MSCM-Net: Multi-Scale Collaborative Mechanism Driven by the Mamba Framework for Advanced Medical Image Segmentation

Peng,Rui[1*] Liu,Ke[2] Peng,Daihong[3] Yang,Sihong[4]

1College of Computer Science and Technology, Huaibei Normal University, Huaibei, Anhui, 235000, China
2Huaibei Institute of Technology, Huaibei, Anhui, 235000, China
3The Third Middle School of Qianshan City, Anqing, Anhui, 246000, China
4Anhui Qianshan Yongan Co, Anqing, Anhui, 246000, China

Abstract: Nowadays, significant progress has been made in Mamba-based medical image segmentation models, however, there is an imbalance between global and local information extraction in these models, we propose a multi-scale synergistic mechanism aiming at the feature extraction stage to be able to extract global and local information effectively. We use a hierarchical receptive field attention strategy in the feature extraction stage to extract global and local information through different receptive fields. In addition, we introduced spatial and channel attention to emphasize on the image recovery stage to ensure that global and local information is not lost. We evaluated the method on ISIC2017, ISIC2018 and Synapse.The Dice similarity coefficient and intersection over union (IoU) metrics on ISIC17, ISIC18 are 89.76 and 81.41, 90.25 and 82.24, respectively. Our model achieves state-of-the-art(sota) results on the ISIC dataset compared to the Mamba approach. On Synapse the Dice and Average Hausdorff distance (HD95) are 80.17 and 14.21. The experimental results show that the proposed method has good generalization ability and robustness, which provides important support for clinical diagnosis and treatment.

Keywords: Mamba; Medical image segmentation; Multi-scale

DOI: 10.62639/sspjiss28.20250203

1. Introduction

In recent years, Convolutional Neural Network (CNN) and Transformer applications have been widely used in medical image segmentation tasks. UNet[1], as a representative of CNN-based models, is well known for its simple and scalable structure, and many of the subsequent improvements are based on this U-shaped architecture. But it lacks efficiency in capturing global context information. Subsequently, in order to solve this problem, researchers applied the Transformer in the field of NLP to the field of medical image segmentation[2], i.e., TransUnet[3], which uses CNN to extract the underlying features of an image, and then uses the Transformer to capture the global contextual information. In order to further utilize CNN and Transformer, Transfuse[4] architecture was born. Unlike TransUnet, Transfuse is a parallel architecture based on CNN and Transformer, which improves the efficiency of the model and alleviates the problems of gradient vanishing and feature decreasing. Nowadays, Vmamba[5] has achieved great success in image classification tasks. VM-Unet[6] introduces Mamba to medical image segmentation for the first time, which is a model based entirely on SSM, but we found that this model is prone to lead to an imbalance between global and local information extraction, which restricts the segmentation effectiveness.

(Manuscript NO.: JISS-25-3-62015)

Corresponding Author

Peng,Rui (2001-), Male, Han, Huaibei Normal University, Huaibei City, Anhui Province, 235000, Graduate Student, Research Direction: Medical Image Segmentation.

About the Author

Liu,Ke (2000-), Female, Han nationality, Huaibei Institute of Technology, Huaibei City, Anhui Province, Bachelor's Degree, research direction: Mathematics and Applied Mathematics.

Peng,Daihong (1974-), Male, Han nationality, Anqing City, Anhui Province, B.S., research direction: Mathematics.

In this study, we propose a segmentation model based on a multi-scale collaborative mechanism improved on the VM-Unet framework, and our main contributions are summarized as.

(1) We propose a new method called MSCM-Net to solve the problem of unbalanced extraction of global and local information.

(2) We propose the Hierarchical Receptive Field Attention Module and introduce the SCSA[7] module.

(3) We conducted three different experiments on ISIC17, ISIC18 and Synapse, which were qualitatively analyzed against traditional segmentation methods, and the results show that our model exhibits excellent performance on all these tasks.

2. Related Works

In this section, we briefly review the related work on commonly used medical image segmentation methods. The existing methods are mainly categorized into two groups, CNN and Transformer based models and Mamba based models.

(1) Based on CNN and transformer models

Traditional medical image segmentation methods are mainly divided into CNN-based methods, Transformer-based methods, and the combination of the two CNN-based medical image segmentation is widely used,Ronneberger and others initially proposed the UNet[1] model, which can effectively capture local and detailed features in the image, and has achieved good results in medical image segmentation , due to its simple structure and extensibility, more subsequent improvements have been based on this framework. For example, the subsequent UNet++[8], which is improved based on the UNet framework, compensate for the feature loss of the encoder by adding jump connections.

With the rapid development of Vision Transformer(ViT), many researchers have started to apply it to segmentation of medical images. For example, TransUnet[3] utilized these Transformer layers to capture remote dependencies in feature maps of low-resolution encoders. In later Transfuse[4] parallelized CNN and Transformer to process global and local features respectively. As well as Liu et al. in 2021 proposed Swin-UNet based on Swin Transformer, which is a combination of CNN and Swin Transformer.

(2) Based on mamba models

Recently, state-space modeling (SSM)[10] has attracted much interest from many researchers. Today's SSM not only establishes long-range dependencies, but also presents linear complexity in the input size. Mamba has been able to become an alternative to CNN and Transformer. U-Mamba[11] presents the first model combining SSM and CNN, and also its first application in medical image segmentation tasks. Meanwhile, SegMamba[12] also demonstrated the potential of SSM-based. Subsequently, VM-UNet[13] introduced visual state space (VSS[14]) as a base block to capture a wide range of contextual information and constructed an asymmetric encoder-decoder structure.

3. Method

This section first introduces the general structure of the MSCM-Net network, and then details our proposed Hierarchical Receptive Field Attention Module (HRFA) and Spatial and Channel Synergistic Attention Module (SCSA) [15].

(1) Overall architecture

As shown in Fig 1(a), the general architecture of MSCM-Net is given. Specifically, MSCM-Net consists of patch

embedding layer, encoder, decoder, final projection layer and jump connections.

The Patch Embedding layer divides the input image $x \in R^{H \times W \times C}$ into non-overlapping patches of size 4x4, and then maps the dimensionality of the image to C, C defaults to 96. This process yields the embedded image $x \in R^{\frac{H}{4} \times \frac{W}{4} \times C}$. Finally, we normalize it and feed it into an encoder for feature extraction. The encoder consists of four stages, we aggregate the attention maps of different granularities by HRFA in the third stage, thus aggregating the global and local information to avoid the loss of information due to feature extraction, and applying patch merge operations at the end of each of the overall three stages in order to reduce the height and width of the input features and increase the number of channels at the same time. We use [2, 2, 2, 2] VSS blocks in the four phases with channel counts of [C, 2C, 4C, 8C] in each phase.

Similarly, the decoder is divided into four stages and patch merge operation is applied in all the last three stages of the decoder to increase the height and width of the feature channels. We in the second stage of the decoder we enhance the feature selection and multi-scale context fusion by SCSA, which enhances the ability of the decoder to recover details and generate high quality output. In four stages we use [2,2,2,1] VSS blocks with channel counts of [8C, 4C, 2C, C] for each stage. After the decoder, a Final Projection layer is used to recover the size of the features to match the segmentation target. Specifically, 4 up-sampling by patch expansion is performed to recover the height and width of the features, and then the number of channels is recovered by the projection layer. Finally, on the jump join, we introduce the SE attention mechanism to enhance the extraction of features to fuse with the decoder.

(2) Hierarchical receptive field attention module

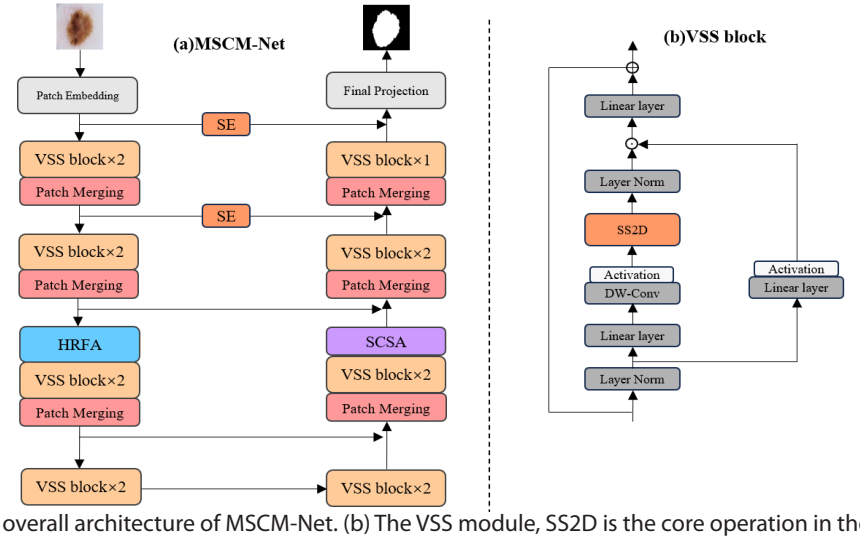

Fig 1: (a) The overall architecture of MSCM-Net. (b) The VSS module, SS2D is the core operation in the VSS block.

As shown in Fig 2, the module is presented in a multi-branch structure, in which the main part of the module first

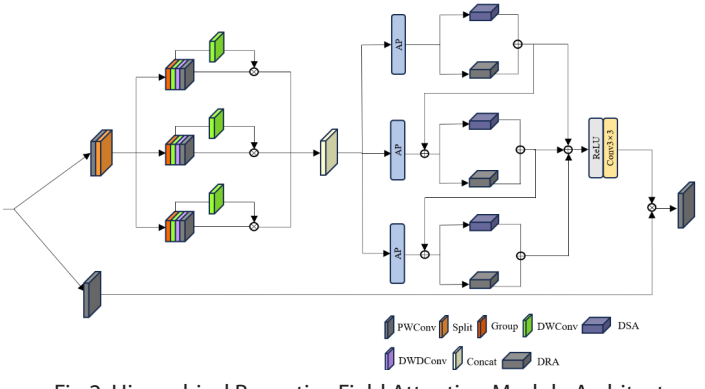

Fig 2: Hierarchical Receptive Field Attention Module Architectur

divides the input features into three parts, and then performs DWConv, DWDConv, and PWConv operations on the respective parts in order to obtain feature maps with different granularities, and fuses the features after the product, and then divides them into three branches, and then performs the parallel Dilated Square Attention(DSA) and Dilated Rectangle Attention(DRA), respectively (as shown in Fig 3), and then outputs the features by Conv fusion, and then combines with the residual connection of the just input feature parts. After that, the features are fused by Conv, and then combined with the residual connection of the just-input feature part to output the features.

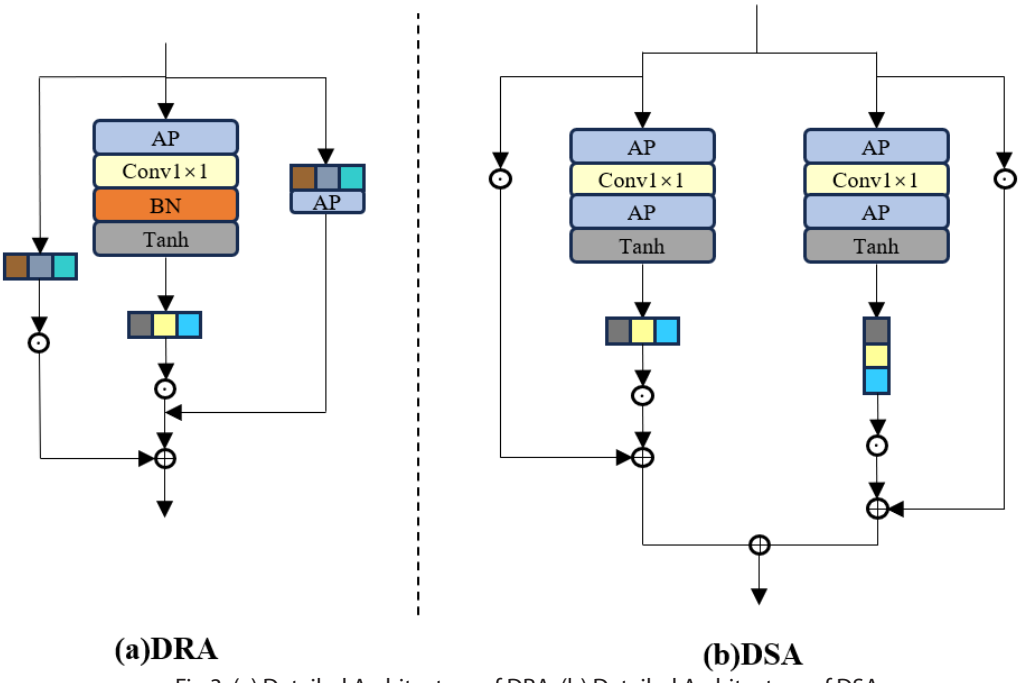

Fig 3: (a) Detailed Architecture of DRA. (b) Detailed Architecture of DSA.

Given the feature input x, the formula is as follows:

$$F_1 = \sum_i f_{DW}\left(f_{DWD}\left(f_{PWD}(x_i)\right)\right), \#(1)$$

$$F_2 = \mathrm{AP}\left(f_{DRA}(F_1) + f_{DSA}(F_1)\right), \#(2)$$

$$F_2^{'} = F_2 + \mathrm{AP}\left(f_{DRA}(F_1) + f_{DSA}(F_1)\right), \#(3)$$

$$F_2^{''} = F_2^{'} + \mathrm{AP}\left(f_{DRA}(F_1) + f_{DSA}(F_1)\right), \#(4)$$

$$F_{out} = Conv\left(F_2 + F_2^{'} + F_3^{''}\right) + f_{PW}(x), \#(5)$$

(3) Spatial and channel synergistic attention module (SCSA)

In order to minimize the loss that global and local feature maps may cause in the image recovery phase of the decoder, we introduce the SCSA[15] module (as shown in Fig 4), which is designed to efficiently select important feature regions and channels, suppress irrelevant information, and contribute to the recovery of details.

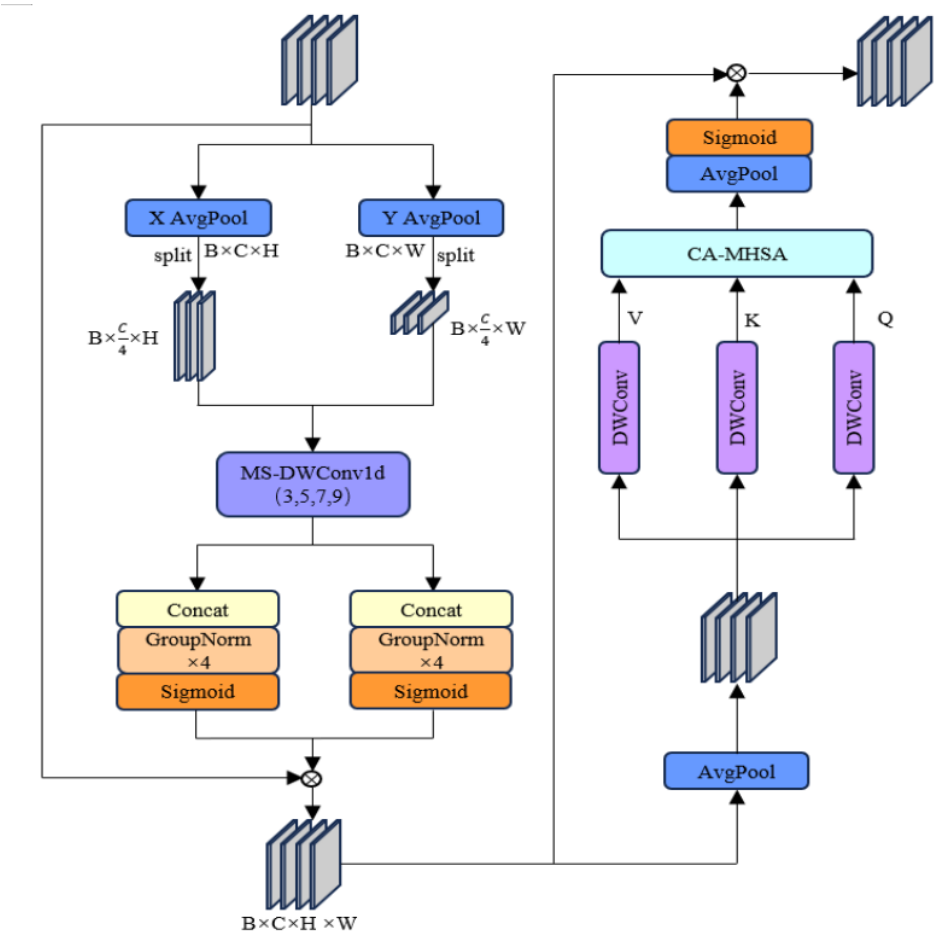

Fig 4: Detailed Architecture of SCSA.

As shown in the figure, the input features are first pooled in x-axis and y-axis in parallel, and then fused into different feature maps by DWConv according to four different convolution kernels [3, 5, 7, 9] respectively, and then the four feature maps are connected two by two by Concat to aggregate different features, and then normalized respectively, which can effectively solve the semantic interference between the features, and then finally the spatial attention is generated using the Sigmoid activation function to generate spatial attention. After that, the features are pooled, and the multi-head attention is computed by the progressive compression method, i.e., three times DWConv, and finally the channel attention is generated by pooling activation. The generated attention is then combined with the spatial attention.

4. Experimentation

In this section, we present the combined experimental results of the skin lesion and polyp segmentation tasks. Specifically, we evaluate the proposed network using the ISIC2017, ISIC2018, and Synapse Abdominal Multi Organ Segmentation (Synapse) datasets.

(1) Datasets

To better evaluate the generalization performance of our method, we used three datasets: (1) ISIC2017 dataset: The International Skin Imaging Collaboration 2018 Challenge (ISIC2018) dataset is a publicly available dataset of skin lesions containing 1,500 training images, 650 testing images, all resized to 256×256 pixels. (2) ISIC2018 dataset: The

International Skin Imaging Collaboration 2018 Challenge (ISIC2018) dataset is a publicly available dataset of skin lesions containing 2,594 training images, 1,000 testing images, all resized to 256×256 pixels. (3) Synapse dataset: The Synapse dataset consists of 30 abdominal CT scans, totaling 3779 axial contrast-enhanced clinical CT images of the abdominal region with a resolution of 512×512. According to the method in reference TransUnet, we randomly divided the Synapse dataset into a training set and a validation set, with 18 cases (2212 axial slices) used for training and the remaining 12 cases used for validation. This dataset is used to evaluate the segmentation performance of our method on eight different abdominal organs (aorta, gallbladder, left kidney, right kidney, liver, pancreas, spleen, and stomach).

(2) Realization details

In this experiment, we implement our model based on the PyTorch deep learning architecture and trained it on the ISIC2017, ISIC2018, and Synapse datasets. The AdamW optimizer was employed for training with a learning rate of 1e-3 and 300 training epochs, while the batch size was configured to 32. To ensure a valid comparison, all experimental parameters were kept consistent across trials, and all experiments were conducted on an NVIDIA RTX A6000 GPU. Evaluation indicators: For ISIC207 and 2018 datasets, We primarily utilize the Dice similarity coefficient and intersection over union (IoU) for evaluating the discrepancies of the model. The Dice similarity coefficient measures the similarity between standard segmentation and predicted segmentation—A higher Dice value indicates better model segmentation performance. A higher IoU indicates a more accurate bounding box. For the Synapse dataset, the average Dice similarity coefficient (DSC) and the average Hausdorff distance (HD) are used as evaluation indicators.

(3) Assessment of results

We conducted experiments on ISIC17 and ISIC18 datasets and compared them with the state-of-the-art methods in recent years, and the DSC results for each method are shown in Table 1 and Fig 5. From these results, it can be seen that our proposed method outperforms the compared methods on both ISIC17 and ISIC18 datasets, and it also proves that our MSCM-Net has excellent generalization ability.

Table 1: Quantitative evaluation of various methods was performed on the ISIC-2017 and ISIC-2018 skin lesion segmentation datasets to verify the generalization ability of the model, where ↑ indicates that the higher the better. (red, green, and blue mark the first, second, and third results).

Dataset	Model	mIoU(%) ↑	DSC(%) ↑	Acc(%) ↑	Spe(%) ↑	Sen(%) ↑
ISIC17	UNet[1]	76.98	86.99	95.65	97.43	86.82
	UTNetV2[16]	77.35	87.23	95.84	98.05	84.85
	TransFuse[4]	79.21	88.40	96.17	97.98	87.14
	MALUNet[17]	78.78	88.13	96.18	98.47	84.78
	VM-UNet[6]	80.23	89.03	96.29	97.58	89.90
	Ours	81.41	89.76	96.60	98.09	89.14
ISIC18	UNet[1]	77.86	87.55	94.05	96.69	85.86
	UNet++[8]	78.31	87.83	94.02	95.75	88.65
	Att-UNet[2]	78.43	87.91	94.13	96.23	87.60
	UTNetV2[16]	78.97	88.25	94.32	96.48	87.60
	SANet[18]	79.52	88.59	94.39	95.97	89.46
	TransFuse[4]	80.63	89.27	94.66	95.74	91.28
	MALUNet[17]	80.25	89.04	94.62	96.19	89.74
	VM-UNet[6]	81.35	89.71	94.91	96.13	91.12
	Ours	82.24	90.25	94.49	95.83	91.04

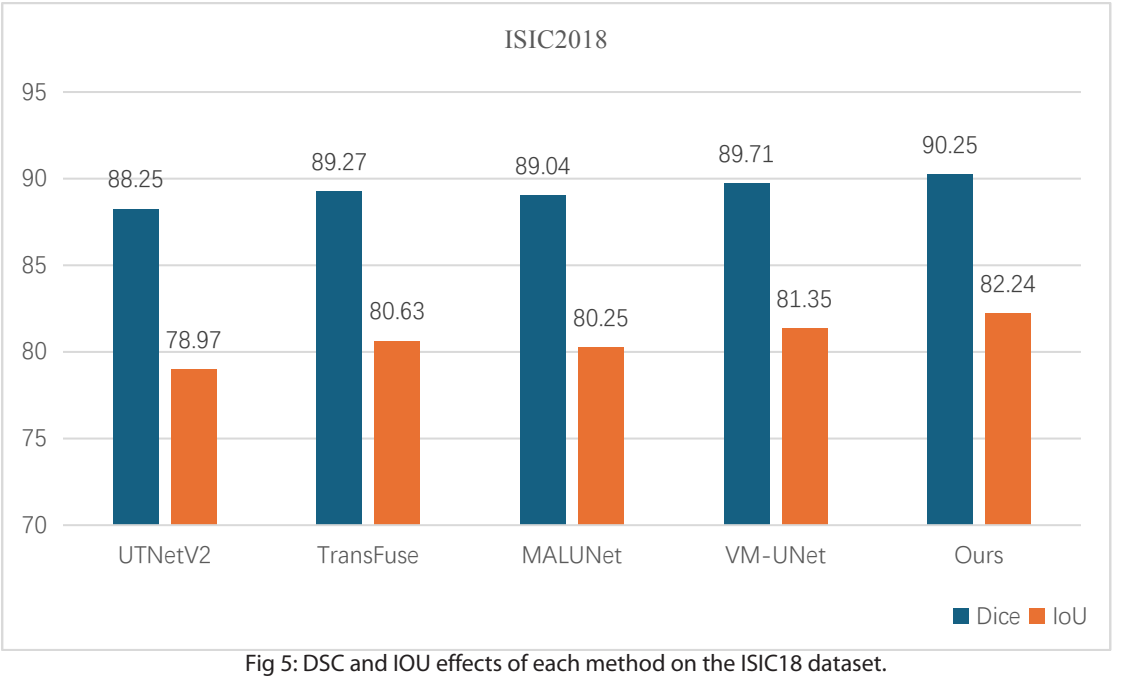

Fig 5: DSC and IOU effects of each method on the ISIC18 dataset.

We have experimented on also on Synapse dataset and compared the proposed method with previous CNN-based, Transformer-based and Mamba-based methods, and our HD metrics reached the optimal results, and the HD and DSC results of MSCM-Net are shown in Table 2, which can be clearly found that our method on Synapse data is also competitive.

Table 2: Various methods are quantitatively evaluated on the Synapse dataset to verify the generalization ability of the model, where ↑ means the higher the better, and ↓ means the lower the better. (red, green, and blue mark the first, second, and third results).

Model	DSC ↑	HD95 ↓	Aorta	Gallbladder	Kidney(L)	Kidney(R)	Liver	Pancreas	Spleen	Stomach
V-Net[19]	68.81	-	75.34	51.87	77.10	80.75	87.84	40.05	80.56	56.98
DARR[20]	69.77	-	74.74	53.77	72.31	73.24	94.08	54.18	89.90	45.96
UNet[1]	76.85	39.70	89.07	69.72	77.77	68.60	93.43	53.98	86.67	75.58
Att-Unet[2]	77.77	36.02	89.55	68.88	77.98	71.11	93.57	58.04	87.30	75.75
TransUnet[3]	77.48	31.69	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62
TransNorm[21]	78.40	30.25	86.23	65.10	82.18	78.63	94.22	55.34	89.50	76.01
Swin U-Net[22]	79.13	21.55	85.47	66.53	83.28	79.61	94.29	56.58	90.66	76.60
TransDeepLab[23]	80.16	21.25	86.04	69.16	84.08	79.88	93.53	61.19	89.00	78.40
UCTransNet[24]	78.23	26.75	-	-	-	-	-	-	-	-
MT-UNet[25]	78.59	26.59	87.92	64.99	81.47	77.29	93.06	59.46	87.75	76.81
VM-UNet[6]	81.08	19.21	86.40	69.41	86.16	82.76	94.17	58.80	89.51	81.40
Ours	80.17	14.21	87.25	67.15	86.13	81.41	94.39	58.12	88.40	78.80

(4) Ablation studies

In the first row of Table 3, it indicates the experimental effect of our replication of VM-Unet. The second and third rows show the different effects of different pre-training weights on our model, and it can be seen that Vmamba-S has the best results.

Table 4 represents the effect of different attentions on the performance of the model, the first row represents the reproduction results, rows 2 to 4, are the experimental effects of adding different attentions individually, we are able to find out that our proposed HRFA has a good help on the feature extraction of the model by comparing the second and first rows, the last row is the combination of three different attentions together, which is constructed as MSCM-Net, and the state-of-the-art results were obtained.

Table 3: Effect of different weights on model accuracy. where ↑ indicates that the higher the better

Init. Weight	ISIC17		ISIC18	
	mIoU(%) ↑	DSC(%) ↑	mIoU(%) ↑	DSC(%) ↑
Baseline	79.95	88.86	80.63	89.27
VMamba-T	77.06	87.05	81.08	89.55
VMamba-S	**81.41**	**89.76**	**82.24**	**90.25**

Table 4: The effect of the proposed combination of attention and other attention on the accuracy of the model. where ↑ indicates that the higher the better

HRFA	SCSA	SE	ISIC17		ISIC18	
			mIoU(%) ↑	DSC(%) ↑	mIoU(%) ↑	DSC(%) ↑
×	×	×	79.95	88.86	80.63	89.27
√	×	×	80.19	89.01	81.90	90.06
×	√	×	80.99	89.49	81.83	90.00
×	×	√	80.07	88.93	81.23	89.64
√	√	√	**81.41**	**89.76**	**82.24**	**90.25**

(5) Visualization charts

As in Fig. 6, GT denotes really labels, and we compare it with the other two state-of-the-art models, we can clearly see in the second row where the red markers are located, and our model visualization is closer to GT.

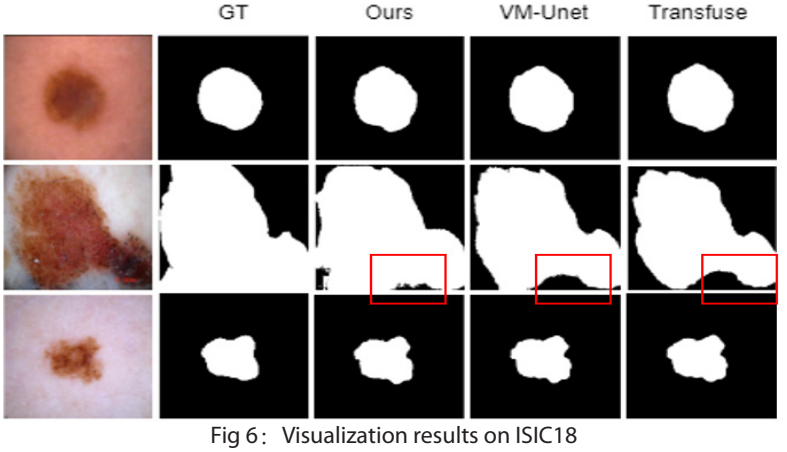

Fig 6： Visualization results on ISIC18

5. Conclusion

In this paper, we propose a new method called MSCM-Net. The method effectively overcomes the problem of global and local information imbalance through our proposed Hierarchical Receptive Field Attention Module and our introduced Spatial and Channel Attention Mudule. We evaluated our method on three different datasets and the results show its clear superiority. In future research, we will consider the issue of modeling computational resources and explore more efficient models and architectures that ensure segmentation accuracy while reducing the demand for computational resources, thus better aiding clinical diagnosis and treatment decisions.

References

[1] Ronneberger O, Fischer P, Brox T: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015, Munich, Germany: Springer International Publishing, 234-241, 2015. https://doi.org/10.1007/978-3-319-24574-4_28.

[2] Vaswani A, Shazeer N, Parmar N, et al: Attention is all you need. Adv Neural Inf Process Syst 30:5998-6008, 2017. https://doi.org/10.48550/arXiv.1706.03762.

[3] Chen J, Lu Y, Yu Q, et al: TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers. Medical Image Analysis, 97, 103280, 2024. https://doi.org/10.1016/j.media.2024.103280.

[4] Zhang Y, Liu H, Hu Q: TransFuse: Fusing Transformers and CNNs for medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021, Strasbourg, France: Springer International Publishing, 14-24, 2021. https://doi.org/10.1007/978-3-030-87193-2_2 .

[5] Liu Y, Xiao T, Gong M, et al: Vmamba: Visual state space model. arXiv preprint arXiv:2401.10166, January 18, 2024. https://doi.org/10.48550/arXiv.2401.10166.

[6] Ruan J, Xiang S: VM-UNet: Vision mamba UNet for medical image segmentation. arXiv preprint arXiv:2402.02491, February 5, 2024. https://doi.org/10.48550/arXiv.2402.02491.

[7] Si Y, Xu H, Zhu X, et al. SCSA: Exploring the synergistic effects between spatial and channel attention. arXiv preprint arXiv:2407.05128, 2024. https://doi.org/10.48550/arXiv.2407.05128.

[8] Zhou Z, Rahman Siddiquee MM, Tajbakhsh N, et al: UNet++: A nested u-net architecture for medical image segmentation. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Granada, Spain: Springer International Publishing, 3-11, 2018. https://doi.org/10.1007/978-3-030-00889-5_1.

[9] Cao, H. et al. Swin-Unet: Unet-Like Pure Transformer for Medical Image Segmentation. In: Karlinsky, L., Michaeli, T., Nishino, K. (eds) Computer Vision – ECCV 2022 Workshops. ECCV 2022. Lecture Notes in Computer Science, vol 13803. Springer, Cham. https://doi.org/10.1007/978-3-031-25066-8_9.

[10] Gu, Albert, Karan Goel, and Christopher Ré. "Efficiently modeling long sequences with structured state spaces." arXiv preprint arXiv:2111.00396 (2021). https://doi.org/10.48550/arXiv.2111.00396.

[11] Ma, Jun, Feifei Li, and Bo Wang. "U-mamba: Enhancing long-range dependency for biomedical image segmentation." arXiv preprint arXiv:2401.04722 (2024). https://doi.org/10.48550/arXiv.2401.04722.

[12] Xing, Zhaohu, et al. "Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation." International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer Nature Switzerland, 2024. https://doi.org/10.1007/978-3-031-72111-3_54.

[13] Ruan J, Xiang S: VM-UNet: Vision mamba UNet for medical image segmentation. arXiv preprint arXiv:2402.02491, February 5, 2024. https://doi.org/10.48550/arXiv.2402.02491 .

[14] Liu Y, Xiao T, Gong M, et al: Vmamba: Visual state space model. arXiv preprint arXiv:2401.10166, January 18, 2024. https://doi.org/10.48550/arXiv.2401.10166.

[15] Si, Yunzhong, et al. "SCSA: Exploring the synergistic effects between spatial and channel attention." arXiv preprint arXiv:2407.05128 (2024). https://doi.org/10.48550/arXiv.2407.05128.

[16] Gao, Yunhe, et al. "A data-scalable transformer for medical image segmentation: architecture, model efficiency, and benchmark." arXiv preprint arXiv:2203.00131 (2022). https://doi.org/10.48550/arXiv.2203.00131.

[17] Ruan J, Xiao C, Tong H, et al: MALUNet: A multi-attention and light-weight UNet for skin lesion segmentation. In: 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Las Vegas, NV: IEEE, 264-269, 2022. https://doi.org/10.1109/BIBM55620.2022.9995040.

[18] Wei J, Hu Y, Zhang R, et al: Shallow attention network for polyp segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021, Strasbourg, France: Springer International Publishing, 699-708, 2021. https://doi.org/10.1007/978-3-030-87193-2_66 .

[19] Milletari, Fausto, Nassir Navab, and Seyed-Ahmad Ahmadi. "V-net: Fully convolutional neural networks for volumetric medical image segmentation." 2016 fourth international conference on 3D vision (3DV). Ieee, 2016. https://doi.org/10.1109/3DV.2016.79.

[20] Alom, Md Zahangir, et al. "Recurrent residual U-Net for medical image segmentation." Journal of medical imaging 6.1 (2019): 014006-014006.

[21] Azad, Reza, et al. "Transnorm: Transformer provides a strong spatial normalization mechanism for a deep segmentation model." IEEE access 10 (2022): 108205-108215. https://doi.org/10.1109/ACCESS.2022.3211501.

[22] Liu Z, Lin Y, Cao Y, et al: Swin Transformer: Hierarchical vision Transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, Canada: IEEE, 10012-10022, 2021. https://doi.org/10.48550/arXiv.2103.14030 .

[23] Azad, Reza, et al. "Transdeeplab: Convolution-free transformer-based deeplab v3+ for medical image segmentation." International Workshop on PRedictive Intelligence In MEdicine. Cham: Springer Nature Switzerland, 2022. https://doi.org/10.1007/978-3-031-16919-9_9.

[24] Wang, Haonan, et al. "Uctransnet: rethinking the skip connections in u-net from a channel-wise perspective with transformer." Proceedings of the AAAI conference on artificial intelligence. Vol. 36. No. 3. 2022. https://doi.org/10.1609/aaai.v36i3.20144.

[25] Wang, Hongyi, et al. "Mixed transformer u-net for medical image segmentation." ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2022. https://doi.org/10.1109/ICASSP43922.2022.9746172.