Metadata of the chapter that will be visualized in SpringerLink

Book Title	Advanced Intelligent Computing Technology and Applications	
Series Title		
Chapter Title	TMU: Transmission-Enhanced Mamba-UNet for Medical Image Segmentation	
Copyright Year	2024	
Copyright HolderName	The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd.	
Author	Family Name	**Yang**
	Particle	
	Given Name	**Xiongfeng**
	Prefix	
	Suffix	
	Role	
	Division	
	Organization	Nankai University
	Address	Tianjin, China
	Email	
Author	Family Name	**Luo**
	Particle	
	Given Name	**Ziyang**
	Prefix	
	Suffix	
	Role	
	Division	
	Organization	Nankai University
	Address	Tianjin, China
	Email	
Author	Family Name	**Wu**
	Particle	
	Given Name	**Yanlin**
	Prefix	
	Suffix	
	Role	
	Division	
	Organization	Nankai University
	Address	Tianjin, China
	Email	
Corresponding Author	Family Name	**Xie**
	Particle	
	Given Name	**Xueshuo**
	Prefix	
	Suffix	

	Role	
	Division	
	Organization	Haihe Lab of ITAI
	Address	Tianjin, China
	Email	xueshuoxie@nankai.edu.cn
Author	Family Name	**Nan**
	Particle	
	Given Name	**Li**
	Prefix	
	Suffix	
	Role	
	Division	
	Organization	Tianjin Eye Hospital
	Address	Tianjin, China
	Email	
Author	Family Name	**Li**
	Particle	
	Given Name	**Tao**
	Prefix	
	Suffix	
	Role	
	Division	
	Organization	Nankai University
	Address	Tianjin, China
	Division	
	Organization	Haihe Lab of ITAI
	Address	Tianjin, China
	Email	

Abstract	In the field of medical image segmentation, the Mamba-UNet is seen as a diamond in the rough due to its robust capability in capturing long-range interactions within images while maintaining linear computational complexity. However, the existing Mamba-based U-shaped networks utilize direct skip connections, which limit the exploration of features at different scales. To address this issue, we propose the Transmission-Enhanced Mamba-UNet (TMU) by incorporating a DCA (Double Cross Attention) block between the encoder and decoder, aiming to enhance skip connections in mamba-based networks. This design elevates segmentation performance by introducing an attention mechanism into the skip connection, effectively fusing features from different layers and improving the model's capacity to capture intricate details and contextual information. We also explored the performance of DCA blocks with different inputs and outputs to find the best combination. Experimental results on the publicly available ACDC dataset and ISIC dataset demonstrate that TMU outperforms the Mamba-UNet model across all evaluation metrics when utilizing pretrained models. Especially in IoU, the TMU model with pretrained weights saw improvements of 1.56% on the ACDC dataset and 0.68% on the ISIC dataset. Identically, without pretrained weights, it exhibited improvements of 4.73% and 0.67%.

Keywords (separated by '-')	Medical Image Segmentation - Mamba - State Space Models - U-Net

TMU: Transmission-Enhanced Mamba-UNet for Medical Image Segmentation

Xiongfeng Yang[1], Ziyang Luo[1], Yanlin Wu[1], Xueshuo Xie[2](✉), Li Nan[3], and Tao Li[1,2]

[1] Nankai University, Tianjin, China
[2] Haihe Lab of ITAI, Tianjin, China
xueshuoxie@nankai.edu.cn
[3] Tianjin Eye Hospital, Tianjin, China

Abstract. In the field of medical image segmentation, the Mamba-UNet is seen as a diamond in the rough due to its robust capability in capturing long-range interactions within images while maintaining linear computational complexity. However, the existing Mamba-based U-shaped networks utilize direct skip connections, which limit the exploration of features at different scales. To address this issue, we propose the Transmission-Enhanced Mamba-UNet (TMU) by incorporating a DCA (Double Cross Attention) block between the encoder and decoder, aiming to enhance skip connections in mamba-based networks. This design elevates segmentation performance by introducing an attention mechanism into the skip connection, effectively fusing features from different layers and improving the model's capacity to capture intricate details and contextual information. We also explored the performance of DCA blocks with different inputs and outputs to find the best combination. Experimental results on the publicly available ACDC dataset and ISIC dataset demonstrate that TMU outperforms the Mamba-UNet model across all evaluation metrics when utilizing pretrained models. Especially in IoU, the TMU model with pretrained weights saw improvements of 1.56% on the ACDC dataset and 0.68% on the ISIC dataset. Identically, without pretrained weights, it exhibited improvements of 4.73% and 0.67%.

Keywords: Medical Image Segmentation · Mamba · State Space Models · U-Net

1 Introduction

Convolutional Neural Networks (CNNs) are foundational in deep learning for image processing, with the U-Net architecture [18] achieving notable success in medical image segmentation through its encoder-decoder structure that captures both global and local contextual information.

The Transformer architecture [20], initially designed for natural language processing (NLP), has been adapted to computer vision with the Vision Transformer (ViT) [8]. ViT processes images as sequence data, allowing it to capture global visual dependencies that are often missed by CNNs. Building on ViT, Transformer-based architectures such as TransUNet and SwinUNet [6] have been proposed to enhance image segmentation performance.

D.-S. Huang et al. (Eds.): ICIC 2024, LNCS 14871, pp. 1–11, 2024.
https://doi.org/10.1007/978-981-97-5609-4_33

Mamba introduces a novel architecture to address the challenges of computational complexity and large-scale data training requirements, featuring selective processing and a simplified State Space Model (SSM) for effective context compression and dynamic behavior adjustment. [10] The Vision Mamba (VMamba) model [16] extends Mamba's capabilities to computer vision, incorporating bidirectional SSM and place embedding techniques for precise visual recognition. VMamba has shown superior performance in benchmarks, particularly in computational and memory efficiency. Mamba-UNet [21] combines the strengths of U-Net and Mamba, ensuring seamless connectivity and information flow between the encoder and decoder paths. To enhance performance further, we propose TMU, which introduces a Double Cross-Attention (DCA) block to the Mamba-UNet. This block integrates channel Cross-Attention (CCA) and spatial Cross-Attention (SCA) [2] between the Vision State Space (VSS) blocks of Mamba-UNet. The aim is to capture channel and spatial dependencies between multi-scale encoder features, thereby enhancing the model's ability to process image features.

We summarize the contributions below:

- We introduced a DCA block to Mamba-UNet, which significantly enhances image segmentation performance through a dual mechanism of double cross attention, thereby improving feature extraction and fusion capabilities. Comprising Channel Cross-Attention (CCA) and Spatial Cross-Attention (SCA), the DCA block refines the accuracy of image segmentation tasks, particularly for complex medical images.

- We conducted extensive exploration into different configurations of the DCA block's input and output to optimize the model further. This focused exploration aimed to identify the most effective strategies to reinforce the skip connection within the model, aiming for greater performance gains and improved precision in handling complex segmentation tasks.

- Our model has demonstrated superiority over traditional U-Net, Swin-UNet, and Mamba-UNet on both the ACDC and ISIC datasets, particularly excelling in medical image segmentation tasks. This success underscores its potential as a valuable tool in medical imaging analysis.

2 Approach

2.1 Overview

The architecture of TMU, as shown in Fig. 1(a), begins with $1 \times H \times W$ image data that is initially processed through a patch partition layer used in Vision Transformers and Vision Mamba. This layer prepares the data for subsequent linear embedding, where the dimensions are refined to $H/4 \times W/4 \times 16$ and then reshaped into $H/4 \times W/4 \times c$ to optimize feature representation. The data then flows through a pair of VSS blocks dedicated to extracting features without resizing the tensor. This is followed by a series of three stages, each of which includes a patch merging layer that downsamples the data before it is processed by another pair of VSS blocks. After these transformations, the model's input adopts the configuration of $H/32 \times W/32 \times 8c$ as it enters the decoder. Within the decoder, the data undergoes three additional stages, each characterized by a patch expanding operation and the application of two VSS blocks for feature extraction.

The final stage involves one last patch expansion and a linear projection, culminating in the generation of a detailed segmented image with dimensions H × W × Class.

In parallel with the initial VSS block processing, the data is also channeled through a DCA block. Here, it undergoes a series of transformations to match the shape of the VSS block output. It then experiences channel cross-attention and spatial cross-attention mechanisms designed to enhance feature integration. The output from these attention mechanisms is subjected to various up-sampling techniques before being integrated into the corresponding stages of the model. This innovative approach of replacing traditional U-Net feature extraction blocks with CNNs or Transformers (in the form of VSS blocks) and incorporating a DCA block enriches the model's capability to capture multi-scale dependencies in features. The DCA block plays a pivotal role in bridging the semantic gap between encoder and decoder features by continuously focusing on the interplay of channel and spatial relationships across the encoding process. This refined attention to feature connectivity leads to an enhanced skip connection, which is essential for the model's overall performance and accuracy in image segmentation tasks.

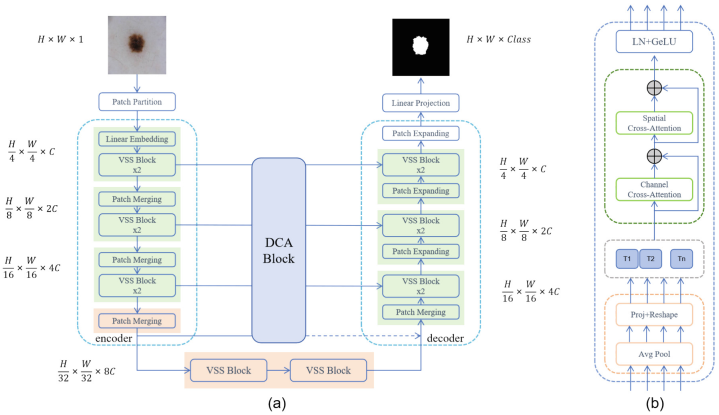

Fig. 1. (a) The architecture of TMU, which is composed of encoder, bottleneck, decoder and skip connections with a DCA block. (b) The architecture of DCA block, which is composed of average pooling layer, linear projection layer, channel cross-attention layer and spatial cross-attention layer.

2.2 DCA Block

In the DCA block architecture shown in Fig. 1(b), the initial stage involves the extraction of patches from four distinct multi-scale encoder stages, commonly referred to as skip connection layers. To accommodate varying scales of encoders, the patches are derived using 2D average pooling with a pool size and stride denoted by Ψ (Psi). Subsequently, these 2D patches undergo a projection process utilizing 1 × 1 depth-wise convolutions.

Subsequent to this extraction, the features are bifurcated into two pathways. The first pathway involves an additive combination with the output from the channel cross-attention block before proceeding to the subsequent stage. The second pathway channels of the features were input to the channel cross-attention block. Within this block, distinct tokens are generated along the channel dimension to serve as keys and values, while the queries are constructed using Ti. Notably, while linear projections are conventionally employed for self-attention mechanisms, recent advancements have seen the successful integration of convolutional operations. This integration not only introduces an element of locality to the self-attention process but also significantly diminishes the computational complexity. Specifically, the utilization of depth-wise convolutions for self-attention is advantageous as it captures local information with minimal additional computational overhead. Leveraging these insights, the Depth-wise Block Attention (DBA) block in our model supplants all linear projections with 1 × 1 depth-wise convolutional projections.

The process then advances to the spatial cross-attention module, which similarly splits the features into two streams. One stream merges with the output and feeds into the next stage, while the other stream enters the spatial cross-attention block. This block performs layer normalization [3] and concatenation along the channel dimension.

$$T_i = DConv1D_{E_i}(\mathrm{Re}shape(AvgPool2D_{E_i}(E_i))) \tag{1}$$

$$Q_i = DConv1D_{Q_i}(T_i) \tag{2}$$

$$K = DConv1D_K(T_C) \tag{3}$$

$$V = DConv1D_V(T_C) \tag{4}$$

$$CCA(Q_i, K, V) = Soft\max(\frac{Q_i^T K}{\sqrt{C}})V^T \tag{5}$$

In contrast to the Channel Cross-Attention (CCA) module, the spatial cross-attention block utilizes the concatenated tokens as both queries and keys, assigning each individual token i as the corresponding value. The queries, keys, and values are all subjected to 1 × 1 depth-wise projections.

$$Q = DConv1D_Q(\overline{T}_C) \tag{6}$$

$$K = DConv1D_K(\overline{T}_C) \tag{7}$$

$$V_i = DConv1D_{V_i}(\overline{T}_i) \tag{8}$$

where $Q \in R^{P \times C_c}$, $K \in R^{P \times C_c}$, $V_i \in R^{P \times C_i}$ are the projected queries, keys and values, respectively. Then, SCA can be expressed as:

$$SCA(Q, K, V_i) = Soft\max(\frac{QK^T}{\sqrt{d_k}})V_i \tag{9}$$

This double cross-attention mechanism is a pivotal component of our model, enhancing its ability to process and integrate multi-scale features effectively.

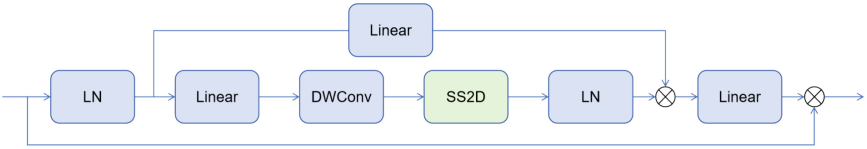

Fig. 2. Flowchart of the VSS Block structure

2.3 VSS Block

The VSS block, as shown in Fig. 2, originating from VMamba, serves as the cornerstone module of our model, depicted in the accompanying figure. Post-layer normalization [3], the input bifurcates into dual branches. The inaugural branch input through a linear layer, culminating in an activation function. Concurrently, the secondary branch subjects the input to a linear layer, followed by depthwise separable convolution and an activation function, before ushering it into the 2D Selective Scanning (SS2D) module for augmented feature extraction. Ensuing this, features undergo normalization via layer normalization, and an element-wise operation is executed with the first branch's output to amalgamate the two trajectories. In the final stage, a linear layer amalgamates the features, which, in conjunction with the residual connection, constitutes the VSS block's output. For activation purposes, this paper adopts SiLU by default.

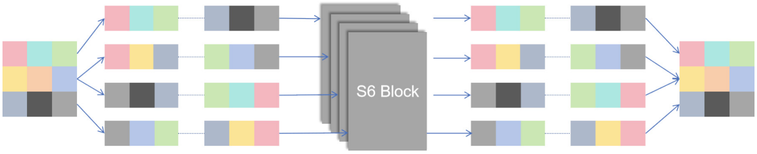

Fig. 3. Illustration of the 2D-Selective-Scan on an image. We commence by scanning an image using CSM (scan expand). The four resulting features are then individually processed through the S6 block, and the four output features are merged (scan merge) to construct the final 2D feature map.

The SS2D Component The SS2D component is a tripartite structure consisting of the ScanExpanding operation, the S6 block, and the ScanMerging operation. Figure demonstrates the ScanExpanding operation, which systematically unfolds the input image into sequences across four distinct trajectories: from top-left to bottom-right, bottom-right to top-left, top-right to bottom-left, and bottom-left to top-right. These sequences are subsequently processed by the S6 block, which performs comprehensive feature extraction to ensure that information from each direction is meticulously scanned, thereby capturing a spectrum of features. Following this, as depicted in Fig. 3, the ScanMerging operation consolidates the sequences from all four directions, summing and amalgamating them to revert the output image to its original dimensions. Originating from Mamba [16], the S6 module enhances the model by incorporating a selection mechanism over S4 [17], which fine-tunes the SSM parameters in response to the input. This refinement allows the model to discern and preserve pertinent information while discarding what is extraneous. The pseudocode for the S6 block is delineated as Algorithm 1.

Algorithm 1 SSM + Selection(S6)

```
Input: x:(B,L,D)
Output: y:(B,L,D)
  $\boldsymbol{A}$: (D,N) ← Parameter
      ▹ Represents structured $N \times N$ matrix
  $\boldsymbol{B}$: (B,L,N) ← $S_B(x)$
  $\boldsymbol{C}$: (B,L,N) ← $S_C(x)$
  $\boldsymbol{\Delta}$: (B,L,D) ← $\tau_\Delta$(Parameter+$s_\Delta(x)$)
  $\overline{A}$, $\overline{B}$: (B,L,D,N) ← discretize($\Delta, \boldsymbol{A}, \boldsymbol{B}$)
  y ← SSM($\overline{A}$, $\overline{B}$,C)(x)
      ▹ Time-varying: recurrence(scan) only
  return y
```

3 Expriment and Results

3.1 Datasets

ACDC Dataset: It contains 100 short-axis MR-cine T1 3D volumes of cardiac anatomy acquired using 1.5T and 3T scanners. The expert annotations are provided for three structures: right ventricle, myocardium, and left ventricle. It was hosted as part of the MICCAI ACDC challenge 2017 [4]. To comply with the input requirements of the segmentation backbone networks, all images were resized to 224 × 224.

ISIC Dataset: The ISIC dataset is an extensively recognized public medical imaging repository, dedicated to the diagnosis and investigation of dermatological conditions. It features an extensive array of skin disease imagery, covering a broad spectrum of skin afflictions and abnormalities. The dataset encompasses a diverse range of skin pathologies [11], including but not limited to melanoma, squamous cell carcinoma, and basal cell carcinoma. Comprising a total of 2,694 dermatological images accompanied by their respective masks, the dataset has been meticulously partitioned into a training set and a validation set with a proportional distribution of 4:1. To comply with the input requirements of the segmentation backbone networks, all images were resized to 224 × 224.

3.2 Implementation Details

Environment: The task was undertaken within an environment configured on Ubuntu 20.04, utilizing Python 3.11 and PyTorch 2.2.1, alongside CUDA 12.2, leveraging the capabilities of an Nvidia GeForce RTX 3090 GPU and an Intel(R) Xeon(R) Gold 6133 CPU. The entire process, including data handling, model training, and inference, was completed in about 3 h. The dataset was prepared for 2D image segmentation tasks, and the Mamba-UNet model was trained for 10,000 iterations using a batch size of 32. We have opted for a hybrid loss function that combines the Cross-Entropy Loss (CE Loss) and the Dice Loss, with each contributing equally to the overall weight of the loss function. An SGD optimizer was selected [5], configured with a learning rate of 0.01, momentum at 0.9, and weight decay of 0.0001. The model's performance was monitored on a validation set after every 50 iterations, with the best-performing model weights being saved to ensure continual improvement and refinement of the model.

Baseline: To ensure a fair assessment, U-Net, Swin-UNet, and Mamba-UNet models were all trained using the same set of hyperparameters. This standardized approach facilitated a direct comparison with the TMU and other established baseline methods.

3.3 Evaluation Metrics

To measure the performance of the model, we have employed three metrics: Intersection over Union (IoU), Dice coefficient, and Sensitivity. The performance of the model is considered better as these metrics approach a value of 1. The formulas for calculating these metrics are as follows:

Intersection over Union (IoU):

$$IoU = \frac{TP}{TP + FP + FN} \tag{10}$$

where TP (True Positives) is the number of correctly predicted positive pixels, FP (False Positives) is the number of incorrectly predicted positive pixels, and FN (False Negatives) is the number of actual positive pixels that were not predicted correctly.

Dice Coefficient:

$$Dice = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{11}$$

The Dice coefficient is similar to the IoU but gives twice the weight to the number of true positives, which can be beneficial in cases where the number of false positives and false negatives are high.

Sensitivity (also known as recall or true positive rate):

$$Sensitivity = \frac{TP}{TP + FN} \tag{12}$$

Sensitivity measures the proportion of actual positives that were correctly identified by the model, which is particularly important in cases where missing positive instances can have significant consequences.

3.4 Comparative Studies

The results indicate that on the ACDC dataset, when using pre-trained models, our model outperforms both U-Net, Swin-UNet, and Mamba-UNet across all evaluated metrics—mean Dice, IoU, and sensitivity. The pre-trained Mamba-UNet achieves a mean Dice of 0.9019, an IoU of 0.8272, and a sensitivity of 0.9098, which are significant improvements over the non-pre-trained versions of U-Net (mean Dice of 0.7717, IoU of 0.6424, and sensitivity of 0.7679) and Swin-UNet (mean Dice of 0.8858, IoU of 0.8026, and sensitivity of 0.8968). The enhanced performance of the pre-trained model demonstrates the benefits of transferring knowledge from large datasets, which allows the model to better generalize and adapt to the specific characteristics of the medical imaging data. The pre-trained model is able to leverage the learned features and representations from a vast amount of annotated data, leading to a more robust and accurate segmentation model (Tables 1 and 2).

Table. 1. Performance analysis of different models under different datasets

Type	Model	ACDC			ISIC		
		Dice	IoU	Sensitivity	Dice	IoU	Sensitivity
Pretrained	U-Net	0.9005	0.8264	0.9065	0.8649	0.7679	0.9355
	Swin-UNet	0.8858	0.8026	0.8968	0.8744	0.7846	0.9329
	Mamba-UNet	0.8922	0.8132	0.9008	0.8875	0.8061	0.9285
	TMU(ours)	**0.9019↑**	**0.8272↑**	**0.9098↑**	**0.8916↑**	**0.8121↑**	**0.9519↑**
Unpretrained	Swin-UNet	0.7717	0.6424	0.7679	0.8435	0.7471	0.8625
	Mamba-UNet	0.8014	0.6859	0.7990	0.8329	0.7438	0.8311
	TMU(ours)	**0.8322↑**	**0.7237↑**	**0.8356↑**	**0.8437↑**	**0.7485↑**	**0.8681↑**

And on the ISIC dataset, our model also demonstrates superior performance compared to U-Net, Swin-UNet, and Mamba-UNet when equipped with pre-trained weights. This enhancement underscores the significance of utilizing pre-trained weights to achieve optimal results. In scenarios where pre-trained weights are not loaded, our model's performance remains competitive, outpacing Mamba-UNet and only marginally lagging behind Swin-UNet. These findings suggest that while our model's efficacy can rival that of Mamba-UNet in certain contexts, it may exhibit a slight deficit in comparison to Swin-UNet. The importance of Mamba's pre-trained weights cannot be overstated, as they play a pivotal role in enhancing the model's performance and facilitating successful data transfer.

The original image, ground truth, and segmentation results of different networks are shown in Fig. 4.

3.5 Studies on Skip Connection Strategies

In previous variations of the U-Net architecture, there were only three skip connections connecting the encoder to the decoder [15]. However, just before entering the bottleneck layer, an additional downsampling step was introduced. Curiously, we decided to pass the result of this downsampling through a DCA block, and to our surprise, this modification improved the overall performance.

Our investigation didn't stop there. We continued experimenting and discovered that when we applied four downsampling steps and fed only the outputs from the last three to the decoder, the segmentation results were optimal. This modified architecture strikes a delicate balance between semantic information and fine-grained details, ultimately leading to more accurate segmentation outcomes.

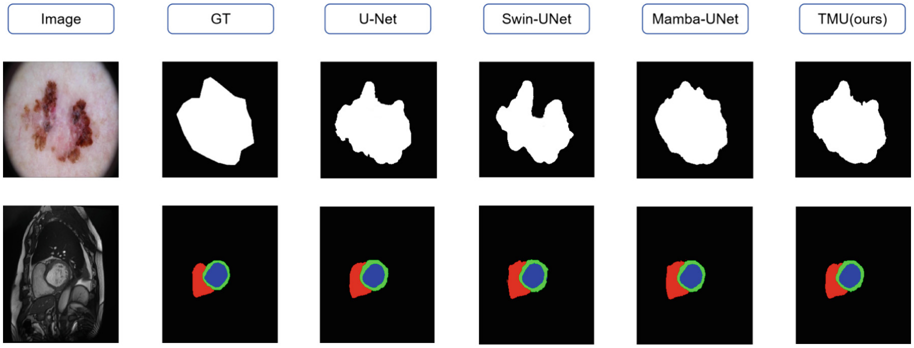

Fig. 4. Results of segmentation on ISIC and ACDC

Table. 2. Results on Skip Connection Strategies

Features In	Features Out	Dice	IoU	Sensitivity
3	3	0.8957	0.8156	0.8967
4	4	0.8960	0.8186	0.9044
4	**3(ours)**	**0.9019**	**0.8272**	**0.9098**

In previous variations of U-Net, there were only three skip connections from the encoder to the decoder. However, just before entering the bottleneck layer, an additional downsampling step was introduced. We then experimented by passing the result of this downsampling through a DCA block. Surprisingly, this improved the performance. Continuing our investigation, we found that when we applied four downsampling steps and fed only the outputs from the last three to the decoder, the results were optimal.

The modified architecture, which incorporates these insights, strikes a balance between semantic information and fine-grained details, leading to better segmentation outcomes.

4 Conclusion

Our contributions extend beyond the immediate improvements to the Mamba-UNet model. The TMU's adaptability to various medical image segmentation tasks and its computational efficiency make it a promising candidate for broader applications in the medical imaging field. The model's architecture serves as a foundation for future research, opening new avenues for the development of advanced segmentation models that can handle the complexities of medical imaging data with greater precision and reliability. In summary, TMU stands as a significant advancement in the realm of medical image segmentation, offering a robust and efficient solution that leverages the strengths of both U-Net and Mamba models, while introducing a novel approach to feature extraction and integration through the double-cross-attention mechanism. The model's exceptional performance and potential for adaptation to diverse tasks position it as a valuable tool in the medical imaging community, contributing to improved diagnostics and patient care.

Acknowledgement. During the journey of researching and writing this thesis, I have been fortunate to receive invaluable assistance and support from many individuals, for which I am deeply grateful. First and foremost, I extend my most sincere gratitude to Professor Tao Li. As my mentor, he not only provided direction for my research with his profound professional knowledge and rigorous academic approach, but also offered selfless guidance and encouragement during times of difficulty and challenge. Professor Li's wisdom and patience have been a crucial cornerstone of my achievements to date. I would also like to express my deep appreciation to Doctor Yanlin Wu. At every stage of experimental design, data analysis, and thesis writing, he offered valuable help and advice. Doctor Wu's experience and insights have significantly enhanced the quality of my research work. I must also extend my special thanks to Teacher Xueshuo Xie, who provided me with valuable guidance both academically and in the revision and refinement of my thesis. Furthermore, I am deeply grateful to the Nankai University Eye Research Institute Open Fund and Doctor Li Nan for their generous support of this research project. Without the substantial support of them, the smooth progress of this research would have been unimaginable. Lastly, I wish to convey my heartfelt thanks to all the teachers and classmates who participated in this research. Their encouragement, support, and understanding have been an indispensable source of motivation for me to complete this work.

References

1. Alom, Md.Z., Hasan, M., Yakopcic, C., Taha, T.M., Asari, V.K.: Recurrent residual convolutional neural network based on U-Net (R2u-Net) for medical image segmentation. arXiv preprint arXiv:1802.06955 (2018)
2. Can Ates, G., Mohan, P.P., Çelik, E.: Dual cross-attention for medical image segmentation. Eng. Appl. Artif. Intell. **126**, 107139 (2023)
3. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)
4. Bernard, O., et al.: Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE Trans. Med. Imaging **37**(11), 2514–2525 (2018)
5. Bottou, L.: Stochastic gradient learning in neural networks. In: Proceedings of Neuro-Nîmes 91, Nimes, France, EC2 (1991)
6. Cao, H., et al.: Swin-Unet: Unet-like pure transformer for medical image segmentation. In: ECCV Workshops (2021)
7. Chen, C.-F., Fan, Q., Panda, R.: CrossViT: cross-attention multi-scale vision transformer for image classification. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 347–356 (2021)
8. Dosovitskiy, A., et al.: An image is worth 16 × 16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
9. Gheini, M., Ren, X., May, J.: Cross-attention is all you need: adapting pretrained transformers for machine translation. In Conference on Empirical Methods in Natural Language Processing (2021)
10. Gu, A., Dao, T.: Mamba: linear-time sequence modeling with selective state spaces. ArXiv, abs/2312.00752 (2023)

11. Gutman, D., et al.: Skin lesion analysis toward melanoma detection: a challenge at the international symposium on biomedical imaging (ISBI) 2016, hosted by the international skin imaging collaboration (ISIC). arXiv preprint arXiv:1605.01397 (2016)
12. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin unetr: swin transformers for semantic segmentation of brain tumors in MRI images. In: International MICCAI Brainlesion Workshop, pp. 272–284. Springer (2021). https://doi.org/10.1007/978-3-031-08999-2_22
13. Huang, Z., et al.: CCNet: criss-cross attention for semantic segmentation. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 603–612 (2018)
14. Li, Z., Liu, F., Yang, W., Peng, S., Zhou, J.: A survey of convolutional neural networks: analysis, applications, and prospects. IEEE Trans. Neural Netw. Learn. Syst. **33**(12), 6999–7019 (2021)
15. Liu, F., Ren, X., Zhang, Z., Sun, X., Zou, Y.: Rethinking skip connection with layer normalization in transformers and resnets. arXiv preprint arXiv:2105.07205 (2021)
16. Liu, Y., et al.: VMamba: visual state space model. ArXiv abs/2401.10166 (2024)
17. Ma, J., Li, F., Wang, B.: U-Mamba: enhancing long-range dependency for biomedical image segmentation. ArXiv abs/2401.04722 (2024)
18. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. ArXiv abs/1505.04597 (2015)
19. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning, pp. 10347–10357. PMLR (2021)
20. Vaswani, A.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
21. Wang, Z., Zheng, J.-Q., Zhang, Y., Cui, G., Li, L.: Mamba-UNet: UNet-like pure visual mamba for medical image segmentation. ArXiv, abs/2402.05079 (2024)
22. Wei, X., Zhang, T., Li, Y., Zhang, Y., Wu, F.: Multi-modality cross attention network for image and sentence matching. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13–19, 2020, pp. 10938–10947. Computer Vision Foundation/IEEE (2020)
23. Wu, H., et al.: CVT: introducing convolutions to vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 22–31 (2021)
24. Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., Wang, X.: Vision Mamba: efficient visual representation learning with bidirectional state space model. ArXiv abs/2401.09417 (2024)