Swin-UMamba: Rethinking Mamba-based UNet with ImageNet-based pretraining

Abstract

Accurate medical image segmentation demands the integration of multi-scale information, spanning from local features to global dependencies. However, it is challenging for existing methods to model long-range global information, where convolutional neural networks (CNNs) are constrained by their local receptive fields, and vision transformers (ViTs) suffer from high quadratic complexity of their attention mechanism. Recently, Mamba-based models have gained great attention for their impressive ability in long sequence modeling. Several studies have demonstrated that these models can outperform popular vision models in various tasks, offering higher accuracy, lower memory consumption, and less computational burden. However, existing Mamba-based models are mostly trained from scratch and do not explore the power of pretraining, which has been proven to be quite effective for data-efficient medical image analysis. This paper introduces a novel Mamba-based model, Swin-UMamba, designed specifically for medical image segmentation tasks, taking advantage of ImageNet-based pre-training. Our experimental results reveal the vital role of ImageNet-based training in improving the performance of Mamba-based models. Swin-UMamba demonstrates superior performance with a large margin compared to CNNs, ViTs, and latest Mamba-based models. In particular, on AbdomenMRI, endoscopy, microscopy and thyroid datasets, Swin-UMamba outperforms its closest counterpart U-Mamba_Enc by an average score of 2.72%.Finally we investigated improvements in the direction of scanning patterns from the Mamba structure and multiscale feature extraction from the model, using Kvasir-SEG and Thyroid datasets to experimentally demonstrate the competitive performance of our improvements.

Keywords: Medical image segmentation, ImageNet-based pretraining, Long-range dependency modeling.

1 Introduction

Medical image segmentation plays an important role in modern clinical practice such as assisting in diagnoses, formulating treatment plans, and implementing therapies [2, 27, 35]. A typical segmentation process relies on experienced doctors, which is both labor-intensive and time-consuming. Besides, the segmentation consistency between experts can vary due to subjective interpretations and inter-observer variability [17, 18]. This highlights the need for automated segmentation methods to enhance efficiency, accuracy, and consistency in medical image analysis to make accurate and rapid diagnoses [19, 34].

In recent years, deep learning has made significant advancements in the field of medical image segmentation [7,15,31,43]. However, accurate medical image segmentation requires integrating local features with their corresponding global dependencies [32]. It is still challenging to efficiently capture complex and long-range global dependencies from image data. Two prevalent approaches, Convolutional neural networks (CNNs) and vision transformers (ViTs), have their own limitations in long-range dependencies modeling. CNNs such as SegResNet [28], U-Net [31], and nnU-Net [15], are commonly employed in medical image segmentation. They are effective at extracting local features but may struggle with capturing global context and long-range dependencies. This is because CNNs are inherently limited by their local receptive fields [23], which restrict their ability to capture information from distant regions in the image. On the other hand, ViTs have shown the capability to handle global context and long-range dependencies [12, 30]. However, ViTs are constrained by their attention mechanism, suffering from high quadratic complexity for long sequences modeling [4], where high-resolution images are not rare in the medical domain (e.g. whole-slide pathology images [38], high-resolution MRI/CT scans [14]). Despite the complexity, transformers are prone to overfitting when dealing with limited datasets [21], indicating their data-hungry nature.

Recently, structured state space sequence models (SSMs) [5, 6] demonstrated their efficiency and effectiveness in long sequence modeling, potentially becoming the solution for long-range dependency modeling in vision tasks. Compared with transformers, they scale linearly or near-linearly with sequence length while maintaining the capability of modeling long-range dependencies, obtaining cutting-edge performance in continuous long-sequence data analysis such as natural language processing and genomic analysis [4]. Several latest studies have preliminarily explored the effectiveness of Mamba in the vision domain [8, 22, 24, 40, 45]. For instance, Vim [45] proposed a generic vision backbones with bidirectional Mamba blocks. In contrast, VMamba [22] builds up a Mamba-based vision backbone with hierarchical representations. Additionally, VMamba introduced a cross-scan module to solve the direction-sensitive problem due to the difference between 1D sequences and 2D images. For medical image segmentation, U-Mamba [24] and SegMamba [40] proposed a task-specific architecture with the Mamba block based on nnUNet and Swin-UNETR, respectively. These models have achieved promising results in various vision tasks, demonstrating the potential of SSMs in vision.

However, existing Mamba-based models are mostly trained from scratch. The impact of pretraining for the Mamba-based model in medical image segmentation tasks remains unclear, which has been proven to be quite effective for data-efficient medical image analysis with CNNs [7] and ViTs [9]. This is particularly important in the field of medicine, where medical image datasets are often limited in size and diversity [36,37]. Understanding the effectiveness of pretraining Mamba-based models in medical image segmentation can offer valuable insights into enhancing the performance of deep learning models in medical imaging applications.

There are several challenges that need to be addressed. First, existing Mamba-based models for medical image segmentation have not taken into account the transferability from ImageNet pretrained models. Consequently, the network structure requires redesigning to integrate the pretrained model. Given the fact that the application of Mamba blocks in the vision domain is relatively new, further experimental evaluation is required for medical image segmentation tasks. Third, there is a need for the scalability and efficiency of Mamba-based models for real-world deployment [44], particularly in resource-constrained environments, which is commonly found in medical practice.

In this paper, we proposed a Mamba-based network Swin-UMamba for 2D medical image segmentation. Swin-UMamba uses a generic encoder to integrate the power of the pretrained vision model with a well-designed decoder for medical image segmentation tasks. In addition, we proposed a variant structure Swin-UMamba† with a Mamba-based decoder, providing fewer parameters and lower FLOPs for efficient applications. Our contribution can be summarized as follows:

- To the best of our knowledge, we are the first attempt to discover the impact of pretrained Mamba-based networks in medical image segmentation. Our experiment verified that ImageNet-based pretraining plays an important role in medical image segmentation for Mamba-based networks, which sometimes is crucial.
- We propose a new Mamba-based network named Swin-UMamba for medical image segmentation, which is particularly designed to unify the power of pretrained models. Additionally, we proposed a variant structure Swin-UMamba† with fewer network parameters and lower FLOPs while maintaining competitive performance.
- Our results show that both Swin-UMamba and Swin-UMamba† can outperform previous segmentation models including CNNs, ViTs, and the latest Mamba-based models with notable margin, highlighting the effectiveness of ImageNet-based pretraining and proposed architecture in medical image segmentation tasks.

2 Related works

2.1 Preliminaries

In contemporary SSM-based models, namely, Structured State Space Sequence Models (S4) and Mamba [8, 22], both depend on a traditional continuous system that maps a one-dimensional input function or sequence, represented as $x(t) \in R$, through intermediary implicit states $h(t) \in R^N$ to an output $y(t) \in R$. This process can be depicted as a linear Ordinary Differential Equation (ODE):

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

where $\mathbf{A} \in R^{N\times N}$ represents the state matrix, while $\mathbf{B} \in R^{N\times 1}$ and $\mathbf{C} \in R^{N\times 1}$ denote the projection parameters.

S4 and Mamba discretize this continuous system to adapt it better for deep learning contexts. Specifically, they incorporate a timescale parameter $\mathbf{\Delta}$ and convert $\mathbf{A}$ and $\mathbf{B}$ into discrete parameters $\overline{\mathbf{A}}$ and $\overline{\mathbf{B}}$ using a consistent discretization rule. The zero-order hold (ZOH) is typically utilized as the discretization rule and can be outlined as follows:

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

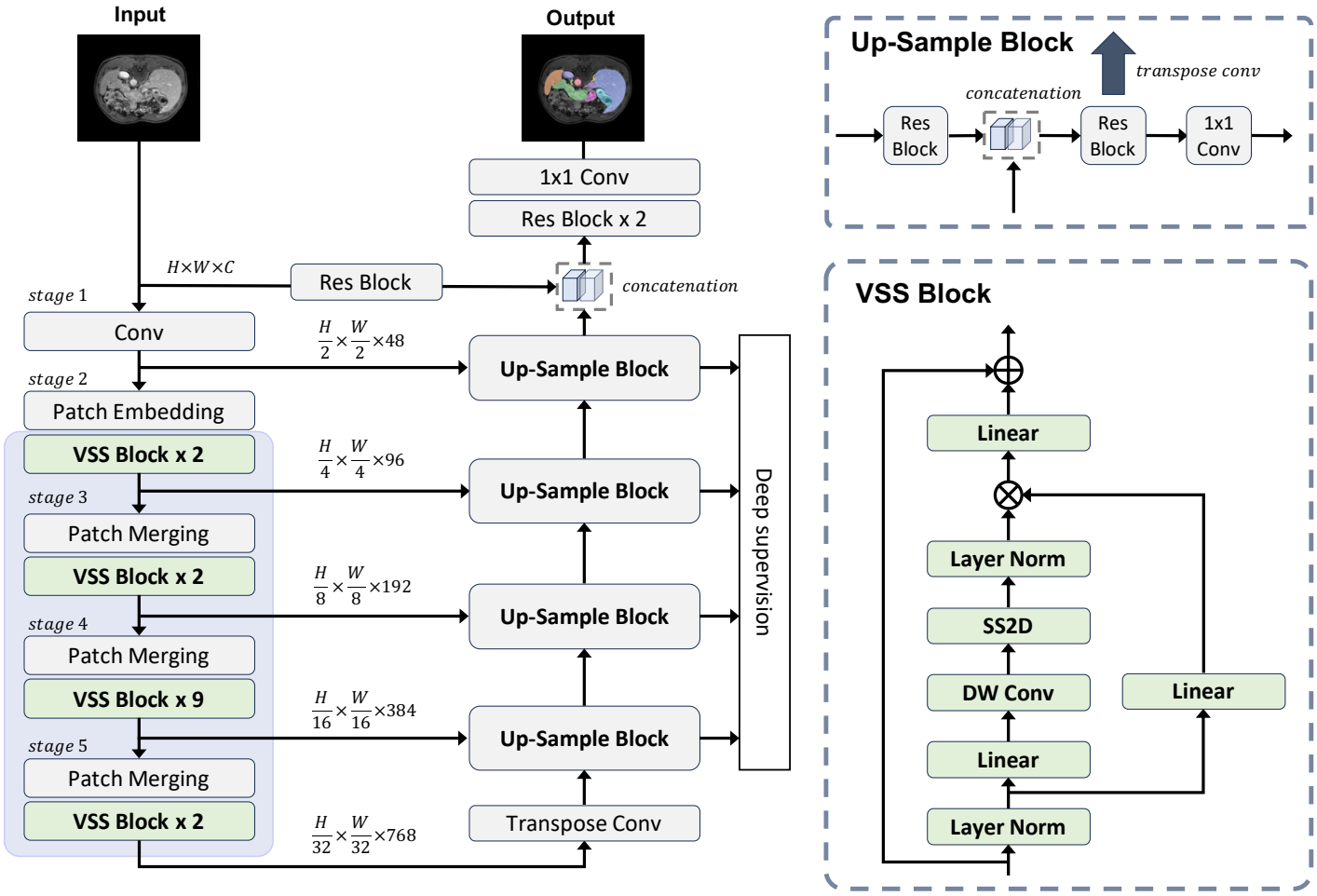

Figure 1. The overall architecture of Swin-UMamba. Swin-UMamba can leverage the power of vision foundation models by loading the weights of pretrained models. Each block within the blue box was initialized with the ImageNet pretrained weights.

Following discretization, SSM-based models can be calculated in two distinct methods: linear recurrence or global convolution, which are denoted as equations (3) and (4), respectively.

$$\begin{aligned} h'(t) &= \overline{\mathbf{A}} h(t) + \overline{\mathbf{B}} x(t) \\ y(t) &= \mathbf{C} h(t) \end{aligned} \tag{3}$$

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

where $\overline{\mathbf{K}} \in R^L$ represents a structured convolutional kernel, and L denotes the length of the input sequence x.

3 Method

3.1 Overview

We illustrate the overall architecture of Swin-UMamba in Fig. 1. It is mainly composed of 1) a Mamba-based encoder that was pretrained on the large-scale dataset (i.e. ImageNet) to extract features at different scales, 2) a decoder with several up-sample blocks for predicting segmentation results, and 3) skip connections to bridge the gap between low-level details and high-level semantics. We will introduce the detailed structure of Swin-UMamba in the following sections.

3.2 Integrating ImageNet-based pretraining

The primary challenge lies in effectively integrating generic pretrained models into medical image segmentation tasks. Prior research [24] typically employs a specific architecture with Mamba blocks, which fails to consider the transferability from generic vision models. To address this issue, we construct an encoder that shares a similar structure with the latest Mamba-based approach in vision, namely, VMamba-Tiny [22]. This model, pretrained on the extensive ImageNet dataset with multi-scale features, allows us to integrate the power of the generic vision model to extract information with long-range modeling capability, mimic the risk of overfitting, and establish a robust initialization for Swin-UMamba.

The encoder of Swin-UMamba can be divided into 5 stages. The first stage is the stem stage. It contains a convolution layer for $2\times$ down-sampling with a 7×7 kernel, a padding size of 3, and a stride size of 2. 2D instance normalization was adopted after the convolution layer. The first stage of Swin-UMamba is different from VMamba because we prefer a gradual down-sampling process where each stage takes $2\times$ down-sampling. This strategy aims to retain low-level details, which is important for medical image segmentation [31, 33]. The second stage uses a patch embedding layer with a 2×2 patch size, maintaining the feature resolution at $\frac{1}{4}\times$ of the original image, which is the same as embedded features in VMamba. Subsequent stages follow the design of VMamba-Tiny, where each stage is composed of a patch merging layer for $2\times$ down-sampling and several VSS blocks for high-level feature extraction. Unlike ViTs, we did not adopt the position embedding in Swin-UMamba due to the causal nature of VSS block [22]. The number of VSS blocks at stage-2 to stage-5 are $\{2, 2, 9, 2\}$, respectively. The feature dimensions after each stage are quadratically increased w.r.t. the stages, resulting in $D = \{48, 96, 192, 384, 768\}$. We initialize the VSS blocks and patch merging layers using the ImageNet pretrained weights from VMamba-Tiny, as illustrated in Fig. 1. Notably, the patch embedding block is not initialized with pretrained weights due to differences in patch size and input channels.

3.3 Swin-UMamba decoder

We follow the commonly used U-shaped architecture with dense skip connections to construct Swin-UMamba. U-Net and its variations have demonstrated remarkable efficiency in medical image segmentation tasks. This architecture leverages skip connections for the recovery of low-level details and employs an encoder-decoder structure for high-level information extraction. 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 [20].

The structure of the up-sample block was illustrated in Fig. 1. Given skip-connected feature z_l' from stage-l and feature z_{l+1} from the last up-sample block, the output feature 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:

$$\hat{z}_l = Res_l^{(2)}(Cat(z_{l+1}, Res_l^{(1)}(z_l'))) \tag{5}$$

$$z_l = DeConv_l(\hat{z}_l), \quad y_l = Conv_l(\hat{z}_l) \tag{6}$$

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 $Res(\cdot)$ was composed of two convolution layers with LeakyRELU activation. In addition to the skip connections between encoding stages and up-sample blocks, we add an extra skip connection from the input with $Res(\cdot) - Cat(\cdot) - Res(\cdot)$ operations. We use a 1×1 convolution to get the final segmentation output.

3.4 Swin-UMamba†: Swin-UMamba with Mamba-based decoder

To further explore the potential of Mamba in medical semantic segmentation, we proposed a variant of Swin-UMamba with a Mamba-based decoder, namely Swin-UMamba†. We will show that Swin-UMamba† can obtain competitive results compared to Swin-UMamba while utilizing fewer network parameters and imposing lower computational burdens. Furthermore, our findings reveal the important role of large-scale pretraining in medical image segmentation tasks regardless of the decoder structure.

Several modifications were made on Swin-UMamba†. First, the up-sample blocks in Swin-UMamba were replaced by patch expanding [3] and two VSS blocks. We found that many parameters and computation burdens were caused by the heavy CNN-based decoder. Second, we use a 4×4 patch embedding layer that directly projects input image from $H \times W \times C$ into feature maps of shape $\frac{H}{4} \times \frac{H}{4} \times 96$ follow VMamba [22]. Notably, the last patch expanding block in Swin-UMamba† is a $4\times$ up-sample operation, mirroring the $4\times$ patch embedding layer. The residual patch expanding layers were $2\times$ up-sampling operation. Skip connections originating from the input image and $2\times$ down-sampled features in Swin-UMamba were removed, as there were no corresponding decoding blocks for them. Additionally, deep supervision was applied at resolutions of $1\times$, $\frac{1}{4}\times$, $\frac{1}{8}\times$, and $\frac{1}{16}\times$, incorporating additional segmentation heads (i.e., 1×1 convolutions mapping high-dimensional features to K) for each scale. Combining all these modifications, the number of network parameters was reduced from 60M to 28M, and the FLOPs were decreased from 68.0G to 18.9G on the AbdomenMRI dataset. Further statistics regarding the number of network parameters and FLOPs are provided in Table 1, Table 2, and Table 3. The structure of Swin-UMamba† is illustrated in Fig. 2.

3.5 Loss function

For our medical image segmentation tasks, we primarily employ basic Cross-Entropy and Dice loss as the loss function cause all of our dataset masks comprise two classes, which are a singular target and the background.

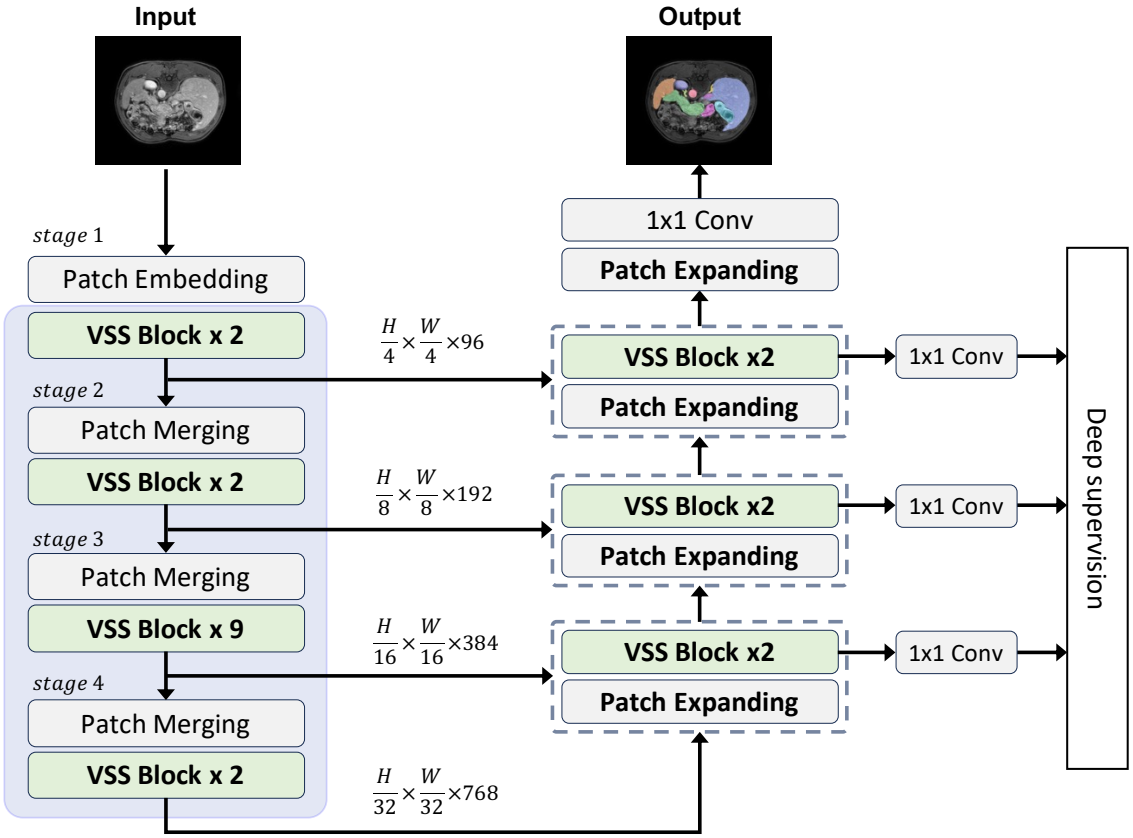

Figure 2. The overall architecture of Swin-UMamba†.

$$
\begin{aligned}
L_{\text{BceDice}} &= \lambda_1 L_{\text{Bce}} + \lambda_2 L_{\text{Dice}} \\
L_{Bce} &= -\frac{1}{N}\sum_{1}^{N}\left[y_i log\left(\hat{y}_i\right) + \left(1-y_i\right) log\left(1-\hat{y}_i\right)\right] \\
L_{\text{Dice}} &= 1 - \frac{2|X \cap Y|}{|X| + |Y|}
\end{aligned}
\tag{7}
$$

(λ_1, λ_2) are constants, with $(1, 1)$ often selected as the default parameters.

4 Experiments

4.1 Datsets

We evaluate the performance and scalability of Swin-UMamba across three distinct medical image segmentation datasets, encompassing organ segmentation, instrument segmentation, and cell segmentation. These datasets are selected across various resolutions and image modalities, providing insights into the model's efficacy and adaptability in diverse medical imaging scenarios.

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 from the MICCAI 2022 AMOS Challenge [26]. We followed the settings in [24] with additional 50 MRI scans for testing. There

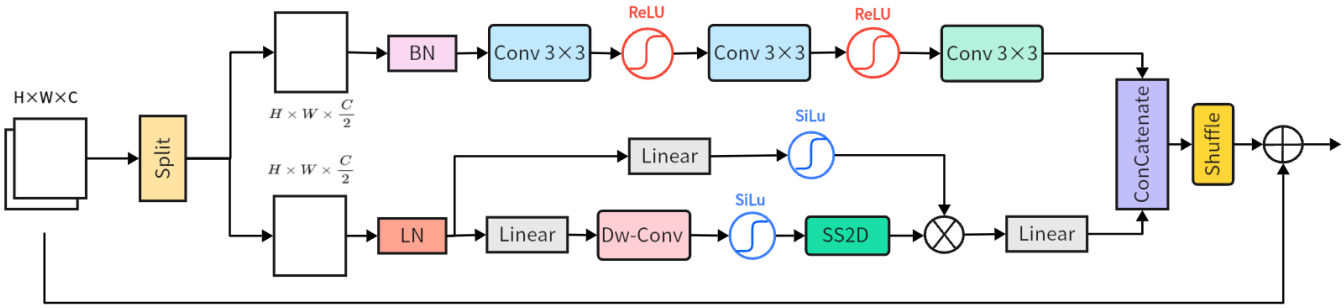

Figure 3. The overall architecture of the Conv-SSM with is a simple two-branch module.The local feature extraction capability of the convolutional layers is combined with the ability of the SSM to capture long-range dependencies.The BN, LN, Linear and DWConv represent Batch normalize,Layer normalization,Linearlayer and Depth-wiseconvolution,respectively.

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 nnUNet framework [15].

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 [1]. 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.

Microscopy images (Microscopy) This dataset is focused on cell segmentation in various microscopy images from the NeurIPS 2022 Cell Segmentation Challenge [25]. It consists of 1000 images for training and 101 images for evaluation.We employed the same data processing strategy as described in [24] for this dataset.

Thyroid images (Thyroid) This data set is focused on the segmentation of thyroid nodules in various thyroid images from the subject group. It consists of 800 images for training, 100 images for evaluation and 100 images for testing.

4.2 Implemetation details

We implemented Swin-UMamba on top of the well-established nnU-Net framework [15]. It's self-configuring feature enabled us to focus on network design rather than other trivial details. The loss function is the sum of Dice loss and cross-entropy loss and we perform deep supervision [20] at each scale. We use an AdamW optimizer with weight decay $= 0.05$ following [22]. A cosine learning rate decay was adopted with an initial learning rate $= 0.0001$. We use the pretrained VMamba-Tiny model[1] to initialize Swin-UMamba for all datasets. During training, we froze parameters from the pretrained model for the first 10 epochs to align other modules. Hyperparameters were kept consistent across all three datasets, except for the number of training epochs and data-specific settings (e.g., image patch size). Swin-UMamba was trained for 100 epochs on the AbdomenMRI dataset,200 epochs on the Thyroid dataset,350 epochs on the Endoscopy dataset,and 450 epochs on the Microscopy dataset, respectively. Following [24], we disabled the testing time argumentation for a more

[1] The pretrained model can be found at: https://github.com/MzeroMiko/VMamba

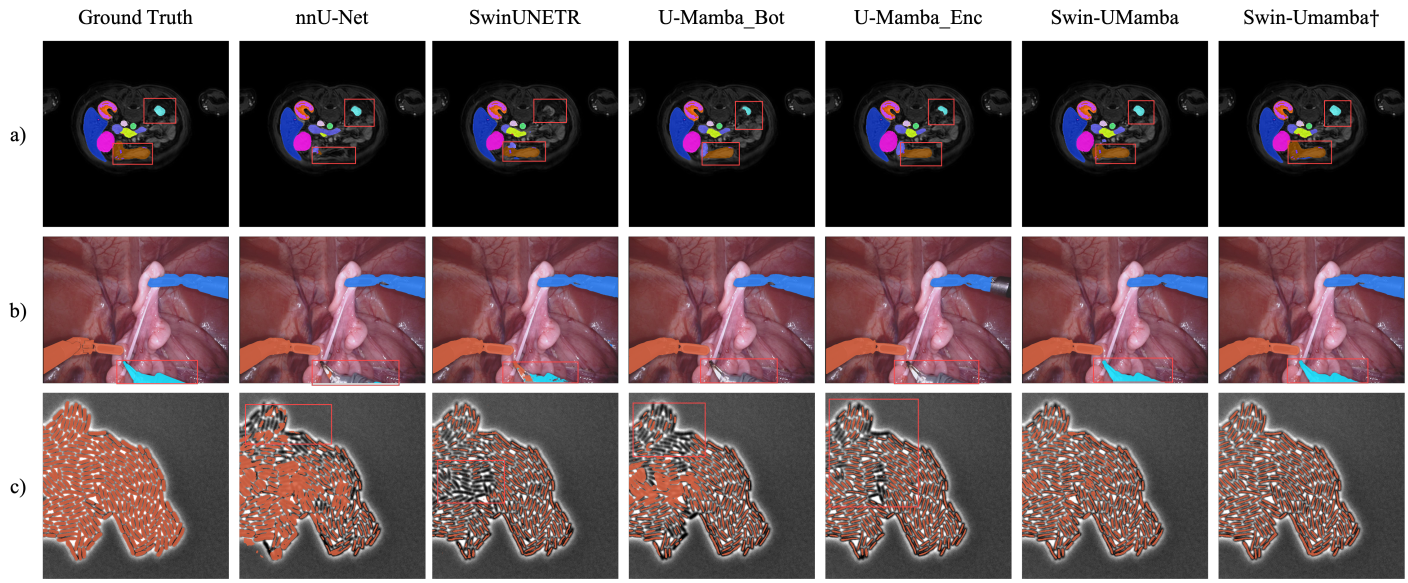

Figure 4. Result visualization on a) AbdomenMRI, b) Endoscopy, and c) Microscopy. Swin-UMamba accurately recognizes the shape and type of the segmented targets.

streamlined and efficient evaluation. It's worth noting that further improvement is possible with additional training and proper hyperparameter tuning, which we leave for future work. Our primary goal is to assess the impact of the pretrained models on medical image segmentation rather than solely aiming for state-of-the-art (SOTA) performance.

4.3 Baselines and evaluation metrics

We select three types of methods as baseline methods for comprehensive evaluation, including CNN-based (nnU-Net [15], SegResNet [28]), transformer-based (UNETR [11], Swin-UNETR [10], nnFormer [43]), and the latest Mamba-based segmentation network U-Mamba [24]. 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 of U-Mamba_Bot and U-Mamba_Enc. It's worth noting that adopting the pretrained model into U-Mamba is not straightforward due to structural differences from the pretrained model [22].

Dice similarity coefficient (DSC) and normalized surface distance (NSD) were used to evaluate segmentation performance on the AbdomenMRI and Endoscopy datasets. For the Microscopy dataset, we use the F1 score for evaluation because it is an instance segmentation task. Furthermore, we compute the number of network parameters (#param) and floating-point operations (FLOPs) using the *fvcore* package to assess the scale and computational burden of each model. The baseline results for DSC, NSD, and F1 score were referenced from [24] except nnFormer. We report the results of nnFormer based on official implementation.

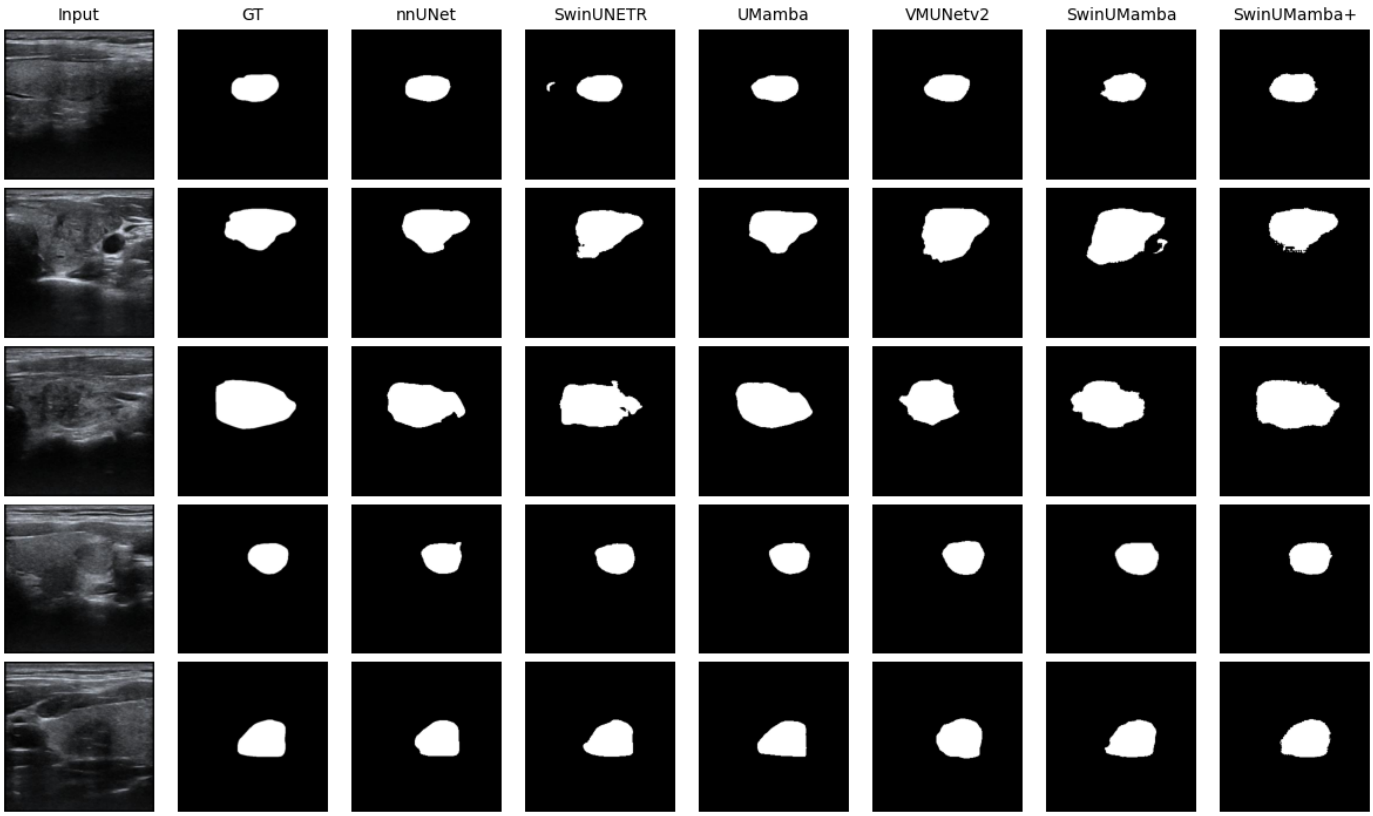

Figure 5. Result visualization on Thyroid. Swin-UMamba accurately recognizes the shape and type of the segmented targets.

5 Results and analysis

5.1 Comparisons on AbdomenMRI dataset

Table 1 presents the segmentation performance on the AbdomenMRI dataset. Both Swin-UMamba and Swin-UMamba† outperform all baseline methods, including CNN-based networks, transformer-based networks, and Mamba-based networks. The superior result demonstrates the great potential of the Mamba-based network in medical image segmentation. Notably, all Mamba-based networks outperform CNN-based and transformer-based baselines by at least 1% on both DSC and NSD. Swin-UMamba exhibits a remarkable 1.34% improvement in DSC over U-Mamba_Enc, which is the previous SOTA model on this dataset. As illustrated in Fig. 4a, Swin-UMamba can recognize the shape and type of target organs, whereas baseline methods fail to accurately identify all target regions.

ImageNet-based pretraining plays a crucial role in our experiments, leading to a significant 7.06% improvement in DSC and a notable 7.74% improvement in NSD for Swin-UMamba. Moreover, leveraging ImageNet-based pretraining facilitates faster and more stable training, requiring merely one-tenth of the training iterations compared to baseline methods. A drastic phenomenon is observed with Swin-UMamba†. Without ImageNet-based pretraining, Swin-UMamba† fails to converge properly on this dataset with default settings. To address this issue, we disable the deep supervision of Swin-UMamba† and extend its training epochs to 200. Despite that, Swin-UMamba† outperforms all baseline methods when utilizing the ImageNet pretrained weights. This improvement is particularly noteworthy considering that Swin-UMamba† has less than half of

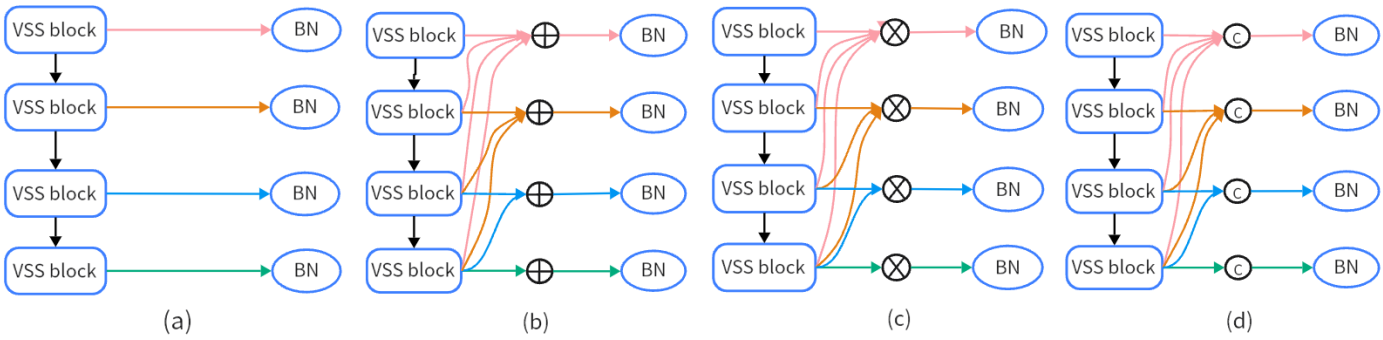

Figure 6. Different styles of feature fusion. (a) Simple skip connection is used, but the features of different scales are not fused. (b) Dense addition connection (DAC) module: the features of different scales adopt the dense composition method of addition. (c) DMC module: the dense composition method of multiplication is adopted between the features. (d) Dense concatenation connection (DCC) module: the features of different scales adopt the dense concatenation method of addition. The composite connections are lines representing some simple operations, i.e., upsampling and 1×1 convolution.

the network parameters and FLOPs compared to the previous SOTA model U-Mamba.

We also observed a disparity in parameter numbers and FLOPs between Swin-UMamba† and Swin-UMamba. This discrepancy is primarily attributed to the CNN-based decoder, as Swin-UMamba† and Swin-UMamba share almost identical structures in the encoder part. We opted to retain the CNN-based decoder to evaluate the impact of pretraining for different models, and it did take better results in this dataset.

5.2 Comparisons on Endoscopy dataset

Table 2 presents the segmentation performance of each model on the Endoscopy dataset. Swin-UMamba† outperforms U-Mamba_Bot over 2.43% on DSC and 2.41% on NSD. The visualized result of Swin-UMamba on Endoscopy is shown in Fig. 4b. Notably, we observed an impressive performance gain of 12.84% on DSC and 12.90% on NSD with the pretrained model for Swin-UMamba. One possible explanation is 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 pretrained model is an effective strategy for mitigating overfitting in such small datasets. In addition, we found that Swin-UMamba† performs better than Swin-UMamba on this dataset, possibly benefiting from its fewer parameters to avoid overfitting.

5.3 Comparisons on Microscopy dataset

Table 3 presents the segmentation performance on the Microscopy dataset. Swin-UMamba and Swin-UMamba† continue to outperform all baseline methods by margins ranging from 1.99% to 20.15%. In contrast to previously mentioned datasets, the Microscopy dataset features higher image resolution, fewer samples, and greater visual difference. This imposes greater demands on the model's capacity for long-range information modeling and data-efficiency. As shown in Fig. 4c, Swin-UMamba can accurately segment target cells while baselines missing some. Moreover, we observe that Swin-UMamba† and Swin-UMamba benefit from the ImageNet pretraining by 12.45% and 7.96% respectively. This once again demonstrates the importance of the pretraining especially for small datasets.

Table 1. Results of organ segmentation on the AbdomenMRI dataset. †: using a Mamba-based decoder. $*$: Deep supervision was disabled and we extend the training epochs to 200. The results of nnU-Net, SegResNet, UNETR, SwinUNETR, and U-Mamba were referenced from [24].

Methods	#param	FLOPs	DSC	NSD
CNN-based				
nnU-Net	33M	23.3G	0.7350	0.7953
SegResNet	6M	24.5G	0.7391	0.7934
Transformer-based				
UNETR	87M	42.1G	0.6105	0.6309
SwinUNETR	25M	27.9G	0.6225	0.7613
nnFormer	60M	50.2G	0.7279	0.7963
Mamba-based				
U-Mamba_Bot	63M	45.7G	0.7588	0.8048
U-Mamba_Enc	67M	49.9G	0.7625	0.8134
w/o ImageNet-based pretraining				
Swin-UMamba	60M	68.0G	0.7011	0.7590
Swin-UMamba†$*$	28M	18.9G	0.6589	0.7234
w/ ImageNet-based pretraining				
Swin-UMamba	60M	68.0G	**0.7760**	**0.8426**
Swin-UMamba†	28M	18.9G	<u>0.7705</u>	<u>0.8376</u>

5.4 Comparisons on Thyroid dataset

Table 4 presents the segmentation performance on the Thyroid dataset. Swin-UMamba and Swin-UMamba† continue to outperform all baseline methods by margins ranging from 1.24% to 18.87%. In contrast to previously mentioned datasets, the Thyroid dataset features higher image resolution, fewer samples, and greater visual difference. This imposes greater demands on the model's capacity for long-range information modeling and data-efficiency. As shown in Fig. 5, Swin-UMamba can accurately segment target cells while baselines missing some.

6 Improvement strategies

6.1 scanning

In contrast, scanning operations generally have linear time complexity, making them more efficient for long sequences. The scan operation involves calculating an array, like the prefix sum, where each value is determined by using the previously calculated value and the current input. Similarly, the recurrent form of SSM can be viewed as a scan operation. Scanning is a crucial component in mamba, especially when handling multidimensional inputs. The selection of the scanning mechanism in Mamba models is crucial as it enhances

Table 2. Results of instruments segmentation on the Endoscopy dataset. † means using a Mamba-based decoder. The results of nnU-Net, SegResNet, UNETR, SwinUNETR, and U-Mamba were referenced from [24].

Methods	#param	FLOPs	DSC	NSD
CNN-based				
nnU-Net	33M	55.9G	0.6264	0.6412
SegResNet	6M	58.9G	0.5498	0.5968
Transformer-based				
UNETR	88M	111.5G	0.5017	0.5168
SwinUNETR	25M	67.1G	0.5528	0.5683
nnFormer	60M	125.5G	0.6122	0.6228
Mamba-based				
U-Mamba_Bot	63M	109.7G	0.6158	0.6280
U-Mamba_Enc	67M	119.8G	0.6310	0.6362
w/o ImageNet-based pretraining				
Swin-UMamba	60M	163.6G	0.5183	0.5310
Swin-UMamba†	28M	45.3G	0.6319	0.6446
w/ ImageNet-based pretraining				
Swin-UMamba	60M	163.6G	<u>0.6804</u>	<u>0.6944</u>
Swin-UMamba†	28M	45.3G	**0.6728**	**0.6908**

efficiency and provides important information.We combined different scanning methods and tested them on the Swin-UMamba architecture, and found that the original Selective 2D Scanning strategy worked better 1.2% than the other scanning strategies.

Selective Scan 2D:

SS2D [22] performs scanning operations in three directions: top to bottom, left to right, and in reverse direction. Each mamba block is placed to work independently within these directions. SS2D mirrors the self-attention process seen in transformers. It overcomes the limitations of bidirectional scan in ViM, but it also leads to a loss of patch continuity. To address this, SS2D incorporates a scan merge step, where representations from each scan direction are combined into a unified output.

Continuous 2D Scan:

Continuous 2D scan [41] resolves the issue which is experienced in SS2D. It involves integrating direction-aware parameters into cross-scan mechanism and organizing patches accordingly. This approach ensures the preservation of patch continuity and maintains the contextual understanding of images.

Table 3. Results of cell segmentation on the Microscopy dataset. † means using a Mamba-based decoder. The results of nnU-Net, SegResNet, UNETR, SwinUNETR, and U-Mamba were referenced from [24].

Methods	#param	FLOPs	F1
CNN-based			
nnU-Net	46M	60.1G	0.5077
SegResNet	6M	62.8G	0.5402
Transformer-based			
UNETR	88M	120.1G	0.4031
SwinUNETR	25M	71.7G	0.3560
nnFormer	60M	136.7G	0.5332
Mamba-based			
U-Mamba_Bot	86M	117.8G	0.5342
U-Mamba_Enc	92M	128.7G	0.5607
w/o ImageNet-based pretraining			
Swin-UMamba	60M	174.4G	0.4456
Swin-UMamba†	27M	48.2G	0.4342
w/ ImageNet-based pretraining			
Swin-UMamba	60M	174.4G	<u>0.5667</u>
Swin-UMamba†	27M	48.2G	**0.5834**

Local Scan:

Local Scanning [13] overcomes limitations of scanning methods in ViM and VMamba by preserving local dependencies in images through distinct local windows. This technique maintains the global context of the image without compromise. The authors suggest using 7×7 and 2×2 local windows to capture the local context while alternating the scan direction. Vertical and horizontal scans with direction flipping are used to grasp the global context of image tokens.

Efficient 2D Scan:

Efficient Scan 2D (ES2D) [29] emphasizes efficient image scanning by skipping scan patches with a step size p It partitions selected spatial dimension features into m and n using sine and cosine functions to determine the patch location.

6.2 Multiplicative Connection Module

As shown in Fig. 6,Dense addition connection (DAC) module and Convolutional Block Attention Module [39], are proposed to aggregate the multiscale context information of medical images. The DMC module cascades the multiscale semantic feature information through dense multiplicative feature fusion, which mini-

Table 4. Results of nodule segmentation on the Thyroid dataset. † means using a Mamba-based decoder. The results of nnU-Net, SegResNet, UNETR, SwinUNETR,VM-UNetV2 [42] and U-Mamba were referenced from [24].

Methods	#param	FLOPs	DSC	NSD
CNN-based				
nnU-Net	33M	23.3G	0.8210	0.8795
SegResNet	6M	24.9G	0.8239	0.8863
Transformer-based				
UNETR	88M	42.1G	0.8206	0.8610
SwinUNETR	25M	27.9G	0.8431	0.8976
nnFormer	60M	50.5G	0.8412	0.8897
Mamba-based				
U-Mamba_Bot	63M	45.7G	0.8567	0.9182
U-Mamba_Enc	67M	49.8G	0.8533	0.9111
VM-UNet-V2	18M	7.56G	0.8624	<u>0.9231</u>
w/o ImageNet-based pretraining				
Swin-UMamba	60M	68.0G	0.7913	0.8376
Swin-UMamba†	28M	18.9G	0.8025	0.8463
w/ ImageNet-based pretraining				
Swin-UMamba	60M	68.0G	**0.8639**	0.8979
Swin-UMamba†	28M	18.9G	<u>0.8626</u>	**0.9247**

mizes the interference of shallow background noise to improve the feature expression and solves the problem of excessive variation in lesion size and type.

6.3 Conv-SSM

Conv-SSM is a simple two-branch module.As show in Fig.3.First, the module input is split into two sub-inputs with the same size using a channel split operation. Next,the two sub-inputs are fed to two branching (Conv-branch and SSM-branch) modules. In the Conv-branch, successive convolutional layers are simply used to model the local features of the input. In the SSM branch, the input is first processed using the layer Normalization layer. Notably, the input is divided into two branches after layer normalization. In the first branch of the SSM, the input passes through the linear layer and the activation function. The second branch is the familiar VSS module.Therefore we replace the VSS module in stage1-4 of the encoder with this module and the number of layers are $[2, 2, 4, 2]$.

Table 5. Ablation studies on Multiplicative Connection and Conv-SSM of Swin-UMamba†

Skip Connection	Conv-SSM	Kvasir-SEG		Thyroid	
		mloU(%)↑	DSC(%)↑	mloU(%)↑	DSC(%)↑
simple connection	TRUE	82.93	90.67	75.02	86.29
	FALSE	82.59	90.47	74.45	86.26
addition	TRUE	**85.23**	**92.03**	**75.21**	**86.64**
	FALSE	84.15	91.39	<u>75.15</u>	86.27
multiplication	TRUE	82.90	90.64	75.03	<u>86.31</u>
	FALSE	79.57	88.63	74.83	86.12
concatenation	TRUE	<u>82.97</u>	<u>90.88</u>	74.99	86.12
	FALSE	79.53	88.61	74.74	86.06

6.4 Ablation studies

In this section, we conduct ablation experiments on the initialization of Multiplicative Connection and the Conv-SSM operation of Encoder and Decoder using the Kvasir-Seg [16] and Thyroid datasets. As stated in the VMamba [22] paper, the depth of the Encoder and the number of channels in the feature map determine the scale of VMamba. In this paper, only uses the pre-trained weights of VMamba on ImageNet-1k for the Encoder part. Therefore, when conducting the model scale ablation experiment in this study, we only vary the depth of the Encoder, as shown in the Table 5. when the depth of the Encoder is set to $[2, 2, 4, 2]$ the segmentation evaluation metrics are relatively better. Therefore, when using Swin-UMamba†, there is no need to choose a particularly large depth. In most cases where the Conv-SSM module is used, the segmentation evaluation metrics are relatively better.

7 Conclusion and future work

This study aims to reveal the impact of ImageNet-based pretraining for Mamba-based models in 2D medical image segmentation. We proposed a novel Mamba-based model, Swin-UMamba, and its variant, Swin-UMamba†, both capable of leveraging the power of pretrained models for segmentation tasks. Our experiments on various medical image segmentation datasets suggest that ImageNet-based pretraining for Mamba-based models offers several advantages, including superior segmentation accuracy, stable convergence, mitigation of overfitting issues, data efficiency, and lower computational resource consumption. We believe that our findings highlight the importance of pretraining in enhancing the performance and efficiency of Mamba-based models in vision tasks.In the future, I will conduct more in-depth research to improve the contribution of the Mamba structure to medical image segmentation.

References

[1] Max Allan, Alex Shvets, Thomas Kurmann, Zichen Zhang, Rahul Duggal, Yun-Hsuan Su, Nicola Rieke, Iro Laina, Niveditha Kalavakonda, Sebastian Bodenstedt, et al. 2017 robotic instrument segmentation

challenge. *arXiv preprint arXiv:1902.06426*, 2019.

[2] Wenjia Bai, Hideaki Suzuki, Jian Huang, Catherine Francis, Shuo Wang, Giacomo Tarroni, Florian Guitton, Nay Aung, Kenneth Fung, Steffen E Petersen, et al. A population-based phenome-wide association study of cardiac and aortic structure and function. *Nature medicine*, 26(10):1654–1662, 2020.

[3] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In Leonid Karlinsky, Tomer Michaeli, and Ko Nishino, editors, *Computer Vision – ECCV 2022 Workshops*, Lecture Notes in Computer Science, pages 205–218. Springer Nature Switzerland.

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

[5] Albert Gu, Karan Goel, and Christopher Re. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*, 2021.

[6] Albert Gu, Isys Johnson, Karan Goel, Khaled Saab, Tri Dao, Atri Rudra, and Christopher Ré. Combining recurrent, convolutional, and continuous-time models with linear state space layers. *Advances in neural information processing systems*, 34:572–585, 2021.

[7] Jiansen Guo, Hong-Yu Zhou, Liansheng Wang, and Yizhou Yu. UNet-2022: Exploring dynamics in non-isomorphic architecture. In Ruidan Su, Yudong Zhang, Han Liu, and Alejandro F Frangi, editors, *Medical Imaging and Computer-Aided Diagnosis*, Lecture Notes in Electrical Engineering, pages 465–476. Springer Nature.

[8] Tao Guo, Yinuo Wang, and Cai Meng. Mambamorph: a mamba-based backbone with contrastive feature learning for deformable mr-ct registration. *arXiv preprint arXiv:2401.13934*, 2024.

[9] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):87–110, 2022.

[10] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI Brainlesion Workshop*, pages 272–284. Springer, 2021.

[11] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 574–584, 2022.

[12] Ali Hatamizadeh, Hongxu Yin, Greg Heinrich, Jan Kautz, and Pavlo Molchanov. Global context vision transformers. In *International Conference on Machine Learning*, pages 12633–12646. PMLR, 2023.

[13] Tao Huang, Xiaohuan Pei, Shan You, Fei Wang, Chen Qian, and Chang Xu. Localmamba: Visual state space model with windowed selective scan. *ArXiv*, abs/2403.09338, 2024.

[14] Juan Eugenio Iglesias, Jean C. Augustinack, Khoa Nguyen, Christopher M. Player, Allison Player, Michelle Wright, Nicole Roy, Matthew P. Frosch, Ann C. McKee, Lawrence L. Wald, Bruce Fischl, and Koen Van Leemput. A computational atlas of the hippocampal formation using ex vivo, ultra-high resolution mri: Application to adaptive segmentation of in vivo mri. *NeuroImage*, 115:117–137, 2015.

[15] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.

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

[17] Leo Joskowicz, D Cohen, N Caplan, and Jacob Sosna. Inter-observer variability of manual contour delineation of structures in ct. *European radiology*, 29:1391–1399, 2019.

[18] Alain Jungo, Raphael Meier, Ekin Ermis, Marcela Blatti-Moreno, Evelyn Herrmann, Roland Wiest, and Mauricio Reyes. On the effect of inter-observer variability for a reliable estimation of uncertainty of medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I*, pages 682–690. Springer, 2018.

[19] Mahendra Khened, Varghese Alex Kollerathu, and Ganapathy Krishnamurthi. Fully convolutional multi-scale residual densenets for cardiac segmentation and automated cardiac diagnosis using ensemble of classifiers. *Medical image analysis*, 51:21–45, 2019.

[20] Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-supervised nets. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, pages 562–570. PMLR. ISSN: 1938-7228.

[21] Tianyang Lin, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. A survey of transformers. *AI Open*, 2022.

[22] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, and Yunfan Liu. Vmamba: Visual state space model. *arXiv preprint arXiv:2401.10166*, 2024.

[23] Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. Understanding the effective receptive field in deep convolutional neural networks. *Advances in neural information processing systems*, 29, 2016.

[24] Jun Ma, Feifei Li, and Bo Wang. U-mamba: Enhancing long-range dependency for biomedical image segmentation. *arXiv preprint arXiv:2401.04722*, 2024.

[25] Jun Ma, Ronald Xie, Shamini Ayyadhury, Cheng Ge, Anubha Gupta, Ritu Gupta, Song Gu, Yao Zhang, Gihun Lee, Joonkee Kim, et al. The multi-modality cell segmentation challenge: towards universal solutions. *arXiv preprint arXiv:2308.05864*, 2023.

[26] Jun Ma, Yao Zhang, Song Gu, Cheng Ge, Shihao Ma, Adamo Young, Cheng Zhu, Kangkang Meng, Xin Yang, Ziyan Huang, et al. Unleashing the strengths of unlabeled data in pan-cancer abdominal organ quantification: the flare22 challenge. *arXiv preprint arXiv:2308.05862*, 2023.

[27] Xueyan Mei, Hao-Chih Lee, Kai-yue Diao, Mingqian Huang, Bin Lin, Chenyu Liu, Zongyu Xie, Yixuan Ma, Philip M Robson, Michael Chung, et al. Artificial intelligence–enabled rapid diagnosis of patients with covid-19. 26(8):1224–1228, 2020.

[28] Andriy 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*, pages 311–320. Springer, 2019.

[29] Xiaohuan Pei, Tao Huang, and Chang Xu. Efficientvmamba: Atrous selective scan for light weight visual mamba. *ArXiv*, abs/2403.09977, 2024.

[30] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in Neural Information Processing Systems*, 34:12116–12128, 2021.

[31] Olaf Ronneberger, Philipp Fischer, and Thomas 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*, pages 234–241. Springer, 2015.

[32] Ashish Sinha and Jose Dolz. Multi-scale self-guided attention for medical image segmentation. *IEEE Journal of Biomedical and Health Informatics*, 25(1):121–130, 2021.

[33] Hui Sun, Cheng Li, Boqiang Liu, Zaiyi Liu, Meiyun Wang, Hairong Zheng, David Dagan Feng, and Shanshan Wang. Aunet: attention-guided dense-upsampling networks for breast mass segmentation in whole mammograms. *Physics in Medicine & Biology*, 65(5):055005, feb 2020.

[34] Hao Tang, Xuming Chen, Yang Liu, Zhipeng Lu, Junhua You, Mingzhou Yang, Shengyu Yao, Guoqi Zhao, Yi Xu, Tingfeng Chen, et al. Clinically applicable deep learning framework for organs at risk delineation in ct images. *Nature Machine Intelligence*, 1(10):480–491, 2019.

[35] Hao Tang, Chupeng Zhang, and Xiaohui Xie. Automatic pulmonary lobe segmentation using deep learning. In *2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019)*, pages 1225–1228. IEEE, 2019.

[36] Ekin Tiu, Ellie Talius, Pujan Patel, Curtis P Langlotz, Andrew Y Ng, and Pranav Rajpurkar. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature Biomedical Engineering*, 6(12):1399–1406, 2022.

[37] Shanshan Wang, Cheng Li, Rongpin Wang, Zaiyi Liu, Meiyun Wang, Hongna Tan, Yaping Wu, Xinfeng Liu, Hui Sun, Rui Yang, et al. Annotation-efficient deep learning for automatic medical image segmentation. *Nature communications*, 12(1):5915, 2021.

[38] Weiwei Wang, Yuanshen Zhao, Lianghong Teng, Jing Yan, Yang Guo, Yuning Qiu, Yuchen Ji, Bin Yu, Dongling Pei, Wenchao Duan, et al. Neuropathologist-level integrated classification of adult-type diffuse gliomas using deep learning from whole-slide pathological images. *Nature Communications*, 14(1):6359, 2023.

[39] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: convolutional block attention module. *CoRR*, abs/1807.06521, 2018.

[40] Zhaohu Xing, Tian Ye, Yijun Yang, Guang Liu, and Lei Zhu. Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation. *arXiv preprint arXiv:2401.13560*, 2024.

[41] Chenhongyi Yang, Zehui Chen, Miguel Espinosa, Linus Ericsson, Zhenyu Wang, Jiaming Liu, and Elliot J. Crowley. Plainmamba: Improving non-hierarchical mamba in visual recognition. *ArXiv*, abs/2403.17695, 2024.

[42] Mingya Zhang, Yue Yu, Limei Gu, Tingsheng Lin, and Xianping Tao. Vm-unet-v2 rethinking vision mamba unet for medical image segmentation. *ArXiv*, abs/2403.09157, 2024.

[43] Hong-Yu Zhou, Jiansen Guo, Yinghao Zhang, Xiaoguang Han, Lequan Yu, Liansheng Wang, and Yizhou Yu. nnformer: Volumetric medical image segmentation via a 3d transformer. *IEEE Transactions on Image Processing*, 32:4036–4045, 2023.

[44] Yongjin Zhou, Weijian Huang, Pei Dong, Yong Xia, and Shanshan Wang. D-unet: A dimension-fusion u shape network for chronic stroke lesion segmentation. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 18(3):940–950, 2021.

[45] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*, 2024.