 2024).

Due to the Transformer's distinctive self-attention mechanism effectively captures long-range dependencies and has demonstrated strong competitiveness in classification tasks (Halder et al., 2024), offering a new direction for medical image segmentation. TransUNet (Chen et al., 2021) is the first model to employ the Transformer for medical image segmentation, achieving better results compared to CNN-based models. However, the quadratic computational complexity of the Transformer and the large number of parameters in Transformer-based models often limit their practical application in medical tasks (Ruan, Xie, Gao, Liu, & Fu, 2023; Valanarasu & Patel, 2022). For instance, TransUNet has 105 M parameters, significantly more than UNet's 4.096 M. Although researchers have developed models with fewer parameters, such as Swin-UNet (Cao et al., 2022), MedT (Valanarasu, Oza, Hacihaliloglu,

* Corresponding author.
E-mail addresses: guangjuli@mail.nwpu.edu.cn (G. Li), qhhuang@nwpu.edu.cn (Q. Huang), wangw73@mail.sysu.edu.cn (W. Wang), liulzh@sysucc.org.cn (L. Liu).

https://doi.org/10.1016/j.eswa.2024.125518
Received 8 July 2024; Received in revised form 22 August 2024; Accepted 5 October 2024
Available online 11 October 2024

& Patel, 2021), AFTer-UNet (Yan et al., 2022), and HiFormer (Heidari et al., 2023), the parameter count remains considerably higher than that of UNet.

Recently, state-space models like Mamba (Gu & Dao, 2023) have paved a new path for developing lightweight models by effectively reducing computational complexity to a linear level while capturing global features. Researchers have applied these models to the field of computer vision, proposing classification models such as ViM (Zhu, Liao et al., 2024) and VMamba (Liu et al., 2024). These models have achieved more accurate predictions than Vision Transformers (Dosovitskiy et al., 2020) with fewer parameters. Inspired by these advancements, a series of medical image segmentation models based on Mamba, such as VM-UNet (Ruan & Xiang, 2024) and VM-UNet V2 (Zhang, Yu, Gu, Lin, & Tao, 2024), have been proposed. Despite their technical innovations, these models still have significantly higher parameter counts compared to traditional UNet models. To further reduce complexity and parameter, researchers proposed the lightweight model UltraLight VM-UNet (Wu, Liu, Liang, & Chang, 2024). However, experiments have shown that it performs poorly on medical datasets with a small number of samples. This is because the lesion areas in medical images can vary greatly in shape and size, requiring models to account for both local and global features to achieve accurate segmentation results (Heidari et al., 2023).

To address the aforementioned issues, this paper proposes a model called SMM-UNet. The core of this model lies in two innovative Mamba structures: Selective Fusion Mamba (SF-Mamba) and Multi-scale Fusion Mamba (MF-Mamba). Unlike existing models that merely combine local and global features, our designed SF-Mamba layer dynamically adjusts the weight distribution of local and global features based on their impact on the model's segmentation accuracy. This mechanism allows the model to flexibly select feature information that benefits the segmentation task, thereby optimizing segmentation performance. In the MF-Mamba module, we use the Mamba structure to fuse global information from feature maps of different scales. This strategy aims to address the challenge of significant morphological changes in lesion areas and improve the model's segmentation robustness for complex morphological lesions. Experiments on three public datasets demonstrate that our SMM-UNet achieves the most accurate segmentation results with fewer model parameters. The main contributions of this paper can be summarized as follows:

(1) We designed a new lightweight model, SMM-UNet (0.038M), which is one of the medical image segmentation models with the fewest known parameters. This model provides a new direction for the design of lightweight models.
(2) Compared to the Mamba, the SF-Mamba we designed dynamically assigns weights to local and global features, thereby flexibly selecting feature information that is more beneficial to segmentation results.
(3) The MF-Mamba module we designed can better capture the features of lesions with varying shapes and sizes by fusing global feature information at different scales, further improving the model's performance.

2. Related work

2.1. CNN-based models

This type of model uses convolutional kernels to learn the features of specific areas in an image (Rao, Rajitha, Srinivasu, Ijaz, & Woźniak, 2024). By stacking multiple convolutional layers, the model can gradually abstract deep and highly generalized features from the image. Since the advent of the UNet (Ronneberger et al., 2015) model, CNN-based architectures have held an important position in the field of medical image segmentation. The success of UNet is mainly attributed to its unique U-shaped network structure, which has provided a valuable reference for subsequent model designs. Building on UNet, researchers have continued to explore and innovate, successfully constructing a series of improved CNN models such as SKUNet (Byra et al., 2020), ESKNet (Chen, Zhou et al., 2024) and MSRF-Net (Srivastava et al., 2021). These improvements were achieved by introducing modules such as selection kernel convolution and multi-scale fusion. While the performance of these models has improved, the development of CNN-based models has been hindered by the limitations of convolution operations (Heidari et al., 2023; Yuan, Zhang, & Fang, 2023).

2.2. Transformer-based models

The unique self-attention mechanism of Transformer (Vaswani et al., 2017) models allows them to effectively capture global dependencies in input sequences. This feature enables Transformers to overcome the limitations of traditional convolution operations and exhibit strong capabilities in image segmentation tasks. In TransUNet (Chen et al., 2021), researchers combined Transformers with the UNet model to achieve better segmentation results. In Swin-UNet (Cao et al., 2022), the authors replaced the convolutional blocks in UNet with Swin-Transformers (Liu et al., 2021), proposing a segmentation model based entirely on Transformers. However, Transformer-based models typically have high computational complexity and a large number of parameters, which hinders their application in practical medical scenarios.

2.3. Mamba-based models

Unlike the quadratic computational complexity of Transformers, Mamba (Gu & Dao, 2023) effectively captures global feature information while maintaining linear complexity (Li et al., 2024; Zhu, Liao et al., 2024). Leveraging this advantage, researchers proposed the ViM (Zhu, Liao et al., 2024) and VMamba (Liu et al., 2024) structures, which reduce model complexity while maintaining competitiveness comparable to Vision Transformers (ViT) (Dosovitskiy et al., 2020), thereby opening up a new path for medical image segmentation research. Building on this foundation, researchers successively introduced the VM-UNet (Ruan & Xiang, 2024) and VM-UNet-V2 (Zhang et al., 2024) models, replacing the commonly used Transformer with Mamba to achieve global feature learning. These models demonstrated that Mamba is suitable for medical image segmentation tasks; however, their parameter counts remain significantly higher than that of the UNet model. Subsequently, a lightweight model called LightM-UNet (Liao et al., 2024) was proposed, containing only 1M parameters. This model, built entirely on the Mamba architecture, further enhances global feature learning through the Residual Vision Mamba layer. The most recent UltraLight VM-UNet (Wu et al., 2024) model drastically reduces the parameter count to 0.038M by processing features in parallel. However, these lightweight models focus primarily on global feature learning and do not fully integrate local features. In medical image segmentation, global features are crucial for locating and contouring lesions, but local features are equally important for refining boundary segmentation. Only by integrating both global and local features can a model achieve more accurate lesion segmentation. To address this, we proposed the SMM-UNet model, which dynamically fuses global and local information through the SF-Mamba module, enabling precise boundary segmentation while accurately delineating lesion areas. Additionally, the MF-Mamba module facilitates the fusion of multi-scale features, effectively handling variations in lesion size. With just 0.038M parameters, SMM-UNet is one of the lightest models currently available and has demonstrated excellent competitiveness across three public datasets.

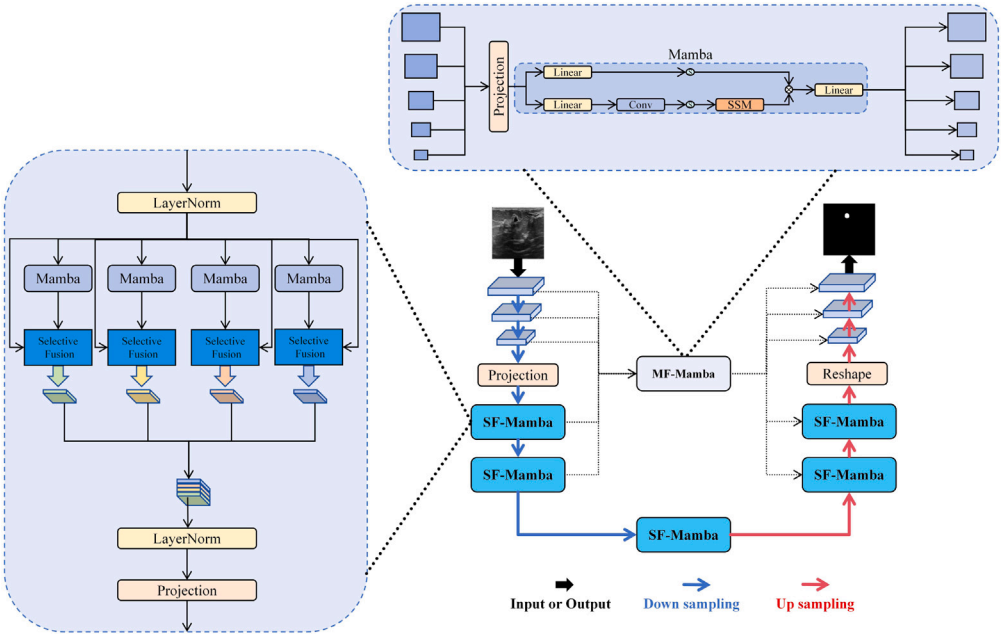

Fig. 1. The SMM-UNet consists of the Selective Fusion Mamba layer (SF-Mamba) and the Multi-scale Fusion Mamba block (MF-Mamba).

3. Method

3.1. Overall architecture

Given the complexity and variability of lesion areas, learning global features is crucial for understanding the shape and contour information of lesions, while local features focus on subtle details, helping to achieve more accurate lesion boundary segmentation. Additionally, fusing feature maps of different scales addresses the uncertainty of lesion size. To this end, we designed the SMM-UNet model (as shown in Fig. 1), which follows the U-shaped network architecture.

In the first three stages, the model uses convolutional layers to learn shallow local features, capturing the edge information of lesions. Subsequently, in the deep feature learning stage, we designed the SF-Mamba layer. This layer captures long-range dependencies of lesions through the Mamba learning mechanism to obtain contour features, while also considering previously extracted local features through a selective fusion strategy.

In the skip connection part of the model, we designed the MF-Mamba module. This module effectively captures lesion areas of different sizes by fusing the global features of feature maps at different scales. As a result, the SMM-UNet model achieves more accurate segmentation in complex and variable lesion areas.

3.2. Selective fusion Mamba

To account for both the fine details of local boundaries and the macro features of global contours, we designed the SF-Mamba layer, as shown in Fig. 2. In this architecture, the feature map is divided into four groups and input into the Mamba branch. The Mamba layer captures global features through its unique learning mechanism.

The specific structure of the Mamba layer, illustrated in Fig. 3, comprises two branches. The first branch projects the input features linearly (Linear), processes the sequence data through convolution operations (Conv), and applies SiLU activation functions to enhance feature representation. These processed features are then fed into the Selective State Space Model (SSM) within Mamba. The SSM enables global modeling of the sequence, allowing effective information propagation across different time steps and capturing the sequence's global characteristics. Additionally, the SSM focuses on the most important parts of the sequence while ignoring less relevant information, thus improving both the efficiency and effectiveness of feature extraction.

$$\begin{aligned} F_1 &= Conv(Linear(F_{in})) \\ F_1 &= SSM(SiLU(F_1)) \end{aligned} \tag{1}$$

The second branch also linearly projects the input features and processes them with an SiLU activation function. Its primary role is to enhance the feature representation of F_{in} preparation for the subsequent merging stage.

$$F_2 = SiLU(Linear(F_{in})) \tag{2}$$

Finally, the output features of the two branches undergo the Hadamard product operation. This merging process helps capture multi-level information, thereby improving the expressiveness of the model.

$$F_{out} = F_1 \otimes F_2 \tag{3}$$

Next, we designed a selective fusion module that assigns weights to local and global features based on their importance, ensuring that the model comprehensively considers both local details and global information when segmenting the lesion area. Specifically (Fig. 4), the features are first dimensionally transformed to obtain the local feature $F_1 \in \mathbb{R}^{H\times W\times C/4}$ and the global feature $F_2 \in \mathbb{R}^{H\times W\times C/4}$, which are then concatenated to obtain the feature $F_c \in \mathbb{R}^{H\times W\times C/2}$. Subsequently, F_c is further processed by fully connected layers to obtain the weight $\beta \in \mathbb{R}^{1\times 1\times C/4}$ for the local feature and the weight $1-\beta \in \mathbb{R}^{1\times 1\times C/4}$ for the global feature.

$$\begin{aligned} F_c &= F_1 \textcircled{c} F_2 \\ \beta, 1-\beta &= Softmax(ReLU(W(F_c))) \end{aligned} \tag{4}$$

These weights are multiplied by F_1 and F_2 respectively and then fused by element-wise addition.

$$\begin{aligned} F_1 &= \beta \otimes F_1 \\ F_2 &= (1-\beta) \otimes F_2 \\ F &= F_1 \oplus F_2 \end{aligned} \tag{5}$$

It is worth mentioning that we verified in subsequent ablation experiments that the design of the SF-Mamba layer did not lead to a significant increase in the number of model parameters, proving its efficiency and practicality.

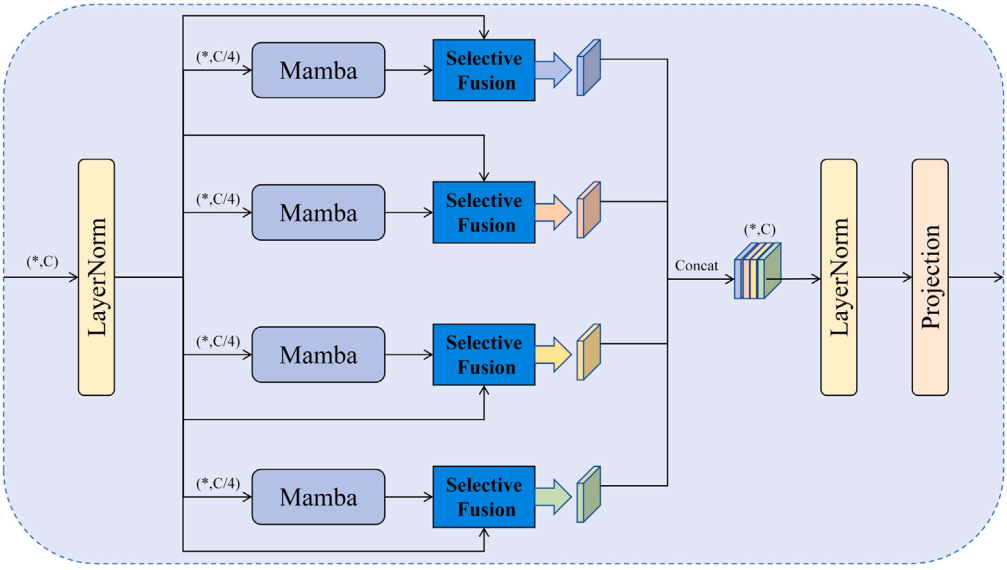

Fig. 2. SF-Mamba: The features first pass through Mamba to capture the global feature information, and then pass through the selective fusion block to fuse local and global features.

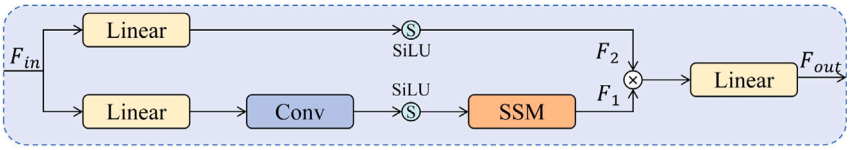

Fig. 3. Mamba: It consists of linear projection, convolution operations, activation functions, and SSM, with SSM being the core component for capturing global features.

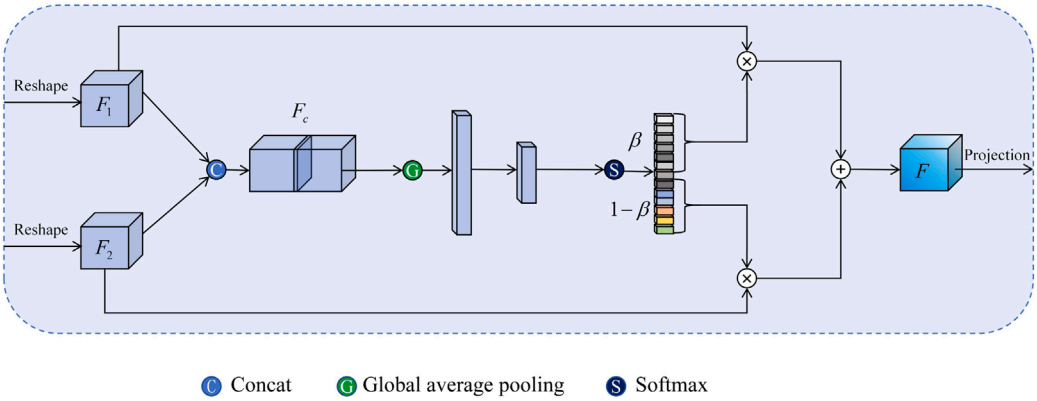

Fig. 4. Selective Fusion: After connecting the local and global features, global average pooling, a fully connected layer, and Softmax processing are performed sequentially to obtain the feature vector. This vector is then divided into two parts, which are used as the weights for the local and global features, respectively.

3.3. Multi-scale fusion Mamba

To address the problem of uncertain lesion area sizes, we designed the MF-Mamba module, as shown in Fig. 5. This module not only integrates features of various scales but also establishes long-range dependencies between feature maps of different scales through the Mamba layer.

Initially, the feature maps from the first five stages (i.e., F1 to F5) are upsampled or downsampled and then concatenated to generate the feature map F. Next, this feature map F is linearly projected and input into Mamba to learn global feature information at different scales. Subsequently, feature fusion occurs along the channel direction through the MLP layer to effectively transfer spatial details and semantic information between the low-resolution and high-resolution feature maps. Consequently, information transfer between feature maps of different scales is achieved, considering the impact of features of different scales on the model's segmentation performance within the global context information framework. Finally, the fused feature map is restored to its original resolution by upsampling or downsampling and then passed to the decoder, enabling the decoder to segment lesions of different sizes more accurately.

$$Loss = 0.5 * Loss_{Dice} + Loss_{BCE} \tag{6}$$

The loss functions utilized in the model training process include Dice loss and Binary Cross-Entropy (BCE) loss. These two loss functions are widely used in medical image segmentation tasks due to their effectiveness in handling class imbalance and capturing fine details in segmentation boundaries. The Dice loss is particularly advantageous for improving the overlap between predicted and actual segmentation

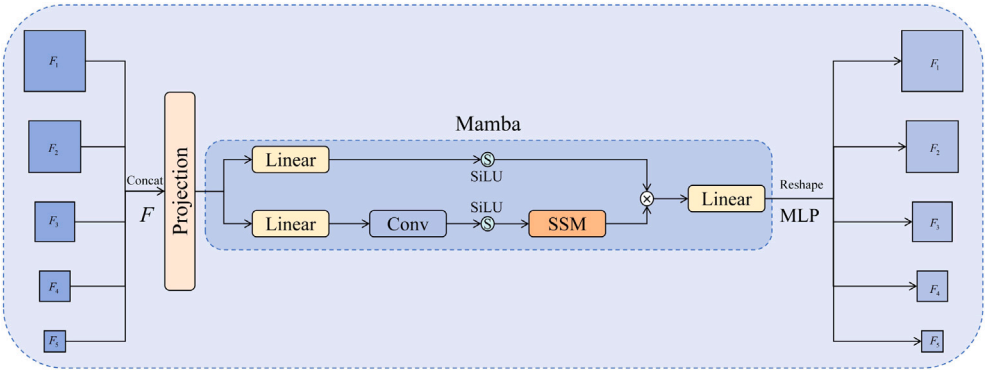

Fig. 5. MF-Mamba: Mamba is employed for learning multi-scale global features, while MLP is utilized for feature fusion.

regions, while the BCE loss provides a robust measure for pixel-wise classification.

4. Experiments

4.1. Implementation details and datasets

The SGD optimizer was used for training with the following parameters: an initial learning rate of 1e–3, momentum of 0.9, weight decay of 1e–4, a minimum learning rate of 1e–5, and a batch size of 8. The model was trained for 400 epochs on an Nvidia A30 GPU with 24 GB of memory. All input images were uniformly resized to 256 × 256 pixels and underwent random cropping and rotation for data augmentation.

To evaluate the model's performance, we used three public datasets: ISIC 2017 (Codella et al., 2018), ISIC 2018 (Codella et al., 2019), and BUSI (Al-Dhabyani, Gomaa, Khaled, & Fahmy, 2020). The ISIC 2017 and ISIC 2018 datasets are for skin disease segmentation, while the BUSI dataset consists of breast ultrasound images. To ensure a fair comparison with existing models, we followed the VM-UNET-V2 (Zhang et al., 2024) strategy for the ISIC 2017 and ISIC 2018 datasets, maintaining a 7:3 ratio between the training and test sets. For the BUSI dataset, we used the UNeXt (Valanarasu & Patel, 2022) strategy, dividing the dataset into 80% training and 20% testing sets. Since UNet, UNeXt, and UltraLight VM-UNet are highly reproducible, these three models were trained and tested using the exact same data splits as our SMM-UNet. For other models, we used the data splits provided in the corresponding publications.

To assess model performance, we utilized two key evaluation metrics: Intersection over Union (IoU) and Dice Similarity Coefficient (DSC). IoU measures the overlap between the predicted segmentation and the ground truth, providing an indication of how well the model identifies the correct regions. DSC, on the other hand, evaluates the similarity between the predicted and actual segmentations, emphasizing the balance between precision and recall. Both metrics are commonly used to quantify the accuracy and effectiveness of segmentation models in medical imaging tasks.

4.2. Comparison with state-of-the-arts (SOTAs)

Table 1 provides a detailed comparison of the segmentation performance of our SMM-UNet model against SOTAs on the ISIC 2017 and 2018 datasets. Although existing models have demonstrated good segmentation results on both datasets, our SMM-UNet model remains significantly competitive. Compared to the classic UNet (4.096M parameters), the number of parameters in SMM-UNet has been drastically reduced to 0.038M, while achieving improvements in DSC and IoU by 2.19%–2.47% and 3.48%–3.91%, respectively, on the two datasets. When compared to VM-UNet (27.42M) and VM-UNet V2 (22.77M), which are also based on the Mamba architecture, SMM-UNet demonstrates significant advantages in both parameter count and segmentation performance. Even in comparison with the latest lightweight model, UltraLight VM-UNet, SMM-UNet shows a 0.41%–0.68% and 0.68%–1.14% improvement in DSC and IoU, respectively, across both datasets, while maintaining a similar number of parameters. Furthermore, the visual comparison of segmentation results in Fig. 6 clearly illustrates the completeness of lesion area segmentation achieved by SMM-UNet. In contrast, other models exhibit varying degrees of omission or incompleteness in segmenting the lesion area.

To comprehensively verify the performance of our SMM-UNet model, we conducted additional experiments on the BUSI dataset. Compared to the ISIC 2017 and 2018 datasets, the BUSI dataset contains higher noise levels and fewer samples, posing a greater challenge to segmentation accuracy. As shown in Table 2, the segmentation accuracy of existing models on the BUSI dataset is generally average. Thus, SMM-UNet demonstrates significant competitiveness. Despite having fewer parameters, SMM-UNet achieves substantial improvements of 1.49%–6.54% in DSC and 2.13%–7.94% in IoU. Additionally, the visual comparison in Fig. 7 indicates that although SMM-UNet occasionally exhibits incomplete segmentation, its results are still more accurate and complete than those of other models. This further confirms that the SMM-UNet model maintains high segmentation performance even when handling medical datasets with high noise content and a limited number of samples.

4.3. Ablation study

We investigated the effects of two core modules and the number of model channels on segmentation performance using the BUSI dataset. Our experiments focused on three main aspects: (1) the placement of selective fusion within the SF-Mamba layer; (2) different methods of channel fusion in the MF-Mamba block; and (3) the number of channels in the entire model. These experiments allowed us to comprehensively understand the impact of various factors on model performance.

Selective Fusion Mamba: From the data in Table 3, we observe that the selective fusion strategy in the SF-Mamba layer significantly improves segmentation accuracy compared to the baseline model based on the original Mamba, while the increase in the number of parameters is almost negligible. Specifically, the SF-Mamba-I strategy performs selective fusion after the four Mamba branches in each layer, allowing the model to consider both local and global feature information. The SF-Mamba-II strategy, on the other hand, selectively fuses the feature maps of the four Mamba branches after they are connected. Although this method also improves segmentation performance, the increase in the number of channels after the feature maps are connected results in a relative increase in the number of parameters. Therefore, when constructing our SMM-UNet model, we chose the SF-Mamba-I method. In addition, it is important to emphasize that Selective Fusion Mamba

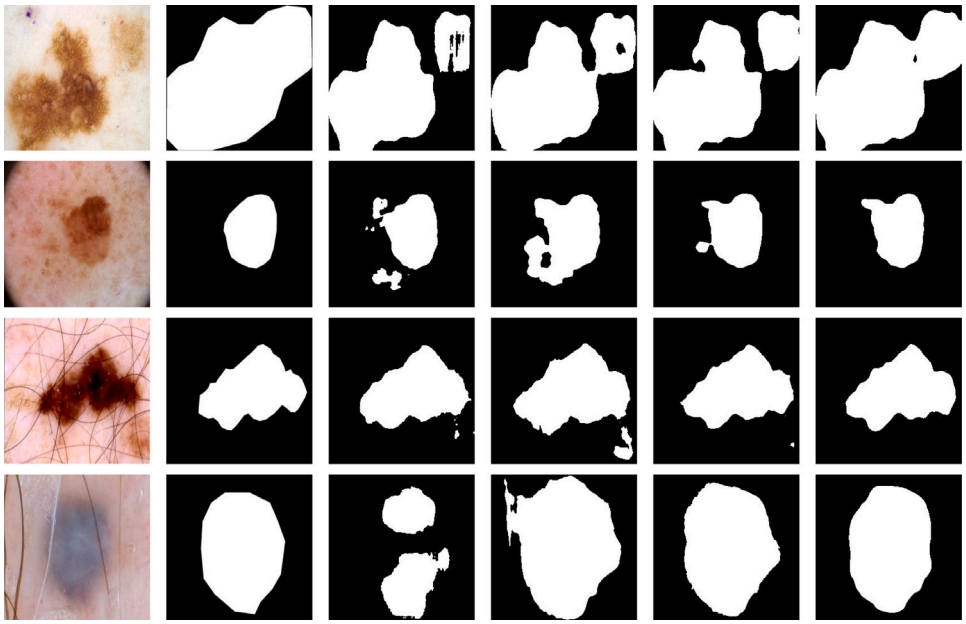

Fig. 6. Segmentation results of UNet, UNeXt, UltraLight VM-UNet and SMM-UNet (ours) on ISIC2017 and 2018 datasets.

Table 1
Segmentation results are compared with the SOTAs on the ISIC 2017 and 2018 datasets.

Datasets	Networks	IoU↑	DSC↑	Param (M)↓
ISIC 2017 (Codella et al., 2018)	UNet (Ronneberger et al., 2015)	81.21	89.38	4.096
	UNet-V2 (Peng, Sonka, & Chen, 2023)	82.18	90.22	25.15
	TransFuse (Zhang, Liu, & Hu, 2021)	79.21	88.40	26.30
	MALUNet (Ruan, Xiang, Xie, Liu, & Fu, 2022)	78.78	88.13	0.175
	UNeXt (Valanarasu & Patel, 2022)	83.31	90.71	1.472
	EGE-UNet (Ruan et al., 2023)	79.81	88.77	0.053
	VM-UNet (2024) (Ruan & Xiang, 2024)	80.23	89.03	27.42
	VM-UNet V2 (2024) (Zhang et al., 2024)	82.34	90.31	22.77
	UltraLight VM-UNet (2024) (Wu et al., 2024)	83.55	90.89	**0.038**
	SMM-UNet (ours)	**84.69**	**91.57**	**0.038**
ISIC 2018 (Codella et al., 2019)	UNet (Ronneberger et al., 2015)	78.39	87.59	4.096
	UNet-V2 (Peng et al., 2023)	80.71	89.32	25.15
	TransFuse (Zhang et al., 2021)	79.21	88.40	26.30
	MALUNet (Ruan et al., 2022)	80.25	89.04	0.175
	UNeXt (Valanarasu & Patel, 2022)	81.48	89.59	1.472
	EGE-UNet (Ruan et al., 2023)	80.94	89.46	0.053
	VM-UNet (2024) (Ruan & Xiang, 2024)	81.35	89.71	27.42
	VM-UNet V2 (2024) (Zhang et al., 2024)	81.37	89.73	22.77
	UltraLight VM-UNet (2024) (Wu et al., 2024)	81.62	89.65	**0.038**
	SMM-UNet (ours)	**82.30**	**90.06**	**0.038**

Table 2
Segmentation results are compared with the SOTAs on the BUSI datasets.

Datasets	Networks	IoU↑	DSC↑	Param (M)↓
BUSI (Codella et al., 2019)	UNet (Ronneberger et al., 2015)	62.40	76.33	4.096
	UCTransNet (Wang, Cao, Wang, & Zaiane, 2022)	60.86	74.58	67.23
	UNeXt (Valanarasu & Patel, 2022)	66.16	79.03	1.472
	MixUNet-L (2024) (Chen, Zhang et al., 2024)	63.45	76.51	5.95
	MSS-UNet (2024) (Zhu, Tian, Chen, Chen and Chen, 2024)	66.67	79.63	0.33
	UltraLight VM-UNet (2024) (Wu et al., 2024)	65.61	78.68	**0.038**
	SMM-UNet (ours)	**68.80**	**81.12**	**0.038**

differs from traditional attention mechanisms. Traditional attention mechanisms typically assign weights to a set of feature maps, guiding the model to focus on the features most relevant to the segmentation results. In contrast, the core objective of Selective Fusion Mamba is to select between local and global feature sets and assign higher weights to the features that best contribute to segmentation outcomes, leveraging their complementary nature. To validate the effectiveness of Selective Fusion Mamba, we compared it with the current state-of-the-art attention mechanisms. As shown in Table 3, SF-Mamba demonstrates greater competitiveness in terms of both accuracy and parameters.

Multi-scale Fusion Mamba: Based on the framework integrating SF-Mamba-I, we explored the impact of MF-Mamba block on model performance. As shown in Table 4, the segmentation accuracy of the model improves further after integrating MF-Mamba. This improvement is

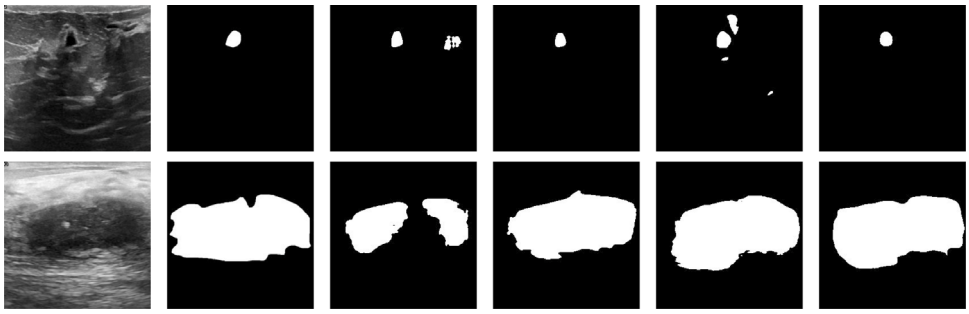

Fig. 7. Segmentation results of UNet, UNeXt, UltraLight VM-UNet and SMM-UNet (ours) on the BUSI datasets.

Table 3
The impact of selective fusion positions in SF-Mamba layers on model segmentation performance.

Methods	IoU ↑	DSC ↑	Param (M)↓
Baseline	63.88	77.50	**0.021**
+EMA (Ouyang et al., 2023)	65.20	78.38	0.023
+ELA (Xu & Wan, 2024)	64.82	77.83	**0.021**
+SF-Mamba-I (ours)	**66.06**	**79.23**	0.022
+SF-Mamba-II (ours)	65.23	78.59	0.023

Table 4
The impact of the channel fusion method in the MF-Mamba block on the model segmentation performance.

Methods	IoU ↑	DSC ↑	Param (M)↓
Baseline+SF-Mamba+MF-Mamba-I (MLP)	**68.80**	**81.12**	0.038
Baseline+SF-Mamba+MF-Mamba-II (1×1 *Conv*)	68.18	80.43	0.038

Table 5
The impact of the number of channels at each stage of the model on the segmentation performance of the model.

Methods	IoU ↑	DSC ↑	Param (M)↓
4-8-16-24-32-48	66.48	79.51	**0.017**
8-16-24-32-48-64	**68.80**	**81.12**	0.038
16-24-32-48-64-128	67.55	80.20	0.083

primarily due to MF-Mamba's ability to efficiently fuse global information from feature maps of different scales. Since these feature maps have receptive fields of varying sizes, their fusion allows the model to capture lesion areas of different sizes, providing the decoder with richer and more comprehensive lesion location information. Additionally, after learning global features through Mamba, we designed two channel fusion methods. The first method uses MLP to perform channel transformation on the linearly processed feature map, referred to as MF-Mamba-I. The second method restores the linearly processed feature map to its original dimension and then reduces the number of channels through 1×1 convolution, referred to as MF-Mamba-II. Both methods have nearly the same number of parameters, but because MLP can fully utilize information from all positions in the feature map, it achieves better segmentation results compared to the local nature of convolution operations.

The Overall Number of Channels in the Model: Generally speaking, increasing the number of channels allows the model to capture more features, which can improve segmentation results. However, having too many channels may cause the model to overfit, ultimately reducing the accuracy of the segmentation results. To address this, we tested three different combinations of channel numbers. As shown in Table 5, as the number of channels increases, the model's parameters also increase, and the segmentation accuracy improves. However, if the number of channels continues to increase, it eventually leads to negative effects.

5. Conclusion

The SMM-UNet medical image segmentation model proposed in this paper has demonstrated significant advantages in the field. Its core lies in the synergy of the two modules, SF-Mamba and MF-Mamba. SF-Mamba achieves accurate positioning of the lesion area and detailed segmentation of the boundary by flexibly adjusting the weights of local and global features. MF-Mamba effectively addresses the challenge of variable lesion sizes by integrating global information across scales. Experimental results show that SMM-UNet not only significantly outperforms traditional CNN and Transformer models in terms of parameter efficiency but also achieves notable improvements in segmentation accuracy. Compared to other Mamba-like models, SMM-UNet demonstrates superior segmentation performance due to its flexible processing of local and global information. Notably, even on the BUSI ultrasound dataset, which has fewer samples, SMM-UNet has shown strong competitiveness, although its performance still has room for further optimization. In the future, we will focus on further optimizing its segmentation performance on small-sample medical image datasets.

CRediT authorship contribution statement

Guangju Li: Conceptualization, Methodology, Data curation, Formal Analysis, Visualization, Writing – original draft. **Qinghua Huang:** Supervision, Writing – review & editing. **Wei Wang:** Investigation. **Longzhong Liu:** Validation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was partially supported by the National Natural Science Foundation of China under Grants 62071382, 12326609, 82272020, 82371983 and 82030047, as well as by the Guangdong Basic and Applied Basic Research Foundation under Grant 2023A1515011357.

Data availability

Data will be made available on request.

References

Al-Dhabyani, W., Gomaa, M., Khaled, H., & Fahmy, A. (2020). Dataset of breast ultrasound images. *Data in Brief*, *28*, Article 104863.

Byra, M., Jarosik, P., Szubert, A., Galperin, M., Ojeda-Fournier, H., Olson, L., et al. (2020). Breast mass segmentation in ultrasound with selective kernel U-Net convolutional neural network. *Biomedical Signal Processing and Control*, *61*, Article 102027.

Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., et al. (2022). Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision* (pp. 205–218). Springer.

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., et al. (2021). Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306.

Chen, Y., Zhang, X., He, Y., Peng, L., Pu, L., & Sun, F. (2024). MixUNet: A lightweight medical image segmentation network capturing multidimensional semantic information. *Biomedical Signal Processing and Control*, *96*, Article 106513.

Chen, G., Zhou, L., Zhang, J., Yin, X., Cui, L., & Dai, Y. (2024). ESKNet: An enhanced adaptive selection kernel convolution for ultrasound breast tumors segmentation. *Expert Systems with Applications*, *246*, Article 123265.

Codella, N. C., Gutman, D., Celebi, M. E., Helba, B., Marchetti, M. A., Dusza, S. W., et al. (2018). 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). IEEE.

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

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Gu, A., & Dao, T. (2023). Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752.

Halder, A., Gharami, S., Sadhu, P., Singh, P. K., Woźniak, M., & Ijaz, M. F. (2024). Implementing vision transformer for classifying 2D biomedical images. *Scientific Reports*, *14*(1), 12567.

Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., et al. (2022). Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 574–584).

Heidari, M., Kazerouni, A., Soltany, M., Azad, R., Aghdam, E. K., Cohen-Adad, J., et al. (2023). Hiformer: Hierarchical multi-scale representations using transformers for medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 6202–6212).

Huang, Q., Zhao, L., Ren, G., Wang, X., Liu, C., & Wang, W. (2023). NAG-Net: Nested attention-guided learning for segmentation of carotid lumen-intima interface and media-adventitia interface. *Computers in Biology and Medicine*, *156*, Article 106718.

Ibtehaz, N., & Rahman, M. S. (2020). MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation. *Neural Networks*, *121*, 74–87.

Li, K., Li, X., Wang, Y., He, Y., Wang, Y., Wang, L., et al. (2024). Videomamba: State space model for efficient video understanding. arXiv preprint arXiv:2403.06977.

Liao, W., Zhu, Y., Wang, X., Pan, C., Wang, Y., & Ma, L. (2024). Lightm-unet: Mamba assists in lightweight unet for medical image segmentation. arXiv preprint arXiv:2403.05246.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., et al. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 10012–10022).

Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., et al. (2024). Vmamba: Visual state space model. arXiv preprint arXiv:2401.10166.

Lu, Z., She, C., Wang, W., & Huang, Q. (2024). LM-Net: A light-weight and multi-scale network for medical image segmentation. *Computers in Biology and Medicine*, *168*, Article 107717.

Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., et al. (2018). Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999.

Ouyang, D., He, S., Zhang, G., Luo, M., Guo, H., Zhan, J., et al. (2023). Efficient multi-scale attention module with cross-spatial learning. In *ICASSP 2023-2023 IEEE international conference on acoustics, speech and signal processing* ICASSP, (pp. 1–5). IEEE.

Peng, Y., Sonka, M., & Chen, D. Z. (2023). U-Net v2: Rethinking the skip connections of U-net for medical image segmentation. arXiv preprint arXiv:2311.17791.

Rao, G. E., Rajitha, B., Srinivasu, P. N., Ijaz, M. F., & Woźniak, M. (2024). Hybrid framework for respiratory lung diseases detection based on classical CNN and quantum classifiers from chest X-rays. *Biomedical Signal Processing and Control*, *88*, Article 105567.

Ronneberger, O., Fischer, P., & Brox, T. (2015). 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.

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

Ruan, J., Xiang, S., Xie, M., Liu, T., & Fu, Y. (2022). 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). IEEE.

Ruan, J., Xie, M., Gao, J., Liu, T., & Fu, Y. (2023). Ege-unet: an efficient group enhanced unet for skin lesion segmentation. In *International conference on medical image computing and computer-assisted intervention* (pp. 481–490). Springer.

Srivastava, A., Jha, D., Chanda, S., Pal, U., Johansen, H. D., Johansen, D., et al. (2021). MSRF-Net: a multi-scale residual fusion network for biomedical image segmentation. *IEEE Journal of Biomedical and Health Informatics*, *26*(5), 2252–2263.

Srivastava, A., Jha, D., Keles, E., Aydogan, B., Abazeed, M., & Bagci, U. (2023). An efficient multi-scale fusion network for 3D organs at risk (OARs) segmentation. In *2023 45th annual international conference of the IEEE engineering in medicine & biology society* EMBC, (pp. 1–4). IEEE.

Valanarasu, J. M. J., Oza, P., Hacihaliloglu, I., & Patel, V. M. (2021). 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* (pp. 36–46). Springer.

Valanarasu, J. M. J., & Patel, V. M. (2022). Unext: Mlp-based rapid medical image segmentation network. In *International conference on medical image computing and computer-assisted intervention* (pp. 23–33). Springer.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Wang, H., Cao, P., Wang, J., & Zaiane, O. R. (2022). Uctransnet: rethinking the skip connections in u-net from a channel-wise perspective with transformer. *Vol. 36*, In *Proceedings of the AAAI conference on artificial intelligence* (pp. 2441–2449).

Wu, R., Liu, Y., Liang, P., & Chang, Q. (2024). Ultralight vm-unet: Parallel vision mamba significantly reduces parameters for skin lesion segmentation. arXiv preprint arXiv:2403.20035.

Xu, W., & Wan, Y. (2024). ELA: Efficient local attention for deep convolutional neural networks. arXiv preprint arXiv:2403.01123.

Yan, X., Tang, H., Sun, S., Ma, H., Kong, D., & Xie, X. (2022). After-unet: Axial fusion transformer unet for medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 3971–3981).

Yuan, F., Zhang, Z., & Fang, Z. (2023). An effective CNN and transformer complementary network for medical image segmentation. *Pattern Recognition*, *136*, Article 109228.

Zhang, Y., Liu, H., & Hu, Q. (2021). Transfuse: Fusing transformers and cnns 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* (pp. 14–24). Springer.

Zhang, M., Yu, Y., Gu, L., Lin, T., & Tao, X. (2024). Vm-unet-v2 rethinking vision mamba unet for medical image segmentation. arXiv preprint arXiv:2403.09157.

Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., & Liang, J. (2018). 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* (pp. 3–11). Springer.

Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., & Wang, X. (2024). Vision mamba: Efficient visual representation learning with bidirectional state space model. arXiv preprint arXiv:2401.09417.

Zhu, W., Tian, J., Chen, M., Chen, L., & Chen, J. (2024). MSS-UNet: A multi-spatial-shift MLP-based UNet for skin lesion segmentation. *Computers in Biology and Medicine*, *168*, Article 107719.