Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 9, No. 3, 2024

DCSS-UNet: UNet based on State Space Model for Polyp Segmentation

Xiuwei Wang [1], Biyuan Li [1, 2, *]

[1] School of Electronic Engineering, Tianjin University of Technology and Education, Tianjin, 300222, China
[2] Tianjin Development Zone Jingnuohanhai Data Technology Co., Ltd, Tianjin, China

[*] **Corresponding author**: Biyuan Li (Email: lby@tute.edu.cn)

Abstract: Early and accurate segmentation of medical images can provide valuable information for medical treatment. In recent years, the automatic and accurate segmentation of polyps in colonoscopy images has received extensive attention from the research community of artificial intelligence and computer vision. Many researchers have conducted in-depth research on models based on CNN and Transformer. However, CNN have limited ability to model remote dependencies, which makes it challenging to fully utilize semantic information in images. On the other hand, the complexity of the secondary computation poses a challenge to the transformer. Recently, state-space models (SSMS), such as Mamba, have been recognized as a promising approach. They not only show superior performance in remote interaction, but also maintain linear computational complexity. Inspired by Mamba, we propose DCSS-UNet, where we utilize visual state space (VSS) blocks in VMamba to capture a wide range of contextual information. In the Skip connection phase, we propose Skip Connects Feature Attention modules(SFA) to better communicate information from the encoder. In the decoder stage, we innovatively combined the Temporal Fusion Attention Module(TFAM) to effectively fuse the feature information. In addition, we introduced a custom Loss calculation method, Tversky Loss, for the model to achieve faster convergence and improve segmentation along polyp boundaries. Our model was trained on the Kvasir-SEG and CVC-ClinicDB datasets, and validated on datasets Kvasir-SEG, CVC-ColonDB, CVC-300, and ETIS. The results show that the model achieves good segmentation accuracy and generalization performance with a low number of parameters. We are 6.1% ahead in the Kavirs-SEG dataset and 3.1% ahead in the CVC-ClinicDB dataset compared to VM-UNet.

Keywords: CNN; Mamba; Medical Image Segmentation; Vision State Space Models.

1. Introduction

Semantic segmentation is a key processing step used to determine the position, shape, and size of objects in an image. Image semantic segmentation is one of the most popular research topics in computer vision and deep learning. The object segmentation model can handle both 2D images and videos, and can even be extended to 3D scenes. As a result, these models may appear in different practical fields, such as medical imaging, intelligent transportation, and autonomous driving.

In the field of medicine, semantic segmentation, such as polyp segmentation, CT segmentation, blood vessel segmentation, etc., has achieved remarkable success in assisting healthcare applications to segment objects such as tumors, cancers, and lung cancer[1]. Colorectal cancer (CRC) is a leading cause of cancer mortality globally[2]. Most colorectal cancers evolve from adenomatous polyps, making early detection and removal of polyps critical for CRC prevention and treatment[3]. Colonoscopy is the gold standard for detecting and removing polyps before they develop into CRC[5]. Although colonoscopy is an effective screening technique, polyps can be visually identified. But due to differences in the location, size, color and texture of polyps, it can be very draining for medical professionals and can lead to missed or misdiagnosed polyps, which can seriously harm the patient's health. Therefore, it is necessary to use image segmentation technology to develop an automatic and accurate application of polyp segmentation.

Machine learning (ML) algorithms, especially convolutional neural networks (CNNs), have shown promising results in medical image segmentation and have been applied to the detection and segmentation of polyps. While deep learning (DL) algorithms can achieve segmentation with high accuracy, they often require large amounts of labeled data, which can be expensive and very time consuming[6]. To improve the accuracy and efficiency of polyp segmentation, researchers have developed various CNN-based deep learning architectures that employ different techniques to solve this complex task. Deep learning architectures for polyp segmentation include U-Net[7], U-Net++[8], ResUNet++[9]. However, the current model is still limited by some problems: first, the data sets of similar polyps have high similarity in color, contour and background, resulting in low precision of polyp boundary model segmentation. Second, there were significant differences between the polyp image datasets from different classes, including variations in imaging equipment, lighting environment, and image type, which limited the model's ability to generalize[10]. Later, Transformer[11] began to be applied from Natural Language Processing (NLP) to Computer Vision (CV). Visual Transformer (ViT)[12] is a model proposed by Google team in 2020 to apply Transformer to image classification. The model based on ViT has also achieved very good results in the field of polyp segmentation. For example, TransUnet was the first model to use a ViT. It is used for feature extraction in the encoding phase and CNN in the decoding phase, showing strong global information acquisition capability. Although CNN and Transformer based methods can obtain accurate segmentation results, they require a large number of parameters and consume a lot of computing resources. In recent years, the

State Space Model[13] (SSM) has attracted great interest from researchers. On the basis of classical SSM research, modern SSM models such as Mamba[14] not only establish long-distance dependence, but also show linear complexity of input size. In addition, SSM-based models have been extensively studied in many fields, including language understanding, general vision, and so on. Influenced by the success of VMamba[15] in image classification tasks, Ruan et al.[16] proposed a VM-UNet model based on VMamba. This is a pure SSM-based model whose encoder consists of a VSS block in VMamba for feature extraction and a patch merge operation for downsampling. The decoder includes VSS block and patch extension operations to restore the size of the split result.

Based on the above work, we propose DCSS UNet, which uses Mamba encoder as the backbone, uses spatial and channel attention to extract high-level and low-level features, and uses bitemporal features for fusion. We have tested on multiple datasets and achieved excellent results with low reference counts. Our key contributions include:

1) We propose DCSS UNet, which is a potential application of the SSM-based model in medical image segmentation.

2) We propose an improved skip connection module SFA to improve model performance through attention mechanism.

3) A new feature fusion module TFAM is introduced to improve the performance of the decoder structure

4) We combine the structure of Unet in the decoder part, and combine the SSM model with CNN.

2. Related work

In the past decade, with the rapid development of computing power, researchers have invested a lot of energy in establishing various computer-assisted clinical support systems. In terms of polyp segmentation, accurate polyp segmentation is crucial for cancer prevention and early screening. For high-risk polyps, medical image segmentation is needed to accurately formulate preoperative plans and determine the most appropriate surgical methods and treatment strategies.

At present, the mainstream medical image segmentation task usually adopts convolutional neural network (CNN), and some widely used architectures have been applied to polyp segmentation. One such architecture is U-Net[7], an encoder-decoder model originally developed for biomedical image segmentation. U-Net[7] has the advantages of relative simplicity and efficiency, and still has good performance on various medical image segmentation tasks. In 2020, PratNet[17] was proposed by Fan, Deng-Ping, et al. This is a CNN architecture specifically designed for automatic polyp segmentation in colonoscopy images. It uses a parallel partial decoder to extract high-level features from images and generate a global map as a preliminary guide for the following processing steps. The boundary clues are mined by the reverse attention module, and the relationship between different regions and the boundary of the image is established. The above studies do not effectively balance the challenges of achieving high accuracy and generalization in polyp segmentation. In 2018, with the introduction of Transformer, many excellent models have also been proposed in the aspect of polyp segmentation, such as TransUNet[18] and SwinUNet[19]. They get good results, but require a lot of computation. In 2023, with the introduction of Mamba, we have a new breakthrough point in the balance between computation and performance. Based on the above work, we propose the DCSS UNet, which uses the Mamba encoder as the backbone and achieves a good balance between precision and the number of calculated parameters. Below we will take a detailed look at the various parts of the network.

3. Method

3.1. Architecture Overview

The overall architecture of DCSS UNet is shown in Figure 1. It consists of three main modules: VMamba encoder, SFA module and improved decoder.

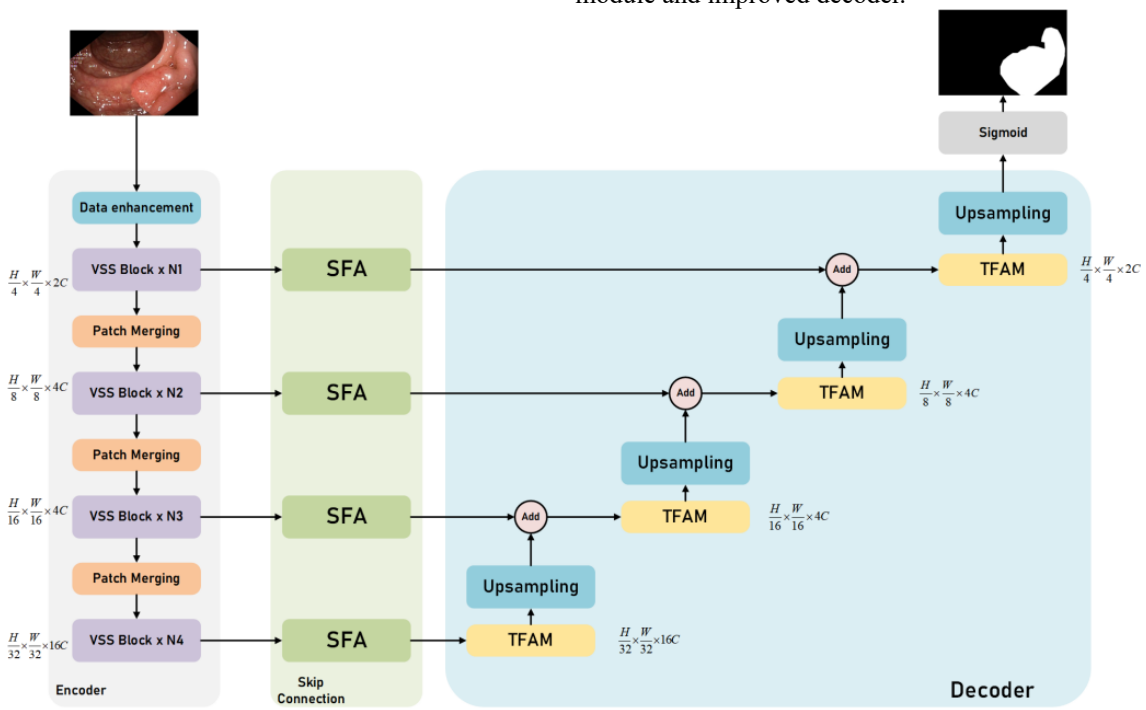

Figure 1. The architecture of DCSS UNet consists of encoder part, jumper connection part and decoder part. The skip connection part uses our proposed SFA to get information from the encoder and pass it to the decoder part. Upgraded upsampling and convolution modules serve as the final level of the network

The VMamba encoder consists of four stages, with a patch merge operation applied at the end of the first three stages to reduce the height and width of input features while increasing the number of channels. We used a [N1, N2, N3, N4] VSS block spanning four stages of the encoder, with channel counts of [C, 2C, 4C, 8C] for each stage. Given an input image I,among $X \in R^{H \times W \times 3}$. It's a feature generated by the encoder at level M. At level i_{th} is f_i^o, where $1 \le i \le M$. These accumulated features $\{f_1^o, f_2^o, ..., f_M^o\}$ are then forwarded to the SFA module for further enhancement. As shown in the figure, the encoder output channel of f_i is $2^i \times C$, which enters the SFA module for feature fusion. The different values of N3 and c are an important factor in distinguishing between minimal, small, and basic framework specifications. According to the VMamba specification, we let C take the value 96, N1 and N2 take the value 2 respectively, and N3 take the value from the set [2,9,27].

Finally, the features are input into our improved decoder network, and the feature fusion is carried out by TFAM, and the original feature map size is restored by upsampling. The sections are described in detail fig 1.

3.2. VSS Block

The encoder core of DCSS UNet uses VMamba's VSS Block. The structure is shown in Figure 2. After layer normalization, the input is divided into two branches. In the first branch, the input passes through a linear layer, followed by an activation function. In the second branch, the input is processed by linear layers, depth-separable convolution, and activation functions before being fed into a two-dimensional selective Scan (SS2D) module for further feature extraction. The features are then normalized using layer normalization, followed by element-level generation using the output from the first branch to merge the two paths. Finally, these features are mixed using a linear layer, and this result is combined with the residual connection to form the output of the VSS block. This article uses SiLU as the activation function by default.

SS2D is the core module in the VSS block, which consists of three parts: the scan expansion operation, the S6 block, and the scan merge operation. The effect of the scan expansion operation is to expand the input image into a sequence in four different directions (top left to bottom right, bottom right to top left, top right to bottom left, bottom left to top right). These sequences are then feature extracted by the S6 block, ensuring that information from different directions is thoroughly scanned to capture different features. Subsequently, the scan merge operation and merge sequence from these four directions restore the output image to the same size as the input image.

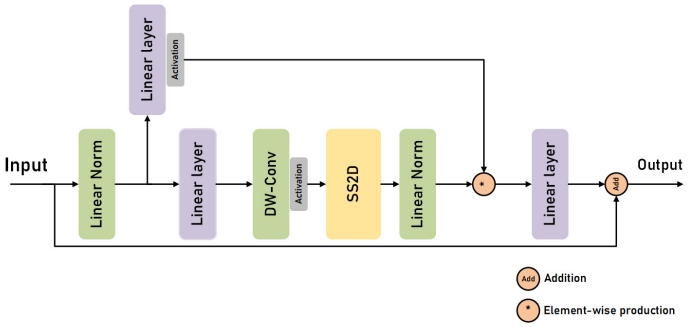

Figure 2. VSS Block as the backbone of DCSS UNet encoder

3.3. SFA Block

In the Skip connect phase, we use the SFA module (Skip Connects Feature Attention modules) to extract the features from the trunk. Based on the idea of CBAM[28], we embed it into the SPA module, whose structure is shown in Figure 3. (a). CBAM (Convolutional Block Attention Module) is a lightweight and versatile module that can obtain more information from the encoder at a very low computational cost, as shown in Figure 3. (b). SFA can be used to calculate spatial and channel attention scores. Given an intermediate feature graph, the SFA module will infer the attention graph along two independent dimensions, channel and space, and then multiply the attention graph by the input feature graph for adaptive feature thinning. The SFA module enhances the network's ability to express local information, extracts more meaningful features, and suppresses unimportant regional responses in a shorter period of time. Helps to locate the target area more precisely. In skip connection, lower feature and higher feature can be integrated better, so as to improve the precision of segmentation. It can effectively suppress the interference of the complex background and ensure the feature of the target area is more prominent.

The process of deriving $CAM(Mc(X))$ and $SAM(Ms(X))$ can be summarized using the following equations:

$$M_c(X) = \sigma(MLP(AvgPool(X)) + MLP(MaxPool(X))) \quad (1)$$

$$M_s(X) = \sigma(Conv([AvgPool(X), MaxPool(X)])) \quad (2)$$

The sigmoid function is represented by σ, and the Input Map is X. SFA Block enhances the network's ability to express local information, extract more meaningful features, and suppress responses from less important regions.

3.4. Improved Upsampling Structure

For the Upsampling block, there are 2 stages in this block. In the first stage, we use TFAM[26](Temporal Fusion Attention Module) to extract feature information from SFA module. We introduced a TFAM to make use of time information for effective feature fusion, as shown in Figure 4. It uses attention and time information to determine what is important between dual temporal features. The relevant calculation formula is as follows:

$$S_C = Concat(Avg(T_1), Max(T_1), Avg(T_2), Max(T_2)) \quad (3)$$

$$W_{c1}, W_{c2} = Conv_1(S_c), Conv_2(S_c) \quad (4)$$

$$W_{c1}^{'}, W_{c2}^{'} = \frac{e^{W_{c1}}}{e^{W_{c1}} + e^{W_{c2}}}, \frac{e^{W_{c2}}}{e^{W_{c1}} + e^{W_{c2}}} \quad (5)$$

$$Output = (W_{c1}^{'} + W_{s1}^{'}) * T_1 + (W_{c2}^{'} + W_{s2}^{'}) * T_2 \quad (6)$$

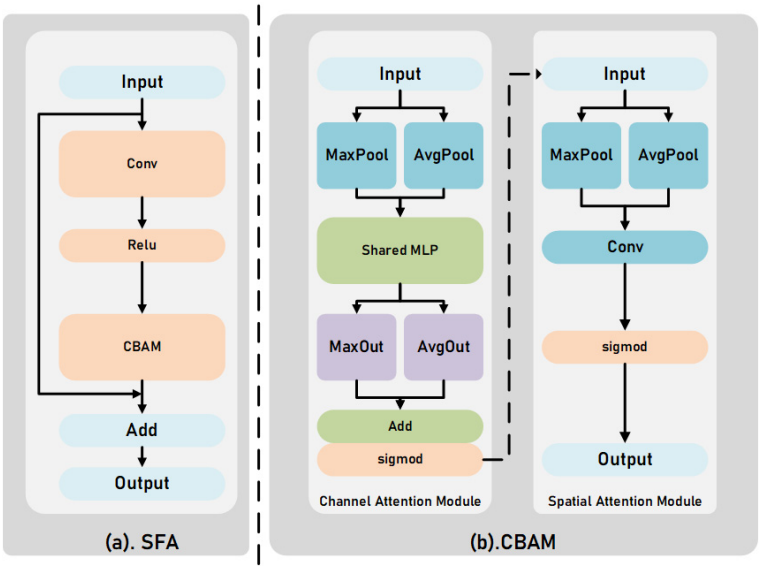

Figure 3. (a). SFA Block. It is embedded with the CBAM module as a new whole for implementing skip connections. **(b).** CBAM Block. It includes channel attention module and spatial attention module

The equation 3 shows that in the channel branch, the dual temporal features of the input aggregate spatial information through global pooling across spatial dimensions. Where S_c denotes the aggregated spatial features, T_1 and T_2 denote bitemporal features, and Avg(·) and Max(·) denote global average pooling and global max-pooling across spatial dimension, respectively. The equation 4 and equation 5 show the aggregate calculation of bitemporal channel weights and bitemporal spatial weights. The equation 6 shows when the sum of bitemporal features is 1, the useful part between the two-time features is retained, and the useless part is discarded, so as to achieve effective feature fusion output. In the second stage, we carry out up-sampling and enlarge the extracted Feature Map to display images with higher resolution.

3.5. Loss function

Dice is the most frequently used metric in medical image competitions. It is a set similarity metric, usually used to calculate the similarity of two samples, and the value threshold is [0, 1]. The number of pixels in the polyp and background is usually unbalanced. The Dice loss function directly measures the degree of overlap between the predicted segmentation result and the real label. It minimizes segmentation errors by maximizing the intersection between predictions and actual labels, and is especially effective for tasks that require precise segmentation. However, it still has some drawbacks, one of which is that it also weighs False Positive (FP) and False Negative (FN) tests. In practical applications, this results in high precision of segmentation maps, but low recall rate.

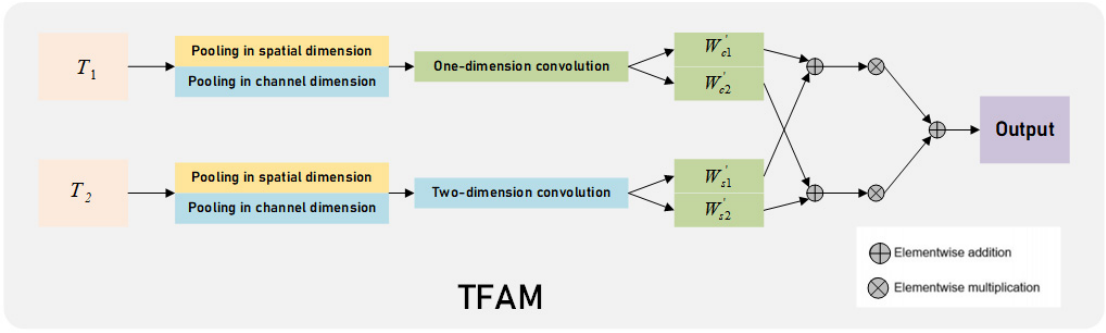

Figure 4. Temporal Fusion Attention Module

To solve similar problems, we introduce a new loss function. The Tversky[27] similarity index is a generalization of the Dice score which allows for flexibility in balancing FP and FNs:

$$TI_C = \frac{\sum_{i=1}^{N} p_{ic} g_{ic} + \varepsilon}{\sum_{i=1}^{N} p_{ic} g_{ic} + \alpha \sum_{i=1}^{N} p_{i\bar{c}} g_{ic} + \beta \sum_{i=1}^{N} p_{ic} g_{i\bar{c}} + \varepsilon} \quad (7)$$

Where, p_{ic} is the probability that pixel i is of the lesion class c and $p_{i\bar{c}}$ is the probability pixel i is of the non-lesion class, $\bar{c}$ . . The same is true for g_{ic} and $g_{i\bar{c}}$, respectively. Hyper parameters α and β can be tuned to shift the emphasis to improve recall in the case of large class imbalance. The Tversky [27] index is adapted to a loss

function (TL) in by minimizing $\sum_C 1 - TI_c$. By using this loss function, we can achieve faster training speeds and higher training accuracy.

4. Experiments

4.1. Dataset

Our experiment dataset followed the experimental setup of parnet: it contains 900 images from the Kvasir-SEG[20] dataset and 550 images from the CVC-ClinicDB[21] dataset. For these datasets, we provide detailed evaluations on several metrics, including Mean Intersection over Union(mIoU), Dice Similarity Coefficient (DSC), Accuracy (Acc), Precision (Pre). In the benchmark testing, we selected four popular datasets: We use the remaining part of the Kvasir-SEG[20], CVC-ClinicDB[21], CVC-ColonDB[22], and ETIS[23] datasets for benchmark testing. To ensure fairness of the data and facilitate accurate evaluation of the model's generalization, all test data is not included in the training data.

Table 1. The description of datasets contains the number of images and the Image's resolution.

Dataset	Resolution	Samples
Kvasir-SEG[20]	332×487~1920×1072	1000
CVC-ClinicDB[21]	384×288	612
CVC-ColonDB[22]	574×500	380
ETIS[23]	1225×966	196

4.2. Data Enhancement

Data enhancement is a key technique to improve the robustness and generalization ability of machine learning models. In this paper, we implement data enhancement on the training set, which significantly improves the generalization ability of the model. The enhanced technologies used include:

- Horizontal and vertical flips.
- Color jitter with a brightness factor uniformly sampled from [0. 6, 1. 6], a contrast of 0. 2, a saturation factor of 0. 1, and a hue factor of 0. 01.
- Affine transforms with rotations of an angle sampled uniformly from [−180°, 180°], horizontal and vertical translations each of a magnitude sampled uniformly from [−0. 125, 0. 125], scaling of a magnitude sampled uniformly from [0. 5, 1. 5] and shearing of an angle sampled uniformly from [−22. 5°, 22. 5°].
- All regions with pixel values of 0 were randomly added to the image, and noise was also added to the regions of the mask at the same location.

4.3. Implementation Details

We use PyTorch as frameworks. All networks were trained with described augmentation methods. We used Adam optimization with an initial learning rate of $1\times e^{-4}$, with a weight decay of $1\times e^{-6}$. Our training took place on an Ubuntu operating system computer with an RTX4090D GPU for accelerated computation.

4.4. Results

We compared some of the most advanced networks available today and achieved the most advanced results. Compared with VM-UNet, which also uses V-Mamba as the backbone, we performed data enhancement and training in the same environment, and achieved a 6.1% lead in the Kavirs-SEG dataset and a 3.1% lead in the CVC-ClinicDB dataset. Compared with other networks of different architectures, we have achieved the same excellent results using a lower computational cost, which is enough to prove the superiority of our network in the architecture and still has the potential for development in the future.

Our network segmentation prediction diagram is shown in Figure 5. In Figure 5. a and b, it can be found that our network can accurately recognize shapes for small defects and small prominencies in the pictures. In Figure 5. c, it can be found that two polyps with very similar color to the background can also be accurately identified at the boundary.

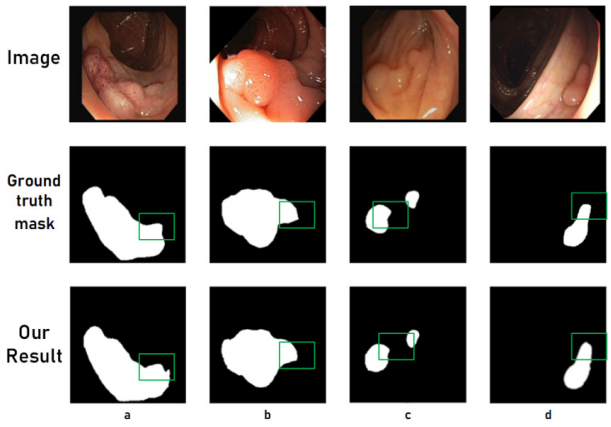

Figure 5. Network segmentation prediction graph

4.5. Ablation Studies

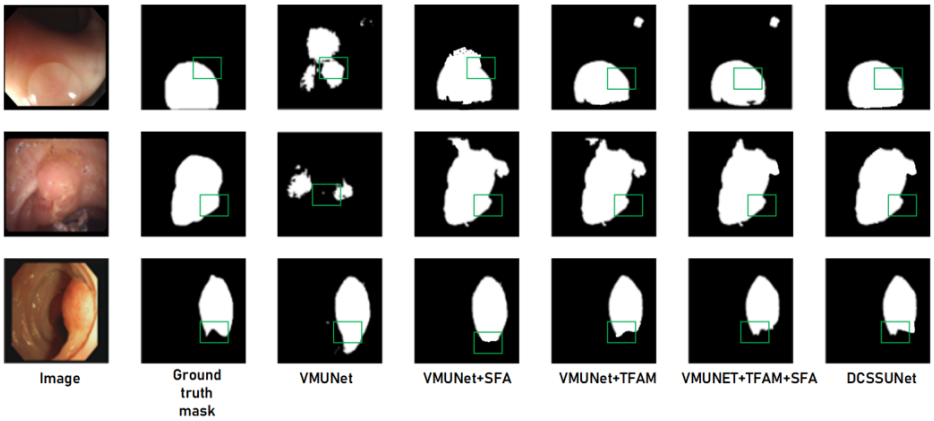

Figure 6. Ablation experiment segmentation prediction map

In order to prove the effectiveness of our work, we verify the role of each module by gradually adding an improved structure on the basis of VMUnet. The results of the ablation experiment are shown in Table 7. The segmentation prediction diagram of ablation experiment is shown in Figure 6.

Table 2. Results of Kavirs-SEG, the best results are shown in bold.

Data	Kavirs-SEG			
Methods	mDice(%)↑	mIoU(%)↑	mPre(%)↑	mRec(%)↑
U-Net(2015)[7]	86. 5	76. 3	85. 9	87. 1
HarDNet(2022)[24]	86. 3	75. 8	93. 5	80. 0
HRNetV2(2019)[25]	85. 3	74. 4	87. 8	83. 0
VM-UNet(2024)[16]	81. 8	69. 2	85. 6	94. 8
DCSS-UNet(Our)	**87. 9**	**78. 4**	**94. 6**	**96. 6**

Table 3. Results of CVC-ClinicDB, the best results are shown in bold.

Data	CVC-ClinicDB			
Methods	mDice(%)↑	mIoU(%)↑	mPre(%)↑	mRec(%)↑
U-Net(2015)[7]	76. 3	61. 6	79. 9	73. 0
HarDNet(2022)[24]	72. 8	57. 2	89. 5	61. 4
HRNetV2(2019)[25]	77. 8	63. 6	82. 6	73. 5
VM-UNet(2024)[16]	86. 8	76. 7	85. 7	97. 9
DCSS-UNet(Our)	**89. 9**	**81. 7**	**93. 3**	**98. 4**

Table 4. Results of ETIS-LaribPolypDB, the best results are shown in bold.

Data	ETIS-LaribPolypDB			
Methods	mDice(%)↑	mIoU(%)↑	mPre(%)↑	mRec(%)↑
U-Net(2015)[7]	79. 8	69. 7	83. 2	77. 2
HarDNet(2022)[24]	**86. 6**	**76. 4**	**97. 0**	78. 2
HRNetV2(2019)[25]	47. 2	30. 9	46. 5	48. 0
VM-UNet(2024)[16]	57. 8	40. 7	50. 0	95. 0
DCSS-UNet(Our)	77. 3	63. 0	74. 7	**97. 9**

Table 5. Results of CVC-ColonDB, the best results are shown in bold.

Data	CVC-ColonDB			
Methods	mDice(%)↑	mIoU(%)↑	mPre(%)↑	mRec(%)↑
U-Net(2015)[7]	**80. 3**	**70. 4**	81. 0	82. 7
HarDNet(2022)[24]	74. 0	58. 7	**95. 0**	60. 6
HRNetV2(2019)[25]	63. 8	46. 9	58. 6	70. 1
VM-UNet(2024)[16]	60. 4	43. 3	64. 4	94. 4
DCSS-UNet(Our)	68. 7	52. 3	82. 8	**96. 0**

Table 6. Comparison of computational complexity and GPU memory usage, using an NVIDIA 4090D GPU (Bold indicates the best).

Model	Input size	Params(M)↓	FLOPs(G)↓
UNet[7]	(3, 256, 256)	28. 15	8. 43
VM-UNet[16]	(3, 256, 256)	34. 62	7. 56
PraNet[17]	(3, 256, 256)	30. 50	13. 15
DCSS-UNet(Our)	(3, 256, 256)	**17. 91**	**4. 43**

Table 7. The role of each module in the ablation experiment,bold indicates the best.

Dataset	Kavirs-SEG	
Model	mDice	mIou
VMUNet	81. 8	69. 2
VMUNet+SFA	83. 6	74. 1
VMUNet+TFAM	85. 5	76. 8
VMUNet+SFA+TFAM	**87. 1**	**77. 9**

We find that simply adding SFA helps the network to find out which places are more worthy of attention, while adding

TFAM makes it easier to integrate the acquired features into the network. Both works together to improve the performance of the network, the increase in computing costs associated with performance improvements is minimal.

5. Conclusion

Based on the results of this paper, the DCSSUNet neural network architecture can achieve advanced performance in the polyp segmentation task of colonoscopy images at a low computational cost. DCSSUNet uses VMamba's encoder as the backbone and improves the decoder structure, the SFA module allows it to focus on the right place, and the TFAM can capture more image information. At the same time, data enhancement techniques help improve its overall performance. The DCSSUNet model has strong generalization ability and can achieve good results even under limited training data. All in all, DCSSUNet architecture shows great application potential in polyp segmentation task. As a practical application of VMamba model using spatial state equation, it is worthy of further study.

Acknowledgments

This work is supported by Natural Science Research Project of Tianjin Education Commission (grant 2020KJ124).

References

[1] Asgari Taghanaki, Saeid, Kumar Abhishek, Joseph Paul Cohen, Julien Cohen-Adad, and Ghassan Hamarneh. (2021) Deep semantic segmentation of natural and medical images: a review. Artificial Intelligence Review 54: 137-178.

[2] Siegel, Rebecca L., Kimberly D. Miller, Hannah E. Fuchs, and Ahmedin Jemal. (2022) Cancer statistics, 2022. CA: a cancer journal for clinicians72, no. 1.

[3] Rock, Cheryl L., Cynthia Thomson, Ted Gansler, Susan M. Gapstur, Marjorie L. McCullough, Alpa V. Patel, Kimberly S. Andrews et al. (2020)American Cancer Society guideline for diet and physical activity for cancer prevention. CA: a cancer journal for clinicians 70, no. 4: 245-271.

[4] Cheng, Jie-Zhi, Dong Ni, Yi-Hong Chou, Jing Qin, Chui-Mei Tiu, Yeun-Chung Chang, Chiun-Sheng Huang, Dinggang Shen, and Chung-Ming Chen. (2016) Computer-aided diagnosis with deep learning architecture: applications to breast lesions in US images and pulmonary nodules in CT scans. Scientific reports 6, no. 1: 24454.

[5] Aasma, Shaukat, Charles J. Kahi, Carol A. Burke, Linda Rabeneck, Bryan G. Sauer, and Douglas K. Rex. (2021)ACG Clinical Guidelines: Colorectal Cancer Screening 2021. The American Journal of Gastroenterology 116, no. 3: 458-479.

[6] Pacal, Ishak, Dervis Karaboga, Alper Basturk, Bahriye Akay, and Ufuk Nalbantoglu. (2020) A comprehensive review of deep learning in colon cancer. Computers in Biology and Medicine 126: 104003.

[7] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. (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 International Publishing.

[8] Zhou, Zongwei, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang., (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 International Publishing.

[9] Jha, Debesh, et al. (2019) Resunet++: An advanced architecture for medical image segmentation. 2019 IEEE international symposium on multimedia (ISM). IEEE.

[10] Nguyen, Dinh Cong, and Hoang Long Nguyen. (2024) PolyPooling: An accurate polyp segmentation from colonoscopy images. Biomedical Signal Processing and Control 92: 105979.

[11] Vaswani, Ashish, et al. (2017) Attention is all you need. Advances in neural information processing systems 30.

[12] Dosovitskiy, Alexey, et al. (2020) An image is worth 16x16 words: Transformers for image recognition at scale."arXiv preprint arXiv:2010.11929.

[13] Kalman, Rudolph Emil. A new approach to linear filtering and prediction problems. (1960): 35-45.

[14] Gu, Albert, and Tri Dao. (2023) Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752.

[15] Zhu, Lianghui, et al. (2024) Vision mamba: Efficient visual representation learning with bidirectional state space model." arXiv preprint arXiv:2401.09417.

[16] Ruan, Jiacheng, and Suncheng Xiang. (2024) Vm-unet: Vision mamba unet for medical image segmentation. arXiv preprint arXiv:2402.02491.

[17] Fan, Deng-Ping, et al. (2020) Pranet: Parallel reverse attention network for polyp segmentation. International conference on medical image computing and computer-assisted intervention. Cham: Springer International Publishing.

[18] Chen, Jieneng, et al. (2021) Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306.

[19] Cao, Hu, et al. (2022) Swin-unet: Unet-like pure transformer for medical image segmentation. European conference on computer vision. Cham: Springer Nature Switzerland.

[20] Jha, Debesh, Pia H. Smedsrud, Michael A. Riegler, Pål Halvorsen, Thomas De Lange, Dag Johansen, and Håvard D. Johansen. (2020) Kvasir-seg: A segmented polyp dataset. In MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part II 26, pp. 451-462. Springer International Publishing.

[21] Bernal, Jorge, F. Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño (2015) WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. Computerized medical imaging and graphics 43: 99-111.

[22] Tajbakhsh N, Gurudu SR, Liang J (2016) Automated polyp detection in colonoscopy videos using shape and context information. IEEE Trans Med Imaging (TMI) 35(2):630–644. https://doi.org/10.1109/TMI.

[23] Silva, Juan, Aymeric Histace, Olivier Romain, Xavier Dray, and Bertrand Granado (2014) Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. International journal of computer assisted radiology and surgery 9 : 283-293.

[24] Liao, Ting-Yu, et al. (2022) HarDNet-DFUS: An enhanced harmonically-connected network for diabetic foot ulcer image segmentation and colonoscopy polyp segmentation. arXiv preprint arXiv:2209.07313.

[25] Sun, Ke, et al. (2019)Deep high-resolution representation learning for human pose estimation.Proceedings of the IEEE/ CVF conference on computer vision and pattern recognition.

[26] Zhao, Sijie, et al. (2023) Exchanging dual-encoder–decoder: A new strategy for change detection with semantic guidance and spatial localization. IEEE Transactions on Geoscience and Remote Sensing 61: 1-16.

[27] Abraham, Nabila, and Naimul Mefraz Khan. (2019) A novel focal tversky loss function with improved attention u-net for lesion segmentation. 2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019). IEEE.

[28] Woo, Sanghyun, et al. (2018) Cbam: Convolutional block attention module." Proceedings of the European conference on computer vision (ECCV).