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. S1. 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. S1, 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.

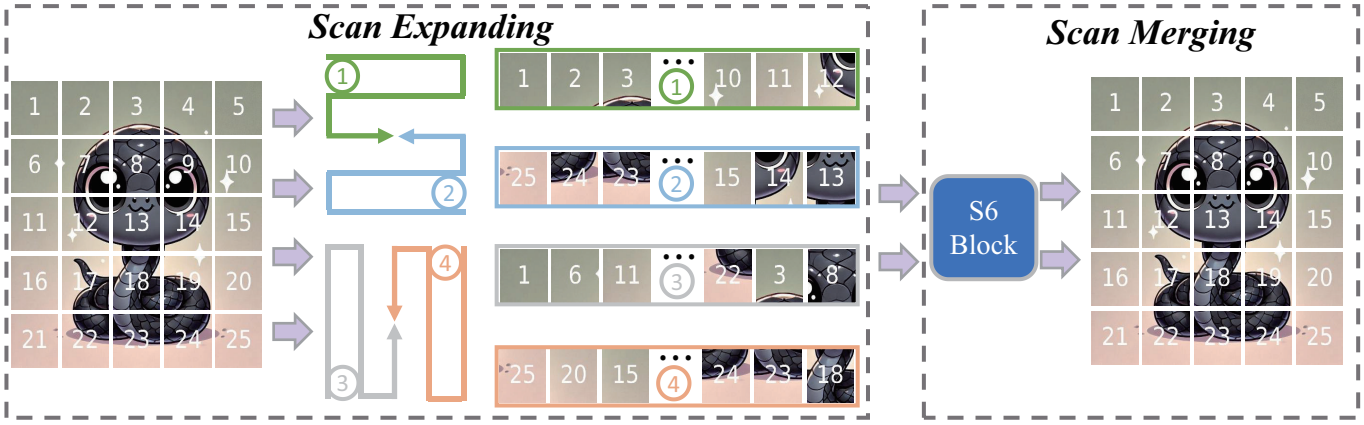

Fig. S1. The image description of SS2D operation.

II. Complete Experimental Results

Figs. S6 – S9, 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.

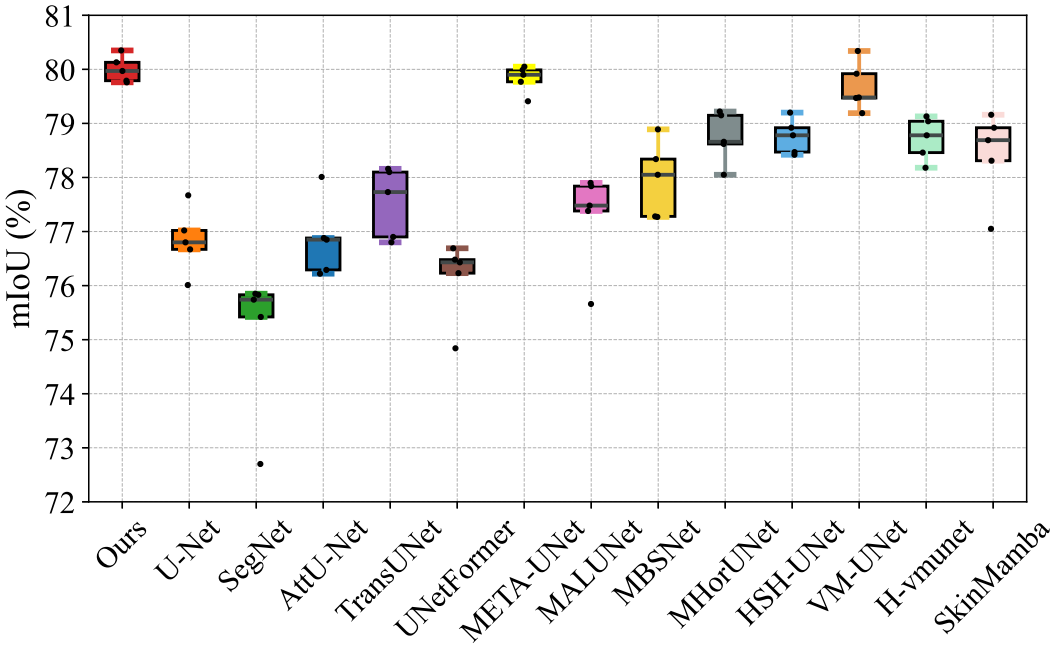

Fig. S2. The box-with-whisker plots of mIoU coefficient compared with all models on ISIC2017 dataset.

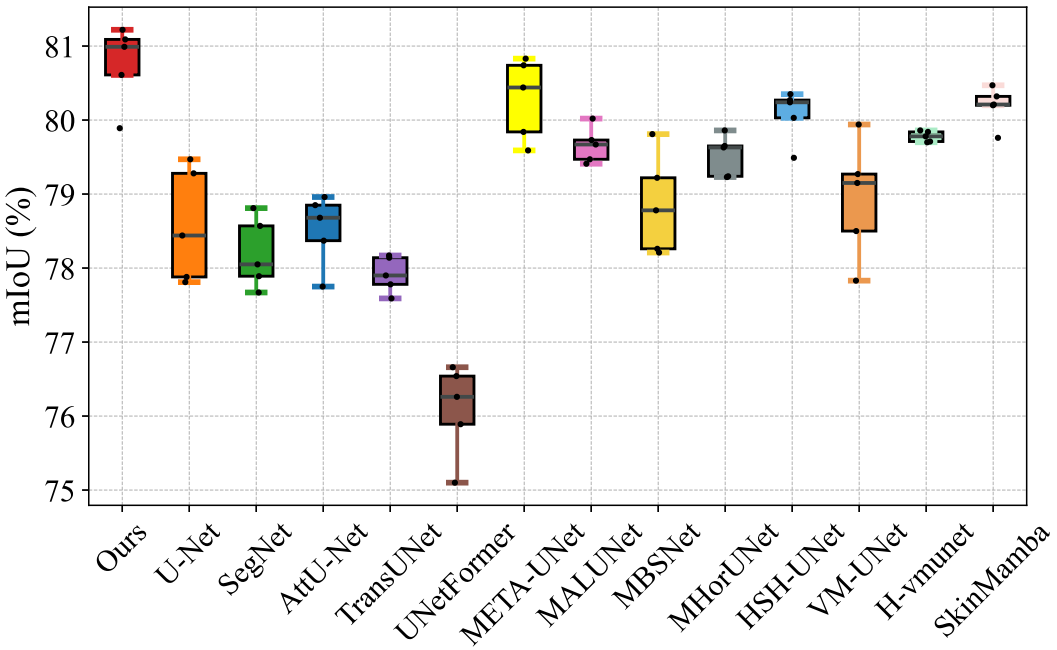

Fig. S3. The box-with-whisker plots of mIoU coefficient compared with all models on ISIC2018 dataset.

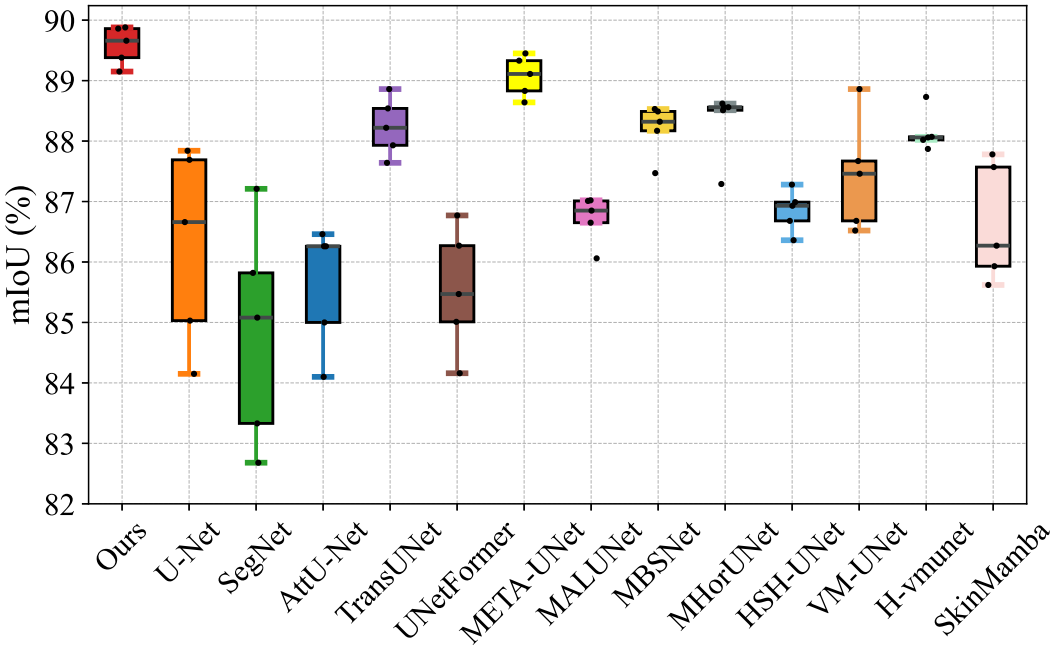

Fig. S4. The box-with-whisker plots of mIoU coefficient compared with all models on PH^2 dataset.

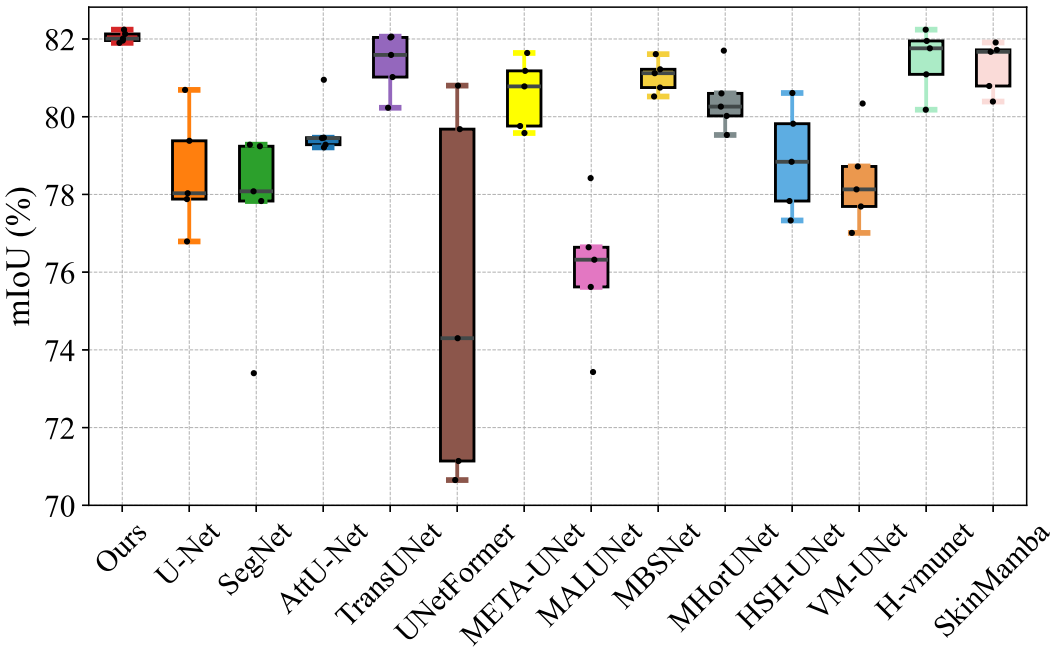

Fig. S5. The box-with-whisker plots of mIoU coefficient compared with all models on STU dataset.

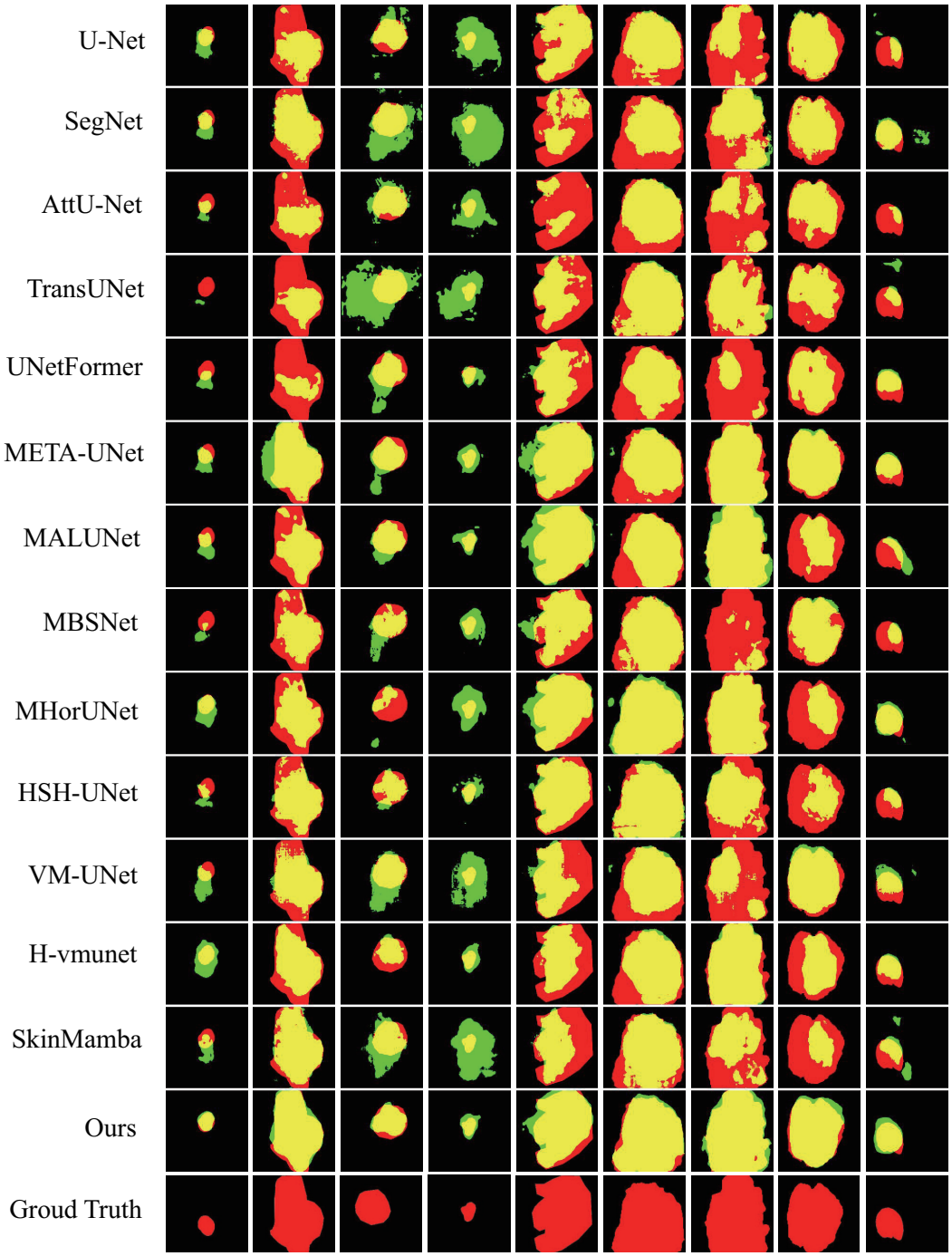

Fig. S6. Visual comparison with different state-of-the-art methods on ISIC2017 dataset.

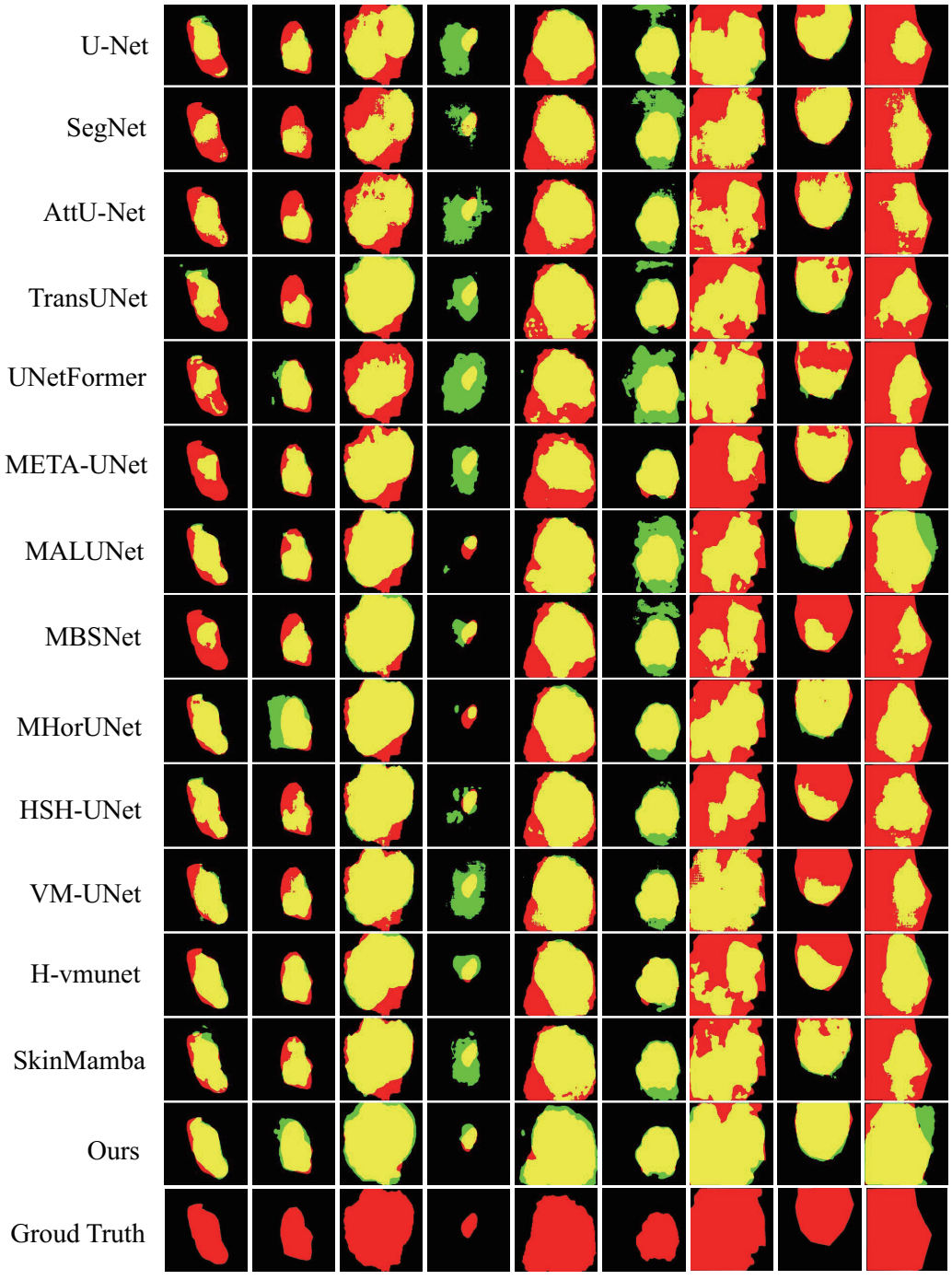

Fig. S7. Visual comparison with different state-of-the-art methods on ISIC2018 dataset.

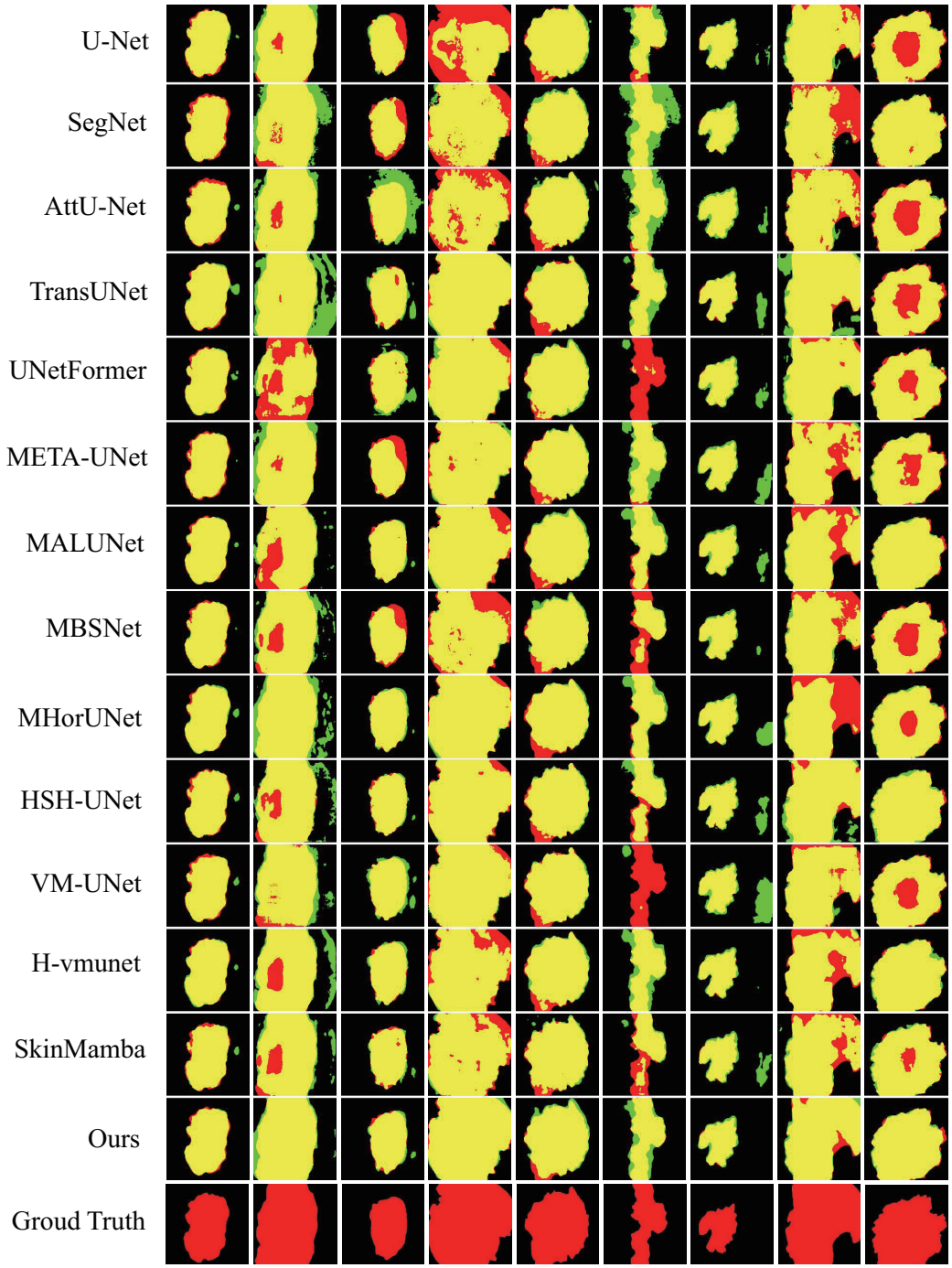

Fig. S8. Visual comparison with different state-of-the-art methods on PH^2 dataset.

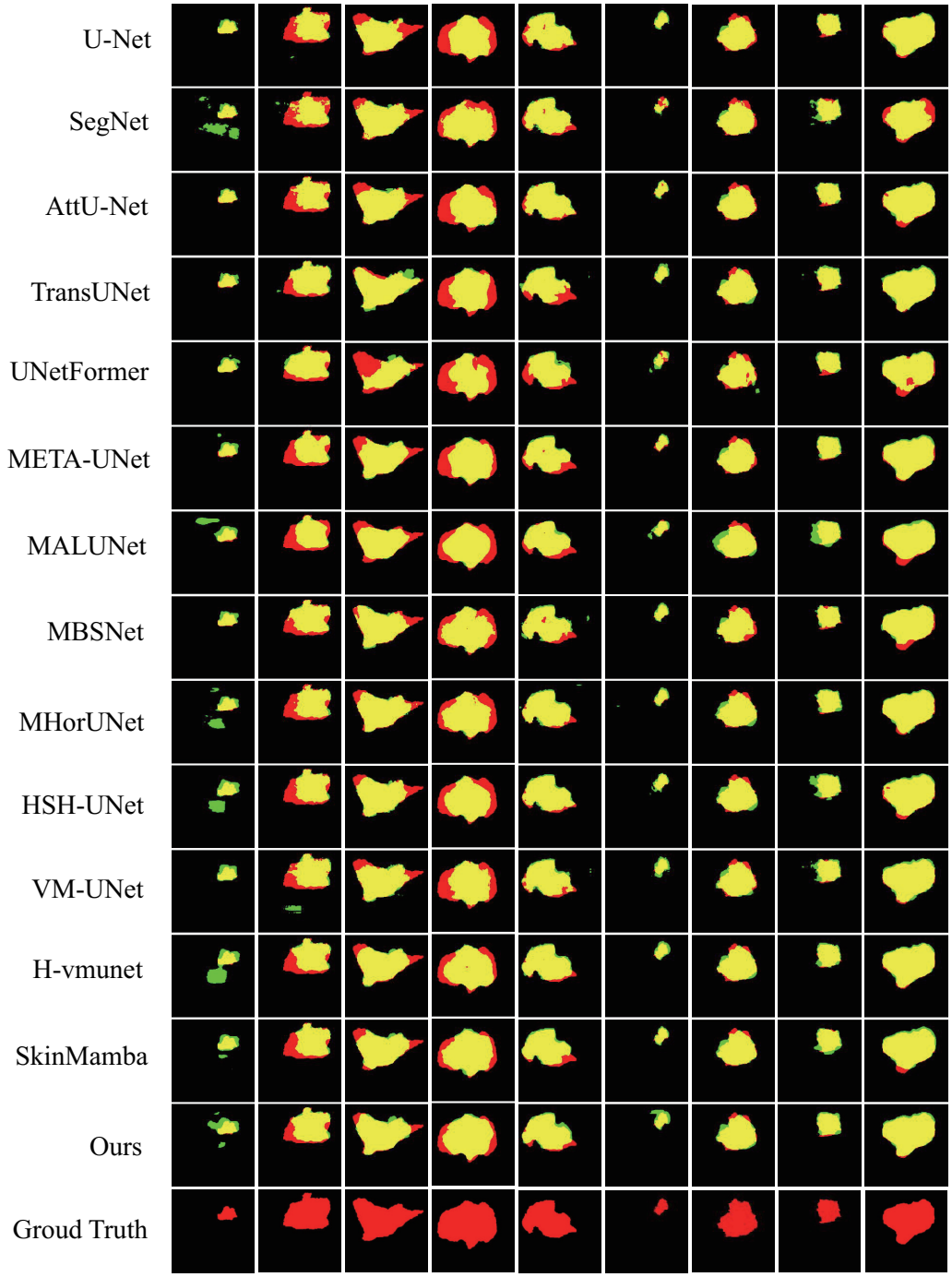

Fig. S9. Visual comparison with different state-of-the-art methods on STU dataset.

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.