High-order Interaction and Low-order Parallelization of Features Fusion with Novel Mamba-UNet Architecture for Medical Image Segmentation

Qianhang Du, Zhenyu Lei, *Member, IEEE*, Jiujun Cheng, Masaaki Omura, *Member, IEEE*, Hideyuki Hasegawa, *Member, IEEE*, and Shangce Gao, *Senior Member, IEEE*

***Abstract*—Medical image segmentation is an essential method for computer-aided diagnosis. Although image segmentation models based on convolutional neural networks (CNNs) and vision transformers (ViTs) have achieved significant advancements, CNNs struggle to effectively capture long-range dependencies, while ViTs face limitations in local information extraction and are hindered by quadratic computational complexity. Recently, their inherent issues have been successfully addressed by the state-space models in Mamba and 2D-selective-scan in Vision Mamba. However, the presence of noise and excessive redundant information in medical images limits the practicality of these methods. To address these challenges, we propose a highly effective and accurate high-low-order feature fusion visual state space module, named HL-VSS. This module primarily consists of two core components: multi-scale spatial convolution (MSC) and high-low-order feature fusion (HLFF). The former component preliminarily suppresses noise and captures multi-scale feature information from medical images, accurately extracting edge and detail features for the fusion component. The latter processes these features, further reducing redundant information through high-order interaction with 2D-selective-scan, and fuses the local features obtained by low-order parallel Mamba, ultimately extracting deeper medical image features. We incorporate HL-VSS into a U-shaped architecture, named high-low-order feature fusion visual Mamba UNet (V-UNet). Comparison experiments and ablation studies are conducted on four publicly available medical image datasets to validate the strong competitiveness of V-UNet in medical image segmentation tasks. The code is available at https://github.com/ai-dqh0106/V-UNet_Code.**

***Index Terms*—state-space models, 2D-selective-scan, Mamba, high-low-order feature fusion, deep learning, medical image segmentation.**

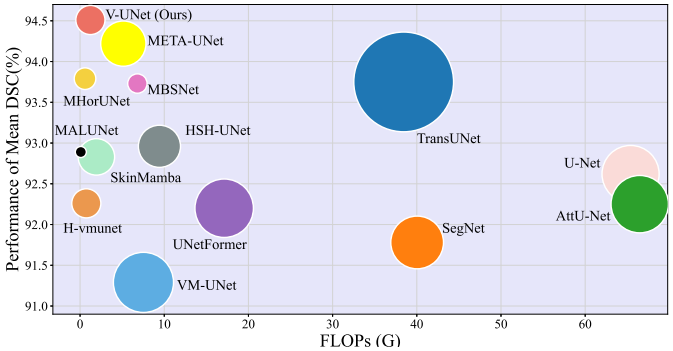


Fig. 1: The comparison with different methods in FLOPs, parameters, and DSC on PH^2 dataset. The X-axis represents FLOPs, the Y-axis represents the evaluation metric of DSC, and the area of the circles indicates the size of the parameters.

I. Introduction

MAKING anatomical or pathological structural alterations in images more visible is the aim of medical image segmentation [1]. To effectively reduce diagnostic errors caused by the visual diagnostic fatigue of medical personnel in large amounts of medical work, medical image segmentation technology has a significant impact on computer-aided medicine [2], [3]. Nowadays, liver tumor, brain tumor, skin lesion, lung, and cardiac image segmentation are the mainly popular medical image segmentation tasks [4]–[6]. With the widespread use of medical imaging technology, X-rays, magnetic resonance imaging, computed tomography, and ultrasound, these have become crucial methods for clinicians to diagnose diseases. However, the diversity of medical image types makes it increasingly difficult to segment key lesions in medical images and extract effective features from the segmented regions [7]. Meanwhile, medical image segmentation has become a major challenge due to noise interference from redundant information, blurry boundaries of target images, the diversity of lesions, and the scarcity of annotated data [8]–[10].

Manually extracting image features as the foundation of early classical machine learning segmentation methods, their performance far from the demands of clinical medicine [11], [12]. With the development of deep learning, various challenges in medical imaging have been effectively addressed [13]. Convolutional neural networks (CNNs) have demonstrated notable performance in image segmentation [14]. It has the ability to automatically learn the features of hierarchical structure from original images and their insensitivity to variations in image noise. The U-shaped architecture based on U-Net [15] has garnered widespread

This research was partially supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI under Grants JP25K21298 and JP25K03179, and Japan Science and Technology Agency (JST) Support for Pioneering Research Initiated by the Next Generation (SPRING) under Grant JPMJSP2145. *(Corresponding authors: Zhenyu Lei; Jiujun Cheng; Shangce Gao.)*

Q. Du, Z. Lei, M. Omura, H. Hasegawa, and S. Gao are with the Faculty of Engineering, University of Toyama, Toyama-shi, 930-8555, Japan. (E-mail: m24C1049@ems.u-toyama.ac.jp; leizg@eng.u-toyama.ac.jp; momura@eng.u-toyama.ac.jp; hasegawa@eng.u-toyama.ac.jp; gaosc@eng.u-toyama.ac.jp).

J. Cheng is with the School of Electronics and Information; Key Laboratory of Embedded System and Service Computing, Ministry of Education, Tongji University, Shanghai, 200092, China. (E-mail: chengjj@tongji.edu.cn).

attention and inspired the development of numerous image segmentation methods [16]–[18]. However, U-shaped methods based on CNNs have difficulty in capturing long-range information [19]. To alleviate the limitation, methods based on the vision transformers (ViTs) are proposed [20]. They are capable of modeling long-range dependencies and have rapidly gained popularity in various computer vision tasks, benefiting medical image segmentation. The methods, such as SegFormer [21], SETR [22], Segmenter [23], TransUNet [24], and UNetFormer [25], have been proposed for medical image segmentation and achieve notable successes. However, the quadratic complexity of self-attention mechanism in ViTs is directly correlated with image size. Although ViTs capture more long-range information, they also pose a challenge of huge computational complexity, due to the increase in the number of model parameters. [26]–[28].

Nowadays, the proposal of state-space models (SSM) has effectively addressed the quadratic computational complexity in ViTs and the challenges faced by CNNs in capturing long-range dependencies, generating significant interest among researchers [29], [30]. It excels at capturing global contextual information and allows for parallel training, which significantly reduces the number of model parameters [31], [32]. Gu et al. firstly added time-varying factors to SSM to propose the Mamba [33]. U-Mamba and Vision Mamba U-Net applied Mamba to medical image segmentation, respectively [34], [35], achieving remarkable competitiveness and further demonstrating the great potential of Mamba. Therefore, leveraging the linear complexity characteristics of Mamba, we design a parallel Mamba module that minimizes the parameters while maximizing the potential of Mamba to extract local detail features. Notably, although the Mamba-based methods exhibit strong performance in segmentation accuracy, they are limited by inadequate feature extraction capabilities due to the interference of noise and redundant information.

Recently, high-order spatial interaction methods have attracted extensive attention from researchers because they can effectively enhance the ability of models to perceive structural contextual information, establish global spatial dependencies, and reduce the impact of redundant information [36]–[38]. However, existing high-order spatial interaction methods often place too much emphasis on global features, which limits their capacity to capture fine boundaries and small structural details, ultimately reducing segmentation accuracy. Vision Mamba introduced the 2D-selective-scan (SS2D) mechanism into image recognition for the first time, significantly expanding the application scope of Mamba in computer vision [39]. To address the challenges faced by both Mamba and high-order spatial interaction methods, we integrate high-order spatial interaction methods into Mamba to establish spatial dependencies and suppress redundant information. In addition, by utilizing the SS2D module to perform horizontal and vertical feature scanning on images, we construct a hierarchical representation that preserves boundary information and small structural details in high-order spatial interaction as much as possible.

To reduce the interference of noise in different types of medical images and extract effective features from segmented regions, we expand the applicability of Mamba in medical image segmentation and propose an efficient and high-precision module, named high-low-order feature fusion visual state-space (HL-VSS). The core components of HL-VSS are multi-scale spatial convolution (MSC) and high-low-order feature fusion (HLFF). MSC is used to initially suppress the noise and capture multi-scale feature information in medical images, accurately extracting effective features to provide for the HLFF. HLFF gradually reduces the redundant information by utilizing high-order interaction SS2D (HI) and employs low-order parallel Mamba (LP) to capture local feature information during feature extraction for the first time. Then, the effective features extracted from HLFF to get deeper medical image features. Inspired by U-Net, we integrate HL-VSS into a U-shaped architecture to construct the high-low-order feature fusion vision Mamba UNet (V-UNet). To validate the performance of V-UNet, we conduct comparison experiments and ablation studies on four publicly available medical image datasets. As shown in Fig. 1, V-UNet achieves strong medical image segmentation performance with relatively fewer parameters, FLOPs, and significant potential.

The contributions of this work can be described as follows:

1) It proposes a novel HL-VSS module to enhance medical image feature extraction capabilities including MSC to suppress image noise and HLFF to effectively reduce redundant information and extract local features. They significantly improve the performance of our method in medical image segmentation.
2) Building upon the U-shaped architecture, it proposes a novel V-UNet based on the HL-VSS module for medical image segmentation, which addresses noise interference and redundant information in segmentation performance. This is the first attempt to tackle the challenge of redundant information through high-low-order feature fusion.
3) It further discusses ablation studies on different numbers of branches in LP, different orders in HI and various components in V-UNet to study their impact on medical image segmentation performance. The study shows deeper insights into MSC and HLFF, providing better performance and interpretability.

The remainder of this paper is summarized as follows: Section II reviews the related work. Section III gives the description of V-UNet and HL-VSS module. The detailed experimental results, ablation studies and analysis are provided in Section IV. Finally, Section V concludes this paper and explores the future work.

II. Related Work

In this section, image segmentation methods based on CNNs, ViTs, and Mamba are first reviewed. Subsequently, the studies and progress of higher-order spatial interaction methods are discussed.

A. Image Segmentation

1) CNNs-based Methods: The performance of image segmentation has been greatly enhanced by using CNNs-based methods. Long et al. firstly proposed the FCN and show

outstanding performance in image segmentation [40]. Subsequently, Ronneberger et al. proposed the U-Net, which has the capability to combine low-level and high-level features, making it highly influential in image segmentation [15]. Based on U-Net architecture, numerous researchers have proposed a number of variant models in image segmentation. For instance, Badrinarayanan et al. proposed SegNet to retain the symmetric structure of U-Net and utilized pooling indices for non-linear upsampling, thereby accurately recovering spatial information [16]. Meanwhile, Diakogiannis et al. proposed ResUNet to integrate residual learning modules into its model architecture, facilitating deeper layer training and resolving the vanishing gradient issue successfully [17]. Additionally, Oktay et al. introduced an attention mechanism to proposed AttU-Net to adaptively focus on important feature regions and suppress irrelevant redundant information [18]. Currently, models based on CNNs and encoder-decoder structures are widely used in image segmentation. MBSNet [41], MHorUNet [42], and HSH-UNet [43] have achieved commendable performance. However, although CNNs-based methods can capture local feature information quickly, they struggle to acquire long-range information. Even with the use of dilated convolutions to expand the receptive field for handling long-range information, the results are still unsatisfactory [44].

2) Transformer-based Methods: The Vision Transformers apply the global attention mechanism of the Transformer to the image domain directly, achieving remarkable results in the ImageNet classification task [20], [45]. Inspired by this, Chen et al. proposed TransUNet to utilize the Transformer architecture in the field of medical image segmentation [24]. It provides a novel approach for medical image segmentation challenges by ingeniously fusing the feature extraction capabilities of CNNs with the long-range information extraction ability of Transformer. Subsequently, Cao et al. proposed the SwinUNet, replacing the traditional Transformer with Swin Transformer [46]. It reduces resolution layer by layer, gradually aggregates features, and uses a window-based attention mechanism to capture features in different scales, aiming to restore the spatial resolution of feature maps. Similarly, Zhang et al. combined Transformer and CNNs as the main branches in a parallel manner to propose TransFuse [47]. Additionally, Wang et al. introduced a pure Transformer U-shaped architecture, named UNetFormer for medical image segmentation [25]. However, although Transformer-based models can capture long-range information in image segmentation, the self-attention method with quadratic complexity is a serious challenge. When it captures more long-range information, it also introduces more parameters and increases computational demand [26]–[28]. In summary, existing image segmentation models based on CNNs and Transformer have limited in long-range dependency and computational complexity, respectively.

B. Mamba

Recently, researchers have sparked a renewed interest in the Mamba [33]. The Mamba effectively addresses the challenges of capturing long-range dependencies in CNNs and the self-attention mechanism with quadratic computational complexity in ViTs. Zhou et al. proposed MDNet to achieve superior prediction performance with fewer parameters and computational complexity by proposing a novel Mamba-efficient fusion module and a diffusion self-distillation strategy [48]. Meanwhile, Ju et al. proposed FMamba to narrow the information gap between network layers through the multi-scale progressive fusion module based on Mamba, achieving global context modeling effects [49]. Additionally, Liu et al. designed a new spectral-spatial adaptive Mamba by adaptively scanning spatial domain pixels and dynamically enhancing spectral bands of spectral scanning [50]. Liu et al. proposed Vision Mamba by developing 2-D Mamba modules for horizontal and vertical image scanning, forming a hierarchical visual model, named SS2D [39]. The emergence of Vision Mamba provides a new direction for computer vision, gaining significant attention in the image classification and segmentation. Moreover, Ruan et al. proposed VM-UNet, that is a classic method to integrate Vision Mamba into a U-shaped architecture for medical image segmentation and demonstrates strong performance [51]. Subsequently, Wu et al. proposed a method for processing features in lightweight with Vision Mamba, achieving efficient performance with minimal time complexity and memory usage [52]. Furthermore, Xing et al. proposed SegMamba to effectively capture long-range dependencies across multi-scale full convolution features and maintains excellent processing speed. It offers a new direction for Mamba in 3D image segmentation [53]. Notably, although Mamba-based methods exhibit strong segmentation performance, they are limited by inadequate feature extraction capabilities due to the interference of noise and redundant information [8]–[10].

C. High-order Spatial Interactions

Traditional CNNs primarily rely on local receptive fields for feature extraction, which may limit their capacity to adequately capture complex structural information. Moreover, medical images typically exhibit high noise levels, subtle pixel intensity variations, and ambiguous boundaries within segmentation regions. To address these challenges, recent studies have introduced high-order spatial interactions to incorporate broader, and even global, spatial dependencies. This approach effectively enhances the ability of model to perceive structural context, thereby enabling more precise discrimination of different tissues or lesion areas. Rao et al. proposed HorNet to further advance this paradigm by integrating efficient, scalable, and plug-and-play high-order spatial interaction modules via gated convolutions and recursive designs, thereby effectively enhancing both CNNs and ViTs-based models and offering novel directions for visual modeling [38]. Additionally, Zheng et al. represented each pixel using high-order features computed from pixel-level affinities, clustering pixels into distinct semantic groups to ensure robustness and stability under feature distortions [54]. Furthermore, Sun et al. embedded early-stage high-order information into the final layers of network, leveraging contextual high-order features extracted across multiple stages to preserve boundary correlations in low-level representations, substantially improving segmentation performance [55]. Zhang et al. proposed a deep network framework

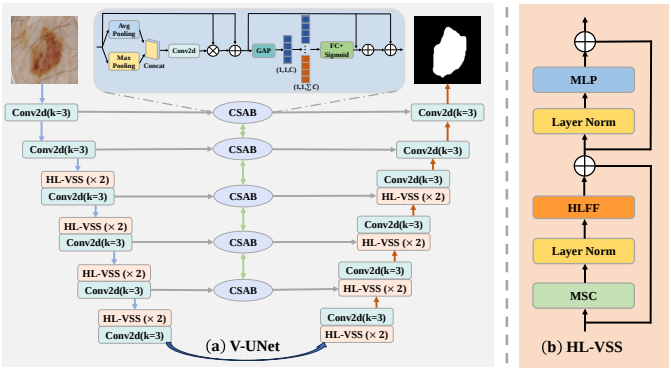


Fig. 2: (a) The overall architecture of the proposed V-UNet, which mainly comprises HL-VSS and CSAB module. (b) HL-VSS module contains MSC, HLFF components to suppress noise and reduce redundant information, respectively. As our HLFF considers two aspects (high-order and low-order), V-UNet can achieve wonderful segmentation performance. CSAB is used to accelerate model fitting in the skip connections.

with adaptive spatial and high-order semantic modulation. By introducing a high-order semantic modulator as a supervisory module and applying shift and scale operations to process multi-scale features, the framework enables the features to focus on common object regions [56]. However, existing high-order spatial interaction methods tend to emphasize global feature information, often resulting in insufficient capture of fine boundaries and small structural details, which negatively impacts segmentation accuracy [42], [57].

III. The proposed Method

In this section, we begin with an overview of the overall architecture of V-UNet. We then present a detailed analysis of each component in the HL-VSS module. Finally, the implement process and advantages of the proposed method are summarized.

A. The Overall Architecture of V-UNet

The complete structure of V-UNet is illustrated in Fig. 2 (a). V-UNet adopts a 6-layer U-shaped architecture, primarily composed of encoders, skip connections, and decoders. It enables efficient feature fusion by progressively reducing the image size while increasing the number of channels. The channel configuration in the 6-layer architecture is set to $[8, 16, 32, 64, 128, 256]$. The first and second layers mainly use standard convolution with a kernel size of $k = 3$ for feature extraction. To reduce the redundant information and enhance the feature extraction capability, we propose the HL-VSS. As the core module in V-UNet, HL-VSS runs through both the encoders and decoders of V-UNet. In Fig. 2 (b), the MSC is employed to suppress noise and capture multi-scale feature information. Subsequently, HLFF is utilized to capture local features and minimize the introduction of redundant information. Finally, a multilayer perceptron (MLP) performs nonlinear mapping and extract diversified feature information. By combining HL-VSS with standard convolution, further extracts features from the third to the sixth layers. In Fig. 2 (a), numerous advanced studies have demonstrated that incorporating channel attention bridge and spatial attention bridge (CSAB) in skip connections provides stable performance for multi-level and multi-scale fusion, and can accelerate model fitting, where GAP is global average pooling [42], [43], [58].

As for the structure of the HL-VSS module, we design an architecture similar to a standard Transformer block, replacing the multi-head self-attention and feed-forward network block with our proposed HLFF and the existing MLP. Each component is surrounded by residual connections, and layer normalization is applied before the HLFF and MLP components. To effectively model the dependencies of the image features, the MSC is designed to initially capture effective features before the HLFF component.

B. Multi-scale Spatial Convolution

The interference of noise in different types of medical images often random and inconsistent makes it increasingly difficult to extract effective features from segmented regions [8], [9]. Moreover, due to the different sizes of segmented regions in various datasets, constructing multi-scale information to accurately segment target regions is crucial for medical image segmentation methods. Inspired by the seamless integration of multi-scale information through the human visual system to interpret the external environment [59], our method is trained by leveraging complementary multi-scale information and employing cross-scale fusion of effective features, thereby effectively suppressing noise and minimizing loss [60].

As illustrated in Fig. 3, the MSC is designed to suppress noise and capture multi-scale feature information [61]. The input features are processed through four convolution blocks, each comprising a normalization layer, a convolution layer (with kernel size of 3×3 or 1×1), and a nonlinear layer. The input $x \in \mathbb{R}^{H\times W\times C}$, where H, W, and C represents the height, width and channel dimension, respectively. The formula of MSC can be represented below:

$$\mathrm{MSC}(x) = x + G^{1\times1}(G^{3\times3}(G^{3\times3}(x)) + G^{1\times1}(x)), \qquad (1)$$

where $G^{k\times k}$ denotes the convolution block and k indicates the size of kernel. The proposed MSC can precisely extracting edges and detailed features, providing effective features for the HLFF component. To verify the effectiveness of MSC, the detailed analysis and visualization results are provided in Section IV-E2.

C. High-low-order Feature Fusion

Existing image segmentation methods based on CNNs and ViTs are limited in terms of long-range dependency and computational complexity, respectively. SS2D is capable of achieving a global receptive field while maintaining linear complexity, demonstrating excellent feature extraction capabilities [39]. However, the more global receptive field SS2D focuses on, the more redundant information is introduced. Inspired by high-order methods can minimize redundant information [36]–[38], we propose the HI, which progressively

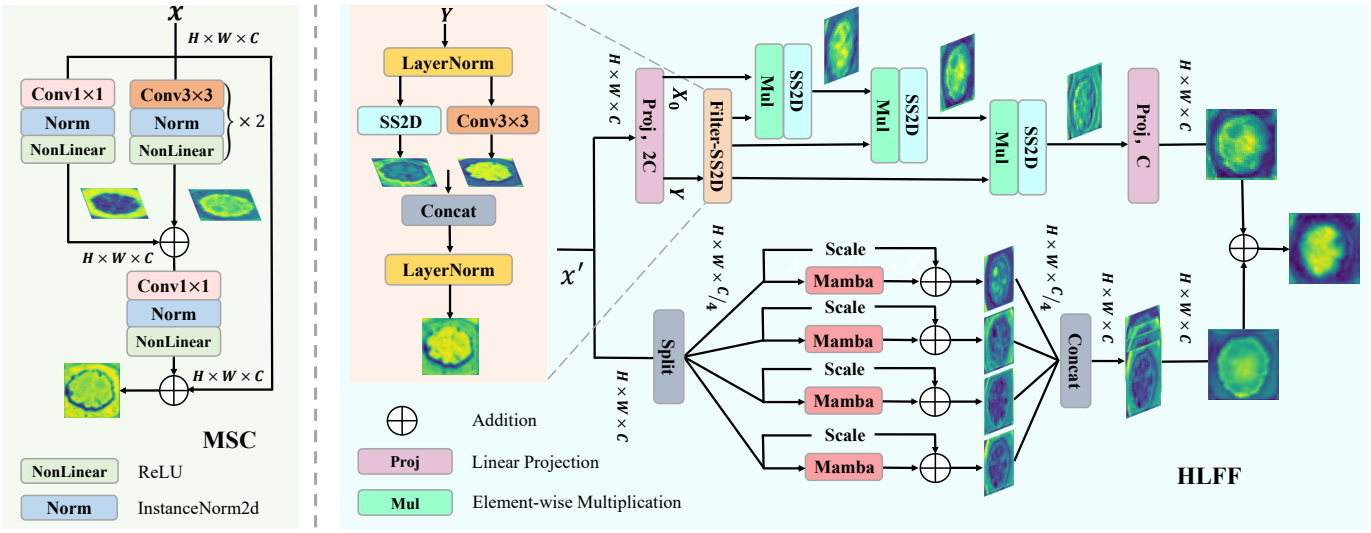


Fig. 3: The implement framework of MSC and HLFF in HL-VSS module.

increases channel width through a recursive design. This allows for higher-order spatial interactions to be achieved with limited complexity, thereby effectively reducing redundant information and enhancing the segmentation performance of the model. Additionally, considering that Mamba can effectively provide local features for the global receptive field, we propose the LP. Through HLFF component, we further enhance the capability of medical image segmentation.

HLFF takes the MSC noise-processed feature sequence as input x'. The HLFF mainly consists of HI and LP. Among them, the LP captures local feature information, ensuring high-resolution texture detection. The HI progressively reduces interference from redundant information, ensuring greater robustness in complex background images [36], [37], [54]. By adopting the high-low-order feature fusion approach, the HLFF can capture more relevant feature information in complex medical images, thereby improving segmentation performance. As shown in Fig. 3, the structure of HI and LP in HLFF are described, and we will introduce them in detail below.

1) High-order Interaction SS2D Component: HI consists of the Hadamard product operation, SS2D operation, Filter-SS2D (FS), layer normalization (LN(·)), and so on. Through Proj(·), a linear projection operation, the total number of channels will be expanded from C to $2C$, resulting in a feature $X_0^{H\times W\times C_0}$ and a set of features $Y = [Y_0^{H\times W\times C_0}, Y_1^{H\times W\times C_1}, \ldots, Y_i^{H\times W\times C_i}]$. The number of channels can be computed for each feature order through the formula $C_i = \frac{C}{2^{n-i-1}}$, where $0 \le i \le n-1$, n is the number of orders. The detailed operation can be represented as:

$$x' = \mathrm{LN}(\mathrm{MSC}(x)), \tag{2}$$

$$[X_0, Y_0, Y_1, \ldots, Y_i] = \mathrm{Proj}(x') \in \mathbb{R}^{H\times W\times 2C}. \tag{3}$$

Subsequently, the FS is proposed to enhance high-order interaction learning capabilities. From Fig. 3, it primarily consists of two channels, after the feature sequence undergoes the Split(·) operation, one half of the channels performs convolution operations (with kernel size of 3×3), and the other half conducts SS2D operations. Finally, the two channels are merged through Concat(·, ·) operation. By combining the feature of these two channels, the high-order interaction learning capabilities are improved. The FS can be formulated as:

$$y_1, y_2 = \mathrm{Split}(\mathrm{LN}(Y_i)), \tag{4}$$

$$\mathrm{FS}(Y_i) = \mathrm{LN}(\mathrm{Concat}(\mathrm{Conv}(y_1), \mathrm{SS2D}(y_2))). \tag{5}$$

Next, for each interaction sequence, the output features of the FS are element-wise multiplied with the corresponding channel-specific feature X_i through a Hadamard product to improve the weight of the feature information. The SS2D operation is performed to refocus on the global receptive field, thereby reducing the introduction of redundant information. This process utilizes a portion of the channel count at each order to execute the operations described above, continuing until the final n-th order is completed. The specific operation is expressed by the following equations:

$$X_{i+1} = \mathrm{SS2D}(X_i \odot \mathrm{FS}(Y_i)), \tag{6}$$

$$\mathrm{HI} = \mathrm{Proj}(X_n) \in \mathbb{R}^{H\times W\times C}, \tag{7}$$

where $\odot$ is the Hadamard product operation. Finally, as seen in (7), the output features of HI are obtained by mapping the feature information through the Proj(·), which reduces the total channels from $2C$ to C.

2) Low-order Parallel Mamba Component: LP divides the total number of channels C into multiple parts $Z = [Z_1^{C/K}, Z_2^{C/K}, Z_3^{C/K}, \ldots, Z_j^{C/K}]$, through the Split(·) operation first. Then, each channel feature is input into the Mamba, and a residual-level joint adjustment factor is designed to enhance the capability of acquiring effective information. The features are processed through parallel Mamba module to ensure that the total processed channels remain constant. The LP can maximize efficiency, maintain high accuracy and minimize parameters. The detailed formula is expressed as follows:

$$Z_1, Z_2, Z_3, \ldots, Z_j = \mathrm{Split}(\mathrm{LN}(x')), \tag{8}$$

$$Z_j = \text{Mamba}(Z_j) + \theta \cdot Z_j \quad (j = 1, 2, 3, \ldots, K), \tag{9}$$

where Mamba$(\cdot)$ is the Mamba module, and θ is the trainable factor. Next, by concatenating multiple feature channels K, the dispersed channel count is merged back to the initial total of C, ensuring that the input and output channels remain consistent.

$$\text{LP} = \text{Concat}(Z_1, Z_2, Z_3, \ldots, Z_j), \tag{10}$$

where Concat$(\cdot, \cdot)$ is the concat operation. LP effectively deepens the processing of the feature information from the dispersed channels, enhancing the diversity of the features.

To ensure the proposed HLFF component captures more feature information while reducing the introduction of redundant information, we simply sum the elements of LP and HI to achieve high-low-order feature fusion. Thus, the whole process of our HLFF component can be summarized as follows:

$$\text{HLFF}(x) = \text{LP}(x) + \text{HI}(x). \tag{11}$$

Without introducing excessive additional parameters, the LP is employed to extract diverse local features in parallel at a low-order. By leveraging high-order interactive feature information through HI, the redundant information is minimized. HLFF effectively captures more medical image features effectively, thereby improving segmentation performance.

D. Multilayer Perceptron

The MLP applies the feature sequence obtained from the HLFF as the input [62]. It performs nonlinear mapping and transformation on the feature sequence to extract diversified features, thereby enhancing the segmentation of fine edges and structures in medical images. The MLP mainly consists of two linear layers: the channel dimensions of the input are expanded from C to $4C$ in the first layer, and then decreased from $4C$ back to C in the second layer.

$$\text{MLP}(x) = \sigma(\text{LN}(\text{HLFF}(x) + x) \cdot w_1 + b_1) \cdot w_2 + b_2, \tag{12}$$

where w_1 and w_2 are the respective weights of the two linear layers. b_1 and b_2 are the bias terms, and $\sigma(\cdot)$ is the activation function GELU.

E. The Implement Process of V-UNet

V-UNet is a novel image segmentation method based on a U-shaped architecture. In the encoders, two 3×3 standard convolution layers are applied to initially capture the feature maps of medical images and increase the number of channels. Meanwhile, the captured initial features are fed into four combination blocks of HL-VSS module and 3×3 standard convolution. Within the HL-VSS module, MSC captures multi-scale feature information to suppress noise from the medical images and provides effective image features for HLFF. HLFF is used to capture local features and reduce redundant information, thereby obtaining deeper features. Finally, an MLP is applied to perform nonlinear mapping and transformation of the feature sequence, extracting diverse features to enhance the segmentation of fine edges and structure. Moreover, the decoders as the inverse of the encoders, gradually restoring the low-resolution feature map to the original image size, preserving the spatial structure of the image and better segmenting abnormal areas. In the skip connections, the CSAB module is employed to link feature information across different stages. It generates spatial and channel attention maps to enhance the richness of decoder features and accelerate model fitting.

F. The Advantages of V-UNet

To better highlight the core strengths of V-UNet, we present the advantages of the proposed V-UNet from the perspectives of its problem-solving capabilities, modules, and architecture:

1) Image segmentation methods based on CNNs and ViTs variants perform poorly in capturing long-range dependencies and suffer from quadratic computational complexity, respectively. However, V-UNet is capable of achieving a global receptive field while maintaining linear complexity, demonstrating excellent feature extraction capabilities.
2) The novel HL-VSS module primarily consists of two core components: MSC and HLFF. The former is used to suppress noise and capture multi-scale feature information, while the latter achieves a global receptive field through high-order interactions and captures local detail features through low-order parallelism to suppress redundant information.
3) Based on the U-shaped architecture and the novel HL-VSS module, we propose V-UNet, which captures long-range dependency information while extracting detailed local features and suppressing redundant information. This is the first attempt to address the challenge of redundant information through high-low-order feature fusion.

IV. Experiments and Analysis

In this section, we first present a detailed description of the dataset, experimental details, and evaluation metrics. We then validate the effectiveness and potential of the proposed method through comparative experiments and visualization analyses. Finally, the superiority of V-UNet is further demonstrated via ablation studies.

A. Datasets Description

To evaluate our proposed method, we conduct extensive experiments on four popular medical image datasets, including ISIC2017 [65], ISIC2018 [66], PH^2 [67], and STU [68]. The detail information of four datasets are listed in Table I.

TABLE I: The different sizes and types of four medical image datasets for image segmentation in detail.

Datasets	Size	Type	#Total	# of train	# of test	Pixels
ISIC2017	Large	Skin	2150	1500	650	256×256
ISIC2018	Large	Skin	2694	1886	808	256×256
PH^2	Medium	Skin	200	140	60	256×256
STU	Small	Breast	42	33	9	128×128

TABLE II: The Parameters, FLOPs, Inference Time, and FPS of comparison models.

Methods	Year	Param.(M)	FLOPs(G)	Inf. Time (ms)	FPS
U-Net [15]	2015	34.53	65.39	6.78	153.94
SegNet [16]	2016	29.44	40.04	9.87	235.42
AttU-Net [18]	2018	34.88	66.48	8.17	139.97
MALUNet [58]	2022	0.18	0.09	9.50	88.59
MBSNet [41]	2023	3.98	6.81	12.32	124.02
MHorUNet [42]	2024	4.96	0.59	65.72	15.94
HSH-UNet [43]	2024	18.8	9.43	69.46	18.24
TransUNet [24]	2024	105.32	38.44	21.90	48.98
UNetFormer [25]	2022	36.17	17.12	18.68	65.05
META-UNet [63]	2024	22.21	5.13	12.83	101.10
VM-UNet [51]	2024	44.27	7.53	37.55	27.01
H-vmunet [57]	2025	8.97	0.74	129.18	7.28
SkinMamba [64]	2024	14.08	1.94	22.67	48.82
V-UNet(Ours)	-	13.46	1.23	81.79	13.24

The ISIC2017 and ISIC2018 datasets are two important collections from the Skin Lesion Identification Challenge organized by the International Skin Imaging Collaboration. In contrast to ISIC2017, ISIC2018 includes a broader range of skin diseases, making the segmentation task more challenging. Unlike the two large skin disease datasets, the PH^2 dataset is created by the Pedro Hispano Hospital in Portugal for research on skin lesion detection. Although PH^2 is a medium-sized dataset, it contains image features similar to those in ISIC2017 and ISIC2018. To better validate the generalization ability of our proposed method, we use the STU dataset, which is collected by the radiology department of the First Affiliated Hospital of Shantou University through GE Voluson E10 ultrasound diagnostic system. Compared to the skin disease datasets, the image features in the STU dataset are completely different. Notably, each dataset represents a specific medical image segmentation task with distinct types, allowing them to verify the universality and robustness of V-UNet.

B. Experimental Details

During the experiments, the following hardware and software configurations are employed: Intel(R) Xeon(R) Silver 4110 CPU @ 2.10 GHz, NVIDIA GeForce RTX 3090, and PyTorch 2.1.0 as the backend. Each method is trained 5 times on each datasets separately with different random seeds and 300 epochs to avoid the impact of randomness on the experimental results. Batch size and initial learning rate are set to 8 and $1e{-}3$, respectively. The learning rate continuously reduces through weight decay in AdamW optimizer [69]. To ensure equitable comparisons in training assessments, the images are reshaped to the same resolution of 256×256. In addition, we apply three standard strategies for data augmentation, including horizontal flip, vertical flip, and random rotation.

We adopt the AdamW optimizer to accelerate convergence and calculate the loss through a combination of binary cross-entropy [63] and dice loss [70]. The combined segmentation loss function is expressed as follows:

$$L_{Seg} = L_{BCE} + L_{Dice}, \tag{13}$$

where L_{Dice} denotes the dice loss and L_{BCE} is the binary cross-entropy loss. The L_{BCE} emphasizes the classification accuracy of each pixel, while the L_{Dice} coefficient focuses on the overlap of the overall segmentation regions. Therefore, L_{Seg} provides stable and effective gradient updates at different stages of training, thereby improving both convergence speed and segmentation performance. To ensure a fair comparison, all methods are trained in the same computing environment.

C. Evaluation Metrics

In this paper, we evaluate our method by using five different evaluation metrics on four types of medical image datasets, including mean intersection over union (mIoU), dice similarity coefficient (DSC), sensitivity (SE), specificity (SP), and accuracy (ACC) [71]. Higher scores of these metrics indicate better segmentation performance.

Among them, mIoU and DSC are the primary metrics used to evaluate the medical image segmentation performance in experimental analysis. mIoU evaluates the proportion of overlap between the actual and predicted areas relative to the total area, representing the ratio of the intersection over the union of the predicted and actual segmentation areas. In contrast, DSC emphasizes the proportion of the overlap in the total area, representing the ratio of the intersection area to the sum of the two areas. SE measures the proportion of all positive ground truth pixels to true positive predictions. SP quantifies the proportion of all negative ground truth pixels to true negative predictions, and ACC calculates the proportion of correctly segmented pixels. By utilizing these five specific metrics, we can conduct a more comprehensive analysis of the segmentation performance and the potential of V-UNet.

D. Comparison with State-of-the-art Methods

To validate our method, we conduct a comparative analysis against various image segmentation models on four classic medical image datasets. From Table II, we list the year, parameters, FLOPs, inference time and FPS of comparison models and our method. Among them, U-Net [15] and SegNet [16] are traditional methods. AttU-Net [18] is classic method with attention mechanisms. TransUNet [24], UNetFormer [25], and META-UNet [63] are classical and state-of-the-art (SOTA) methods based on Transformers. MALUNet [58], MBSNet [41], MHorUNet [42] and HSH-UNet [43] are SOTA methods based on CNNs. VM-UNet [51], H-vmunet [57], and SkinMamba [64] are SOTA methods based on Mamba. Notably, although AttU-Net and SkinMamba are preprint versions, they serve as representative models of the CNNs and Mamba, respectively, and thus retain substantial comparative significance. Additionally, we highlight the best results are in bold and the second-best results are underlined to give a more intuitive depiction of the segmentation performance in Tables III and IV. The p-value is a crucial metric for assessing the statistical significance of results. The result is typically considered statistically significant when the p-value is less than or equal to a predefined threshold, commonly set at 0.05. As shown in Tables III and IV, the p-value based

TABLE III: Experimental results (mean±std) of models for ISIC2017 and ISIC2018 datasets.

Methods	ISIC2017						ISIC2018					
	mIoU(%)	DSC(%)	SE(%)	SP(%)	ACC(%)	*p*-value	mIoU(%)	DSC(%)	SE(%)	SP(%)	ACC(%)	*p*-value
U-Net [15]	76.64±0.61	86.90±0.32	83.92±1.30	98.16±0.23	95.76±0.10	6.90E-05	78.58±0.70	88.00±0.43	86.26±0.73	96.86±0.27	94.28±0.20	1.33E-03
SegNet [16]	75.08±1.20	85.80±0.81	82.78±2.31	97.94±0.29	95.40±0.15	1.34E-03	78.20±0.42	87.76±0.26	85.68±1.02	96.90±0.36	94.20±0.11	3.77E-05
AttU-Net [18]	76.84±0.64	86.90±0.38	83.60±1.17	98.24±0.14	95.80±0.11	2.20E-04	78.52±0.45	87.98±0.26	85.14±1.43	97.28±0.44	94.34±0.10	1.15E-04
TransUNet [24]	77.54±0.59	87.34±0.37	84.02±1.79	98.34±0.34	95.92±0.07	5.21E-04	77.92±0.21	87.58±0.12	85.08±0.80	97.08±0.29	94.13±0.08	6.22E-05
UNetFormer [25]	76.12±0.68	86.44±0.43	81.58±1.41	**98.56±0.25**	95.72±0.12	1.69E-04	76.10±0.57	86.42±0.35	82.68±1.44	97.22±0.40	93.66±1.10	1.88E-06
META-UNet [63]	79.74±0.22	88.72±0.14	86.02±0.43	98.44±0.10	96.33±0.05	6.93E-02	80.08±0.37	88.94±0.23	86.54±1.64	97.44±0.55	94.76±0.05	6.12E-02
MALUNet [58]	77.26±0.80	87.18±0.50	83.70±1.48	98.34±0.19	95.90±0.15	1.78E-03	79.66±0.21	88.68±0.13	88.06±3.32	95.24±4.07	94.64±0.05	6.40E-03
MBSNet [41]	77.96±0.61	87.64±0.39	84.98±1.17	98.20±0.18	95.98±0.13	1.98E-03	78.86±0.59	88.18±0.38	85.52±1.34	97.28±0.39	94.42±0.13	1.52E-03
MHorUNet [42]	78.74±0.43	88.12±0.26	86.44±1.17	98.06±0.22	96.10±0.06	2.08E-03	79.50±0.27	88.60±0.17	86.04±0.34	97.36±0.10	94.58±0.07	3.45E-03
HSH-UNet [43]	78.76±0.29	88.10±0.19	85.64±1.27	98.26±0.23	96.14±0.05	2.19E-04	80.06±0.30	88.94±0.19	87.10±0.85	97.18±0.29	94.70±0.09	5.02E-02
VM-UNet [51]	79.68±0.38	88.70±0.24	**88.12±1.34**	97.90±0.34	96.24±0.08	1.89E-01	78.92±0.72	88.22±0.44	87.08±1.55	96.70±0.36	94.38±0.16	3.80E-03
H-vmunet [57]	78.72±0.33	88.06±0.23	86.10±1.04	98.12±0.19	96.08±0.04	6.73E-04	79.78±0.07	88.74±0.05	85.98±1.04	**97.50±0.40**	94.70±0.06	1.25E-02
SkinMamba [64]	78.42±0.77	87.88±0.47	86.14±1.02	98.02±0.07	96.02±0.12	1.06E-02	80.20±0.70	89.01±0.43	86.83±1.02	97.33±0.27	94.80±0.17	7.97E-02
V-UNet(Ours)	**80.02±0.19**	**88.89±0.12**	87.65±0.51	98.08±0.16	**96.33±0.05**	-	**80.76±0.48**	**89.36±0.30**	**89.34±1.21**	96.58±0.44	**94.82±0.15**	-

TABLE IV: Experimental results (mean±std) of models for PH^2 and STU datasets.

Methods	PH^2						STU					
	mIoU(%)	DSC(%)	SE(%)	SP(%)	ACC(%)	*p*-value	mIoU(%)	DSC(%)	SE(%)	SP(%)	ACC(%)	*p*-value
U-Net [15]	86.28±1.45	92.60±0.83	89.78±1.67	**97.60±0.14**	94.68±0.55	8.82E-03	78.56±1.35	88.00±0.82	80.66±1.62	**99.54±0.12**	96.66±0.22	6.38E-03
SegNet [16]	84.82±1.64	91.78±0.98	88.82±1.86	97.20±0.69	94.08±0.67	4.66E-03	77.56±2.16	87.34±1.37	81.10±1.41	99.18±0.25	96.46±0.40	1.49E-02
AttU-Net [18]	85.64±0.94	92.24±0.54	89.52±0.73	97.31±0.54	94.42±0.42	5.82E-04	79.66±0.63	88.66±0.42	81.94±0.79	98.98±1.09	**97.36±1.02**	1.87E-03
TransUNet [24]	88.22±0.45	93.74±0.24	93.62±0.61	96.40±0.17	95.36±0.19	1.17E-03	81.38±0.71	89.72±0.42	84.52±0.43	99.32±0.13	97.08±0.13	1.17E-01
UNetFormer [25]	85.56±0.92	92.20±0.53	89.42±0.82	97.30±0.24	94.38±0.39	6.35E-04	73.32±7.12	84.42±4.88	76.64±6.22	99.16±0.36	95.74±1.22	7.90E-02
META-UNet [63]	89.04±0.30	94.22±0.16	93.06±0.57	97.30±0.28	95.78±0.12	1.28E-02	80.15±1.23	88.97±0.77	82.60±1.47	99.45±0.08	96.92±0.18	2.32E-02
MALUNet [58]	86.74±0.34	92.88±0.21	91.60±0.99	96.68±0.44	94.78±0.10	8.27E-06	76.06±1.62	86.42±1.04	82.06±1.80	98.60±0.26	96.08±0.27	1.98E-03
MBSNet [41]	88.20±0.37	93.72±0.22	92.22±0.84	97.32±0.41	95.40±0.15	6.18E-04	81.02±0.39	89.52±0.25	83.68±1.26	99.42±0.24	97.06±0.05	5.86E-03
MHorUNet [42]	88.32±0.51	93.78±0.29	93.16±0.59	96.76±0.41	95.44±0.22	4.24E-03	80.42±0.74	89.14±0.44	85.82±1.23	98.80±0.18	96.82±0.12	9.79E-03
HSH-UNet [43]	86.86±0.30	92.96±0.16	91.90±1.08	96.56±0.53	94.76±0.19	7.53E-07	78.86±1.22	88.20±0.78	84.44±2.08	99.00±0.22	96.62±0.23	7.49E-03
VM-UNet [51]	87.46±0.85	91.28±3.60	92.24±1.02	96.72±0.25	95.08±0.32	1.44E-01	78.36±1.12	87.86±0.72	83.04±1.46	98.96±0.22	96.56±0.19	2.99E-03
H-vmunet [57]	88.16±0.28	93.68±0.17	92.20±0.72	97.26±0.50	95.40±0.11	7.31E-05	81.46±0.73	89.78±0.44	86.00±0.36	99.00±0.20	97.04±0.14	1.78E-01
SkinMamba [64]	86.64±0.90	92.82±0.53	90.90±1.13	97.08±0.32	94.78±0.35	2.34E-03	81.30±0.59	89.68±0.37	86.02±1.22	98.96±0.19	97.02±0.10	6.72E-02
V-UNet(Ours)	**89.59±0.27**	**94.51±0.14**	**94.82±0.45**	96.67±0.23	**95.90±0.12**	-	**82.05±0.11**	**90.10±0.08**	**86.90±0.99**	98.98±0.23	97.13±0.05	-

on mIoU demonstrates that our method significant superiority over 13 SOTA methods across four medical image datasets.

ISIC2017 and ISIC2018. From Table III, we objectively evaluate two large skin lesion datasets, and the experimental results demonstrate that our method has outstanding performance based on evaluation metrics. In contrast to other methods, although our method performs slightly worse in SP, this single metric alone cannot provide a comprehensive evaluation of segmentation performance. Notably, V-UNet shows a significant advantage in the other key metrics: mIoU, DSC, SE, and ACC. On the ISIC2017 dataset, our method is equal to META-UNet in ACC and is lower 0.36% in SP, but it improves 0.28% and 0.17% in the critical segmentation metrics of mIoU and DSC, respectively, and also increases 1.63% in SE. Although META-UNet enhances the accuracy of feature extraction by adaptively fusing global and local features, it lacks a dedicated noise processing mechanism before feature fusion. In contrast, V-UNet incorporates the MSC component to effectively leverage multi-scale feature information, thereby mitigating the impact of noise in medical images and improving segmentation performance. On the ISIC2018 dataset, our method outperforms the SkinMamba by 0.56%, 0.35%, 2.51%, and 0.02% in mIoU, DSC, SE, and ACC, respectively. This is because the ISIC2018 dataset presents a diverse range of skin disease types, complex lesion structures, and higher levels of noise in medical images. While SkinMamba employs a sophisticated model architecture to capture more informative features, it also introduces redundant information, which degrades segmentation performance. However, V-UNet utilizes a simpler design and leverages high-order interactions to progressively mitigate the impact of redundancy. The comparison experiment results indicate that our method considerably enhances segmentation performance on these two large datasets.

Unlike the two large skin lesion datasets mentioned above, the PH^2, as a medium-sized dermatological dataset, has fewer ground truth annotations and category labels. However, it retains nearly identical image characteristics to ISIC2017 and ISIC2018, and the model training speed is relatively fast. Therefore, in the subsequent analysis of model stability, hyperparameter analysis, and ablation experiments, we will primarily focus on the PH^2 dataset.

PH^2 and STU. From Table IV, it is important to note that mIoU and DSC are significant metrics in image segmentation tasks. On the PH^2 dataset, the segmentation performance of U-Net highlights this point. Despite the U-Net achieves the best in SP, it performs moderately in other metrics, underscoring the importance of considering overall segmentation quality. U-Net captures local features solely through CNNs, which limits its ability to process medical images and often leads to feature loss. In contrast, our method leverages HLFF to establish long-range dependencies, effectively preserving feature information, achieving the best results in other metrics, showcasing its superior capability. Compared to the second-best results in

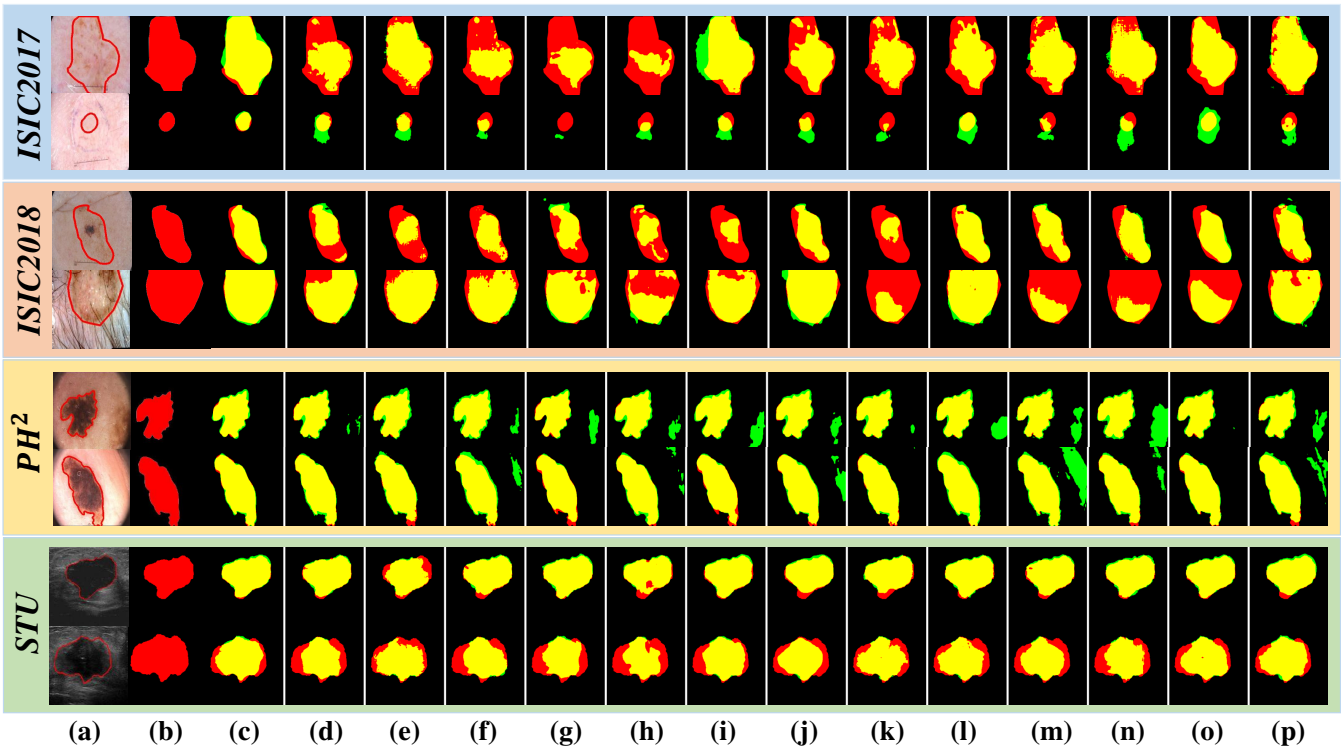


Fig. 4: Visual comparison with different SOTA methods on different datasets. Green pixels indicate the predictions and red pixels represent ground truth. Yellow pixels represent the overlap regions between the prediction and ground truth. (a) Input image. (b) Ground Truth. (c) V-UNet (Ours). (d) U-Net. (e) SegNet. (f) AttU-Net. (g) TransUNet. (h) UNetFormer. (i) META-UNet. (j) MALUNet. (k) MBSNet. (l) MHorUNet. (m) HSH-UNet. (n) VM-UNet. (o) H-vmunet. (p) SkinMamba.

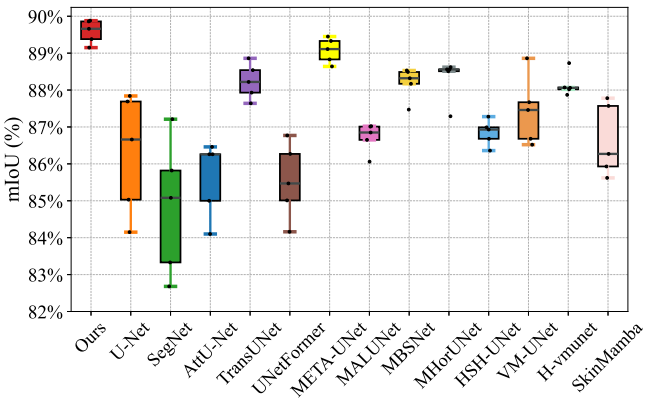


Fig. 5: The box-with-whisker plots of mIoU coefficient compared with all models on PH^2 dataset.

mIoU, DSC, SE, and ACC, our method excels in other metrics, with improvements of 0.55%, 0.29%, 1.2%, and 0.12%. The STU dataset, which belongs to a completely different category from the aforementioned medical image segmentation datasets. Although our method slightly lags behind AttU-Net in ACC, it surpasses the second-best results by 0.59%, 0.32%, and 0.88% in mIoU, DSC, and SE, respectively. Although STU dataset has a fewer number of medical image samples with more redundancy information, our method employs high-order interactions and low-order parallel to effectively reduce redundancy while capturing more effective features. These experimental results demonstrate that V-UNet possesses strong generalization ability and has significant competitiveness.

As shown in Fig. 4, we observe that the segmentation results closely resemble the original ground truth, confirming the effectiveness and great potential of our method. Consistent with the quantitative analysis results, the V-UNet shows superior segmentation performance in boundary details compared to other SOTA models. On the ISIC2017 and ISIC2018 datasets, V-UNet provides accurate lesion localization with smooth boundaries and reduces spurious detections. In contrast, SkinMamba is competitive but occasionally over-segments irregular regions, which results in less stable contours. Meanwhile, on the PH^2 dataset, V-UNet sustains high accuracy across both lesion cores and boundaries. H-vmunet achieves near-perfect alignment with ground truth, but its excessive sensitivity to fine boundary details introduces artifacts that compromise stability. Furthermore, on the STU dataset, V-UNet generates compact masks with fewer spurious detections under noisy ultrasound conditions. Although META-UNet achieves superior boundary refinement, its reliance on fine-grained features leads to slight degradation in blurred regions. Notably, our method improves the stability of prediction performance and demonstrates significant advantages in boundary detail segmentation. Visual comparisons with more visualized images on four medical image datasets are provided in the supplementary files.

As illustrated in Fig. 5, we present a box-with-whisker plot to verify the robustness of V-UNet on the PH^2 dataset. The

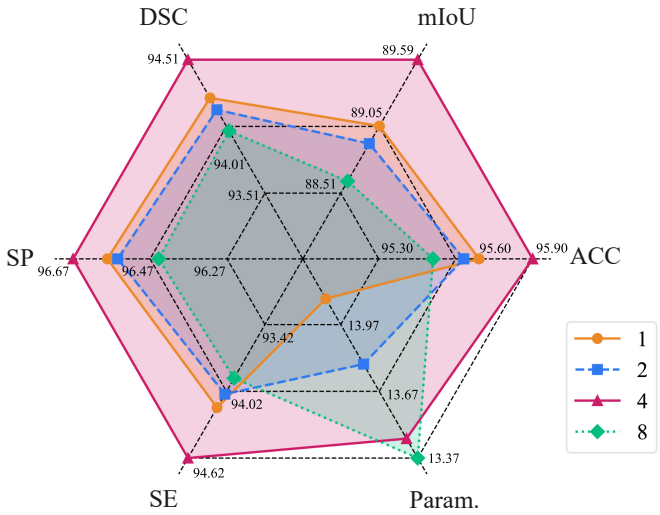


Fig. 6: Rader chart of low-order parallel hyperparameter K on PH^2 dataset.

plot is based on the results of five independent experiments to evaluate thirteen models using the mIoU metric. Additionally, the experimental results of our method are stable and outstanding, demonstrating excellent robustness and strong generalization ability. The box-with-whisker plots for other datasets are provided in the supplementary files.

E. Ablation Study

1) Analysis of Hyperparameter: The HLFF component enables V-UNet achieve superior experimental performance across different types of medical image segmentation tasks. Apart from some predefined parameters, the number of branches K in LP and the different orders in HI, ranging from the third to the sixth layers, also impact the segmentation performance of V-UNet. To maximize the segmentation performance, we conduct hyperparameter discussion experiments on the PH^2 dataset and give a detailed experimental analysis.

The number of branches in LP influences both the model parameters and local detail features. To balance model parameters and local detail features, we discuss the effect of different values of K in LP to maximize the image segmentation performance of V-UNet. As illustrated in Fig. 6, radar chart is used to present the average segmentation performance across different numbers of branches (K = 1, 2, 4, 8). The six axes represent five standard image segmentation evaluation metrics along with the number of model parameters. The more extensive the area enclosed by the radar chart, the better the overall segmentation performance it indicates. Notably, the evaluation metric axes extend outward from the origin in ascending order, whereas the parameter axis extends in descending order, allowing for a more intuitive visualization of comprehensive segmentation performance. As the number of branches increases, the area enclosed by the radar charts initially expands and subsequently contracts. Although $K = 8$ has the smallest number of parameters, it also exhibits a significant drop in segmentation performance. When $K = 4$, the proposed method achieves an optimal trade-off between model complexity and the preservation of local feature details, while effectively mitigating the redundant information. Furthermore, the radar chart demonstrates that $K = 4$ consistently delivers superior segmentation performance across all evaluation metrics. Therefore, we adopt $K = 4$ in the proposed method to maximize the performance of medical image segmentation.

The different orders in HI, ranging from the third to the sixth layers, significantly affect both the model parameters and the introduction of redundant information. We select the same type of order setting based on prior research [38]. As shown in Table V, it presents the detailed experimental results about different orders of HI in V-UNet. Specifically, our method uses orders of $[2, 2, 2, 2]$, which indicates that the spatial interaction order of four HL-VSS modules in the encoder and decoder input stream directions is 2, 2, 2, and 2, respectively. Notably, when the orders are set to $[2, 2, 2, 2]$, the capability of capturing complex spatial dependencies while effectively suppressing image noise in medical imaging has been thoroughly demonstrated, yielding a mean segmentation performance that is significantly superior to other order settings. Besides, increasing the order generally introduces a larger number of parameters and more intricate spatial dependencies, which ultimately lead to a degradation in segmentation performance. Therefore, we adopt the orders of $[2, 2, 2, 2]$ for our method to maximize medical image segmentation performance while ensuring relatively low parameters and FLOPs.

In summary, when the number of low-order parallel branches in LP is set to $K = 4$ and the orders of HI from the third to the sixth layers are configured as $[2, 2, 2, 2]$, V-UNet achieves the best segmentation performance without introducing excessive parameters and redundant information.

2) Analysis of Different Components: To further validate the advantages of our method, we conduct the ablation experiments on the PH^2 dataset. Apart from the variations in the components, the experimental setup remains consistent with previous experiments. As depicted in Table VI, our objective is to validate the effectiveness of different components in V-UNet, considering the ablation results of MSC, HI, and LP.

MSC component. In the HL-VSS module, we utilize the MSC to suppress noise and capture multi-scale feature information from medical images, providing effective features for HLFF component. From Table VI and Fig. 7, MSC significantly suppress noise and minimizing loss by leveraging complementary multi-scale information and employing cross-scale fusion of effective features when HI and LP fuse the features processed by MSC (MSC+HI, MSC+LP). This validates the effectiveness of the MSC component, demonstrating its substantial impact on medical image segmentation. In contrast, when the features extracted by HI and LP are simply fused without MSC (LP+HI), the segmentation performance tends to decline due to interference from noise and the limitation of single feature information. As illustrated in the 4th and 7th columns of Fig. 7, we can find that the feature maps generated by the (LP+HI) exhibit relatively blurred target structure contours, with insufficiently defined texture details along image boundaries. Moreover, the highlighted regions tend to suffer from saturation, while shadowed areas experience a loss of

TABLE V: The experimental results (mean±std) of different orders in V-UNet on PH^2 dataset.

Settings	Param.(M)	FLOPs(G)	mIoU	DSC	SE	SP	ACC	*p*-value
[1, 2, 3, 4]	23.65	1.69	87.86±0.37	93.58±0.22	94.18±1.20	95.74±0.77	95.18±0.18	2.65E-04
[2, 2, 2, 2]	**13.46**	**1.23**	**89.51±0.30**	**94.45±0.16**	94.38±0.52	**96.75±0.15**	**95.89±0.13**	-
[2, 3, 4, 5]	29.96	2.22	88.96±0.71	94.42±0.67	**94.52±0.67**	96.30±0.76	95.64±0.34	1.78E-01
[3, 3, 3, 3]	19.76	1.74	88.28±0.45	93.78±0.28	94.32±0.43	95.96±0.45	95.36±0.21	3.31E-03
[3, 4, 5, 6]	36.26	2.75	88.28±0.61	93.80±0.34	94.02±0.86	96.14±0.41	95.36±0.24	9.56E-03
[4, 4, 4, 4]	26.07	2.27	88.32±0.30	93.80±0.17	94.08±0.66	96.16±0.52	95.40±0.14	1.54E-04

TABLE VI: Ablation experimental results (mean±std) of different components on PH^2 dataset.

Settings	Param.(M)	FLOPs(G)	mIoU	DSC	SE	SP	ACC	*p*-value
HI	9.10	0.74	87.06±1.52	93.08±0.86	93.34±0.99	95.74±0.88	94.84±0.67	2.69E-02
LP	**2.60**	**0.36**	87.94±0.44	93.56±0.27	94.18±0.57	95.78±0.32	95.20±0.21	8.98E-04
HI+LP (HLFF)	11.85	1.12	88.42±0.54	93.87±0.30	93.81±0.42	96.38±0.52	95.42±0.24	5.62E-04
MSC+HI	13.33	1.20	88.73±0.64	94.04±0.31	94.34±0.73	96.25±0.52	95.52±0.26	3.09E-02
MSC+LP	7.09	0.80	88.92±0.55	94.14±0.32	94.34±0.95	96.40±0.36	95.62±0.20	6.42E-02
MSC+LP+HI (V-UNet)	13.46	1.23	**89.60±0.27**	**94.52±0.16**	**94.61±0.44**	**96.68±0.12**	**95.92±0.10**	-

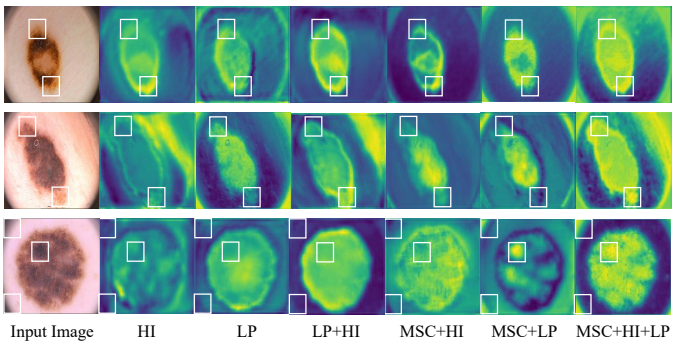


Fig. 7: The feature maps of different components.

fine details, and the pronounced local noise is observed. Notably, the (MSC+LP+HI) effectively suppresses random noise, yielding smoother edge transitions and enhanced local texture fidelity. The ability to capture fine-grained features is notably improved, particularly in the detailed regions highlighted by the white boxes. This is because after MSC suppress noise and captures multi-scale feature information, HI and LP are able to better fuse effective features.

HI and LP components. Although the standalone use of HI or LP can achieve a certain degree of segmentation performance, each exhibits notable limitations. Specifically, HI fails to capture the high-resolution information, whereas LP struggles to effectively suppress redundant information, thereby compromising its robustness in complex background scenarios. From Fig. 7, the integration of LP and HI features (LP + HI) through high-low-order feature fusion significantly enhances the ability of model to extract salient features, effectively mitigating the respective shortcomings of HI and LP. In comparison to performing segmentation tasks using HI and LP individually, the features processed by MSC (MSC+HI, MSC+LP) demonstrate improved feature extraction capabilities and greater resilience to noise interference. However, even after MSC processing, using a single component remains susceptible to the inherent limitations previously identified for HI and LP. Therefore, we propose a fusion strategy that combines MSC-processed HI and LP features (MSC+LP+HI) via high-low-order feature fusion. As illustrated in the 5th, 6th, and 7th columns of Fig. 7, compared to (MSC+HI) and (MSC+LP), the (MSC+LP+HI) employs feature fusion to produce more pronounced hierarchical texture details in the feature regions (as indicated by the white boxes). The feature fusion not only preserves critical structural features but also effectively suppresses background noise, thereby further enhancing the distinction between the target regions and the true boundary information. This method maximally reduces the impact of noise and redundant information, thereby substantially improving segmentation performance and enhancing the potential for clinical applications in medical image segmentation.

A series of ablation experiments provide compelling evidence that our method not only outperforms existing methods but also benefits from MSC, HI, and LP. The comprehensive experiments and comparative results ultimately confirm that V-UNet is a superior model.

V. Conclusion

In this paper, we propose a novel medical image segmentation model named V-UNet. Its aim is to delineate image edge details and reduce noise and redundant information in medical images through the high-low-order feature fusion visual state space (HL-VSS). Specifically, a multi-scale spatial convolution (MSC) component is designed to suppress noise and capture multi-scale feature information, effectively extracting edge and detail features. Then, we propose a high-order interaction and low-order parallel feature fusion (HLFF) component to minimize the introduction of redundant information and enhance local feature extraction capabilities. For the overall architecture, we still employ the classic and efficient U-shaped architecture to construct V-UNet for the medical image segmentation task. Experimental results on four public medical image datasets show that our proposed method is better than other SOTA models in medical image segmentation. This further demonstrates the effectiveness of V-UNet in suppressing various types of noise in medical images, as well as its accuracy and potential for clinical segmentation applications.

However, the limitation of this paper does not consider the impact of image frequency noise on the segmentation performance. Our future work plans to design a novel 2D discrete cosine transform module to generate channel attention maps by extracting frequency statistics, suppressing the influence of noisy channels in images. Besides, it would be interesting to apply our proposed image segmentation method to address other segmentation challenges [72]–[74].

References

[1] J. E. Iglesias and M. R. Sabuncu, "Multi-atlas segmentation of biomedical images: A survey," *Medical Image Analysis*, vol. 24, no. 1, pp. 205–219, 2015.

[2] D.-T. Lin, C.-C. Lei, and S.-W. Hung, "Computer-aided kidney segmentation on abdominal ct images," *IEEE Transactions on Information Technology in Biomedicine*, vol. 10, no. 1, pp. 59–65, 2006.

[3] J. Yanase and E. Triantaphyllou, "A systematic survey of computer-aided diagnosis in medicine: Past and present developments," *Expert Systems with Applications*, vol. 138, p. 112821, 2019.

[4] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[5] N. Salpea, P. Tzouveli, and D. Kollias, "Medical image segmentation: A review of modern architectures," in *European Conference on Computer Vision*. Springer, 2022, pp. 691–708.

[6] M. Antonelli, A. Reinke, S. Bakas, K. Farahani, A. Kopp-Schneider, B. A. Landman, G. Litjens, B. Menze, O. Ronneberger, R. M. Summers *et al.*, "The medical segmentation decathlon," *Nature Communications*, vol. 13, no. 1, p. 4128, 2022.

[7] R. Wang, T. Lei, R. Cui, B. Zhang, H. Meng, and A. K. Nandi, "Medical image segmentation using deep learning: A survey," *IET Image Processing*, vol. 16, no. 5, pp. 1243–1267, 2022.

[8] D. L. Pham, C. Xu, and J. L. Prince, "Current methods in medical image segmentation," *Annual Review of Biomedical Engineering*, vol. 2, no. 1, pp. 315–337, 2000.

[9] M. H. Hesamian, W. Jia, X. He, and P. Kennedy, "Deep learning techniques for medical image segmentation: achievements and challenges," *Journal of Digital Imaging*, vol. 32, pp. 582–596, 2019.

[10] T. Dhar, N. Dey, S. Borra, and R. S. Sherratt, "Challenges of deep learning in medical image analysis—improving explainability and trust," *IEEE Transactions on Technology and Society*, vol. 4, no. 1, pp. 68–75, 2023.

[11] H. Zhang, S. Cholleti, S. A. Goldman, and J. E. Fritts, "Meta-Evaluation of image segmentation using machine learning," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 1. IEEE, 2006, pp. 1138–1145.

[12] T. A. Soomro, L. Zheng, A. J. Afifi, A. Ali, S. Soomro, M. Yin, and J. Gao, "Image segmentation for mr brain tumor detection using machine learning: A review," *IEEE Reviews in Biomedical Engineering*, vol. 16, pp. 70–90, 2022.

[13] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 7, pp. 3523–3542, 2021.

[14] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A survey of convolutional neural networks: Analysis, applications, and prospects," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 12, pp. 6999–7019, 2021.

[15] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-assisted Intervention – MICCAI 2015*. Springer International Publishing, 2015, pp. 234–241.

[16] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[17] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, "ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 162, pp. 94–114, 2020.

[18] 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," *arXiv preprint arXiv:1804.03999*, 2018.

[19] J. Long, M. Li, and X. Wang, "Integrating spatial details with long-range contexts for semantic segmentation of very high-resolution remote-sensing images," *IEEE Geoscience and Remote Sensing Letters*, vol. 20, pp. 1–5, 2023.

[20] 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 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations (ICLR)*, 2021.

[21] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 12 077–12 090, 2021.

[22] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6881–6890.

[23] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 7262–7272.

[24] J. Chen, J. Mei, X. Li, Y. Lu, Q. Yu, Q. Wei, X. Luo, Y. Xie, E. Adeli, Y. Wang *et al.*, "TransUNet: Rethinking the u-net architecture design for medical image segmentation through the lens of transformers," *Medical Image Analysis*, vol. 97, p. 103280, 2024.

[25] L. Wang, R. Li, C. Zhang, S. Fang, C. Duan, X. Meng, and P. M. Atkinson, "UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 190, pp. 196–214, 2022.

[26] H. You, Y. Xiong, X. Dai, B. Wu, P. Zhang, H. Fan, P. Vajda, and Y. C. Lin, "Castling-ViT: Compressing self-attention via switching towards linear-angular attention at vision transformer inference," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14 431–14 442.

[27] F. Babiloni, I. Marras, J. Deng, F. Kokkinos, M. Maggioni, G. Chrysos, P. Torr, and S. Zafeiriou, "Linear complexity self-attention with 3rd order polynomials," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 11, pp. 12 726–12 737, 2023.

[28] M.-H. Guo, Z.-N. Liu, T.-J. Mu, and S.-M. Hu, "Beyond Self-Attention: External attention using two linear layers for visual tasks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 5, pp. 5436–5447, 2022.

[29] A. Gu, *Modeling Sequences with Structured State Spaces*. Stanford University, 2023.

[30] A. Gu, I. Johnson, K. Goel, K. Saab, T. Dao, A. Rudra, and C. Ré, "Combining recurrent, convolutional, and continuous-time models with linear state space layers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 572–585, 2021.

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

[32] K. Goel, A. Gu, C. Donahue, and C. Ré, "It's raw! audio generation with state-space models," in *International Conference on Machine Learning*. PMLR, 2022, pp. 7616–7633.

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

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

[35] Z. Wang, J.-Q. Zheng, Y. Zhang, G. Cui, and L. Li, "Mamba-UNet: Unet-like pure visual mamba for medical image segmentation," *arXiv preprint arXiv:2402.05079*, 2024.

[36] C. Jiang, R. Wu, Y. Liu, Y. Wang, Q. Chang, P. Liang, and Y. Fan, "A high-order focus interaction model and oral ulcer dataset for oral ulcer segmentation," *Scientific Reports*, vol. 14, no. 1, p. 20085, 2024.

[37] R. Wu, Y. Liu, P. Liang, and Q. Chang, "Only Positive Cases: 5-fold high-order attention interaction model for skin segmentation derived classification," *arXiv preprint arXiv:2311.15625*, 2023.

[38] Y. Rao, W. Zhao, Y. Tang, J. Zhou, S. N. Lim, and J. Lu, "HorNet: Efficient high-order spatial interactions with recursive gated convolutions," *Advances in Neural Information Processing Systems*, vol. 35, pp. 10 353–10 366, 2022.

[39] X. Liu, C. Zhang, and L. Zhang, "Vision Mamba: A comprehensive survey and taxonomy," *arXiv preprint arXiv:2405.04404*, 2024.

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

[41] Y. Ye, P. Huang, Y. Sun, and D. Shi, "MBSNet: A deep learning model for multibody dynamics simulation and its application to a vehicle-track system," *Mechanical Systems and Signal Processing*, vol. 157, p. 107716, 2021.

[42] R. Wu, P. Liang, X. Huang, L. Shi, Y. Gu, H. Zhu, and Q. Chang, "MHorUNet: High-order spatial interaction unet for skin lesion segmentation," *Biomedical Signal Processing and Control*, vol. 88, p. 105517, 2024.

[43] R. Wu, H. Lv, P. Liang, X. Cui, Q. Chang, and X. Huang, "HSH-UNet: Hybrid selective high order interactive U-shaped model for automated skin lesion segmentation," *Computers in Biology and Medicine*, vol. 168, p. 107798, 2024.

[44] S. Mehta, M. Rastegari, A. Caspi, L. Shapiro, and H. Hajishirzi, "ESPNet: Efficient spatial pyramid of dilated convolutions for semantic segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 552–568.

[45] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

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

[47] Y. Zhang, H. Liu, and Q. Hu, "TransFuse: Fusing transformers and cnns for medical image segmentation," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*. Springer International Publishing, 2021, pp. 14–24.

[48] W. Zhou, H. Wu, and Q. Jiang, "MDNet: Mamba-effective diffusion-distillation network for rgb-thermal urban dense prediction," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[49] M. Ju, S. Xie, and F. Li, "Improving skip connection in u-net through fusion perspective with mamba for image dehazing," *IEEE Transactions on Consumer Electronics*, 2024.

[50] Q. Liu, J. Yue, Y. Fang, S. Xia, and L. Fang, "HyperMamba: A spectral-spatial adaptive mamba for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[51] J. Ruan, J. Li, and S. Xiang, "VM-UNet: Vision mamba unet for medical image segmentation," *ACM Transactions on Multimedia Computing, Communications and Applications*, 2024.

[52] R. Wu, Y. Liu, P. Liang, and Q. Chang, "Ultralight VM-UNet: Parallel vision mamba significantly reduces parameters for skin lesion segmentation," *arXiv preprint arXiv:2403.20035*, 2024.

[53] Z. Xing, T. Ye, Y. Yang, G. Liu, and L. Zhu, "SegMamba: Long-range sequential modeling mamba for 3D medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2024, pp. 578–588.

[54] C. Zheng, J. Nie, Z. Wang, N. Song, J. Wang, and Z. Wei, "High-order semantic decoupling network for remote sensing image semantic segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–15, 2023.

[55] X. Sun, Y. Zhang, C. Chen, S. Xie, and J. Dong, "High-order paired-aspp for deep semantic segmentation networks," *Information Sciences*, vol. 646, p. 119364, 2023.

[56] K. Zhang, Y. Wu, M. Dong, B. Liu, D. Liu, and Q. Liu, "Deep object co-segmentation and co-saliency detection via high-order spatial-semantic network modulation," *IEEE Transactions on Multimedia*, vol. 25, pp. 5733–5746, 2022.

[57] R. Wu, Y. Liu, P. Liang, and Q. Chang, "H-vmunet: High-order vision mamba unet for medical image segmentation," *Neurocomputing*, vol. 624, p. 129447, 2025.

[58] J. Ruan, S. Xiang, M. Xie, T. Liu, and 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.

[59] S. Nishida, T. Ledgeway, and M. Edwards, "Dual multiple-scale processing for motion in the human visual system," *Vision Research*, vol. 37, no. 19, pp. 2685–2698, 1997.

[60] Z. Liu, D. Chen, W. Pei, Q. Ma *et al.*, "Scale-teaching: Robust multi-scale training for time series classification with noisy labels," *Advances in Neural Information Processing Systems*, vol. 36, pp. 33 726–33 757, 2023.

[61] M. M. Rahman, M. Munir, and R. Marculescu, "EMCAD: Efficient multi-scale convolutional attention decoding for medical image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 11 769–11 779.

[62] I. O. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, J. Uszkoreit *et al.*, "MLP-Mixer: An all-MLP architecture for vision," *Advances in Neural Information Processing Systems*, vol. 34, pp. 24 261–24 272, 2021.

[63] H. Wu, Z. Zhao, and Z. Wang, "META-Unet: Multi-scale efficient transformer attention unet for fast and high-accuracy polyp segmentation," *IEEE Transactions on Automation Science and Engineering*, vol. 21, no. 3, pp. 4117–4128, 2024.

[64] S. Zou, M. Zhang, B. Fan, Z. Zhou, and X. Zou, "SkinMamba: A precision skin lesion segmentation architecture with cross-scale global state modeling and frequency boundary guidance," *arXiv preprint arXiv:2409.10890*, 2024.

[65] 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.

[66] N. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti *et al.*, "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC)," *arXiv preprint arXiv:1902.03368*, 2019.

[67] T. Mendonça, P. M. Ferreira, J. S. Marques, A. R. Marcal, and J. Rozeira, "PH 2-a dermoscopic image database for research and benchmarking," in *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2013, pp. 5437–5440.

[68] Z. Zhuang, N. Li, A. N. Joseph Raj, V. G. Mahesh, and S. Qiu, "An RDAU-NET model for lesion segmentation in breast ultrasound images," *PloS One*, vol. 14, no. 8, p. e0221535, 2019.

[69] P. Zhou, X. Xie, Z. Lin, and S. Yan, "Towards understanding convergence and generalization of AdamW," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 9, pp. 6486–6493, 2024.

[70] A. Kumar, Y. Guo, X. Huang, L. Ren, and X. Liu, "SeaBird: Segmentation in bird's view with dice loss improves monocular 3D detection of large objects," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 10 269–10 280.

[71] T. Eelbode, J. Bertels, M. Berman, D. Vandermeulen, F. Maes, R. Bisschops, and M. B. Blaschko, "Optimization for medical image segmentation: theory and practice when evaluating with dice score or jaccard index," *IEEE Transactions on Medical Imaging*, vol. 39, no. 11, pp. 3679–3690, 2020.

[72] J.-H. Nam, N. S. Syazwany, S. J. Kim, and S.-C. Lee, "Modality-agnostic domain generalizable medical image segmentation by Multi-Frequency in Multi-Scale Attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 11 480–11 491.

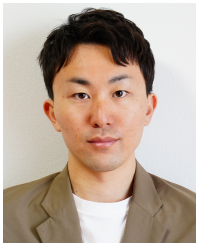
[73] X. Shen, J. Yang, C. Wei, B. Deng, J. Huang, X.-S. Hua, X. Cheng, and K. Liang, "DCT-Mask: Discrete cosine transform mask representation for instance segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8720–8729.

[74] D. Ravì, M. Bober, G. M. Farinella, M. Guarnera, and S. Battiato, "Semantic segmentation of images exploiting dct based features and random forest," *Pattern Recognition*, vol. 52, pp. 260–273, 2016.

Qianhang Du received the B.E. degree in data science and big data technology from University of Bengbu, Anhui, China in 2023. He is currently pursuing his M.E. degree at University of Toyama, Japan. His current research interests include deep learning, neural networks, medical image segmentation.

Zhenyu Lei (Member, IEEE) received the Ph.D. degree in Science and Engineering from the University of Toyama, Toyama, Japan, in 2023. He is currently an Assistant Professor with the Faculty of Engineering, University of Toyama, Japan. His current research interests include evolutionary computation, machine learning, and neural network for real-world applications and optimization problems.

Jiujun Cheng received the Ph.D. degree from the Beijing University of Posts and Telecommunications, in 2006. In 2009, he was a Visiting Professor with Aalto University, Espoo, Finland. He is currently a Professor with Tongji University, Shanghai, China. He has over 80 publications including conference and journal articles. His current research interests include mobile computing, complex networks, Internet of Vehicles, and autonomous cars.



Masaaki Omura (Member, IEEE) received the Ph.D. degree in engineering from Chiba University, Japan, in 2020. While he was a Ph.D. Student, he was supported by the Research Fellowship for Young Scientists (DC1) of the Japan Society for the Promotion of Science (JSPS). From 2020 to 2021, he has been working as a Postdoctoral Researcher with University of Toyama, Japan, granted by the Research Fellowship for Young Scientists (PD) of the JSPS. From 2022 to 2024, he was an Assistant Professor with University of Toyama. Since 2024, he has been an Associate Professor at University of Toyama. He is a member of the Acoustical Society of Japan, a Fellow of the Japan Society of Ultrasound in Medicine.

Hideyuki Hasegawa (Member, IEEE) received the B.S. degree in control engineering from Nanjing University of Science and Technology, Nanjing, China, in 1983, the M.S. degree in automatic control from Beijing Institute of Technology, Beijing, China, in 1986, and the Ph.D. degree from Tohoku University, Sendai, Japan, in 2001. He has been a Professor with the Graduate School of Science and Engineering for Research, University of Toyama, Toyama, Japan, since 2015. His research interests are medical ultrasonics. Dr. Hasegawa received the research fellowships for young scientists (DC1). He serves as a Technical Program Committee Member of the IEEE International Ultrasonics Symposium and an Associate Editor for IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control.

Shangce Gao (Senior Member, IEEE) received his Ph.D. degree in Innovative Life Science from the University of Toyama, Toyama, Japan, in 2011. He is currently a Professor with the Faculty of Engineering at the University of Toyama. His research interests focus on brain-inspired neural networks and their applications in medical diagnosis and drug discovery. He serves as an Associate Editor for several international journals, including IEEE Transactions on Neural Networks and Learning Systems and the IEEE/CAA Journal of Automatica Sinica.

Supplementary File for "High-order Interaction and Low-order Parallelization of Features Fusion with Novel Mamba-UNet Architecture for Medical Image Segmentation"

Qianhang Du, Zhenyu Lei, *Member, IEEE*, Jiujun Cheng, Masaaki Omura, *Member, IEEE*, Hideyuki Hasegawa, *Member, IEEE* and Shangce Gao, *Senior Member, IEEE*

I. Preliminaries

A. State-Space Models

SSM is based on the principles of control theory and provides linear scalability with sequence length for long-range dependency [1]. Inspired by classical continuous systems, SSM has attracted widespread attention for its ability to map sequences $x(t) \in \mathbb{R}^{1\times N}$ to an output $y(t) \in \mathbb{R}^{1\times N}$ through a hidden state $h(t) \in \mathbb{R}^{N\times N}$. It can be formulated by a linear ordinary differential equations as follows:

$$\begin{aligned} h'(t) &= \boldsymbol{A}h(t) + \boldsymbol{B}x(t), \\ y(t) &= \boldsymbol{C}h(t), \end{aligned} \tag{1}$$

where $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ denotes the state matrix, $\boldsymbol{B} \in \mathbb{R}^{N\times 1}$ and $\boldsymbol{C} \in \mathbb{R}^{1\times N}$ represent the projection matrices for a size N. In deep learning, it is necessary to discretize continuous systems. Convert continuous (1) into discrete-time representations, thereby maintaining consistency with the data sampling rate. A common discretization rule is zero-order hold, and the discrete-time SSM can be expressed below:

$$\begin{aligned} \overline{\boldsymbol{A}} &= \exp(\Delta\boldsymbol{A}), \\ \overline{\boldsymbol{B}} &= (\Delta\boldsymbol{A})^{-1}(\exp(\Delta\boldsymbol{A}) - \boldsymbol{I}) \cdot \Delta\boldsymbol{B}, \end{aligned} \tag{2}$$

where Δ is a timescale parameter, $\overline{\boldsymbol{A}}$ and $\overline{\boldsymbol{B}}$ are the transformed discrete parameters.

Following discretization, the SSM-based model can be calculated through global convolution or linear recursion, which are specified by the following equations:

$$\begin{aligned} h(t) &= \overline{\boldsymbol{A}}h(t-1) + \overline{\boldsymbol{B}}x(t), \\ y(t) &= \boldsymbol{C}h(t), \end{aligned} \tag{3}$$

$$\begin{aligned} \overline{\boldsymbol{K}} &= \left(\boldsymbol{C}\overline{\boldsymbol{B}}, \boldsymbol{C}\overline{\boldsymbol{A}\boldsymbol{B}}, \ldots, \boldsymbol{C}\overline{\boldsymbol{A}}^{L-1}\overline{\boldsymbol{B}}\right), \\ y(t) &= x(t) * \overline{\boldsymbol{K}}, \end{aligned} \tag{4}$$

where L denotes the length of the input sequence and $\overline{\boldsymbol{K}} \in \mathbb{R}^{L}$ represents a structured convolution kernel.

B. 2D-selective-scan

We illustrate the details of SS2D in Fig. S.I. It primarily consists of three parts: a scan expansion operation, a structured state-space sequence model with a selection mechanism block, named S6, and a scan merge operation [1], [2]. In Fig. S.I, the supplied image is unfolded into sequences along four distinct directions by the scan expansion operation. Then, the S6 block processes these sequences for feature extraction, making sure the information is captured from various directions to extract directional features. Subsequently, the scan merge operation combines the sequences from the four directions, restoring the output image to the same dimensions as the input. Through real-time filtering out irrelevant data, S6 allows the model to identify and keep essential information.

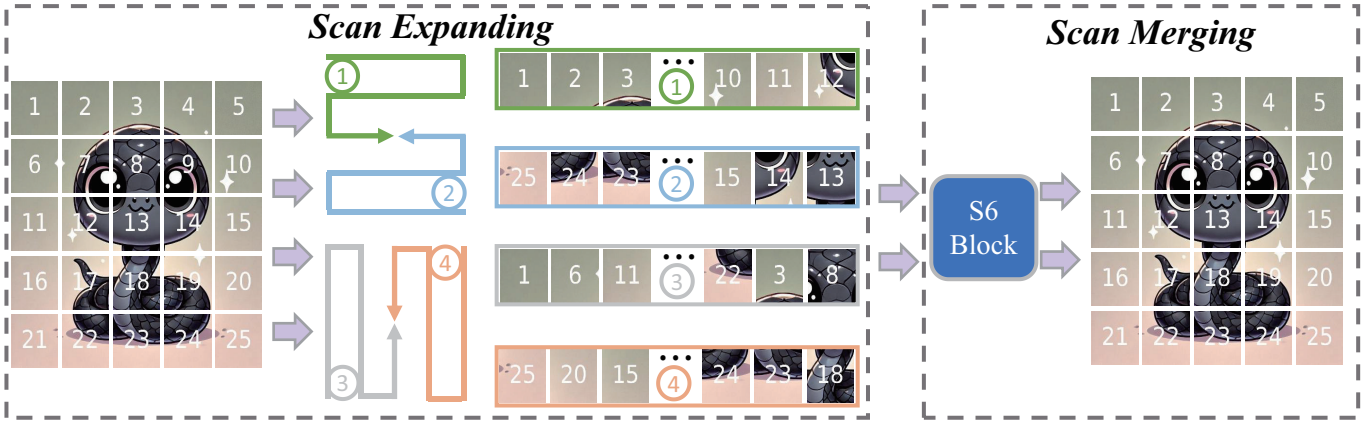


Fig. S.I. The image description of SS2D operation.

II. Complete experimental results

Figs. S.VI – S.IX, we visualize the segmentation results of V-UNet and SOTA models on four public medical image datasets. The visualized images show that the segmentation results are highly similar to the original ground truth, confirming the effectiveness and great potential of our method. Furthermore, it is evident that the SOTA models exhibit excessive predictions and insufficient prediction performance across the four public medical image datasets. Although SkinMamba demonstrates good segmentation performance on images with large lesion areas, it tends to over-predict in regions with small lesions. This issue arises because SkinMamba can establish a global receptive field for global features but its segmentation performance significantly degrades when dealing with small local lesions. In contrast, HSH-UNet excels at segmenting small local lesion areas but exhibits insufficient segmentation performance for larger lesion areas. Compared to the aforementioned methods, the proposed V-UNet first reduces noise impact through MSC and then captures both global and local features via the HLFF component, showcasing excellent image segmentation performance. Notably, V-UNet improves the stability of prediction performance and offers significant advantages in segmenting detailed boundaries.

In summary, V-UNet utilizes the VSS module for medical image processing. The VSS module mainly consists of two core components: multi-scale spatial convolution (MSC) and high-low-order feature fusion (HLFF). The former preliminary filters out noise and capture multi-scale feature information of medical images. The latter obtains the processed features, further reducing the redundant information through high-order interaction with 2D-selective-scan, and fusing the local features obtained by low-order parallel Mamba, thereby capturing deeper medical image features. V-UNet demonstrates strong segmentation performance in separating various anatomical features or abnormalities in medical images, highlighting its reliability and potential for clinical applications.

References

[1] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.
[2] X. Liu, C. Zhang, and L. Zhang, "Vision Mamba: A comprehensive survey and taxonomy," *arXiv preprint arXiv:2405.04404*, 2024.

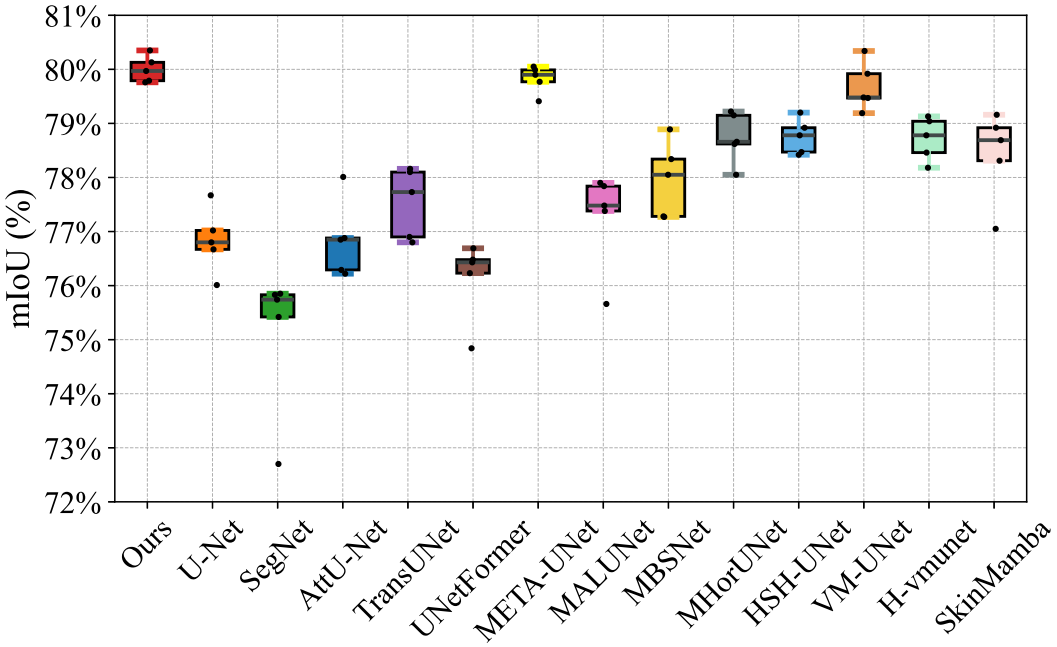


Fig. S.II. The box-with-whisker plots of mIoU coefficient compared with all models on ISIC2017 dataset.

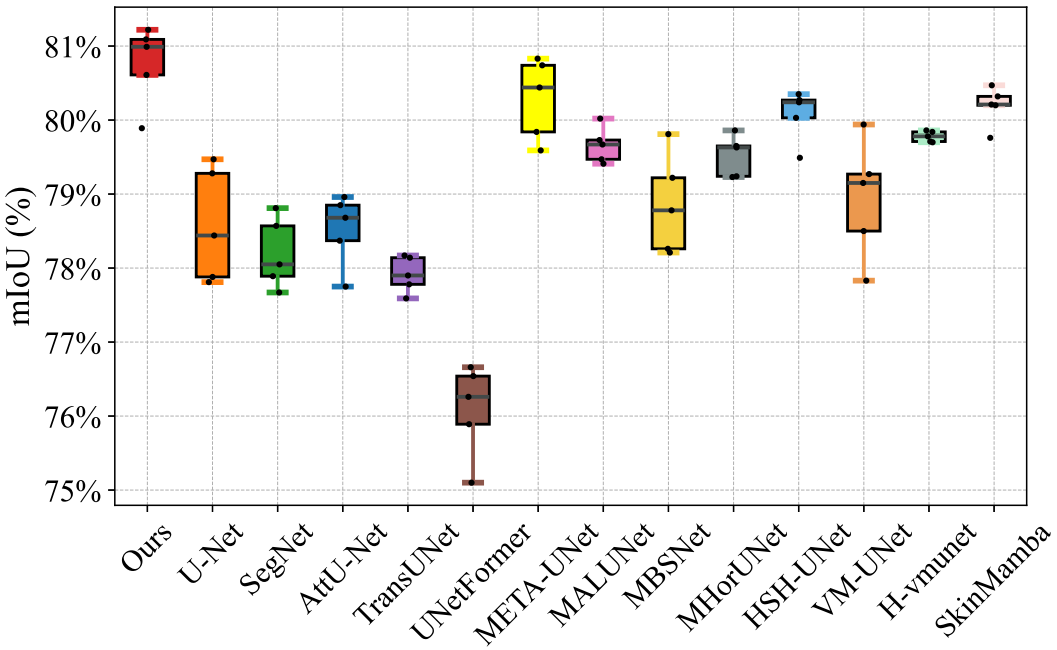


Fig. S.III. The box-with-whisker plots of mIoU coefficient compared with all models on ISIC2018 dataset.

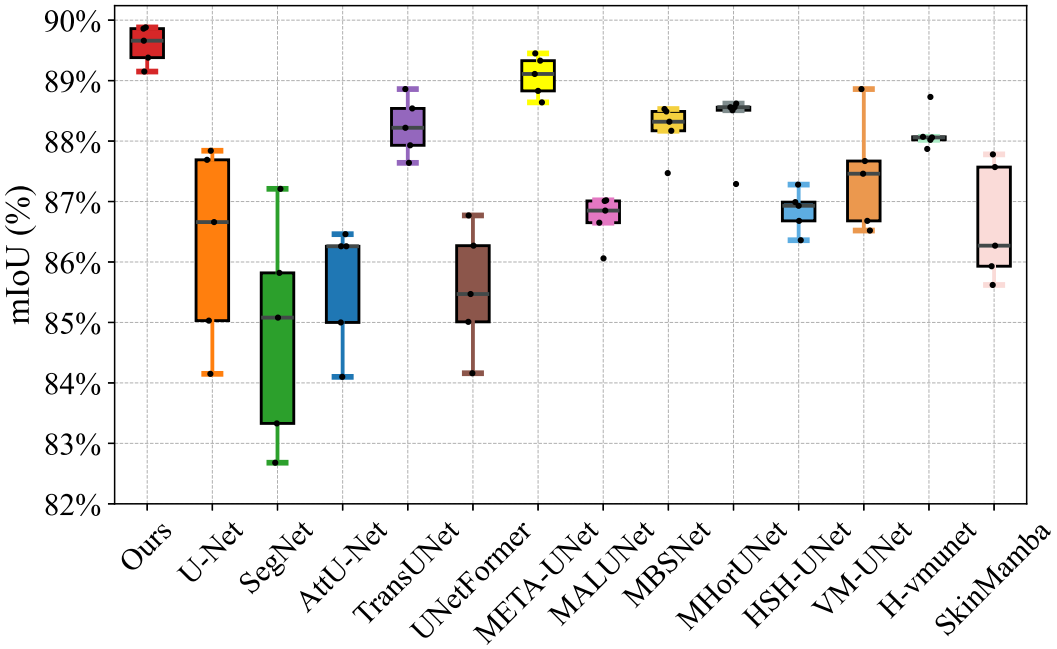


Fig. S.IV. The box-with-whisker plots of mIoU coefficient compared with all models on PH^2 dataset.

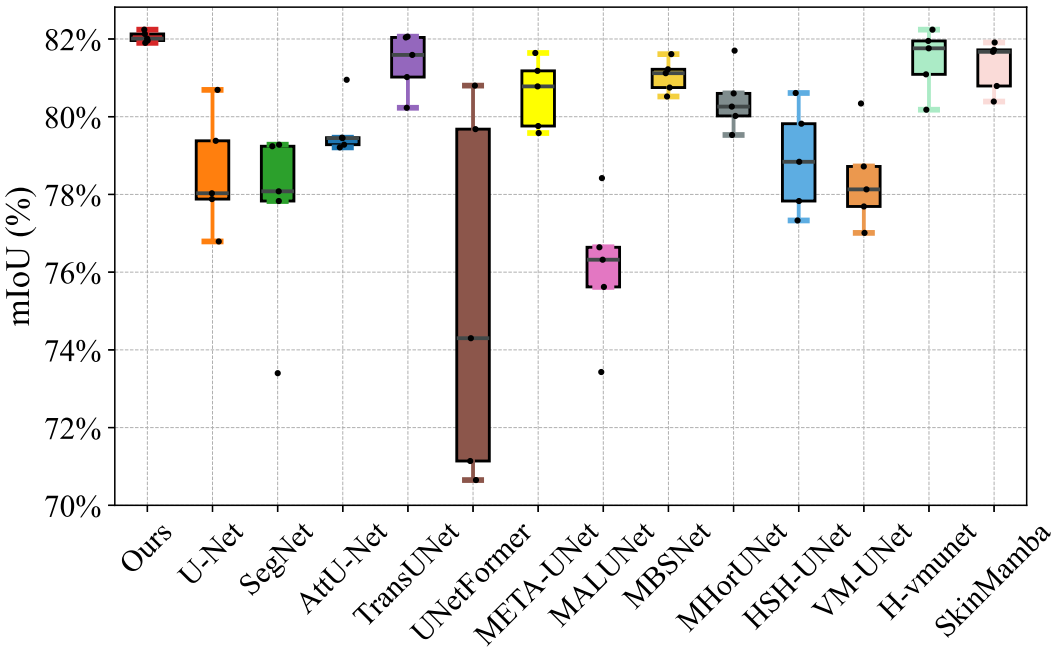


Fig. S.V. The box-with-whisker plots of mIoU coefficient compared with all models on STU dataset.

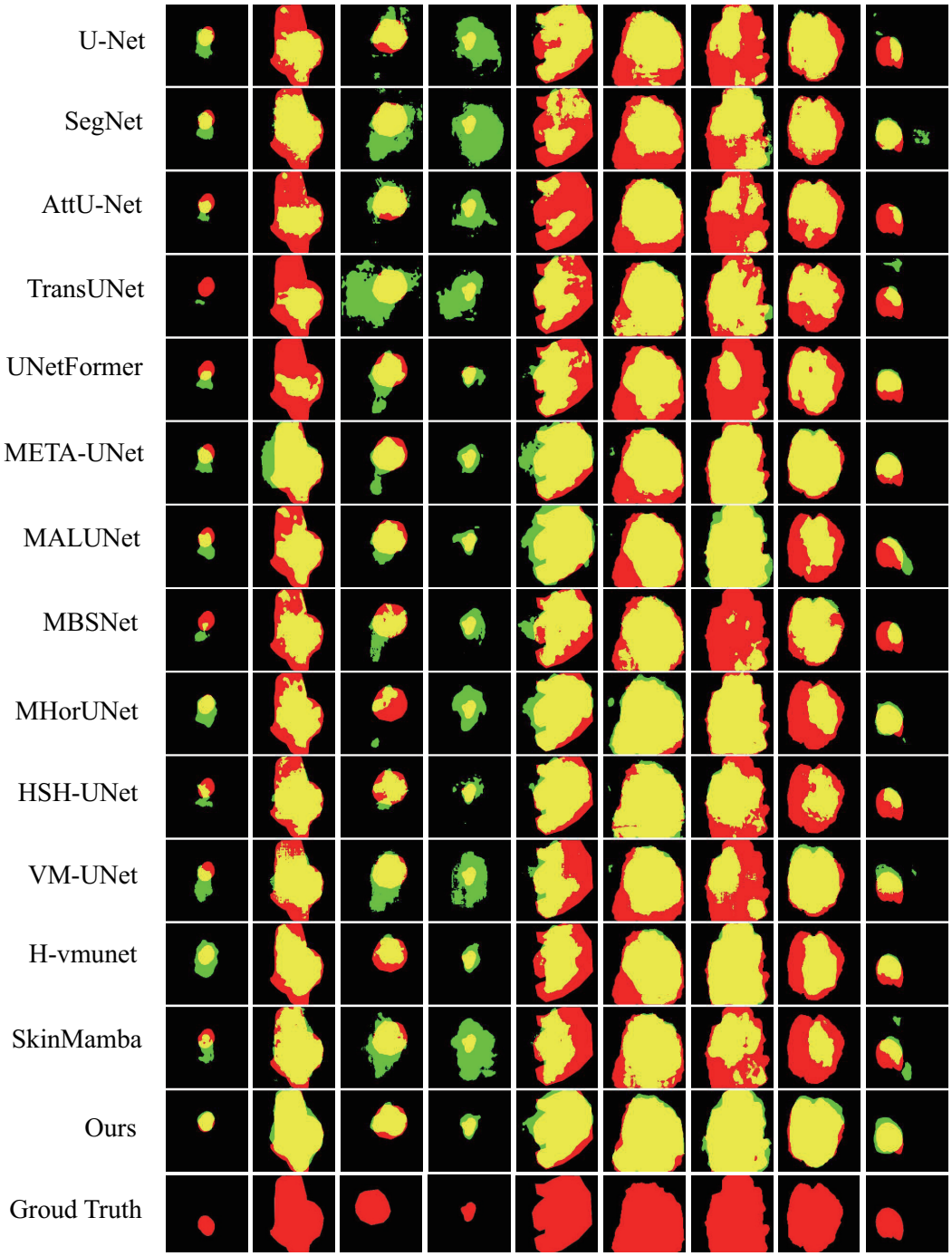


Fig. S.VI. Visual comparison with different state-of-the-art methods on ISIC2017 dataset.

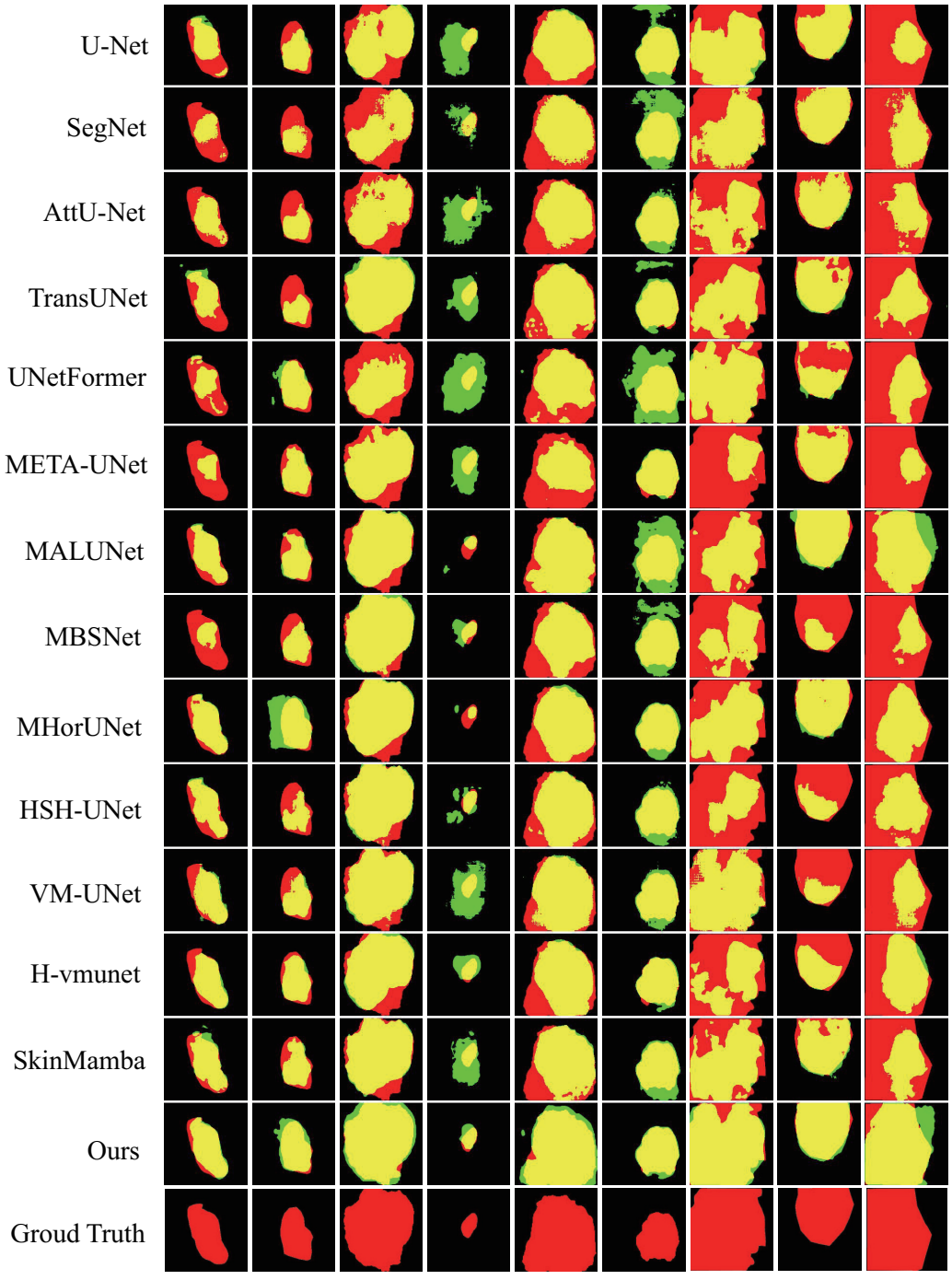


Fig. S.VII. Visual comparison with different state-of-the-art methods on ISIC2018 dataset.

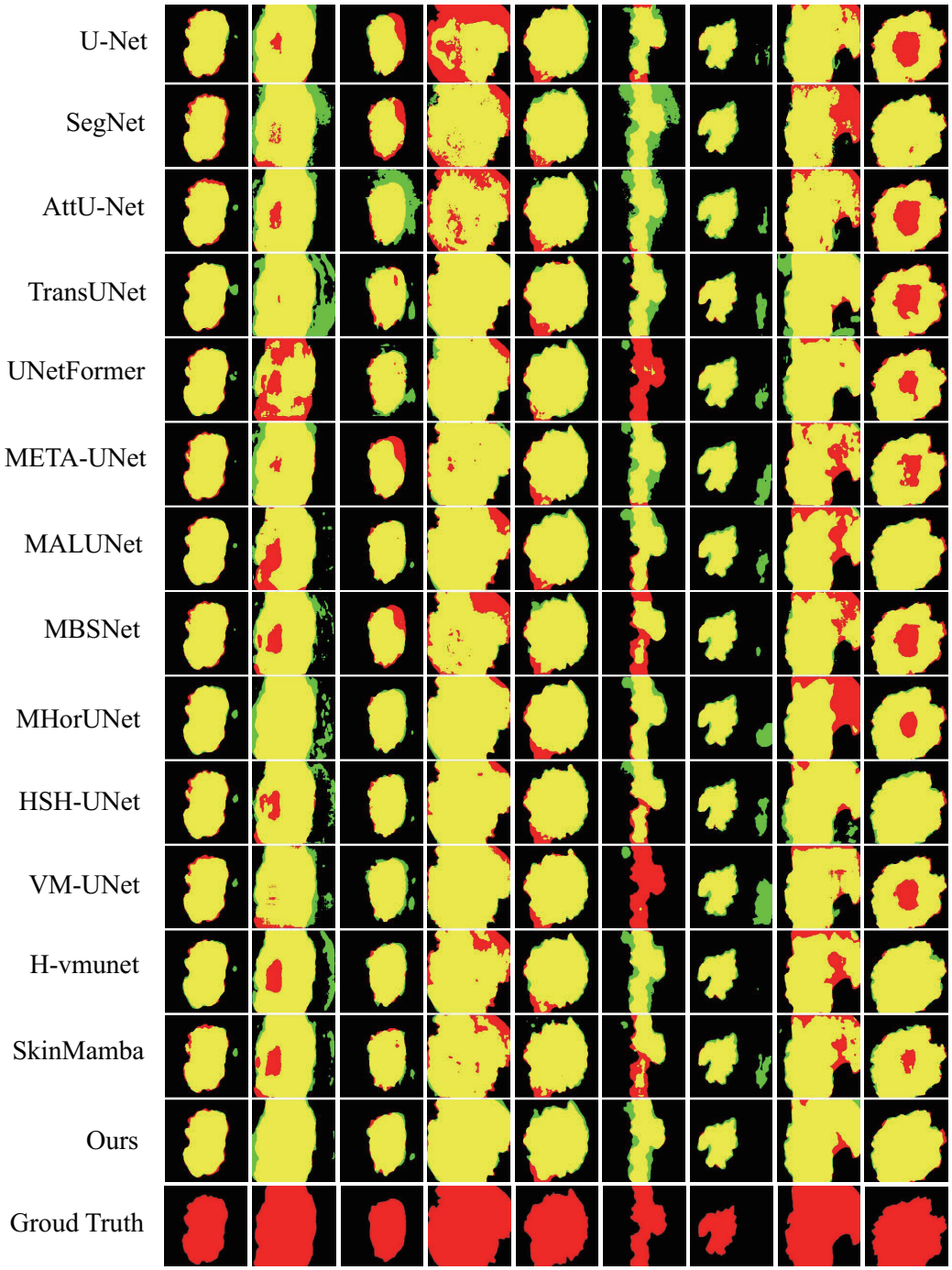


Fig. S.VIII. Visual comparison with different state-of-the-art methods on PH^2 dataset.

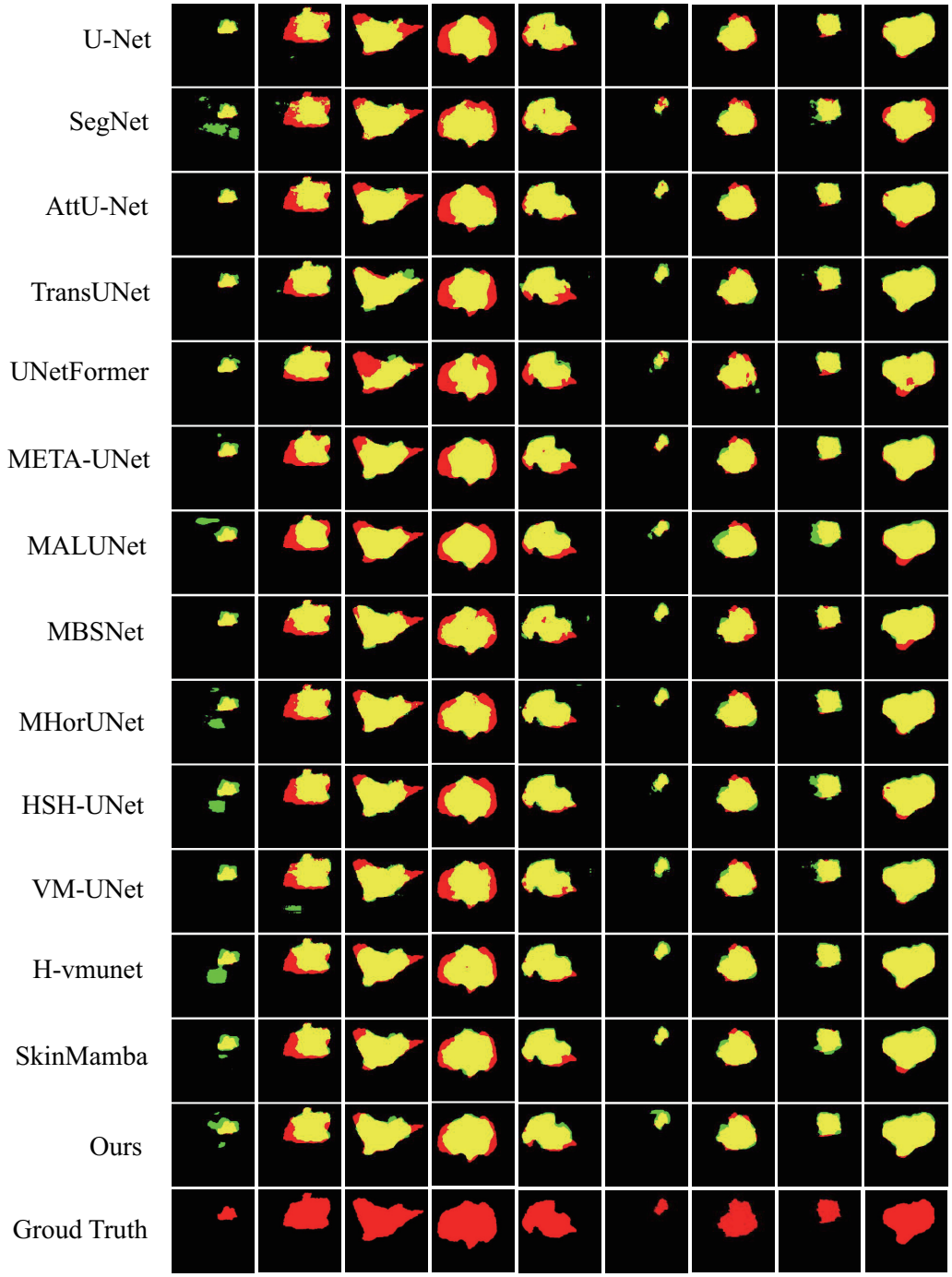


Fig. S.IX. Visual comparison with different state-of-the-art methods on STU dataset.