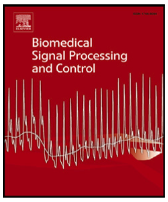
Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

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



SK-VM++: Mamba assists skip-connections for medical image segmentation

Renkai Wu [a,d,1], Liuyue Pan [b,1], Pengchen Liang [a,d], Qing Chang [a,e], Xianjin Wang [c,*], Weihuan Fang [b,*]

[a] Department of General Surgery, Shanghai Key Laboratory of Gastric Neoplasms, Shanghai Institute of Digestive Surgery, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, 200031, China
[b] Department of Radiology, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, 200080, China
[c] Department of Urology, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, 200080, China
[d] School of Microelectronics, Shanghai University, Shanghai, 201800, China
[e] Department of General Medicine, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, 200031, China

ARTICLE INFO

Keywords:
Medical image segmentation
Mamba
UNet
Skip-connections
Deep learning

ABSTRACT

In medical automatic image segmentation engineering, the U-shaped structure is the primary key framework. And the skip-connection operation in it is an important operation for key fusion of high and low features, which is one of the highlights of the U-shaped architecture. However, the traditional U-shaped architecture usually employs direct concatenation or different variants of convolution-based module composition. The recent emergence of Mamba, based on state-space models (SSMs), has shaken up the traditional convolution and Transformers that have long been the foundational building blocks. In this study, we analyze the impact of Mamba on skip-connection operations for U-shaped architectures and propose a novel skip-connection operation (SK-VM++) combining the UNet++ framework and Mamba. Specifically, Mamba is able to refine the fusion of high and low feature information better than traditional convolution. In addition, SK-VM++ leverages the excellent property of Mamba's concatenation, making it significantly less sensitive to changes in computational complexity and parameters caused by changes in the number of channels. In particular, the number of channels increases from 64 to 512, and the convolution-based FLOPs and parameters rise by 8.82 and 6.22 times, respectively, compared to our proposed Mamba-based skip-connection operation. In addition, comparing with the most popular nnU-Net and VM-UNet, the DSC of SK-VM++ improves by 2.01% and 1.10% on the ISIC2017 dataset, 1.59% and 9.10% on the CVC-ClinicDB dataset, 1.23% and 18.94% on the Promise12 dataset and 46.25% and 34.01% improvement on the UWF-RHS dataset. The code is available from https://github.com/wurenkai/SK-VMPlusPlus.

1. Introduction

With the rapid development of computer technology, automatic medical image segmentation technology has been widely used in the medical field [1–4]. Deep learning-based medical image segmentation technology can assist doctors in diagnosis, improve diagnostic accuracy and reduce misdiagnosis rate. It makes excellent medical diagnosis technology available to poor regions at a very low cost [5]. However, current medical image segmentation techniques are usually constructed based on convolution and Vision Transformers (ViT) [6]. Convolution-based models are usually difficult to learn remote feature information in images [7,8]. As for ViT-based methods, researchers proposed Swin-UNet [9], which is the first U-shaped model based on pure Transformers for medical image segmentation. However, ViT-based is limited by the limitation of Transformers, which usually requires more training samples. And the medical field is characterized by sample scarcity, which makes it difficult for ViT-based methods to perform.

Recently, state space models (SSMs) have attracted the interest of many researchers and were introduced to deep learning in [10–12]. SSMs has excellent remote dependency capturing ability, which is inspired by continuous state space models of control systems. In particular, the linear state space layer (LSSL) [10] of SSMs is able to exhibit excellent remote dependency capturing capability. However, its high computational complexity for is usually not applicable to various applications. To address this issue, in S4 [11], researchers proposed to normalize the parameters into a diagonal structure. In addition, S4ND [13] proposed for the first time to apply state-space mechanisms

[*] Corresponding authors.
E-mail addresses: xianjin09@163.com (X. Wang), victorhunter@163.com (W. Fang).
[1] These authors contributed equally.

https://doi.org/10.1016/j.bspc.2025.107646
Received 28 October 2024; Received in revised form 3 January 2025; Accepted 31 January 2025
Available online 11 February 2025

to vision tasks and showed the potential to compete with ViT. In particular, recently, time-varying parameters have been added to SSMs, and Mamba [14] has been proposed, an algorithmic structure claimed to be strongly competitive with current Transformers. In addition, Vision Mamba [15], VMamba [16], and so on, have been proposed based on Mamba for vision infrastructure building blocks and are widely cited in the field of visual recognition.

In the field of medical images, U-Mamba [17] proposed a hybrid convolution and Mamba architecture for medical image segmentation. And in VM-UNet [18], a U-model with pure Vision Mamba was proposed for medical image segmentation for the first time. And in Wu et al. [19], researchers proposed a novel High-order Vision Mamba (H-vmunet) for medical image segmentation. This benefits from the proposed High-order 2D-selective-scan (H-SS2D) significantly reduces the redundant information introduced by SS2D in the global receptive field. In addition, lightweight Mamba-based models [20,21] have been explored for medical image segmentation. In particular, in the present Mamba-based medical image segmentation algorithms are likewise improved based on the UNet framework. This is also the most classical framework for medical image segmentation research. In particular, UNet mainly consists of an encoder, a decoder and a skip-connection. In particular, the skip-connection is able to fuse the high and low feature information well, which makes it able to process the medical image information well. This is due to the fact that in medical images, both high and low feature information are crucial for the target lesions.

However, the importance of the impact of Mamba on skip-connections of U-shaped architectures in the field of medical image segmentation has been overlooked. In this work, we explore the impact of Mamba on skip-connection operations for U-shaped architectures in the field of medical image segmentation and propose a novel Mamba-based skip-connection operation (SK-VM++). Specifically, we combine the Mamba and UNet++ [22] architectures to deeply investigate the impact of Mamba in skip-connections operations. In particular, Mamba has excellent long sequence feature learning capabilities and leverages the powerful high- and low-level context feature collection capabilities of the UNet++ framework to further significantly improve Mamba's performance in skip-connection operations. We conduct detailed experiments on four different types of publicly available medical image segmentation datasets (ISIC2017, CVC-ClinicDB, Promise12, and UWF-RHS) and confirm the effectiveness of Mamba in skip-connection operations. In addition, we also use multi-scale supervised learning in the training process for improving the model's ability to learn different scales of lesions.

Our contributions can be summarized as follows:

- Based on Mamba and U-shaped structures, we explore the effect of Mamba on skip-connections of U-shaped structures.
- A novel Mamba-based skip-connection approach (SK-VM++) is proposed and used for medical image segmentation. We leverage Mamba's excellent long sequence feature learning capability, and the UNet++ framework's robust collection of high and low level contextual features, to further significantly improve Mamba's performance in skip-connection operation.
- The proposed Mamba-based UNet++-like skip-connection approach (SK-VM++) reduces the FLOPs and parameters of conventional UNet++ by 86.90% and 79.01%, respectively. In addition, our approach drastically reduces the impact of the number of channels on the computational load, and achieves a significant improvement in segmentation performance.
- The code of the proposed method will be made public at GitHub.

2. Related work

2.1. Image segmentation

Image segmentation is one of the important tasks in the field of computing. In particular, for computer vision, image segmentation task has been one of the key directions in which a wide range of researchers have been working on. In addition, the application of image segmentation task in various industries is extremely wide. Whereas, more than a decade ago, the research direction of image segmentation techniques was mainly to recognize targets using traditional mathematical methods and processing [23]. They usually used mathematical methods to distinguish regions with distinct thresholds of pixel values. However, traditional methods usually show poor performance when encountering complex image segmentation tasks [24]. And with the breakthrough of computer hardware, the computer computing power has been significantly improved. And at this time, image segmentation techniques based on deep learning came into being. Full Convolution Net (FCN) [25] is the first image segmentation technique using deep learning and has achieved exciting results. The emergence of FCN has made many researchers in image segmentation techniques extremely interested in deep learning. In particular, the emergence of another full convolution net (UNet) [26] has opened up a new direction for medical image segmentation, which combines deep learning and the framework's own leapfrog connectivity, resulting in excellent performance in the field of medical image segmentation.

2.2. Medical image segmentation

In the field of medical image segmentation, UNet has been used as the main framework. Many researchers have made many improvements based on this framework to enhance its performance. In Oktay et al. [27], researchers proposed to add the attention mechanism to the skip-connection operation of the UNet framework to enhance the attention to the lesion during high and low feature fusion. In Aghdam et al. [28], researchers further enhance the sensitivity of Swin-UNet [9] to lesions by introducing an attention mechanism into Swin-UNet. Swin-UNet is the first pure Transformers-based U-model for medical image segmentation. In Azad et al. [2], researchers combine both convolution and Transformers to construct a UNet-like model for medical image segmentation and achieve better performance than both alone. SCR-Net [29] proposes two novel modules for polyp segmentation, which include semantic calibration and semantic refinement modules. In Zhao et al. [30], researchers proposed a subtractive skip-connection method, which is significantly different from the traditional skip-connection using summation. M^{2}SNet [31] proposes to combine different types of convolution kernels for fusion of features during subtractive skip-connection. However, this operation further increases the computational complexity. In Ruan et al. [1], researchers proposed a lightweight medical image segmentation model. They proposed novel Channel Attention Bridge Module and Spatial Attention Bridge Module to fuse the multi-stage information of UNet and achieved better performance at lower number of channels. In Peng et al. [32], the researchers proposed the fusion of multi-stage feature information at different scales using Hadamard multiplication, which operates to better fuse feature information at different scales. C^{2}SDG [33] proposed a feature learning method using novel contrast enhancement. They first augmented the low-level features with a specific number of channels and performed contrast enhancement training. In Wu et al. [34], researchers proposed attentional learning of features under different stages using ViTs with different resolutions. Their experiments confirmed that combining multiple ViTs with different resolutions for attentional learning can effectively improve the generalization ability of the model. MHorUNet [35] proposed a High-order spatial interaction U-model for medical image segmentation. They introduced the novel High-order spatial interaction mechanism into the UNet framework and confirmed its superior generalization ability in an external clinical dataset. Moreover, in some medical image recognition tasks, the robust AI model constructed in [36] can effectively improve the diagnosis of diabetic retinopathy (DR). In [37], researchers developed a novel algorithm to identify patients infected with COVID-19, which novelly combines two-dimensional (2D) curvelet transformation, chaotic salp

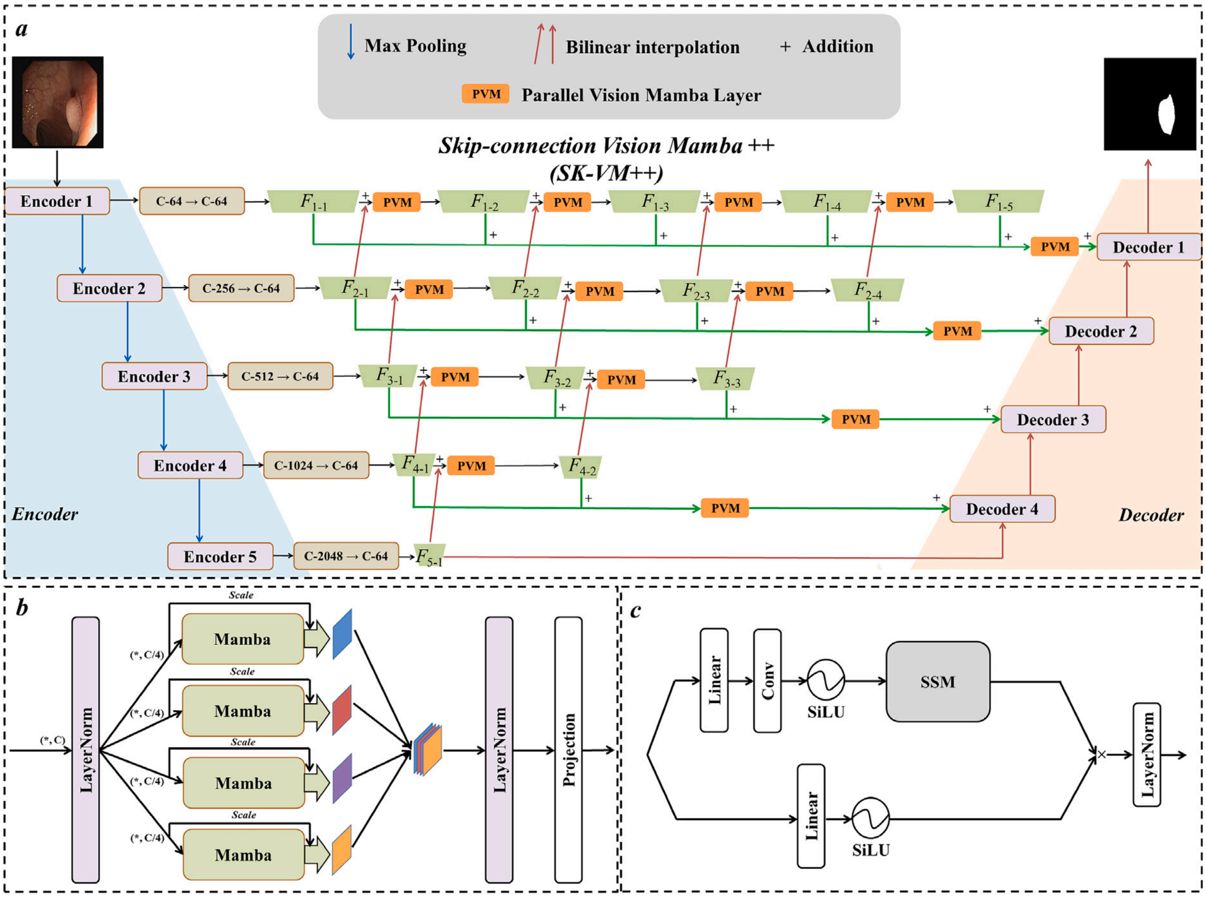


Fig. 1. (a) The proposed Mamba-based skip-connection method (SK-VM++). In particular, the skip-connection part is composed of Parallel Vision Mamba layer (PVM layer) using pure Mamba. The number of channels is standardized to 64 at each layer of encoder feature input to the skip-connection operation to reduce the computational burden. (b) The structure of the Parallel Vision Mamba layer (PVM layer), which performs feature learning after dividing the number of channels equally to reduce the computational burden. (c) The structure of Mamba.

swarm algorithm (CSSA), and deep learning techniques to accurately diagnose COVID-19 cases.

Recently, with the emergence of Vision Mamba [15] and VMamba [16], many researchers have gradually explored the potential of Mamba in medical image segmentation. U-Mamba [17] explores the application of Mamba in medical image segmentation for the first time. It adopts a hybrid approach of convolution and state space for feature extraction. VM-UNet [18] is the first pure Vision Mamba based U-model for medical image segmentation. VM-UNet by pure Vision Mamba shows the strong competitiveness of Mamba with convolution and Transformers in the field of medical image segmentation. In Wu et al. [19], researchers proposed a novel High-order Vision Mamba (H-vmunet) for medical image segmentation. It does so by introducing a High-order 2D-selective-scan (H-SS2D), which guarantees the excellent global sensory field of SS2D while minimizing the introduction of redundant information. In addition, there are some proposals based on Mamba lightweight model [20,21] for medical image segmentation. Numerous previous studies have demonstrated the strong competitiveness of Mamba in the field of medical image segmentation.

However, the importance of Mamba's influence on U-structure skip-connections has been overlooked. In this paper, we explore the impact of Mamba on U-structure skip-connections and propose a novel Mamba-based skip-connection (SK-VM++) operation. The proposed SK-VM++ employs UNet++-like connectivity and uses Mamba to replace the convolution layer of the traditional UNet++ skip-connection. Specifically, we elaborate in the methods section.

3. Method

3.1. Architecture overview

Our proposed model (SK-VM++) for Mamba-based skip-connection operations is shown in Fig. 1(a). Inspired by Zhou et al. [22], we adopt a UNet++-like structure. Among them, we adopt a pure Mamba approach for skip-connection in the skip-connection part. In addition, SK-VM++ consists of encoder, decoder and skip-connection operations, and contains a total of 5 layers of structure. In our study, the SK-VM++ encoder uses the classic Res2Net-50 [38] structure. In order to reduce the computational complexity when the encoder inputs are skip-connection, we unify the number of channels to 64 by a convolution when performing the skip-connection operation. Each layer of the decoder consists of a convolution with a convolution kernel of 3, a batch normalization layer, a Relu activation function, and an up-sampling operation. In addition, to enhance the sensitivity of the model to lesions at different scales, we use LossNet (as in Fig. 2) for multi-scale supervised learning in the training phase.

3.2. Preliminaries

Currently, SSMs-based models, specifically referred to as Mamba (Fig. 1(c)), rely on a linear time-invariant system. Specifically, they are often referred to as linear ordinary differential equations (ODEs). Suppose that the input is $x(t) \in \mathbb{R}^L$, which is mapped to $y(t) \in \mathbb{R}^L$ by means of ODEs. Specifically, ODEs can be expressed by the following equations:

$$h'(t) = Ah(t) + Bx(t) \tag{1}$$

$$y(t) = Ch(t) + Dx(t) \tag{2}$$

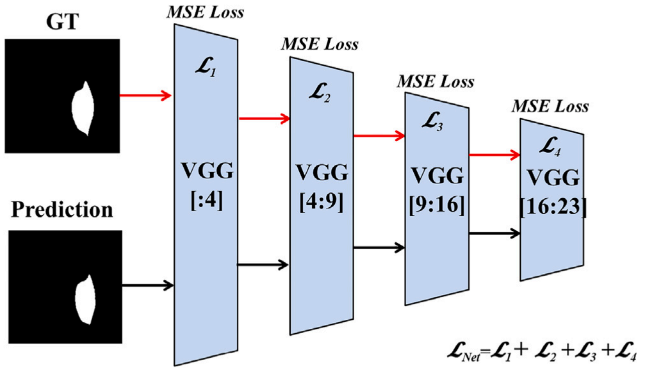


Fig. 2. The structure of LossNet used for multi-scale supervised learning in the training process. The LossNet consists of a VGG-16 partial architecture. It will supervise the refinement of the prediction results from under 4 different scales during the training process.

where $A \in \mathbb{C}^{N\times N}$ denotes the state matrix, $B, C \in \mathbb{C}^{N}$ both denote the projection matrix, N denotes the state size, and $D \in \mathbb{C}$ denotes the skip-connection.

And the above mentioned continuous time based model will be challenging to use with deep learning. So, to overcome this difficulty, people discretize the above continuous time model in Mamba. The discretization enables efficient computational operations during deep learning computations [10]. Specifically, the introduction of time parameters and the discretization rules transform the parameters of the continuous time model into discrete parameters. In particular, continuous parameters A and B become discrete parameters $\bar{A}$ and $\bar{B}$ as in the following equations:

$$\bar{A} = e^{\Delta A} \tag{3}$$

$$\bar{B} = \left(e^{\Delta A} - I\right) A^{-1} B \tag{4}$$

After passing the discretization, the previous ODEs can be discretized using the Zero Order Holding (ZOH) rule:

$$h_k = \bar{A} h_{k-1} + \bar{B} x_k \tag{5}$$

$$y_k = \bar{C} h_k + \bar{D} x_k \tag{6}$$

Deep learning can be efficiently performed by discretizing the system's Mamba, and for more specifics the reader is advised to refer to [14].

3.3. Mamba-based skip-connection: SK-VM++

In this study, we propose a novel skip-connection operation based on Mamba to explore the potential of Mamba in skip-connection. Specifically, we adopt a UNet++-like form of replacing convolution with novel Mamba. In particular, SK-VM++'s skip-connection operation is based on a pure Mamba operation.

In addition, in Wu et al. [21], researchers found that Mamba with parallel operation will greatly reduce the parameter and computational complexity, and proposed the Parallel Vision Mamba Layer (PVM Layer) as shown in Fig. 1(b). The effectiveness of the PVM Layer for visual processing was also confirmed in numerous previous studies [39–41]. According to this, in the proposed SK-VM++, we adopt PVM Layer as a replacement for Mamba. Specifically, the PVM Layer can be expressed as the following equations:

$$x_1, x_2, x_3, x_4 = Sp\left[LN\left(X_{in}\right)\right] \tag{7}$$

$$M_i = Mamba\left(x_i\right) + \theta \cdot x_i \quad i = 1, 2, 3, 4 \tag{8}$$

$$X_{out} = Cat\left(M_1, M_2, M_3, M_4\right) \tag{9}$$

$$y = Pro\left[LN\left(X_{out}\right)\right] \tag{10}$$

where Sp is the splitting operation, which splits the input by the number of channels. x_1, x_2, x_3, x_4 denote the split features, each with one-fourth of the original number of channels. θ denotes the residual factor, which effectively improves the application of Mamba in vision [20,21]. Cat denotes the Concat operation. Pro denotes the projection operation. The basic core of PVM Layer is still Mamba, but it achieves competitive performance with lower parameters and computational complexity of Mamba. Moreover, this advantage mentioned above is more obvious with the increase of channels [21]. This is the reason why our proposed SK-VM++ uses PVM Layer as a carrier for Mamba. This will effectively reduce the computational pressure on the skip-connection part, specifically, as can be derived from our ablation experiments.

The proposed SK-VM++ in the skip-connection part, they will undergo an intensive PVM Layer operation. In particular, the specific number of PVM Layer is determined by the model order. For example, in the first stage, there are 5 PVM Layer operations, where the last PVM Layer is the result of the previous 4 PVM Layer operations. Among them, each of the previous 4 PVM Layer operations is accompanied by a connection layer, which is used to fuse the previous features of the same stage with the up-sampled features of higher stage. The specific operation can be expressed by the following equations:

$$F_{i-j} = \begin{cases} F_{i-1} & j = 1 \\ PVM\left(F_{i-(j-1)} + Up\left(F_{(i+1)-(j-1)}\right)\right) & j > 1 \end{cases} \tag{11}$$

where i denotes the order, j denotes the number of nodes (from left to right for the same stage, with the number of nodes increasing from 1), PVM denotes the use of PVM Layer operation with an batch normalization layer and activation function, and Up denotes the up-sampling operation. Specifically, the first node ($j = 1$) feature of each order is determined only by the input features under the channel transformation of the same stage. And when $j > 1$, each node feature is determined by the previous feature $F_{i-(j-1)}$ of the same stage and the up-sampled feature $F_{(i+1)-(j-1)}$ of the higher stage. In addition, at the end of the skip-connection operation of each stage, the features of all the previous nodes are fused for one PVM Layer operation. By combining UNet++-like with lightweight Mamba blocks (PVM Layer), better performance is achieved with lower computational complexity and parameters of conventional UNet++. Specifically, in the skip-connection part, the SK-VM++ skip-connection part with Mamba has 86.90% and 79.01% lower FLOPs and parameters than the traditional UNet++ with convolution, and the performance is significantly improved (Fig. 5). In particular, this effect will become more pronounced as the number of channels increases (Fig. 6).

In addition, inspired by [30,31], we employ a LossNet for multi-scale supervised learning in the training phase. Specifically, LossNet employs the VGG-16 [42] composition pre-trained in ImageNet [43]. As shown in Fig. 2, is the composition structure of LossNet, which supervised refinement of the prediction results under 4 different scales. In particular, their mean square errors are computed under each scale and finally summed up. Multi-scale supervised learning using LossNet is due to the fact that lesion feature targets in medical images are usually inconsistent in size, for example, there are large area skin lesions and tiny nevus-type lesions in skin lesions. Therefore, multi-scale supervised learning using LossNet can guide the model training from multiple scales.

4. Experiment

4.1. Datasets

We validate the proposed SK-VM++ on three different types of publicly available medical image datasets, including skin lesion, polyp and prostate segmentation datasets.

The skin lesion segmentation dataset was obtained from the ISIC2017 [44] dataset published by the International Skin Imaging Collaboration (ISIC). Specifically, 2000 images and the corresponding labels were acquired from the ISIC2017 dataset. The initial image size is 576 × 767. we randomly divide the training set, validation set and test set into 1250, 150 and 600 images, respectively. The image size is uniformly preprocessed to 256 × 256.

The polyp segmentation dataset was obtained from the CVC-ClinicDB [45] for automated polyp detection provided by MICCAI 2015. Specifically, 612 images were acquired from the CVC-ClinicDB dataset along with the corresponding labels. The initial image size is 576 × 768. we randomly divide the training set, validation set and test set into 429, 62 and 121 images, respectively. The image size is uniformly preprocessed to 256 × 256.

The prostate segmentation dataset was obtained from the ’Prostate MR Image Segmentation’ dataset provided by MICCAI 2012, referred to as Promise12. Specifically, 100 cases along with the corresponding labels were acquired from the Promise12 dataset. Specifically, the training set has 65 cases with a total of 1817 images. The validation set has 15 cases with a total of 355 images. The test set has 20 cases with a total of 554 images. The image size is uniformly preprocessed to 320 × 320.

The Ultra-Wide Fundus Hemorrhage Segmentation dataset was obtained from the large-scale, high-quality Ultra-Wide Fundus Hemorrhage Segmentation dataset (UWF-RHS dataset), which was collected and made publicly available by Wu et al. [39] at the Ruijin Hospital, Shanghai Jiao Tong University School of Medicine. Specifically, the UWF-RHS dataset contains 2580 high-quality retinal hemorrhage images as well as labeled masks. We divided the training set, validation set and test set according to the [39] setting with the ratio 7:1:2. Since the green channel in the fundus image can highlight the fundus hemorrhage feature more strongly and reduce the interference of other factors, we keep the same settings as in [39,46,47] and take only the green channel as input to the model. In addition, the image size is uniformly set to 384 × 384.

4.2. Implementation details

All experiments were implemented based on Python 3.8 and Pytorch 1.13.0. A single NVIDIA V100 GPU with 32 GB of memory supported our experiments. In the training phase, all experiments used the same data enhancement operations, which included vertical and horizontal flipping, as well as random flipping operations. Implementing the same data enhancement operations can help the model adapt to different situations in order to improve the generalization ability of the model, which has been widely verified in numerous previous studies [1,2,18,28,34,35]. Input image sizes were used depending on the dataset, with 256 × 256 size for skin lesion and polyp segmentation and 320 × 320 size for prostate segmentation. Resizing the image can significantly improve computational efficiency, and because neural networks are scale invariant, key features in the image can still be learned efficiently without affecting the semantic key features [48,49]. The loss function consists of IOU loss, BceDice loss and LossNet. The training epoch was set to 250 and the batch size was 8. AdamW was used as the optimizer for training and the initial learning rate was set to 0.001. A cosine annealing learning scheduler was used in the training phase with a minimum learning rate of 0.00001.

4.3. Evaluation metrics

In the field of medical image segmentation, dice similarity coefficient (DSC), sensitivity (SE), specificity (SP) and accuracy (ACC) are the most important evaluation metrics [1,2,35]. DSC is mainly used for evaluating the similarity between predicted and true values. SE mainly measures the percentage of true positives among true positives and false negatives. SP mainly measures the percentage of true negatives among true negatives and false positives. ACC is mainly used for evaluating the percentage of correct classification. The above evaluation metrics can be expressed in the following equations:

$$\mathrm{DSC} = \frac{2\mathrm{TP}}{2\mathrm{TP}+\mathrm{FP}+\mathrm{FN}} \tag{12}$$

$$\mathrm{ACC} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \tag{13}$$

$$\mathrm{Sensitivity} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{14}$$

$$\mathrm{Specificity} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} \tag{15}$$

where TP is true positive, TN is true negative, FP is false positive and FN is false negative.

4.4. Comparison results

In order to demonstrate the superiority of the performance of the proposed SK-VM++, we conducted comparison experiments with 14 most classical and advanced medical image segmentation models. Specifically, they include UNet [26], nnU-Net [50], Att U-Net [27], ATTENTION SWIN U-Net [28], TransNorm [2], SCR-Net [29], MSNet [30], M^{2}SNet [31], MALUNet [1], UNet v2 [32], C^{2}SDG [33], META-Unet [34]. MHorUNet [35] and VM-UNet [18].

In addition, in order to deeply explore the impact of Mamba on skip-connection operations, the proposed SK-VM++ adopts two different forms to accommodate clinically different computational consumption requirements. In particular, the encoder of SK-VM++ (L) utilizes the Res2Net-50 [38] structure without pre-training. The main architecture of SK-VM++ (S) is a pure Mamba model, in which both the encoder and decoder use the ultra-lightweight module (PVM Layer) [21] for feature learning.

Tables 1, 2, 3 and 4 show the results of the comparison experiments in the skin lesion segmentation, polyp segmentation, prostate segmentation and ultra-wide fundus hemorrhage segmentation tasks, respectively. From the table, it can be concluded that the proposed SK-VM++ shows the best performance in all f our medical image segmentation datasets in several evaluation metrics. In particular, the performance excellence of the proposed SK-VM++ (L) is even more evident in the polyp, prostate and ultra-wide fundus hemorrhage segmentation datasets. This is due to the fact that these three tasks are more complex than the skin lesion segmentation task, which can significantly highlight the difference in model performance. Specifically, the DSC values of the proposed SK-VM++ (L) are 9.10%, 18.94% and 34.01% higher than the traditional Mamba-based VM-UNet on the polyp, prostate and ultra-wide fundus hemorrhage segmentation datasets, respectively. In addition, the proposed performance-efficiency trade-off SK-VM++ (S) model also improves 7.24%, 14.93% and 18.47% over VM-UNet on polyp, prostate and ultra-wide fundus hemorrhage segmentation datasets, respectively. As shown in Fig. 3 are the visualization results in the four public datasets. In particular, as can be seen from the figure, the polyp and prostate segmentations have more blurred boundaries, and the ultra-wide fundus hemorrhagic segmentation is characterized by a wide distribution and small lesions. As shown in Fig. 3 for cases A3, B3, C3 and Dx, they all have more ambiguous and tougher segmentation goals. And the proposed SK-VM++ is a leader in both boundary recognition and lesion segmentation completeness. However, the advantages of SK-VM++ in lesion boundary recognition and completeness are more difficult to be demonstrated when dealing with ultra-wide fundus hemorrhage segmentation tasks, especially when the lesions are widely distributed and tiny like D1 and D2 cases. This is also determined by the specificity of the ultra-wide fundus hemorrhage segmentation task, yet SK-VM++ still visually outperforms most of the comparison models for D1 and D2 in Fig. 3. This shows that the skip-connection operation combined with Mamba is able to better fuse the fine features. This is due to the fact that Mamba it is

Table 1
Performance comparison on ISIC2017 dataset.

Methods	Segmentation target	Pramas[↓]	GFLOPs[↓]	DSC[↑]	SE[↑]	SP[↑]	ACC[↑]
UNet [26]	Skin lesions	2.009	3.224	0.8159	0.8172	0.9680	0.9164
nnU-Net [50]	Skin lesions	29.967	16.245	0.8989	0.8871	0.9807	0.9638
Att U-Net [27]	Skin lesions	3.581	8.574	0.8082	0.7998	0.9776	0.9145
ATTENTION SWIN U-Net [28]	Skin lesions	46.910	14.181	0.8859	0.8492	0.9847	0.9591
TransNorm [2]	Skin lesions	117.676	39.284	0.8933	0.8535	0.9859	0.9582
SCR-Net [29]	Skin lesions	0.801	1.567	0.8898	0.8497	0.9853	0.9588
MSNet [30]	Skin lesions	29.741	8.994	0.9067	0.8771	0.9860	0.9647
M^{2}SNet [31]	Skin lesions	29.742	9.026	0.9139	0.8779	**0.9894**	0.9676
MALUNet [1]	Skin lesions	**0.175**	**0.083**	0.8896	0.8824	0.9762	0.9583
UNet v2 [32]	Skin lesions	25.034	5.399	0.9149	**0.9052**	0.9821	0.9670
C^{2}SDG [33]	Skin lesions	22.010	7.972	0.8938	0.8859	0.9765	0.9588
META-Unet [34]	Skin lesions	22.209	5.139	0.9068	0.8801	0.9836	0.9639
MHorUNet [35]	Skin lesions	9.585	0.864	0.9132	0.8974	0.9834	0.9666
VM-UNet [18]	Skin lesions	44.274	7.562	0.9070	0.8837	0.9842	0.9645
SK-VM++ (S, Ours)	Skin lesions	2.528	0.886	0.9076	0.8791	0.9859	0.9651
SK-VM++ (L, Ours)	Skin lesions	29.368	7.699	**0.9170**	0.8881	0.9882	**0.9686**

Table 2
Performance comparison on CVC-ClinicDB dataset.

Methods	Segmentation target	Pramas[↓]	GFLOPs[↓]	DSC[↑]	SE[↑]	SP[↑]	ACC[↑]
UNet [26]	Polyps	2.009	3.224	0.9039	0.8926	0.9914	0.9821
nnU-Net [50]	Polyps	29.967	16.245	0.9154	0.8991	**0.9966**	0.9864
Att U-Net [27]	Polyps	3.581	8.574	0.8697	0.8474	0.9894	0.9760
ATTENTION SWIN U-Net [28]	Polyps	46.910	14.181	0.7526	0.6503	0.9943	0.9631
TransNorm [2]	Polyps	117.676	39.284	0.8845	0.8634	0.9918	0.9801
SCR-Net [29]	Polyps	0.801	1.567	0.8951	0.8701	0.9922	0.9807
MSNet [30]	Polyps	29.741	8.994	0.9050	0.8720	0.9932	0.9832
M^{2}SNet [31]	Polyps	29.742	9.026	0.9092	0.9064	0.9909	0.9829
MALUNet [1]	Polyps	**0.175**	**0.083**	0.8562	0.8475	0.9862	0.9731
UNet v2 [32]	Polyps	25.034	5.399	0.9055	0.8861	0.9930	0.9825
C^{2}SDG [33]	Polyps	22.010	7.972	0.8967	0.8724	0.9923	0.9810
META-Unet [34]	Polyps	22.209	5.139	0.8975	0.8768	0.9919	0.9811
MHorUNet [35]	Polyps	9.585	0.864	0.8930	0.8803	0.9904	0.9801
VM-UNet [18]	Polyps	44.274	7.562	0.8524	0.8370	0.9867	0.9726
SK-VM++ (S, Ours)	Polyps	2.528	0.886	0.9141	0.8888	0.9942	0.9842
SK-VM++ (L, Ours)	Polyps	29.368	7.699	**0.9300**	**0.9091**	0.9952	**0.9871**

Table 3
Performance comparison on Promise12 dataset.

Methods	Segmentation target	Pramas[↓]	GFLOPs[↓]	DSC[↑]	SE[↑]	SP[↑]	ACC[↑]
UNet [26]	Prostates	2.009	5.037	0.8278	0.7874	0.9980	0.9944
nnU-Net [50]	Prostates	29.967	25.383	0.8961	0.8761	0.9986	0.9966
Att U-Net [27]	Prostates	3.581	13.398	0.8635	0.8501	0.9979	0.9955
ATTENTION SWIN U-Net [28]	Prostates	46.945	22.175	0.8112	0.7407	**0.9988**	0.9945
TransNorm [2]	Prostates	117.676	64.406	0.8615	0.8395	0.9982	0.9956
SCR-Net [29]	Prostates	0.801	2.449	0.8808	0.8742	0.9980	0.9960
MSNet [30]	Prostates	29.741	14.054	0.8856	0.8551	0.9987	0.9962
M^{2}SNet [31]	Prostates	29.742	14.103	0.8967	0.8866	0.9984	0.9965
MALUNet [1]	Prostates	**0.178**	**0.130**	0.7840	0.7611	0.9969	0.9929
UNet v2 [32]	Prostates	25.034	8.436	0.8566	0.8252	0.9982	0.9953
C^{2}SDG [33]	Prostates	22.010	12.457	0.8447	0.8230	0.9978	0.9948
META-Unet [34]	Prostates	22.209	8.030	0.7920	0.7550	0.9973	0.9933
MHorUNet [35]	Prostates	9.585	1.350	0.8600	0.8364	0.9981	0.9954
VM-UNet [18]	Prostates	44.274	11.816	0.7627	0.7115	0.9974	0.9925
SK-VM++ (S, Ours)	Prostates	2.528	1.385	0.8766	0.8604	0.9971	0.9951
SK-VM++ (L, Ours)	Prostates	29.368	12.030	**0.9071**	**0.8963**	0.9986	**0.9969**

different from convolution and Transformers in that it has a selective scanning mechanism and is concerned with changes in subtle features in a sequential manner [14]. Whereas traditional skip-connection are more likely to use different variants of convolution modules.

In particular, among the comparison models Att U-Net, ATTENTION SWIN U-NET, MSNet, M^{2}SNet, MALUNet, UNet v2 and so on. They all propose their own advanced skip-connection operations, and comparing our novel Mamba-based skip-connection operations, we outperform all other current comparison models. In addition, we will further confirm the effectiveness of Mamba in skip-connection in the next ablation experiments.

4.5. Ablation experiment

In order to verify the effect of Mamba on skip-connection operations at different scales, we conducted ablation experiments. The experimental setup is shown in Fig. 4(a). We add Mamba into the traditional UNet++ structure stage by stage, where Baseline denotes that the skip-connection is the traditional UNet++ structure, S_1 denotes that the convolution of the skip-connection operation in the first stage of the model is replaced by the Mamba-based PVM Layer, and S_{1-4} denotes that the convolution of the skip-connection operation in the first to fourth stages of the model is replaced by the Mamba-based

Table 4
Performance comparison on UWF-RHS dataset.

Methods	Segmentation target	Pramas[↓]	GFLOPs[↓]	DSC[↑]	SE[↑]	SP[↑]	ACC[↑]
UNet [26]	UWF Hemorrhage	2.009	7.211	0.6196	0.5013	0.9978	0.9886
nnU-Net [50]	UWF Hemorrhage	29.967	36.551	0.4534	0.3023	**0.9994**	0.9868
Att U-Net [27]	UWF Hemorrhage	3.581	19.251	0.5545	0.4431	0.9971	0.9891
ATTENTION SWIN U-Net [28]	UWF Hemorrhage	46.983	31.965	0.3448	0.2359	0.9984	0.9872
TransNorm [2]	UWF Hemorrhage	117.676	88.468	0.3096	0.2052	0.9989	0.9874
SCR-Net [29]	UWF Hemorrhage	0.801	3.484	0.6434	0.5693	0.9963	0.9985
MSNet [30]	UWF Hemorrhage	29.741	20.216	0.5455	0.4099	0.9982	0.9876
M^{2}SNet [31]	UWF Hemorrhage	29.742	20.287	0.6597	0.5834	0.9968	0.9988
MALUNet [1]	UWF Hemorrhage	**0.178**	**0.165**	0.4760	0.3822	0.9958	0.9847
UNet v2 [32]	UWF Hemorrhage	25.034	12.091	0.4644	0.3540	0.9968	0.9852
C^{2}SDG [33]	UWF Hemorrhage	22.010	17.706	0.6487	0.5500	0.9973	0.9892
META-Unet [34]	UWF Hemorrhage	22.209	11.332	0.5631	0.4635	0.9966	0.9869
MHorUNet [35]	UWF Hemorrhage	9.585	1.922	0.4480	0.3224	0.9978	0.9856
VM-UNet [18]	UWF Hemorrhage	44.274	16.987	0.4948	0.3673	0.9978	0.9864
SK-VM++ (S, Ours)	UWF Hemorrhage	2.528	1.973	0.5862	0.4931	0.9969	0.9882
SK-VM++ (L, Ours)	UWF Hemorrhage	29.368	17.302	**0.6631**	**0.5911**	0.9965	**0.9891**

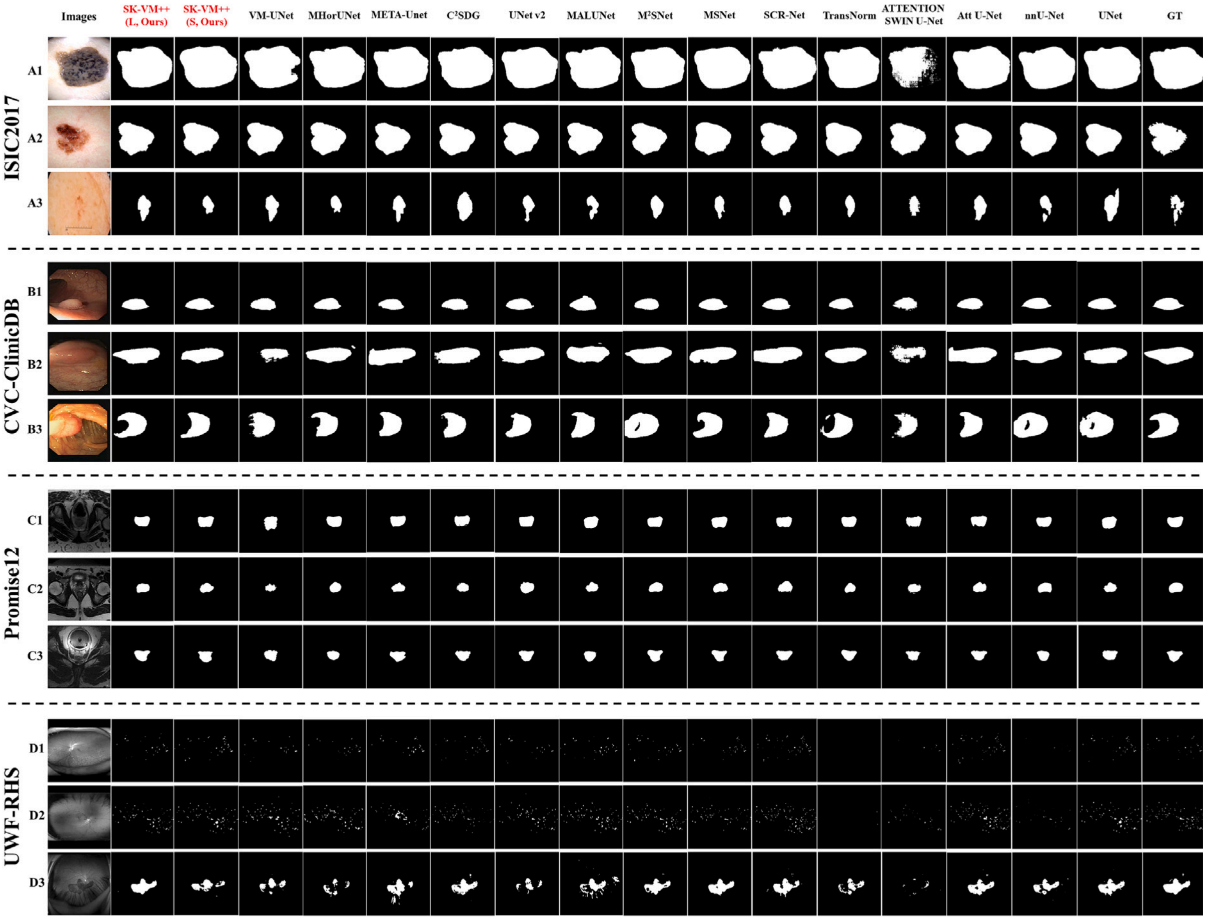


Fig. 3. Visualization of segmentation comparison results on four publicly available medical image segmentation datasets. Ax represent skin lesion segmentation, Bx represent polyp segmentation, Cx represent prostate segmentation, and Dx represent ultra-wide fundus hemorrhage segmentation.

PVM Layer, and so on. It can be learned from Fig. 4(a) that the DSC value is increasing as Mamba is added stage by stage. This shows that Mamba is able to perform the skip-connection operation better than the traditional convolution.

In particular, we also visualize in Fig. 5 the intermediate feature maps using convolution and Mamba in the skip-connection operation. We can see that since Mamba it is based on sequences for scanning selection mechanism to extract features. So, Mamba based skip-connection shows more fine grained feature extraction in multiple intermediate feature maps. Whereas the convolution approach to feature extraction tends to lose the fine-grained features, which for segmentation tasks with fuzzy boundaries, this will result in not being able to handle the boundary information well. In particular, the example skin lesion segmentation in Fig. 5, which also has a blurred boundary. And the Mamba-based skip-connection operation is able to handle the fuzzy boundary better, which once again confirms the effectiveness of our method.

In addition, replacing all Mamba-based PVM Layer with convolution increases the computational complexity (FLOPs) and parameters by 86.90% and 79.01%, respectively. Specifically, since Mamba-based skip-connection operation is our main highlight, we only consider the

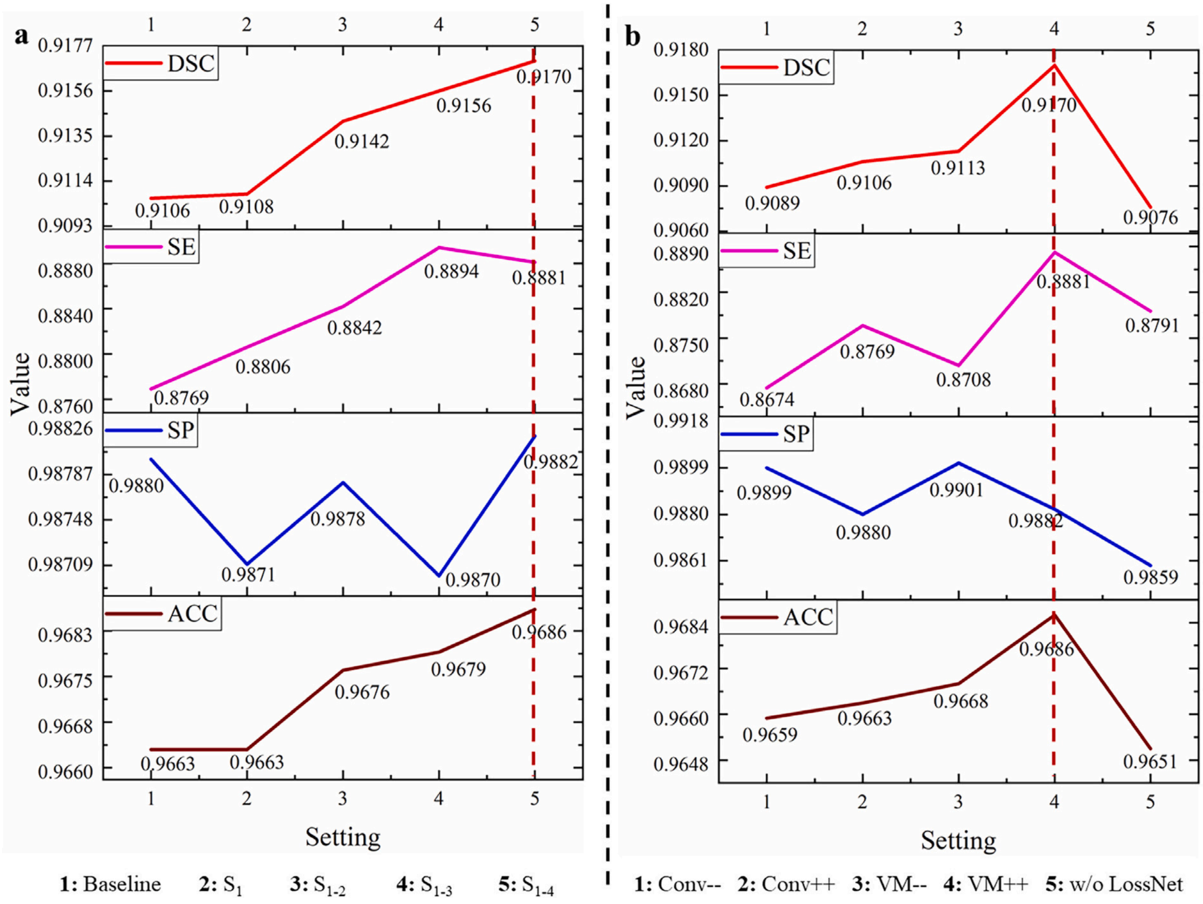


Fig. 4. (a) Ablation experiments on the effect of Mamba on skip-connection operations at different scales. (b) Ablation experiments with different skip-connection operations and settings.

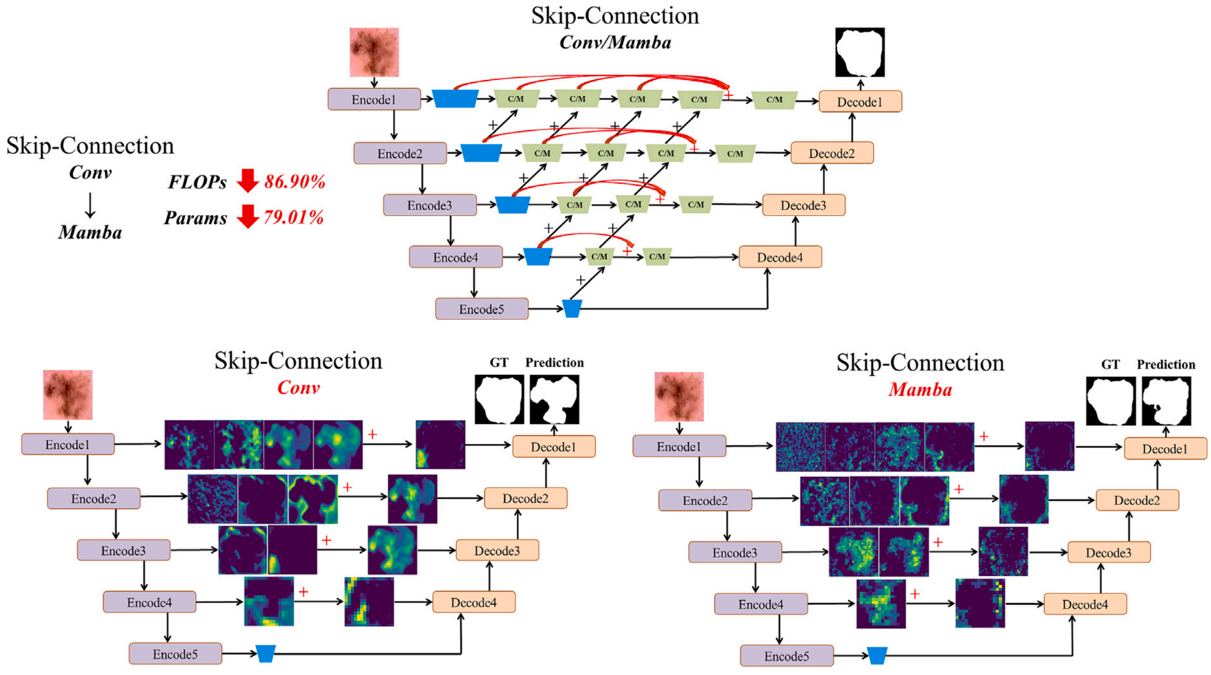


Fig. 5. Visualization of intermediate feature maps for skip-connection operations in UNet++ using convolution and Mamba.

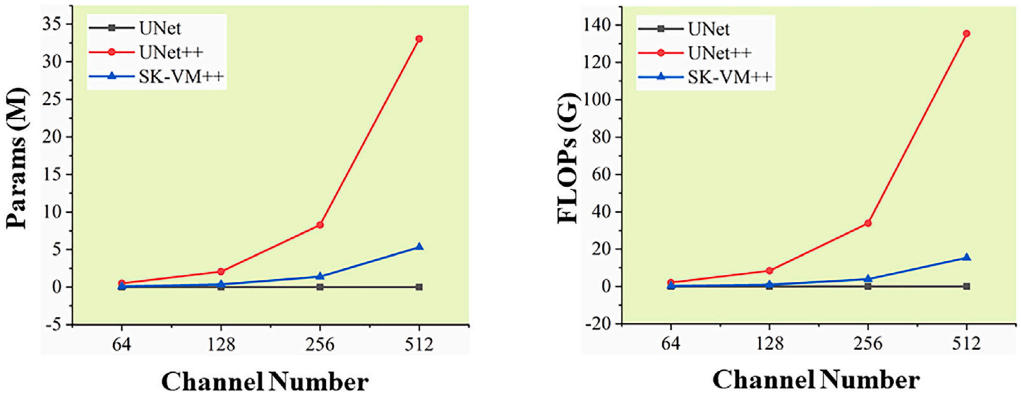


Fig. 6. The effect of channel number variation on the computational complexity (FLOPs) and parameters of Convolution-based and Mamba-based skip-connection operations. In particular, the skip-connection of a conventional UNet does not introduce any computational complexity or parameters, which are all zero.

computational complexity and parameter changes induced by the skip-connection operation separately. In particular, as shown in Fig. 6, the Flops and parameters of the convolution-based UNet++ increase rapidly as the number of channels increases, with the number of channels increasing from 64 to 512, the Flops increasing from 2.1286G to 135.410G, and the parameters increasing from 0.5189M to 33.051M. However, the Mamba-based SK-VM++, with the number of channels increasing from 64 to 512 increases the Flops by only 15.1061G and the parameters by only 5.227M. Comparing with our proposed SK-VM++, the FLOPs and parameters of the convolution-based UNet++ increase by 8.82 times and 6.22 times, respectively.

What is more, ablation experiments were conducted to verify the impact of Mamba's skip-connection operation in different frameworks. In [30,31], researchers proposed that subtraction networks also have better performance. So, as shown in Fig. 4(b), we performed ablation experiments by setting up additive and subtractive networks with convolution and Mamba. Where Conv++ i.e. the traditional UNet++ and VM++ is our proposed SK-VM++. Through Fig. 4(b), the combination of Mamba's additive operation can provide better fusion of high and low level features.

In addition, in order to verify the impact of employing LossNet in the model for multi-scale supervised learning, we also conducted ablation experiments. As shown in Fig. 4(b), the model performance gets significantly degraded after removing LossNet during the training process. This is due to the fact that the target features of medical images are usually irregular in size, and the use of multi-scale supervised learning in the training process can effectively improve the model's ability to learn lesions at different scales.

5. Discussion

In this study, the impact of Mamba's skip-connection operation in U-shaped models is discussed in detail. The skip-connections incorporating Mamba are able to better fuse high and low level features. This is possible due to Mamba's novel state-space model (SSM), which is capable of focusing on subtle feature changes during the fusion of high and low level features from a sequence perspective through a scanning selection mechanism (Fig. 5). The starting point for exploring Mamba in skip-connection operations in comparison to the building blocks of currently existing work (Transformer and Convolution) is that Mamba has the advantage of better processing of long sequences, a more efficient selectivity mechanism and lower computational complexity and parameters [14,16]. Specifically, Mamba is able to optimize the computational process with hardware-aware algorithms during long sequence processing, whereas Transformer usually faces efficiency challenges in long sequence processing. In addition, Mamba allows the model to selectively focus on or ignore certain inputs at each time step when processing features in the long sequence space, whereas Transformer needs to process all time steps at each time step. And Convolution is difficult to learn the remote feature information relationship in the image when processing long sequence feature information. Comparing to natural scene images, medical images are more focused and their long sequence feature information relations in both high and low stages can guide the lesion feature learning, so the processing of long sequence spatial features is critical. For skip-connection operations, in particular, the UNet++-like architecture combines feature information from different high and low stage contexts to achieve superior feature learning, so the processing of long sequence spatial features is critical. To summarize, the exploration of Mamba in skip-connection operation becomes an imperative process. Further, in [21], researchers explored the lightweight impact of Mamba, which possesses lower parameter variation and computational complexity. This is also imperative for exploring in skip-connection operations to mitigate the computational burden of traditional skip-connection operations by introducing Mamba. We validate the effectiveness of the Mamba-based skip-connection operation model (SK-VM++) on four publicly available medical image datasets, outperforming the traditional convolution-based and Transformer-based models on several metrics.

In particular, the Mamba-based skip-connection operation is able to handle boundary ambiguity data well. Specifically, polyp and prostate datasets, they usually have fuzzy boundaries. And the proposed Mamba-based SK-VM++ significantly outperforms other comparative models in several evaluation metrics. We have discussed in detail Fig. 3 in the comparison results subsection in response to the more ambiguous and more daunting segmentation goals. In this subsection, we specifically visualize intermediate feature maps of skin lesions with fuzzy boundaries (Fig. 5). We can clearly see that the convolution-based model with the same framework is less sensitive to fuzzy boundaries. However, the Mamba-based model can be seen from each intermediate feature map, which is based on Mamba from a more refined perspective focusing on the changes in subtle features during fusion. This can significantly improve the model's ability to fuse high and low features, which in turn improves the model's segmentation performance.

In addition, better segmentation performance can provide more accurate quantitative results for clinical diagnosis, which is one of the key points for the model to assist clinical applications, as shown in Fig. 7. The model can also be used in the clinical diagnosis of the patient. Specifically, accurate segmentation of skin lesions, especially for those with ambiguous boundaries, provides a more accurate relative area (percentage). Subsequently, by further introducing a standard scale, the corresponding area of the skin lesion can be calculated. Accurate quantitative calculation of the area of the lesion can be clinically helpful in determining the nature and severity of the skin lesion, and provide quantitative results for follow-up changes in the quantitative size of the lesion and treatment options [51,52]. In addition, accurate segmentation of polyps can help to determine the

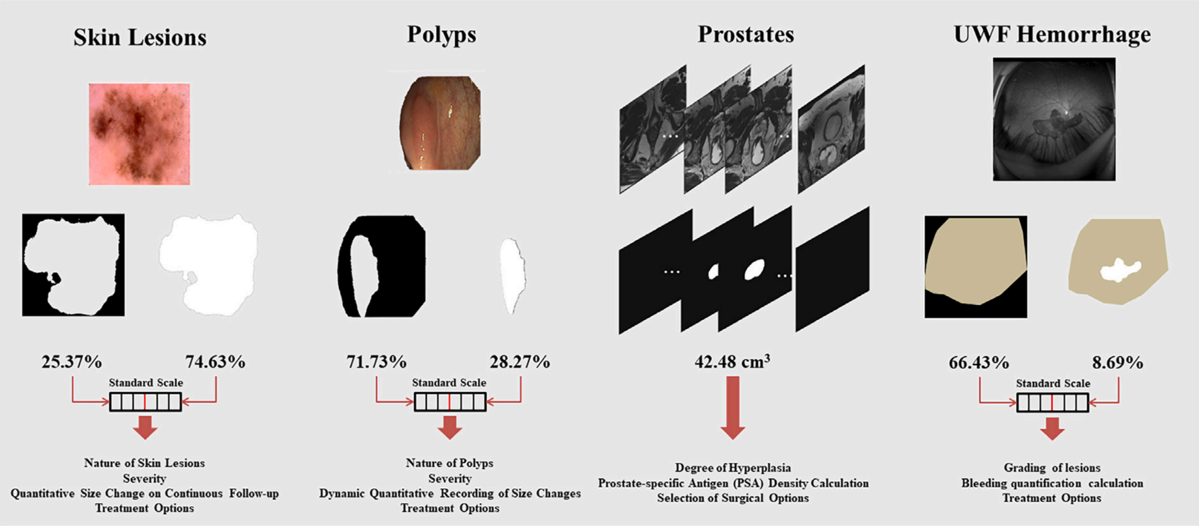


Fig. 7. Accurate segmentation can provide quantitative results for clinical diagnosis and assist in clinical decision making.

severity of polyps in clinical practice, quantify the size of the changes in dynamics, and provide quantitative results for treatment options [53,54]. Furthermore, accurate segmentation of the prostate can provide quantitative results in determining the degree of hyperplasia, calculating prostate-specific antigen (PSA) density, and selecting surgical options [55,56]. In addition, automated segmentation of ultra-wide fundus hemorrhages can reduce ophthalmologist stress [39] due to the wide and fine distribution of ultra-wide fundus hemorrhages. In particular, the automated segmentation of ultra-wide fundus hemorrhages enables an initial screening classification for early nonproliferative DR, based on the number of quadrant closure hemorrhages [57]. Therefore, accurate segmentation of lesions is one of the key points for model-assisted clinical application, and skip-connection operation is one of the key points for U-shaped structures to improve segmentation accuracy.

6. Conclusion

In this study, we analyze the effect of Mamba on the skip-connection operation of U-shaped models and propose a novel skip-connection method (SK-VM++) based on Mamba. Specifically, we combine the Mamba and UNet++ [22] architectures to deeply investigate the impact of Mamba in skip-connections operations. In particular, Mamba has excellent long sequence feature learning capabilities and leverages the powerful high- and low-level context feature collection capabilities of the UNet++ framework to further significantly improve Mamba's performance in skip-connection operations. In addition, the proposed SK-VM++ has lower computational complexity and parameters than the conventional convolution-based skip-connection operation UNet++. In particular, SK-VM++ has lower sensitivity to changes in computational complexity and parameters caused by changes in the number of channels. Specifically, comparing our proposed Mamba-based skip-connection operation, the number of channels increases from 64 to 512, and the convolution-based FLOPs and parameters rise by 8.82 and 6.22 times, respectively, compared to SK-VM++. Furthermore, we demonstrate the effectiveness of our method on four different types of medical image segmentation datasets (ISIC2017, CVC-ClinicDB, Promise12 and UWF-RHS). Through our study, Mamba may become a novel skip-connection base building block in the future.

CRediT authorship contribution statement

Renkai Wu: Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Conceptualization. **Liuyue Pan:** Writing – review & editing, Validation, Methodology, Conceptualization. **Pengchen Liang:** Writing – review & editing, Visualization, Validation, Data curation. **Qing Chang:** Writing – review & editing, Resources. **Xianjin Wang:** Supervision, Project administration, Funding acquisition. **Weihuan Fang:** Supervision, Project administration, Funding acquisition.

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.

Acknowledgments

This work was supported partly by Medicine-engineering Interdisciplinary Project set up by University of Shanghai for Science and Technology (2023-LXY-RUIJIN01Z).

Data availability

The proposed methodology code is provided in the last sentence of the abstract.

References

[1] J. Ruan, S. Xiang, M. Xie, T. Liu, Y. Fu, MALUNet: A multi-attention and light-weight unet for skin lesion segmentation, in: 2022 IEEE International Conference on Bioinformatics and Biomedicine, BIBM, IEEE, 2022, pp. 1150–1156.
[2] R. Azad, M.T. Al-Antary, M. Heidari, D. Merhof, Transnorm: Transformer provides a strong spatial normalization mechanism for a deep segmentation model, IEEE Access 10 (2022) 108205–108215.
[3] R. Gu, G. Wang, J. Lu, J. Zhang, W. Lei, Y. Chen, W. Liao, S. Zhang, K. Li, D.N. Metaxas, et al., CDDSA: Contrastive domain disentanglement and style augmentation for generalizable medical image segmentation, Med. Image Anal. 89 (2023) 102904.
[4] L.R. Soenksen, T. Kassis, S.T. Conover, B. Marti-Fuster, J.S. Birkenfeld, J. Tucker-Schwartz, A. Naseem, R.R. Stavert, C.C. Kim, M.M. Senna, et al., Using deep learning for dermatologist-level detection of suspicious pigmented skin lesions from wide-field images, Sci. Transl. Med. 13 (581) (2021) eabb3652.
[5] X.-X. Yin, L. Sun, Y. Fu, R. Lu, Y. Zhang, [Retracted] U-Net-based medical image segmentation, J. Heal. Eng. 2022 (1) (2022) 4189781.
[6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16 × 16 words: Transformers for image recognition at scale, 2020, arXiv preprint arXiv:2010.11929.
[7] Y. Rao, W. Zhao, Y. Tang, J. Zhou, S.N. Lim, J. Lu, Hornet: Efficient high-order spatial interactions with recursive gated convolutions, Adv. Neural Inf. Process. Syst. 35 (2022) 10353–10366.

[8] Y. Rao, W. Zhao, Z. Zhu, J. Lu, J. Zhou, Global filter networks for image classification, Adv. Neural Inf. Process. Syst. 34 (2021) 980–993.

[9] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, M. Wang, Swin-unet: Unet-like pure transformer for medical image segmentation, in: European Conference on Computer Vision, Springer, 2022, pp. 205–218.

[10] A. Gu, I. Johnson, K. Goel, K. Saab, T. Dao, A. Rudra, C. Ré, Combining recurrent, convolutional, and continuous-time models with linear state space layers, Adv. Neural Inf. Process. Syst. 34 (2021) 572–585.

[11] A. Gu, K. Goel, C. Ré, Efficiently modeling long sequences with structured state spaces, 2021, arXiv preprint arXiv:2111.00396.

[12] J.T. Smith, A. Warrington, S.W. Linderman, Simplified state space layers for sequence modeling, 2022, arXiv preprint arXiv:2208.04933.

[13] E. Nguyen, K. Goel, A. Gu, G.W. Downs, P. Shah, T. Dao, S.A. Baccus, C. Ré, S4nd: Modeling images and videos as multidimensional signals using state spaces, 2022, arXiv preprint arXiv:2210.06583.

[14] A. Gu, T. Dao, Mamba: Linear-time sequence modeling with selective state spaces, 2023, arXiv preprint arXiv:2312.00752.

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

[16] Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, Y. Liu, Vmamba: Visual state space model, 2024, arXiv preprint arXiv:2401.10166.

[17] J. Ma, F. Li, B. Wang, U-mamba: Enhancing long-range dependency for biomedical image segmentation, 2024, arXiv preprint arXiv:2401.04722.

[18] J. Ruan, S. Xiang, Vm-unet: Vision mamba unet for medical image segmentation, 2024, arXiv preprint arXiv:2402.02491.

[19] R. Wu, Y. Liu, P. Liang, Q. Chang, H-vmunet: High-order vision mamba unet for medical image segmentation, 2024, arXiv preprint arXiv:2403.13642.

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

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

[22] Z. Zhou, M.M. Rahman Siddiquee, N. Tajbakhsh, J. Liang, Unet++: A nested u-net architecture for medical image segmentation, in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4, Springer, 2018, pp. 3–11.

[23] S. Maurya, S. Tiwari, M.C. Mothukuri, C.M. Tangeda, R.N.S. Nandigam, D.C. Addagiri, A review on recent developments in cancer detection using machine learning and deep learning models, Biomed. Signal Process. Control. 80 (2023) 104398.

[24] X. Liu, L. Song, S. Liu, Y. Zhang, A review of deep-learning-based medical image segmentation methods, Sustainability 13 (3) (2021) 1224.

[25] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.

[26] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, Springer, 2015, pp. 234–241.

[27] O. Oktay, J. Schlemper, L.L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N.Y. Hammerla, B. Kainz, et al., Attention u-net: Learning where to look for the pancreas, 2018, arXiv preprint arXiv:1804.03999.

[28] E.K. Aghdam, R. Azad, M. Zarvani, D. Merhof, Attention swin u-net: Cross-contextual attention mechanism for skin lesion segmentation, in: 2023 IEEE 20th International Symposium on Biomedical Imaging, ISBI, IEEE, 2023, pp. 1–5.

[29] H. Wu, J. Zhong, W. Wang, Z. Wen, J. Qin, Precise yet efficient semantic calibration and refinement in convnets for real-time polyp segmentation from colonoscopy videos, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, No. 4, 2021, pp. 2916–2924.

[30] X. Zhao, L. Zhang, H. Lu, Automatic polyp segmentation via multi-scale subtraction network, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24, Springer, 2021, pp. 120–130.

[31] X. Zhao, H. Jia, Y. Pang, L. Lv, F. Tian, L. Zhang, W. Sun, H. Lu, M^{2}SNet: Multi-scale in multi-scale subtraction network for medical image segmentation, 2023, arXiv preprint arXiv:2303.10894.

[32] Y. Peng, M. Sonka, D.Z. Chen, U-Net v2: Rethinking the skip connections of U-Net for medical image segmentation, 2023, arXiv preprint arXiv:2311.17791.

[33] S. Hu, Z. Liao, Y. Xia, Devil is in channels: Contrastive single domain generalization for medical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2023, pp. 14–23.

[34] H. Wu, Z. Zhao, Z. Wang, META-Unet: Multi-scale efficient transformer attention Unet for fast and high-accuracy polyp segmentation, IEEE Trans. Autom. Sci. Eng. (2023).

[35] R. Wu, P. Liang, X. Huang, L. Shi, Y. Gu, H. Zhu, Q. Chang, MHorUNet: High-order spatial interaction UNet for skin lesion segmentation, Biomed. Signal Process. Control. 88 (2024) 105517.

[36] Y.B. Özçelik, A. Altan, Overcoming nonlinear dynamics in diabetic retinopathy classification: a robust AI-based model with chaotic swarm intelligence optimization and recurrent long short-term memory, Fractal Fract. 7 (8) (2023) 598.

[37] A. Altan, S. Karasu, Recognition of COVID-19 disease from X-ray images by hybrid model consisting of 2D curvelet transform, chaotic salp swarm algorithm and deep learning technique, Chaos Solitons Fractals 140 (2020) 110071.

[38] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, P. Torr, Res2net: A new multi-scale backbone architecture, IEEE Trans. Pattern Anal. Mach. Intell. 43 (2) (2019) 652–662.

[39] R. Wu, P. Liang, Y. Huang, Q. Chang, H. Yao, Automatic segmentation of hemorrhages in the ultra-wide field retina: Multi-scale attention subtraction networks and an ultra-wide field retinal hemorrhage dataset, IEEE J. Biomed. Heal. Inform. (2024).

[40] G. Li, Q. Huang, W. Wang, L. Liu, Selective and multi-scale fusion Mamba for medical image segmentation, Expert Syst. Appl. 261 (2025) 125518.

[41] Z. Zhang, B. Peng, T. Zhao, An ultra-lightweight network combining Mamba and frequency-domain feature extraction for pavement tiny-crack segmentation, Expert Syst. Appl. 264 (2025) 125941.

[42] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, arXiv preprint arXiv:1409.1556.

[43] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, Ieee, 2009, pp. 248–255.

[44] N.C. Codella, D. Gutman, M.E. Celebi, B. Helba, M.A. Marchetti, S.W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, et al., Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic), in: 2018 IEEE 15th International Symposium on Biomedical Imaging, ISBI 2018, IEEE, 2018, pp. 168–172.

[45] J. Bernal, F.J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, F. Vilariño, WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians, Comput. Med. Imaging Graph. 43 (2015) 99–111.

[46] A. Skouta, A. Elmoufidi, S. Jai-Andaloussi, O. Ouchetto, Hemorrhage semantic segmentation in fundus images for the diagnosis of diabetic retinopathy by using a convolutional neural network, J. Big Data 9 (1) (2022) 78.

[47] P. Xiuqin, Q. Zhang, H. Zhang, S. Li, A fundus retinal vessels segmentation scheme based on the improved deep learning U-Net model, IEEE Access 7 (2019) 122634–122643.

[48] H. Talebi, P. Milanfar, Learning to resize images for computer vision tasks, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 497–506.

[49] S. Saponara, A. Elhanashi, Impact of image resizing on deep learning detectors for training time and model performance, in: International Conference on Applications in Electronics Pervading Industry, Environment and Society, Springer, 2021, pp. 10–17.

[50] F. Isensee, P.F. Jaeger, S.A. Kohl, J. Petersen, K.H. Maier-Hein, Nnu-net: a self-configuring method for deep learning-based biomedical image segmentation, Nature Methods 18 (2) (2021) 203–211.

[51] M. Boniol, J.-P. Verriest, R. Pedeux, J.-F. Doré, Proportion of skin surface area of children and young adults from 2 to 18 years old, J. Invest. Dermatol. 128 (2) (2008) 461–464.

[52] G. Gethin, The importance of continuous wound measuring, WOUNDS UK 2 (2) (2006) 60.

[53] N. Safavian, S.K. Toh, M. Pani, R. Lee, Endoscopic measurement of the size of gastrointestinal polyps using an electromagnetic tracking system and computer vision-based algorithm, Int. J. Comput. Assist. Radiol. Surg. 19 (2) (2024) 321–329.

[54] D.G. Hewett, Measurement of polyp size at colonoscopy: Addressing human and technology bias, Dig Endosc. 34 (7) (2022) 1478–1480.

[55] N.F. Wasserman, E. Niendorf, B. Spilseth, Measurement of prostate volume with MRI (a guide for the perplexed): biproximate method with analysis of precision and accuracy, Sci. Rep. 10 (1) (2020) 575.

[56] A. Bezinque, A. Moriarity, C. Farrell, H. Peabody, S.L. Noyes, B.R. Lane, Determination of prostate volume: a comparison of contemporary methods, Academic Radiol. 25 (12) (2018) 1582–1587.

[57] C.J. Flaxel, R.A. Adelman, S.T. Bailey, A. Fawzi, J.I. Lim, G.A. Vemulakonda, G.-s. Ying, Diabetic retinopathy preferred practice pattern®, Ophthalmology 127 (1) (2020) P66–P145.