Swin-UMamba†: Adapting Mamba-based vision foundation models for medical image segmentation

Jiarun Liu, Hao Yang, Hong-Yu Zhou, Lequan Yu, Yong Liang, Yizhou Yu, Shaoting Zhang, Hairong Zheng, Shanshan Wang

***Abstract*— Vision foundation models have shown great potential in improving generalizability and data efficiency, especially for medical image segmentation since medical image datasets are relatively small due to high annotation costs and privacy concerns. However, current research on foundation models predominantly relies on transformers. The high quadratic complexity and large parameter counts make these models computationally expensive, limiting their potential for clinical applications. In this work, we introduce Swin-UMamba†, a novel Mamba-based model for medical image segmentation that seamlessly leverages the power of the vision foundation model, which is also computationally efficient with the linear complexity of Mamba. Moreover, we investigated and verified the impact of the vision foundation model on medical image segmentation, in which a self-supervised model adaptation scheme was designed to bridge the gap between natural and medical data. Notably, Swin-UMamba† outperforms 7 state-of-the-art methods, including CNN-based, transformer-based, and Mamba-based approaches across AbdomenMRI, Encoscopy, and Microscopy datasets. The code and models are publicly available at: https://github.com/JiarunLiu/Swin-UMamba.**

This research was partly supported by the National Natural Science Foundation of China (62222118, U22A2040), Shenzhen Science and Technology Program (RCYX20210706092104034, JCYJ20220531100213029), Guangdong Provincial Key Laboratory of Artificial Intelligence in Medical Image Analysis and Application (2022B1212010011), the major key project of Pengcheng Laboratory under grant PCL2023AS1-2, and Key Laboratory for Magnetic Resonance and Multimodality Imaging of Guangdong Province (2020B1212060051). *(Jiarun Liu and Hao Yang contributed equally to this work.) (Corresponding authors: Shanshan Wang; Hong-Yu Zhou.)*

Jiarun Liu and Hao Yang are with the Paul C. Lauterbur Research Center for Biomedical Imaging, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China, and with the University of Chinese Academy of Sciences, Beijing 100049, China, and with the Pengcheng Laboratory, Shenzhen 518055, China (e-mail: jr.liu@siat.ac.cn, h.yang1@siat.ac.cn, wj.huang@siat.ac.cn).
Hong-Yu Zhou and Yizhou Yu are with Department of Computer Science, The University of Hong Kong, Hong Kong (e-mail: whuzhouhongyu@gmail.com, yizhouy@acm.org).
Hairong Zheng, and Shanshan Wang are with the Paul C. Lauterbur Research Center for Biomedical Imaging, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China (e-mail: yanxi@first-imaging.com, cheng.li6@siat.ac.cn, hr.zheng@siat.ac.cn, ss.wang@siat.ac.cn).
Lequan Yu is with Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong (e-mail: lqyu@hku.hk).
Yong Liang is with the Pengcheng Laboratory, Shenzhen 518055, China (e-mail: liangy02@pcl.ac.cn, gmshi@xidian.edu.cn).
Shaoting Zhang is with the Shanghai Artificial Intelligence Laboratory, Shanghai 200030, China (e-mail: zhangshaoting@pjlab.org.cn).

***Index Terms*— Medical image segmentation, Mamba-based model, Long-range dependency modeling, Segmentation network, Foundation model adaption.**

I. Introduction

Medical image segmentation plays an important role in modern clinical practice to facilitate accurate diagnoses, disease progress monitoring, and treatment planning [1], [2]. Traditionally, this task heavily relies on the expertise of clinicians, leading to labor-intensive and time-consuming segmentation procedures. Moreover, the inherent subjectivity and inter-observer variability among experts can introduce inconsistencies in segmentation [3]. These highlight the need for automated segmentation methods to enhance efficiency, accuracy, and consistency in medical image analysis, facilitating precise and rapid diagnoses [4].

Deep learning models have witnessed remarkable advancements in automatic medical image segmentation [5]–[9]. However, these methods heavily rely on large-scale, task-specific, and crowd-labeled data to train a deep neural network (DNN) for a specific task [10], which is challenging to generalize to other tasks. Besides, collecting large-scale medical datasets can be challenging due to high annotation costs and privacy concerns [11]. Training deep neural networks on small datasets may lead to limited model performance and poor generalizability for clinical practice [12]. The lack of large and high-quality labeled datasets limits the application of supervised deep learning in medical image analysis.

Recent research has demonstrated vision foundation models [13]–[20] can achieve remarkable performance without relying on large-scale high-quality labeled datasets for various medical tasks. These models are usually pretrained on large-scale datasets to learn robust and generalizable feature representations. Transformer-based architecture is a popular choice for foundation models, as they attain excellent results when pretrained at a sufficient scale, and can be transferred to various tasks with fewer datapoints [21]. However, despite their advantages, foundation models often have high computational costs due to the quadratic complexity of the attention mechanism [22] and the large number of parameters. This limitation hinders their practical use in clinical settings, especially where computational resources are constrained. In contrast, CNNs are more efficient in processing image data but struggle with

modeling long-range dependencies due to their local receptive field imposed by the convolution operation [23]. Therefore, there is an urgent need to develop foundation models that can not only enhance generalizability and data efficiency but also hugely improve the computational efficiency for specialized tasks.

In contrast to transformer-based models, structured state space sequence models (SSMs) [24], [25] have demonstrated efficiency and effectiveness in long sequence modeling, showing potential for next-generation foundation models. Unlike the quadratic complexity of transformers, SSMs scale linearly or near-linearly with sequence length while maintaining the capability of modeling long-range dependencies. The latest work Mamba [22] demonstrated the cutting-edge performance of SSMs in long-sequence data analysis tasks, such as natural language processing and genomic analysis. The efficiency of SSMs also helps in high-resolution image processing, e.g., whole-slide pathology images [26] and high-resolution MRI/CT scans [27]. Several latest studies have preliminarily explored the effectiveness of Mamba in the vision domain [28]–[32]. However, these Mamba-based models are often trained from scratch without leveraging transfer learning from existing foundation models [29], [31], which could enhance data efficiency and generalizability in medical image segmentation tasks [33]. Further research is in need to evaluate and optimize the integration of Mamba-based foundation models into the medical domain.

To effectively adapt Mamba-based models in medical image segmentation tasks, the first challenge lies in the fact that the structure of existing Mamba-based models for medical image segmentation often differs from popular vision foundation models [29], [31], posing challenges in effectively integrating foundation models to improve segmentation performance. Another challenge is that current Mamba-based vision foundation models are primarily trained on natural image datasets, such as ImageNet, where the natural images are distinct from medical images in visual appearance, data modalities, and segment targets. Directly adopting general vision foundation models in medical image segmentation tasks may result in suboptimal performance due to this gap [23]. Given the fact that the application of Mamba blocks in the vision domain is relatively new, further experimental evaluation is required for Mamba-based medical image segmentation. Additionally, there is a need to further enhance the scalability and efficiency of Mamba-based models for real-world deployment [34].

In this paper, we proposed a Mamba-based network Swin-UMamba† leverages the power of vision foundation models with a generic Mamba-based encoder and a Mamba-based decoder. An additional network Swin-UMamba was introduced with a CNN-based decoder to evaluate the impact of the vision foundation models in medical image segmentation tasks under different settings. Moreover, we proposed a self-supervised model adaption scheme to bridge the gap between large-scale pretraining datasets and medical image segmentation datasets. Our main contributions can be summarized as follows:

- We are the first attempt to discover the impact of the Mamba-based vision foundation model for medical image segmentation. The results verified the effectiveness of the Mamba-based foundation model.
- We propose a Mamba-based medical image segmentation network to leverage the power of foundation models.
- To bridge the gap between natural and medical data, we introduce an additional self-supervised model adaption scheme.
- Extensive experiments on AbdomenMRI, Endoscopy, and Microscopy datasets demonstrate that Swin-UMamba† outperforms 7 state-of-the-art segmentation models, including CNNs, ViTs, and the latest Mamba-based models. Our results show notable improvements in segmentation accuracy, along with significantly fewer model parameters and FLOPs.

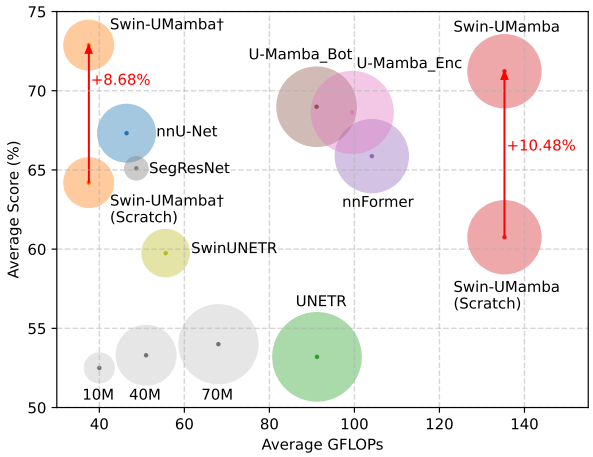


Fig. 1: The segmentation score, GFLOPs versus model size based on the average results across AbdomenMRI, Endoscopy, and Microscopy datasets. The size of each circle indicates the size of the model, and the position of the center point indicates the segmentation score and GFLOPs. *Scratch* denotes the model was trained from scratch, while the red arrow indicates the improvement by fine-tuning with a foundation model. The grey circles in the bottom left indicate the model size with 10 million, 40 million, and 70 million parameters, respectively.

II. Related Works

A. Automatic medical image segmentation

1) CNN-based methods: CNN has become one of the fundamental networks in vision applications [35] that excel at capturing translational invariances features. However, it is challenging for CNNs to capture long-range dependencies due to their intrinsic locality [23]. The limited receptive field can lead to sub-optimal performance when dealing with structures across various shapes and scales [35]. Atrous convolution [36] can enlarge the receptive field, but it causes the loss of detailed information [35]. A widely adopted model in medical image segmentation is U-Net [9], which leverages a U-shape architecture to capture both global context and local details. However, U-Net does not solve the intrinsic locality issue of convolution. nnU-Net [8] further enhances U-Net by enabling automatic configuration of preprocessing, network architecture, training, and post-processing for any new segmentation task. SegResNet

[37] adds an additional variational auto-encoder branch to impose additional constraints. Nevertheless, improving CNN's ability to model long-range dependencies remains an open question.

2) Transformer-based methods: The transformer architecture was originally introduced for natural language processing with the attention mechanism. ViT [21] firstly adopts transformer into vision tasks by transforming the image into a sequence of patches. Compared with CNNs, ViT can capture global context through attention. This is important for accurate medical image segmentation since the organs can spread over a large receptive field. Besides, it helps the model in preventing misclassification when medical images pose high fine-grained variability [23]. Swin-Unet [38] firstly proposed a pure transformer-based architecture based on the Swin-Transformer without convolution layer for medical image segmentation. In contrast, UNETR [39] and TransUNet [40] incorporate a CNN-based decoder for precise localization, where CNNs are sophisticated at local information modeling. Swin-UNETR [41] further improves UNETR with shifted windows [42]. Another hybrid approach is nnFormer [6], which exploits the combination of interleaved convolution and self-attention operations and significantly outperforms previous transformer-based counterparts. Despite their advantages, the transformer poses a high quadratic complexity of attention with respect to the sequence length, leading to a heavy computation burden [22], particularly for high-resolution images [29], [31]. Besides, due to the lack of effective inductive bias, pretraining on large-scale datasets is widely adopted for transformer-based models to avoid overfitting and improve the segmentation performance [23]. Some methods enhance the segmentation performance by incorporating multimodal data [43]–[45], which can provide complementary information beyond images. However, this study focused on optimizing the Mamba-based model using image data to establish a strong baseline before exploring the integration of multimodal data.

B. State space models in vision

State space models (SSMs) have recently emerged as a promising architecture class for sequence modeling in deep learning. This class of models can be computed very efficiently as either a recurrence or convolution, with linear or near-linear scaling in sequence length. Structured state space sequence models (S4) [24] improve SSMs by imposing structured forms on the state matrix, which was crafted and initialized with high-order polynomial projection operator [46]. In addition to S4, Mamba [22] incorporated an input-dependent selection mechanism and an efficient hardware-aware algorithm. It helps the model filter out irrelevant information and improves the ability for efficient long sequence modeling. Vim [31] is one of the preliminary works that adopt Mamba in vision tasks by introducing a generic vision backbone with bidirectional Mamba blocks. Vim demonstrates the advantage of using Mamba in vision tasks by providing higher accuracy, lower computation burden, and less memory consumption. However, the distinction between 2D visual data and 1D language sequences requires careful consideration when adopting Mamba into vision tasks. While 2D spatial information is crucial in vision tasks [29], it is not the primary focus in 1D sequence modeling. Directly adopting Mamba to flattened images would inevitably result in restricted receptive fields, where the relationships against unscanned patches cannot be estimated. VMamba [29] introduced a cross-scan module to solve the direction-sensitive problem due to the difference between 1D sequences and 2D images. Despite their successes in various vision tasks, models like Vim and VMamba are primarily trained on the ImageNet dataset, leaving their potential for medical image segmentation remains unexplored. U-Mamba [28] adopts Mamba in medical image segmentation tasks by incorporating the Mamba block into the nnU-Net framework [8]. Although Mamba-based models have achieved promising results in vision tasks, training a Mamba-based model from scratch on medical image segmentation datasets may yield suboptimal performance. Mamba-based models face the same challenges as transformers in lacking effective inductive bias in image data modeling [47].

III. Method

This paper proposes a Mamba-based network Swin-UMamba† that leverages the power of the vision foundation model for medical image segmentation. It follows the classic U-Net structure with: 1) a Mamba-based encoder to take the power of the Mamba-based vision foundation model, 2) a Mamba-based decoder to predict segmentation masks, and 3) skip connections to bridge the gap between low-level details and high-level semantics. Additionally, we propose Swin-UMamba with a CNN-based decoder to investigate the impact of foundation models with different network structures. Moreover, we introduced a self-supervised model adaption scheme to bridge the gap between natural and medical data. The following sections will introduce the detailed structure of Swin-UMamba† and Swin-UMamba.

A. Preliminary

Recent studies demonstrate the efficiency of SSMs in sequence modeling tasks. It maps 1D sequence $x(t) \in \mathbb{R}^L$ to $y(t) \in \mathbb{R}^L$ with a compressed hidden state $h(t) \in \mathbb{R}^N$, enabling each element in the sequence (e.g., text sequence) to interact with any of the previously scanned samples. SSM can be formulated as linear ordinary differential equations (ODEs):

$$\begin{aligned} h'(t) &= Ah(t) + Bx(t) \\ y(t) &= Ch(t) + \lambda x(t) \end{aligned} \tag{1}$$

where $A \in \mathbb{R}^{N\times N}$, $B \in \mathbb{R}^{N\times 1}$, and $C \in \mathbb{R}^{1\times N}$ are parameters of SSM for a state size N and $\lambda \in \mathbb{R}^1$ is the weight of skip connection. To integrate SSMs into deep learning, S4 [24] discrete the ODEs into a discrete function with the input $x_k \in \mathbb{R}^{L\times D}$:

$$\begin{aligned} h_k &= \bar{A}h_{k-1} + \bar{B}x_k \\ y_k &= \bar{C}h_k + \lambda x(t) \end{aligned} \tag{2}$$

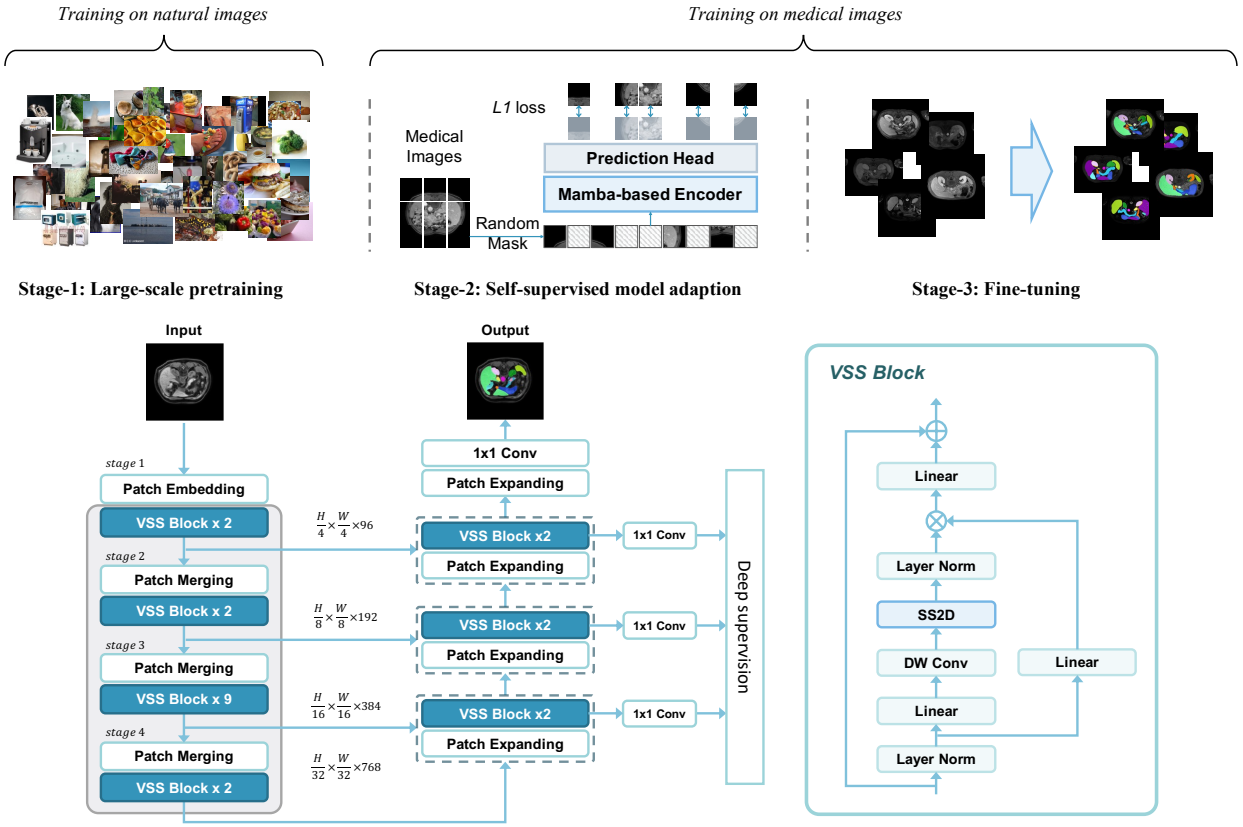


Fig. 2: The overall training scheme and the model architecture of Swin-UMamba†. Top: To fill the gap between natural and medical data, we proposed a self-supervised model adaption scheme. Bottom: The overall architecture of Swin-UMamba†. Swin-UMamba† can leverage the power of vision foundation models, whereas each block within the grey box was initialized with the weights from a foundation model. The structure of the VSS block is illustrated on the right of the model.

The discrete parameters $\bar{A}, \bar{B}, \bar{C}$ can be discretized with the zero-order hold rule:

$$\begin{aligned} \bar{A} &= exp(\Delta A), \\ \bar{B} &= (exp(\Delta A) - I)\, A^{-1} B, \\ \bar{C} &= C \end{aligned} \tag{3}$$

In practice, $\bar{B}$ can be approximated with first-order Taylor series:

$$\bar{B} = (exp(\Delta A) - I)\, A^{-1} B \approx (\Delta A)(\Delta A)^{-1} \Delta B = \Delta B \tag{4}$$

Mamba further extends the S4 operator with a selective scan mechanism, i.e. S6. The SSM matrices B, C, and Δ of S6 are derived from the input data x with three linear functions. Selective scan helps control information that propagates or interacts along the sequence dimension. We refer to [22], [29] for further details about S6.

B. Mamba-based VSS block

Representing visual data is challenging for Mamba-based models since Mamba is only able to estimate the relationship against scanned patches [29], [31], resulting in a restricted receptive fields issue. Drawing from insights presented in [29], we introduce the visual state space (VSS) block as the basic unit in Swin-UMamba† to solve this issue with 2D-selective-scan (SS2D). As illustrated in Fig. 3, SS2D unfolds image patches into feature sequences in four different scan directions. These sequences consist of the same image features but are arranged in different orders. Each sequence will be processed through the same S6 operator and then merged back to form the complete 2D feature map. Given input feature z, the output feature $\bar{z}$ of SS2D can be written as:

$$z_v = expand(z, v) \tag{5}$$

$$\bar{z}_v = S6(z_v) \tag{6}$$

$$\bar{z} = merge(\bar{z}_1, \bar{z}_2, \bar{z}_3, \bar{z}_4) \tag{7}$$

where $v \in V = \{1, 2, 3, 4\}$ represents four different scanning directions. $expand(\cdot)$ and $merge(\cdot)$ correspond to the *scan expand* and *scan merge* operations in [29]. The other components of the VSS block follow the design in [22]. Fig. 2 illustrates the overall structure of VSS block.

C. Integrating Mamba-based vision foundation model

The primary challenge lies in effectively integrating vision foundation models into medical image segmentation tasks. Prior research [28] typically employs a task-specific architecture with Mamba blocks, which fails to consider the transferability from generic vision models. To address this limitation, we develop an encoder that shares a similar structure with the latest Mamba-based foundation model VMamba-Tiny [29].

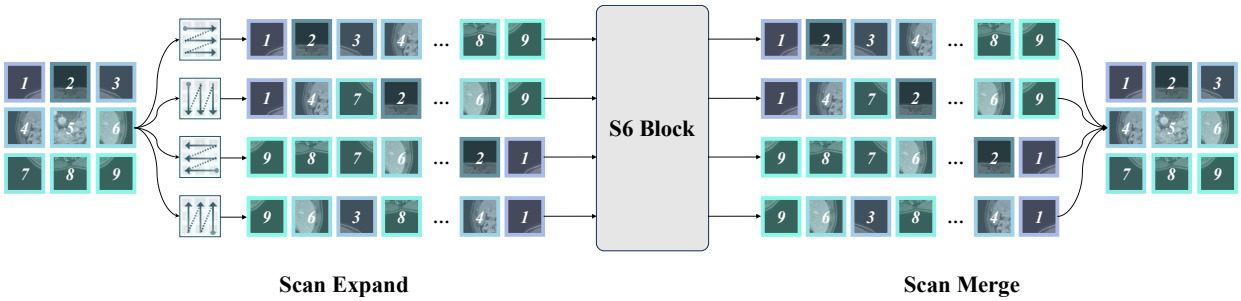


Fig. 3: Illustration of 2D-selective-scan (SS2D). SS2D first expands 2D image features in four different directions and then merges the processed sequences back after S6.

This model, pretrained on the extensive ImageNet dataset with multi-scale features, enables us to integrate the power of the vision foundation model to extract information with long-range modeling capability, mitigate the risk of overfitting, and establish a robust initialization.

As illustrated in Fig. 2, the encoder of Swin-UMamba† consists of 4 stages. The initial stage comprises a 4×4 patch embedding layer and 2 VSS blocks. It maps the input image from $H \times W \times C$ into feature maps of shape $\frac{H}{4} \times \frac{H}{4} \times 96$. Subsequent stages are composed of a patch merging layer [42] for $2\times$ down-sampling and multiple VSS blocks for high-level feature extraction. The numbers of VSS blocks at each stage are $\{2, 2, 9, 2\}$, respectively. Unlike ViT, Swin-UMamba† does not adopt position embedding due to the causal nature of the VSS block [29]. The feature dimensions increase quadratically with the stages, resulting in $D = \{96, 192, 384, 768\}$. The weights of the VSS blocks and patch merging layers are initialized from VMamab-Tiny [29]. Since the number of input channels may differ across datasets, we did not use the weights of the patch embedding layer in this scenario.

D. Mamba-based decoder

Similar to the widely adopted U-shaped structure, Swin-UMamba† uses a decoder based on the VSS block. The overall architecture leverages skip connections to recover low-level details and employs a multi-scale encoder-decoder structure for high-level information extraction. Inspired by [38], we use a patch expanding layer to perform the up-sampling operation at each decoding stage.

Take the first decoding stage as an example. The input features of size $\left\{\frac{H}{32} \times \frac{W}{32} \times 768\right\}$ are up-sampled through the patch expanding layer to the size of $\left\{\frac{H}{16} \times \frac{W}{16} \times 384\right\}$. Subsequently, these up-sampled features are combined with the skip-connected features, followed by a linear projection to ensure consistent feature dimensions before concatenation. Next, the concatenated features are processed through 2 VSS blocks. To enable deep supervision, a 1×1 convolution layer is added after each stage to generate segmentation outputs. Specifically, the patch expanding layer of the final stage performs a $4\times$ up-sampling operation, mirroring the 4×4 patch embedding layer. After the last patch embedding layer, a 1×1 convolution is applied to adjust the feature dimension to match the number of classes K for final segmentation. Besides, a pure Mamba-based structure helps to reduce the number of network parameters and enhance computation efficiency, since Mamba exhibits a linear complexity with sequence length [22]. Swin-UMamba† significantly saves the number of network parameters to 28M and FLOPs to 18.9G, whereas other models may encounter twice or even more computations.

E. Swin-UMamba with CNN-based decoder

To further investigate the impact of the vision foundation model for Mamba-based segmentation with various structures, we proposed Swin-UMamba with a CNN-based decoder. Compared with Swin-UMamba†, Swin-UMamba mainly varies in the type of decoder blocks and the number of encoding/decoding stages.

We add an additional stem stage with a 7×7 convolution layer for $2\times$ down-sampling before the patch embedding layer. The patch size of the patch embedding layer is reduced to 2×2 to keep the shape of features to the VSS blocks. This modification helps to perform a gradual down-sampling process where each stage takes $2\times$ down-sampling. Skip connection is adopted at each scale, including the original image. This scheme aims to retain low-level details, which is important for medical image segmentation [9], [48].

To enhance the native up-sample block in U-Net, we introduce two modifications: 1) an extra convolution block with a residual connection to process skip connection features, and 2) an additional segmentation head at each scale for deep supervision [49]. Given skip-connected features z'_l from stage-l and features z_{l+1} from the last up-sample block, the output features z_l of l-th up-sample block and the segmentation map $y_l \in R^{h_l \times w_l \times K}$ at stage-l can be formulated as follows:

$$\hat{z}_l = Res_l^{(2)}(Cat(z_{l+1}, Res_l^{(1)}(z'_l))) \tag{8}$$

$$z_l = DeConv_l(\hat{z}_l) \tag{9}$$

$$y_l = Conv_l(\hat{z}_l) \tag{10}$$

where $Cat(\cdot)$, $DeConv_l(\cdot)$, $Conv_l(\cdot)$ are the feature concatenation operation, transpose convolution, and a segmentation head with 1×1 convolution that project feature from dimension d_l to the number of class K, respectively. h_l and w_l are the height and width of the feature map at stage-l. $Res_l^{(1)}(\cdot)$ and $Res_l^{(2)}(\cdot)$ are two convolution blocks with residual connection at stage-l. Each $Res(\cdot)$ is composed of two convolution layers with LeakyRELU activation.

F. Self-supervised model adaption

Foundation models can provide a good initialization for various downstream tasks. However, current Mamba-based vision foundation models are mostly pretrained on large-scale natural image datasets, such as ImageNet. The data distributions of ImageNet and medical datasets are different since they have different data modalities and target objects. As illustrated in Fig. 2, to overcome the distribution shift and maximize the potential of foundation models, we introduce an additional self-supervised model adaption scheme with masked image modeling (MIM) [50], [51] on the corresponding medical dataset. Specifically, for each image x, we masked out 60% image patches before inputting it into the model. The model consists of two components: a Mamba-based encoder to process image features and a lightweight one-layer head to map the feature vector to the pixel space. The target of the model is to predict the raw pixel values of the randomly masked patches:

$$L_{MIM} = \frac{1}{\Omega(x_M)} \|\bar{y}_M - x_M\|_1 \tag{11}$$

where $\bar{y}$ is the model predictions. M is the set of masked pixels, $\Omega(x_M)$ is the number of masked pixels. Reconstructing the raw pixel values from masked images encourages the model to learn useful features from the medical images without human annotations.

With the model adaption scheme, the training can be divided into three stages: 1) pretrain the encoder on the ImageNet dataset; 2) adapt the encoder on the target medical segmentation dataset with masked image modeling; 3) fine-tuning with Swin-UMamba† on the segmentation dataset. Our experiment results show that this self-supervised model adaption scheme helps to minimize the gap between the natural and medical datasets, thus improving the segmentation accuracy and training speed.

IV. Experiments

A. Datasets

We assess the performance and scalability of Swin-UMamba† on three different medical image segmentation datasets, covering organ, instrument, and cell segmentation tasks. These datasets vary in resolution and imaging modalities, offering insights into the model's efficacy and adaptability in diverse medical imaging scenarios.

1) Abdomen MRI (AbdomenMRI): This dataset focused on segmenting 13 abdominal organs from MRI scans, including the liver, spleen, pancreas, right kidney, left kidney, stomach, gallbladder, esophagus, aorta, inferior vena cava, right adrenal gland, left adrenal gland, and duodenum. It was originally provided by the MICCAI 2022 AMOS Challenge [52]. We followed the settings in [28] employing additional 50 MRI scans for testing. There are 60 MRI scans with 5615 slices for training and 50 MRI scans with 3357 slices for testing. We cropped the images into patches of size $(320, 320)$ for training and testing with the nnU-Net framework [8].

2) Endoscopy images (Endoscopy): This dataset aims to segment 7 instruments from endoscopy images, including the large needle driver, prograsp forceps, monopolar curved scissors, cadiere forceps, bipolar forceps, vessel sealer, and drop-in ultrasound probe. It was originally from the MICCAI 2017 EndoVis Challenge [53]. It consists of 1800 image frames for training and 1200 image frames for testing. Images were cropped into $(384, 640)$ following the data processing procedure within nnU-Net for both training and testing. It's worth noting that images in this dataset exhibit a unique aspect ratio compared to other datasets.

3) Microscopy images (Microscopy): This dataset focuses on cell segmentation in various microscopy images from the NeurIPS 2022 Cell Segmentation Challenge [54]. It consists of 1000 images for training and 101 images for evaluation. The images in Microscopy were cropped into $(512, 512)$ for training and testing. By default, it is an instance segmentation dataset. We employed the same data processing strategy as described in [28] for this dataset.

B. Implemetation details

We implemented Swin-UMamba† and Swin-UMamba on top of the well-established nnU-Net framework [8]. Its self-configuring feature enabled us to focus on network design rather than other trivial details. The loss function for the segmentation network was the sum of Dice loss and cross-entropy (CE) loss:

$$L_{seg} = L_{Dice} + L_{CE} \tag{12}$$

In practice, we perform deep supervision [49] during the training.

$$L_{all} = \sum \alpha_i L_{seg}^{\Psi_i} \tag{13}$$

where Ψ_i denote the image resolution scale factor i and $\alpha_i = 1/2^i$ is corresponding weight factor which will normalized to 1. For Swin-UMamba, $\Psi = \left\{1\times, \frac{1}{2}\times, \frac{1}{4}\times, \frac{1}{8}\times\right\}$. Swin-UMamba† use a different Ψ because it did not have feature map at $\frac{1}{2}\times$ scale, where $\Psi = \left\{1\times, \frac{1}{4}\times, \frac{1}{8}\times\right\}$. We used an AdamW optimizer with weight decay $= 0.05$ following [29]. A cosine learning rate decay was adopted with an initial learning rate $= 0.0001$. During training, we froze parameters from the foundation model for the first 10 epochs to align other randomly initialized modules. Hyperparameters were kept consistent across all three datasets, except for the number of training epochs and data-specific settings (e.g., image size). Our models were trained for 200 epochs on the AbdomenMRI dataset, 400 epochs on the Endoscopy, and 500 epochs on the Microscopy dataset. Following [28], we disabled the testing time augmentation for a more streamlined and efficient evaluation.

During the self-supervised model adaptation stage, we adopted the latest Mamba-based foundation model VMamba-Tiny [29] that was pretrained on ImageNet to initialize our model. We froze the encoder weights for the first 10 epochs. The overall training epochs in the self-supervised model adaptation stage are 50 epochs for the AbdomenMRI dataset, 200 epochs for the Microscopy dataset, and 800 epochs for the Endoscopy dataset. All images were resized to $192px \times 192px$

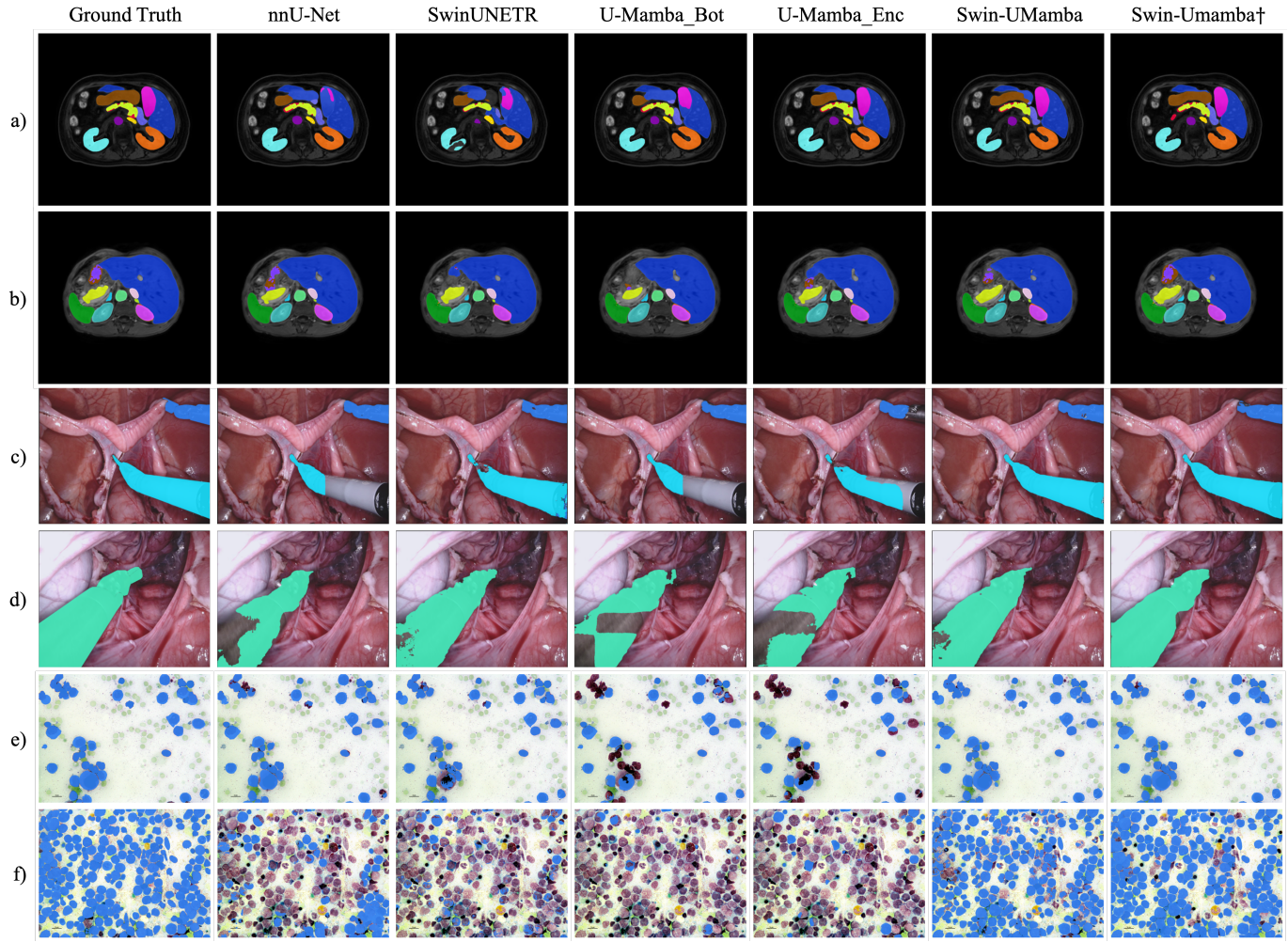


Fig. 4: Example visualization results of the three datasets. a) and b) are visualization examples of the AbdomenMRI dataset, c) and d) are visualization examples of the Endoscopy dataset, and e) and f) are visualization examples of the Microscopy dataset. Swin-UMamba† accurately recognizes the shape and type of the segmented targets.

during this stage. We use the AdamW optimizer with a learning rate of 2e-4. For more details, please refer to our open-source codes[1].

C. Comparison methods and evaluation metrics

We select three types of methods for comparison, including CNN-based (nnU-Net [8], SegResNet [37]), transformer-based (UNETR [39], Swin-UNETR [41], nnFormer [6]), and the latest Mamba-based segmentation network U-Mamba [28]. Specifically, U-Mamba has two variants: U-Mamba_Bot and U-Mamba_Enc. U-Mamba_Bot only adopts the Mamba block in the bottleneck, while U-Mamba_Enc adopts the Mamba block in each encoder stage. We compared Swin-UMamba† with both U-Mamba_Bot and U-Mamba_Enc. It's worth noting that adopting the same foundation model into U-Mamba or other comparison methods is not straightforward due to structural differences from foundation models [29]. The results of nnU-Net, SegResNet, UNETR, Swin-UNETR, and U-Mamba were referenced from [28], where these models were trained from scratch for 1000 epochs using stochastic gradient descent and an unweighted sum of Dice and cross-entropy loss. The results for nnFormer [6] are based on the official implementation. We train nnFormer for 1000 epochs with the default Adam optimizer and the same unweighted sum of Dice and cross-entropy loss.

We use the dice similarity coefficient (DSC) and normalized surface distance (NSD) to assess segmentation performance on the AbdomenMRI and Endoscopy datasets. For the Microscopy dataset, we follow [28] and [54] to use the F1 score for evaluation as it is an instance segmentation task. Furthermore, we compute the number of network parameters (#param) and floating-point operations (FLOPs) with the *fvcore* package[2] to assess the scale and computational costs of each model. If not otherwise specified, we use *w/ foundation model* to indicate the model was trained on the ImageNet pretrained VMamba-Tiny model with self-supervised model adaption. On the contrary, *w/o foundation model* indicates that the model was trained from scratch without pretraining

[1] https://github.com/JiarunLiu/Swin-UMamba

[2] https://github.com/facebookresearch/fvcore

TABLE I: Results of organ segmentation on the AbdomenMRI dataset. The results of nnU-Net, SegResNet, UNETR, SwinUNETR, and U-Mamba were referenced from [28]. *: Deep supervision was disabled.

Methods	#param	FLOPs	DSC	NSD
CNN-based				
nnU-Net	33M	23.3G	0.7450±0.1117	0.8153±0.1145
SegResNet	6M	24.5G	0.7317±0.1379	0.8034±0.1386
Transformer-based				
UNETR	87M	42.1G	0.5747±0.1672	0.6309±0.1858
SwinUNETR	25M	27.9G	0.7028±0.1348	0.7669±0.1442
nnFormer	60M	50.2G	0.7279±0.1486	0.7963±0.1322
Mamba-based				
U-Mamba_Bot	63M	45.7G	0.7588±0.1051	0.8285±0.1074
U-Mamba_Enc	67M	49.9G	0.7625±0.1082	0.8327±0.1087
w/o foundation model				
Swin-UMamba	60M	68.0G	0.7054±0.1387	0.7647±0.1455
Swin-UMamba†*	28M	18.9G	0.6653±0.1123	0.7312±0.1199
w/ foundation model				
Swin-UMamba	60M	68.0G	0.7704±0.0936	0.8354±0.0956
Swin-UMamba†	28M	18.9G	**0.7767±0.0940**	**0.8428±0.0938**

TABLE II: Results of instruments segmentation on the Endoscopy dataset. The results of nnU-Net, SegResNet, UNETR, SwinUNETR, and U-Mamba were referenced from [28].

Methods	#param	FLOPs	DSC	NSD
CNN-based				
nnU-Net	33M	55.9G	0.6264±0.3024	0.6412±0.3074
SegResNet	6M	58.9G	0.5820±0.3268	0.5968±0.3303
Transformer-based				
UNETR	88M	111.5G	0.5017±0.3201	0.5168±0.3235
SwinUNETR	25M	67.1G	0.5528±0.3089	0.5683±0.3123
nnFormer	60M	125.5G	0.6135±0.2763	0.6228±0.2832
Mamba-based				
U-Mamba_Bot	63M	109.7G	0.6540±0.3008	0.6692±0.3050
U-Mamba_Enc	67M	119.8G	0.6303±0.3067	0.6451±0.3104
w/o foundation model				
Swin-UMamba	60M	163.6G	0.5483±0.3047	0.5632±0.3085
Swin-UMamba†	28M	45.3G	0.6402±0.3260	0.6547±0.3301
w/ foundation model				
Swin-UMamba	60M	163.6G	0.6786±0.2962	0.6936±0.3003
Swin-UMamba†	28M	45.3G	**0.6931±0.2976**	**0.7094±0.3019**

or additional adaptation stages.

D. Results on AbdomenMRI dataset

The segmentation results on the AbdomenMRI dataset are presented in Table I. All Mamba-based networks outperform CNN-based and transformer-based baselines. The superior result demonstrates the great potential of the Mamba-based network in medical image segmentation. Notably, Swin-UMamba† outperforms all comparison methods, including CNN-based networks, transformer-based networks, and Mamba-based networks. Swin-UMamba† exhibits 1.42% improvement in DSC over U-Mamba_Enc, which is the previous SOTA model on this dataset. It is particularly noteworthy that Swin-UMamba† has less than half of the network parameters and FLOPs compared to U-Mamba_Enc. As illustrated in Fig. 4a, Swin-UMamba† can recognize the shape and type of target organs, whereas other methods may miss or misclassify some target regions.

The results in Table I also demonstrate the effectiveness of foundation models in medical image segmentation across different model architectures. For instance, it leads to significant increases for both metrics achieved by Swin-UMamba with 6.50% in DSC and 7.07% in NSD. Swin-UMamba can effectively leverage the knowledge from existing foundation models for segmentation tasks. Moreover, leveraging the vision foundation model facilitates faster and more stable training. Compared with other methods, Swin-UMamba† only requires 100 epochs for training vs 1000 epochs for comparison methods. Training from scratch can be very unstable, as we found that Swin-UMamba† fails to converge properly on this dataset with default settings. To address this issue, we disabled the deep supervision of Swin-UMamba† when trained from scratch on the AbdomenMRI dataset. Despite that, Swin-UMamba† outperforms all baseline methods with a foundation model. The improvement is particularly noteworthy considering that Swin-UMamba† has less than half of the network parameters and FLOPs compared to the previous SOTA model U-Mamba. Additionally, the self-supervised model adaption did help the model to improve performance by reducing the gap between ImageNet and medical images. We discuss this in Sec.IV-G

We also observed a disparity in parameter numbers and FLOPs between Swin-UMamba† and Swin-UMamba. This discrepancy is primarily attributed to the decoder, as the Mamba-based decoder of Swin-UMamba† have fewer parameters than the CNN-based decoder of Swin-UMamba.

E. Results on Endoscopy dataset

Table II presents the segmentation performance of each model on the Endoscopy dataset. Remarkably, Swin-UMamba† outperforms U-Mamba_Bot by over 3.91% in DSC and 4.02% in NSD. We noted that the Endoscopy dataset is smaller than the AbdomenMRI dataset, and models are prone to overfitting to the training data. Leveraging the power of a foundation model is an effective strategy for mitigating overfitting in such small datasets. We observed an impressive performance gain of 13.03% in DSC and 13.04% in NSD with the pretrained model for Swin-UMamba, which has much more parameters than Swin-UMamba†. Additionally, Swin-UMamba† outperforms U-Mamba_Enc by 0.99% in DSC when trained from scratch. Compared with U-Mamba_Enc, Swin-UMamba† has much fewer network parameters, which makes it easier to train on a small dataset. The visualized result of Swin-UMamba† on Endoscopy is shown in Fig. 4b.

F. Results on Microscopy dataset

Table III presents the segmentation performance on the Microscopy dataset. Swin-UMamba† and Swin-UMamba continue to outperform all comparison methods by margins ranging from 2.29% to 22.52% in F1 score. In contrast to previously mentioned datasets, the Microscopy dataset features a higher image resolution, fewer samples, and greater visual differences. This imposes greater demands on the

TABLE III: Results of cell segmentation on the Microscopy dataset. The results of nnU-Net, SegResNet, UNETR, Swin-UNETR, and U-Mamba were referenced from [28].

Methods	#param	FLOPs	F1
CNN-based			
nnU-Net	46M	60.1G	0.5383±0.2657
SegResNet	6M	62.8G	0.5411±0.2633
Transformer-based			
UNETR	88M	120.1G	0.4357±0.2572
SwinUNETR	25M	71.7G	0.3967±0.2621
nnFormer	60M	136.7G	0.5332±0.2543
Mamba-based			
U-Mamba_Bot	86M	117.8G	0.5389±0.2817
U-Mamba_Enc	92M	128.7G	0.5607±0.2784
w/o foundation model			
Swin-UMamba	60M	174.4G	0.4561±0.2806
Swin-UMamba†	27M	48.2G	0.5186±0.2727
w/ foundation model			
Swin-UMamba	60M	174.4G	0.5836±0.2396
Swin-UMamba†	27M	48.2G	**0.6219±0.2452**

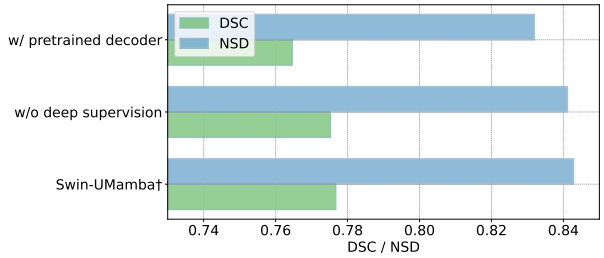


Fig. 5: Ablations of different training procedures on the AbdomenMRI dataset.

TABLE IV: Ablations of different network settings on the AbdomenMRI dataset.

Methods	#param	FLOPs	DSC	NSD
Swin-UMamba†	28M	18.9G	**0.7767±0.0940**	**0.8428±0.0938**
w/ symmetric decoder	36M	23.4G	0.7635±0.0964	0.8326±0.0979
w/ symmetric encoder	19M	11.3G	0.7570±0.1021	0.8264±0.1057

model's capacity for long-range information modeling and data-efficiency. As shown in Fig. 4c, Swin-UMamba† can accurately recognize target cells while baselines may missing some targets. Moreover, we observe that foundation models are more effective for Swin-UMamba as it has more parameters, which makes it harder to train on small datasets. Swin-UMamba† and Swin-UMamba benefit from the foundation model by 10.33% and 12.75% in F1 score respectively. This demonstrates that foundation models can be more useful for larger models with smaller datasets.

G. The impact of self-supervised model adaption

The experimental results show that the vision foundation model significantly enhances medical image segmentation, particularly when using small datasets and larger models. We further investigate the effect of different pretraining strategies by training the Swin-UMamba† with three initialization methods: 1) pretrained on medical images, 2) pretrained on ImageNet, and 3) pretrained on ImageNet followed by self-supervised adaptation to medical images. As shown in TableV, the additional self-supervised adaptation consistently improves segmentation performance across all three datasets compared to either ImageNet or medical pretraining alone.

The degree of improvement from the medical adaptation varies between datasets. For instance, it increases the F1 score by 2.31% on the Microscopy dataset over ImageNet pretraining. On the AbdomenMRI dataset, it improves the DSC by 0.62%. This additional adaption scheme helps to bridge the gap between ImageNet and medical image datasets. Interestingly, direct pretraining on medical data does not outperform ImageNet pretraining. The ImageNet-pretrained model achieves better results than directly pretrained with medical images on all three datasets. A key difference between the two is the scale of available data for pretraining: ImageNet contains approximately 200 times more images than the medical datasets. We hypothesize that pretraining on larger medical image datasets can further enhance segmentation performance, although it requires substantial data and computational resources.

H. Ablations

In this section, we present the results obtained by ablation studies from the perspective of the network structure design, the training procedure, the Mamba blocks, and the effective receptive field.

1) Network structure: As shown in Fig. 2, Swin-UMamba† exhibits an asymmetrical design, with the encoder and decoder having different blocks in stage 3. Here, we demonstrate two symmetric structures: 1) increasing the number of VSS blocks in stage 3 of the decoder to 9 (w/ symmetric decoder), and 2) reducing the number of VSS blocks in stage 3 of the encoder to 2 (w/ symmetric encoder). As shown in Table. IV, adopting the symmetric decoder does not yield improved performance but increases network parameters and computational cost. Using the symmetric encoder has an observable performance drop since the structure of the encoder is changed. The knowledge from the foundation model cannot be effectively transferred to the modified encoder due to structure differences.

2) Training procedure: We compare two different training procedures by 1) disabling the deep supervision from training (w/o deep supervision), and 2) loading the weights of VMamba-Tiny to the decoder of Swin-UMamba† (w/ pretrained decoder). As shown in Fig. 5, loading VMamba-Tiny to the decoder does not lead to better results. In contrast, the segmentation performance becomes worse with 1.21% decrease in DSC and 1.09% decrease in NSD. The structure of the decoder is different from VMamba-Tiny, thus the impact of the model can be ineffective. Besides, the data flow in the decoder is completely reversed to the pretrained backbone. Disable deep supervision leads to a 0.15% drop in DSC and 0.17% in NSD.

3) Mamba blocks: To assess the effectiveness of different Mamba blocks, we replace the basic block (VSS block) in Swin-UMamba† with Vim block [31] and Mamba block

TABLE V: Results of Swin-UMamba† with different model initialization.

Initialization	AbdomenMRI DSC	AbdomenMRI NSD	Endoscopy DSC	Endoscopy NSD	Microscopy F1
Medical pretraining	0.7208±0.1200	0.7905±0.1269	0.5932±0.3235	0.6072±0.3278	0.5287±0.2700
ImageNet pretraining	0.7705±0.0963	0.8376±0.0981	0.6783±0.2969	0.6933±0.3011	0.5982±0.2364
ImageNet pretraining + medical adaption	**0.7767±0.0940**	**0.8428±0.0938**	**0.6931±0.2976**	**0.7094±0.3019**	**0.6219±0.2452**

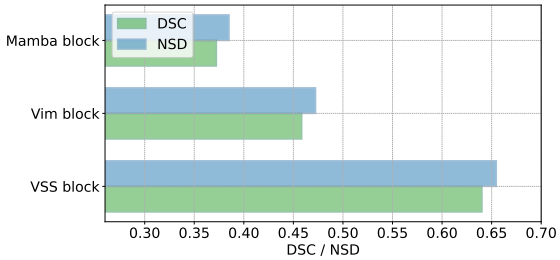
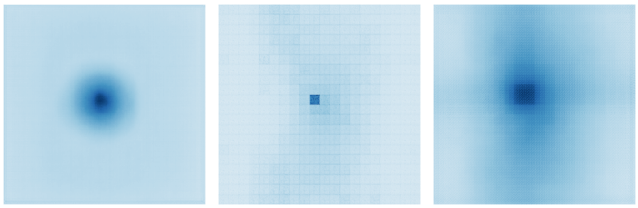


Fig. 6: Ablations of different basic blocks with Swin-UMamba† on the Endoscopy dataset. All models were trained from scratch without vision foundation model adaptation.



(a) SegResNet (b) UNETR (c) Swin-UMamba†

Fig. 7: Visualizations of the effective receptive fields (ERFs) of (a) SegResNet, (b) UNETR, and (c) Swin-UMamba†. A larger ERF is indicated by a more extensively distributed dark area. Swin-UMamba† demonstrates the largest ERF.

[22]. Compared with the VSS block, the Vim block uses a bidirectional state space model, while the Mamba block scans in only one direction. These models were trained from scratch without vision foundation model adaptation. Experiments were conducted with the Endoscopy dataset because more stable training can be achieved on this dataset when the models are trained from scratch. Each model was trained up to 1000 epochs. Fig. 6 presents the segmentation performance of different blocks. VSS block achieves the best performance, with Vim block ranking second. The results demonstrate that the VSS block is more effective than the Vim and Mamba blocks in medical image segmentation. This suggests that adapting Mamba for vision tasks is important to avoid suboptimal performance. Directly applying the Mamba block performs worst because it would inevitably result in restricted receptive fields as the relationships with unscanned patches cannot be estimated [29].

4) Effective receptive field: We compare the effective receptive field (ERF) of Swin-UMamba† with the CNN-based nnU-Net and the transformer-based UNETR. As shown in Fig. 7, Swin-UMamba† demonstrates the largest ERF, represented by the dark area covering the widest region. By having a larger ERF, the model can capture more comprehensive contextual information, enabling it to understand the spatial relationships and dependencies between different parts of the image, thereby reducing the risk of misclassification. Moreover, We also find that the ERF was not uniformly distributed across the entire image. This non-uniform distribution may be correlated to the real data distribution, where organs tend to appear in the center region. The ERF was computed based on the output features of the encoder across 1000 random samples from the AbdomenMRI dataset.

V. Conclusion

This study proposed a novel Mamba-based model Swin-UMamba† for medical image segmentation. Swin-UMamba† can effectively leverage the power of the vision foundation model while maintaining the computation efficiency of Mamba. Extensive experiments suggest that foundation models offer several advantages for the Mamba-based model in medical image segmentation tasks, including lower computational resource consumption, superior segmentation accuracy, stable convergence, mitigation of overfitting issues, and improved data efficiency. Moreover, we propose a self-supervised model adaptation scheme to bridge the gap between natural and medical data domains. In the future, we believe that a Mamba-based foundation model that is trained on large-scale medical image data holds the potential to further improve the performance of medical image analysis.

References

[1] W. Bai *et al.*, "A population-based phenome-wide association study of cardiac and aortic structure and function," *Nature medicine*, vol. 26, no. 10, pp. 1654–1662, 2020.

[2] X. Mei *et al.*, "Artificial intelligence–enabled rapid diagnosis of patients with COVID-19," *Nature medicine*, vol. 26, no. 8, pp. 1224–1228, 2020.

[3] L. Joskowicz, D. Cohen, N. Caplan, and J. Sosna, "Inter-observer variability of manual contour delineation of structures in CT," *European radiology*, vol. 29, pp. 1391–1399, 2019.

[4] H. Tang *et al.*, "Clinically applicable deep learning framework for organs at risk delineation in CT images," *Nature Machine Intelligence*, vol. 1, no. 10, pp. 480–491, 2019.

[5] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Transactions on Medical Imaging*, vol. 39, no. 6, pp. 1856–1867, 2020.

[6] H.-Y. Zhou *et al.*, "nnFormer: Volumetric medical image segmentation via a 3D transformer," *IEEE Transactions on Image Processing*, vol. 32, pp. 4036–4045, 2023.

[7] J. Guo, H.-Y. Zhou, L. Wang, and Y. Yu, "UNet-2022: Exploring dynamics in non-isomorphic architecture," in *Medical Imaging and Computer-Aided Diagnosis*, Lecture Notes in Electrical Engineering, pp. 465–476, Springer Nature, 2023.

[8] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18, no. 2, pp. 203–211, 2021.

[9] 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: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pp. 234–241, Springer, 2015.

[10] F. Chen *et al.*, "Deep semi-supervised ultrasound image segmentation by using a shadow aware network with boundary refinement," *IEEE Transactions on Medical Imaging*, vol. 42, no. 12, pp. 3779–3793, 2023.

[11] S. Wang *et al.*, "Annotation-efficient deep learning for automatic medical image segmentation," *Nature communications*, vol. 12, no. 1, p. 5915, 2021.

[12] R. Fan *et al.*, "One-vote veto: Semi-supervised learning for low-shot glaucoma diagnosis," *IEEE Transactions on Medical Imaging*, vol. 42, no. 12, pp. 3764–3778, 2023.

[13] R. Bommasani *et al.*, "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.

[14] J. Ma *et al.*, "Segment anything in medical images," *Nature Communications*, vol. 15, no. 1, p. 654, 2024.

[15] J. Cheng *et al.*, "SAM-Med2D," *arXiv preprint arXiv:2308.16184*, 2023.

[16] H. Wang *et al.*, "SAM-Med3D," *arXiv preprint arXiv:2310.15161*, 2023.

[17] R. J. Chen *et al.*, "Towards a general-purpose foundation model for computational pathology," *Nature Medicine*, vol. 30, no. 3, pp. 850–862, 2024.

[18] Y. Zhou *et al.*, "A foundation model for generalizable disease detection from retinal images," *Nature*, vol. 622, no. 7981, pp. 156–163, 2023.

[19] Z. Wang *et al.*, "Foundation model for endoscopy video analysis via large-scale self-supervised pre-train," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, vol. 14228, pp. 101–111, Springer Nature Switzerland, 2023. Series Title: Lecture Notes in Computer Science.

[20] X. Wang *et al.*, "Transformer-based unsupervised contrastive learning for histopathological image classification," *Medical Image Analysis*, vol. 81, p. 102559, 2022.

[21] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2020.

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

[23] F. Shamshad *et al.*, "Transformers in medical imaging: A survey," *Medical Image Analysis*, vol. 88, p. 102802, 2023.

[24] A. Gu, K. Goel, and C. Re, "Efficiently modeling long sequences with structured state spaces," in *International Conference on Learning Representations*, 2022.

[25] A. Gu *et al.*, "Combining recurrent, convolutional, and continuous-time models with linear state space layers," *Advances in neural information processing systems*, vol. 34, pp. 572–585, 2021.

[26] W. Wang *et al.*, "Neuropathologist-level integrated classification of adult-type diffuse gliomas using deep learning from whole-slide pathological images," *Nature Communications*, vol. 14, no. 1, p. 6359, 2023.

[27] J. E. Iglesias *et al.*, "A computational atlas of the hippocampal formation using ex vivo, ultra-high resolution MRI: Application to adaptive segmentation of in vivo MRI," *NeuroImage*, vol. 115, pp. 117–137, 2015.

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

[29] Y. Liu *et al.*, "VMamba: Visual state space model," *arXiv preprint arXiv:2401.10166*, 2024.

[30] Z. Xing, T. Ye, Y. Yang, G. Liu, and L. Zhu, " SegMamba: Long-range Sequential Modeling Mamba For 3D Medical Image Segmentation ," in *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, vol. LNCS 15008, Springer Nature Switzerland, October 2024.

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

[32] T. Guo, Y. Wang, and C. Meng, "MambaMorph: a mamba-based backbone with contrastive feature learning for deformable mr-ct registration," *arXiv preprint arXiv:2401.13934*, 2024.

[33] J. Liu, H. Yang, H.-Y. Zhou, Y. Xi, L. Yu, C. Li, Y. Liang, G. Shi, Y. Yu, S. Zhang, H. Zheng, and S. Wang, "Swin-UMamba: Mamba-based unet with imagenet-based pretraining," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024* (M. G. Linguraru, Q. Dou, A. Feragen, S. Giannarou, B. Glocker, K. Lekadir, and J. A. Schnabel, eds.), (Cham), pp. 615–625, Springer Nature Switzerland, 2024.

[34] Y. Zhou, W. Huang, P. Dong, Y. Xia, and S. Wang, "D-UNet: A dimension-fusion u shape network for chronic stroke lesion segmentation," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 18, no. 3, pp. 940–950, 2021.

[35] R. Wang *et al.*, "Medical image segmentation using deep learning: A survey," *IET Image Processing*, vol. 16, no. 5, pp. 1243–1267, 2022.

[36] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[37] A. Myronenko, "3D MRI brain tumor segmentation using autoencoder regularization," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II 4*, pp. 311–320, Springer, 2019.

[38] H. Cao *et al.*, "Swin-Unet: Unet-like pure transformer for medical image segmentation," in *Computer Vision – ECCV 2022 Workshops*, pp. 205–218, Springer Nature Switzerland.

[39] A. Hatamizadeh *et al.*, "UNETR: Transformers for 3d medical image segmentation," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 574–584, 2022.

[40] J. Chen *et al.*, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[41] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu, "Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images," in *International MICCAI Brainlesion Workshop*, pp. 272–284, Springer, 2021.

[42] Z. Liu *et al.*, "Swin Transformer: Hierarchical vision transformer using shifted windows," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9992–10002, IEEE.

[43] C.-M. Feng, " Enhancing Label-efficient Medical Image Segmentation with Text-guided Diffusion Models ," in *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, vol. LNCS 15008, Springer Nature Switzerland, October 2024.

[44] W. Huang, C. Li, H.-Y. Zhou, H. Yang, J. Liu, Y. Liang, H. Zheng, S. Zhang, and S. Wang, "Enhancing representation in radiography-reports foundation model: a granular alignment algorithm using masked contrastive learning," *Nature Communications*, vol. 15, no. 1, p. 7620. Publisher: Nature Publishing Group.

[45] W. Huang, C. Li, H. Yang, J. Liu, Y. Liang, H. Zheng, and S. Wang, "Enhancing the vision–language foundation model with key semantic knowledge-emphasized report refinement," *Medical Image Analysis*, vol. 97, p. 103299.

[46] A. Gu, T. Dao, S. Ermon, A. Rudra, and C. Ré, "Hippo: Recurrent memory with optimal polynomial projections," *Advances in neural information processing systems*, vol. 33, pp. 1474–1487, 2020.

[47] K. Han *et al.*, "A survey on vision transformer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 87–110.

[48] H. Sun *et al.*, "AUNet: attention-guided dense-upsampling networks for breast mass segmentation in whole mammograms," *Physics in Medicine & Biology*, vol. 65, p. 055005, feb 2020.

[49] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, "Deeply-supervised nets," in *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, pp. 562–570, PMLR.

[50] Z. Xie *et al.*, "SimMIM: a simple framework for masked image modeling," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9643–9653, 2022.

[51] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15979–15988. ISSN: 2575-7075.

[52] J. Ma *et al.*, "Unleashing the strengths of unlabeled data in pan-cancer abdominal organ quantification: the flare22 challenge," *arXiv preprint arXiv:2308.05862*, 2023.

[53] M. Allan *et al.*, "2017 robotic instrument segmentation challenge," *arXiv preprint arXiv:1902.06426*, 2019.

[54] J. Ma *et al.*, "The multi-modality cell segmentation challenge: towards universal solutions," *arXiv preprint arXiv:2308.05864*, 2023.