VISION STATE SPACE DUALITY FOR MEDICAL IMAGE SEGMENTATION: ENHANCING PRECISION THROUGH NON-CAUSAL MODELING

Anonymous authors
Paper under double-blind review

ABSTRACT

In medical image analysis, Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) have set significant benchmarks. However, CNNs exhibit limitations in long-range modeling capabilities, whereas Transformers are hampered by their quadratic computational complexity. Recently, State Space Models (SSMs) have gained prominence in vision tasks as they offer linear computational complexity. State Space Duality (SSD), an improved variant of SSMs, was introduced in Mamba2 to enhance model performance and efficiency. Inspired by this, we have tailored the Vision State Space Duality (VSSD) model for medical image segmentation tasks by integrating it within a UNet-like architecture, which is renowned for its effectiveness in the field. Our modified model, named VSSD-UNet, employs skip connections to preserve spatial information and utilizes a series of VSSD blocks for feature extraction. In addition, VSSD-UNet employs a hybrid structure of VSSD and self-attention in the decoder part, ensuring that both local details and global contexts are captured. Finally, we conducted comparative and ablation experiments on two public lesion segmentation datasets: ISIC2017 and ISIC2018. The results show that VSSD-UNet outperforms several types of UNet in medical image segmentation under the same hyper-parameter setting. Our code will be released soon.

1 INTRODUCTION

In the medical imaging domain, segmentation is vital for advancing clinical diagnostics, informing treatment strategies, and enabling a deeper understanding of anatomical and pathological characteristics. The ability to accurately segment images into distinct regions corresponding to different tissues, organs, or abnormalities is crucial for a range of medical applications, from oncology to neurology. The integration of deep learning techniques, particularly Convolutional Neural Networks (CNNs) LeCun & Bengio (1998), has marked a significant leap forward in the accuracy and efficiency of medical image segmentation. CNNs have demonstrated their prowess in capturing local features and spatial hierarchies, leading to significant improvements in segmentation tasks. Furthermore, the advent of Vision Transformers (ViTs) Dosovitskiy et al. (2020) has introduced a new paradigm, harnessing self-attention mechanisms to capture global dependencies and long-range interactions within images, which is particularly beneficial for understanding the complex patterns present in medical imaging data.

Despite the remarkable achievements of CNNs and ViTs, there are inherent challenges that limit their effectiveness in medical image segmentation. CNNs, while excellent at capturing local features, often struggle to model long-range spatial dependencies that are essential for accurately segmenting large or complex anatomical structures. This limitation can result in segmentation inaccuracies. On the other hand, ViTs, despite their ability to provide a more comprehensive view of the image, are hindered by their quadratic computational complexity. This complexity becomes a significant bottleneck when scaling to the high-resolution images that are common in medical imaging, where detailed and precise segmentation is critical for clinical decision-making. The computational demands of ViTs can be prohibitive, particularly in time-sensitive clinical settings where real-time processing is desirable. Consequently, how to efficiently enhance the long-range dependency remains an open question.

Recently, structured state-space models (SSMs) Gu (2023); Gu et al. (2021b) inspired by classical state-space models have garnered significant interest for their computational efficiency and excellent performance in modeling long-range dependencies. Notably, Mamba, a state-of-the-art selective structured state-space model, addresses the inherent limitations of previous SSMs. It successfully demonstrates efficiency and effectiveness in long sequence modeling and achieves cutting-edge performance in continuous long sequence data analysis, such as in natural language processing and genomic analysis. Internally, Mamba integrates time-varying parameters and employs a novel hardware-aware algorithm for highly efficient training and inference, thereby avoiding the high quadratic computational complexity caused by self-attention mechanisms. Recent studies have tentatively delved into the effectiveness of SSMs across a range of visual tasks, including ImageNet classification Zhu et al. (2024b); Liu et al. (2024b), classifying remote sensing images Chen et al. (2024a), image dehazing Zheng & Wu (2024), analyzing point clouds Liang et al. (2024), and segmenting medical images Ruan & Xiang (2024b); Ma et al. (2024). This inspires us to explore the potential of using Mamba blocks to enhance long-range dependency modeling in medical image segmentation tasks.

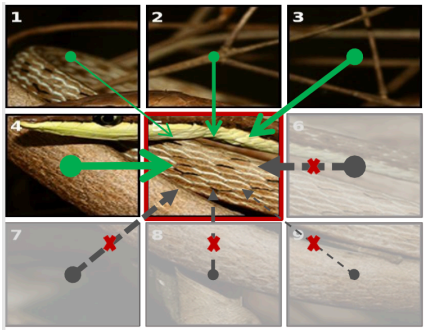

Figure 1: Two challenges when applying SSM/SSD to image data.

However, there exists a major concern regarding the application of SSD/SSMs in vision tasks, where the image data is naturally non-causal while SSD/SSMs have inherent causal properties. While another concern is flattening 2D feature maps into 1D sequences disrupts the inherent structural relationships among patches. We provide an illustration in Fig. 1 to facilitate a more intuitive understanding of these two concerns. In this example, the central token within the flattened 1D sequences is restricted to accessing only previous tokens, unable to integrate information from subsequent tokens. Additionally, the token 1, which is adjacent to the central token in the 2D space, becomes distantly positioned in the 1D sequence, disrupting the natural structural relationships.

In this work, we introduce VSSD-UNet, a model that integrates Vision State Space Duality (VSSD) within a UNet-like architecture, a framework known for its effectiveness in medical image segmentation. VSSD-UNet leverages the non-causal properties of VSSD to capture both local and global features within medical images effectively. It employs skip connections to preserve spatial hierarchies and integrates VSSD blocks for feature extraction, ensuring that the model can extract fine details while maintaining a broader contextual understanding. In addition, we employ a hybrid structure of VSSD and self-attention in the decoder part. Building on these techniques, our model provides superior segmentation performance while maintaining computational efficiency, addressing the limitations of existing models. The VSSD-UNet model represents a significant advancement in the field of medical image segmentation, offering a potential solution to the challenges faced by current deep learning models.

In summary, this paper presents several key contributions to the field of medical image segmentation. Firstly, we introduce VSSD-UNet, a novel model tailored for medical image segmentation that combines the strengths of VSSD and UNet architectures. Secondly, we provide a comprehensive evaluation of VSSD-UNet against existing segmentation models on standardized medical imaging datasets, demonstrating its superior performance. Finally, we offer an in-depth analysis of the model's computational efficiency and accuracy, highlighting its potential for real-world clinical applications.

2 RELATED WORKS

2.1 MEDICAL IMAGE SEGMENTATION

Medical image segmentation is a critical task that entails the pixel-wise classification of various anatomical structures, such as lesions, tumors, or organs, across diverse imaging modalities like endoscopy, MRI, or CT scans Chen et al. (2021). U-shaped networks Ronneberger et al. (2015); Oktay et al. (2018); Zhou et al. (2018); Huang et al. (2020); Lou et al. (2021); Ibtehaz & Kihara (2023); Chen et al. (2022); et al. (2021) have become particularly popular due to their straightforward yet effective encoder-decoder architecture. The UNet Ronneberger et al. (2015), a seminal work in this area, employs skip connections to effectively fuse features at different resolution levels. This design has been further refined by UNet++ Zhou et al. (2018), which introduces nested encoder-decoder pathways with dense skip connections, and UNet 3+ Huang et al. (2020), which presents comprehensive skip pathways for full-scale feature integration. DC-UNet Lou et al. (2021) pushes the envelope by integrating a multi-resolution convolution scheme and residual paths into its skip connections. The DeepLab series, including DeepLabv3 Chen et al. (2017) and DeepLabv3+ Chen et al. (2018), leverages atrous convolutions and spatial pyramid pooling to effectively handle multi-scale information. SegNet Badrinarayanan et al. (2017) utilizes pooling indices for feature map upsampling, ensuring boundary detail preservation. The nnU-Net et al. (2021) automatically tailors hyperparameters based on dataset-specific characteristics and employs standard 2D and 3D UNets. These U-shaped models have collectively set a high benchmark in the field of medical image segmentation.

In recent years, vision transformers have emerged as a powerful force in medical image segmentation, capable of capturing pixel relationships at a global scale Cao et al. (2021); Chen et al. (2021); Dong et al. (2021); Rahman & Marculescu (2023a;b); Wang et al. (2022a); Zhang et al. (2021); Xie et al. (2021). TransUNet Chen et al. (2021) represents a novel fusion of CNNs for local feature extraction and transformers for global context understanding, thereby enhancing the capture of both local and global features. Swin-Unet Cao et al. (2021) further extends this concept by integrating Swin Transformer blocks Liu et al. (2021) into a U-shaped model for both encoding and decoding processes. Drawing on these ideas, MERIT Rahman & Marculescu (2023b) introduces a multi-scale hierarchical transformer that employs self-attention across various window sizes, thereby enhancing the model's ability to capture multi-scale features that are crucial for medical image segmentation. These advances demonstrate the potential of transformers to significantly impact the field of medical image analysis.

2.2 VISION TRANSFORMERS

The emergence of Vision Transformers (ViTs) Dosovitskiy et al. (2020); Liu et al. (2021); Wang et al. (2021); Dong et al. (2022); Touvron et al. (2021) has reinvigorated the computer vision domain, a field that was once predominantly governed by Convolutional Neural Networks (CNNs) Krizhevsky et al. (2012); Simonyan & Zisserman (2014); He et al. (2016); Xie et al. (2017); Huang et al. (2019); Howard et al. (2017); Tan & Le (2019); Liu et al. (2022b). However, the self-attention mechanism in ViTs, which entails quadratic computational complexity, presents considerable difficulties when dealing with high-resolution imagery, necessitating substantial computational resources. To surmount this challenge, various strategies have been introduced, such as hierarchical model structures Liu et al. (2021; 2022a); Dong et al. (2022); Wang et al. (2021; 2022c); Han et al. (2021), windowed attention techniques Liu et al. (2021); Hassani et al. (2023); Tu et al. (2022); Zhu et al. (2023), and alternative forms of self-attention mechanisms Wang et al. (2022b); Xia et al. (2023); Yu et al. (2022). Additionally, linear attention methods Katharopoulos et al. (2020); Choromanski et al. (2020); Qin et al. (2022); Han et al. (2024) have managed to scale down the computational complexity to a linear rate by reordering the self-attention's query, key, and value operations. Yet, despite these improvements, the efficacy of linear attention still lags behind that of the quadratic self-attention Vaswani et al. (2017) and its derivatives Hassani et al. (2023); Fan et al. (2024); Zhu et al. (2023).

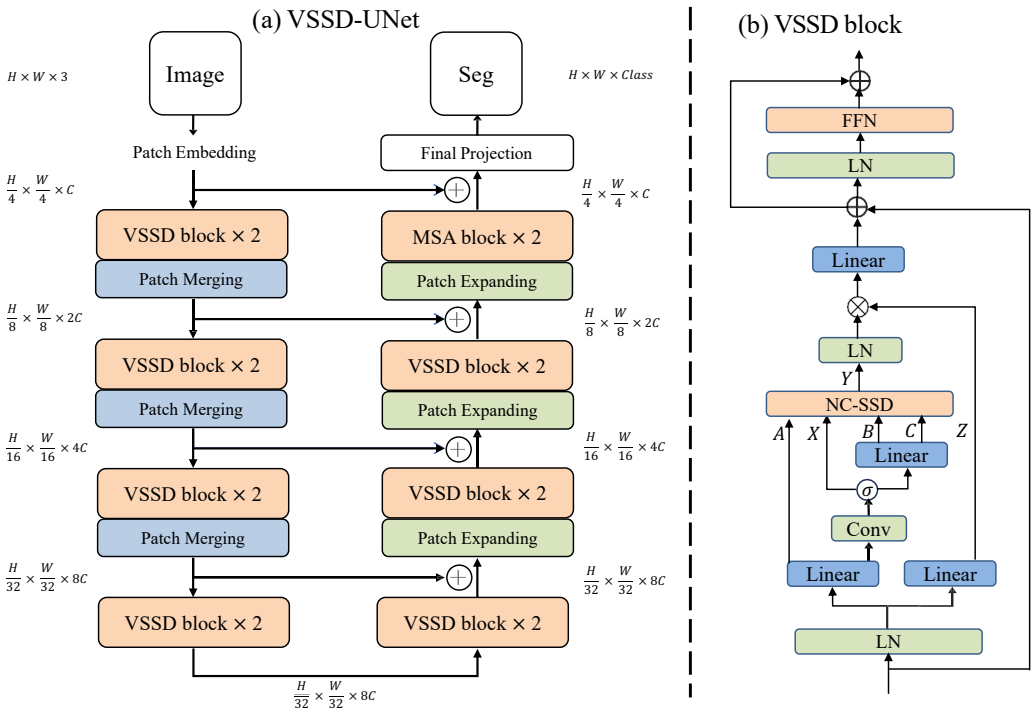

Figure 2: The architecture of VSSD-UNet, which is composed of encoder, bottleneck, decoder and skip connections. The encoder, bottleneck and decoder are all constructed based on Visual Mamba block.

2.3 State Space Models

State Space Models (SSMs) Gu et al. (2020; 2021c;a); Smith et al. (2022); Fu et al. (2022); Gu & Dao (2023) have garnered significant research interest due to their expansive receptive fields and linear computational complexity. A notable SSM, Mamba Gu & Dao (2023), has introduced the S6 block, which has demonstrated comparable or superior performance to transformers in Natural Language Processing (NLP) tasks. This has spurred further explorations Pei et al. (2024); Huang et al. (2024); Du et al. (2024); Yang et al. (2024a); Chen et al. (2024b); Li et al. (2024); Yang et al. (2024b); Ruan & Xiang (2024b) into adapting the S6 block for visual tasks, with studies showing it can compete with both CNNs and Vision Transformer (ViT) models. However, a key challenge in developing Mamba-based models for computer vision lies in aligning the model's causal nature with the non-causal aspects of image data. A common strategy to overcome this is to flatten 2D feature maps into 1D sequences using various scanning methods before processing them through the S6 block. These diverse scanning approaches have been proven effective across multiple studies Zhu et al. (2024a); Liu et al. (2024a); Huang et al. (2024); Pei et al. (2024); Shi et al. (2024). Recently, Mamba2 Dao & Gu (2024) has identified a close relationship between SSMs and structured masked attention, establishing them as dual concepts and introducing State Space Duality (SSD). We extend this work to show that SSD can be adapted into a non-causal model through a simple transformation, eliminating the need for specific scanning routes.

3 Methods

3.1 Architecture Overview

The architecture of the proposed VSSD-UNet is outlined in Figure 2. The input images are first divided into patches similar to ViT and VMamba and transformed into sequences. An initial linear embedding layer adjusts feature dimensions to an arbitrary size denoted as C. These patch tokens are processed by several VSSD blocks and patch merging layers to generate hierarchical features.

The patch merging layers are responsible for reducing the image scale and boosting the feature dimensions, whereas the VSSD blocks concentrate on learning feature representations. The encoder produces outputs with resolutions of $\frac{H}{4} \times \frac{W}{4} \times C$, $\frac{H}{8} \times \frac{W}{8} \times 2C$, $\frac{H}{16} \times \frac{W}{16} \times 4C$, and $\frac{H}{32} \times \frac{W}{32} \times 8C$, respectively. The decoder includes VSSD and patch expanding layers to restore the feature size. It replaces the NC-SSD block with self-attention module exclusively in the last stage, recovering spatial details lost during downsampling through skip connections. Both the encoder and decoder use two VSSD blocks each. The details of VSS block, patch merging of encoder, and patch expanding of decoder is discussed in the following subsections.

3.2 VSSD BLOCK

3.2.1 PRELIMINARIES OF MAMBA

The SSM is a concept derived from modern control theory's linear time-invariant system which maps the continuous stimulation $x(t) \in \mathcal{R}$ to response $y(t) \in \mathcal{R}$. This process can be formulated through the subsequent linear ordinary differential equation (ODE),

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

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

Structured State Space Sequence Model (S4) and Mamba discretize this continuous system to make it more suitable for deep learning scenarios. Specifically, they introduce a timescale parameter $\mathbf{\Delta}$ and transform $\mathbf{A}$ and $\mathbf{B}$ into discrete parameters $\overline{\mathbf{A}}$ and $\overline{\mathbf{B}}$ using a fixed discretization rule. Typically, the zero-order hold (ZOH) is employed as the discretization rule and can be defined as follows:

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

After discretization, Eq. 1 can be rewritten as,

$$\begin{aligned} h_k &= \bar{\mathbf{A}}h_{k-1} + \bar{\mathbf{B}}x_k \\ y_k &= \mathbf{C}h_k \end{aligned} \tag{3}$$

At last, the output can be calculated in a convolution representation, as follows,

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

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

3.2.2 VSSD BLOCK

State Space Duality (SSD) is an enhancement over traditional State Space Models (SSMs), offering improved performance and efficiency in processing sequence data. However, SSD inherently operates in a causal manner, which limits its applicability to non-causal vision tasks where information from future steps is just as relevant as past steps. To address this, we utilize Non-Causal SSD (NC-SSD), which modifies the role of the state transition matrix $\mathbf{A}$ to enable non-causal processing.

In the traditional SSD framework, the model updates the hidden state $h(t)$ and computes the output $y(t)$ as follows:

$$h(t) = A_t h(t-1) + \mathbf{B}_t x(t), y(t) = \mathbf{C}_t h(t). \tag{5}$$

where A is the state transition matrix, B is the input matrix. C is the output matrix.

In NC-SSD, it transforms the role of A from a matrix to a scalar to facilitate non-causal processing. The key equation becomes:

$$h(t) = h(t-1) + \frac{1}{A} \cdot \mathbf{B}_t \cdot x(t) \tag{6}$$

This equation shows that the current state $h(t)$ is influenced by the previous state $h(t-1)$, the input matrix B_t, and the current input $x(t)$, with the influence weighted by $\frac{1}{A}$.

To fully achieve non-causality, NC-SSD employs bidirectional scanning, which involves processing the data in both forward and reverse sequences. The combined hidden state H from bidirectional scanning is given by:

$$\mathbf{H}_i = \sum_{j=1}^{i} \frac{1}{A_j} \mathbf{Z}_j + \sum_{j=-L}^{-1} \frac{1}{A_{i+j}} \mathbf{Z}_{i+j}. \tag{7}$$

where $Z_j = B_j \cdot x(j)$ is the transformed input for the j-th token in the sequence.

By integrating the results from both directions, it can be ensured that each token has access to global information, not just the tokens before it in the sequence. Assuming each token's contribution can be considered independently, the hidden state H can be simplified to:

$$\mathbf{H} = \sum_{j=1}^{L} \frac{1}{A_j} \mathbf{Z}_j. \tag{8}$$

This equation shows that all tokens contribute equally to the hidden state H, effectively removing the causal constraint and allowing the model to process information in a non-linear sequence.

To implement VSSD efficiently, we revise the tensor contraction algorithm to:

1.Expand the input $\mathbf{X}$ using $\mathbf{B}$:

$$\mathbf{Z} = \text{contract}(\text{LD}, \text{LN} \rightarrow \text{LND})(\mathbf{X}, \mathbf{B}) \tag{9}$$

2.Unroll scalar SSM recurrences to create a global hidden state $\mathbf{H}$:

$$\mathbf{H} = \text{contract}(\text{LL}, \text{LDN} \rightarrow \text{ND})(\mathbf{M}, \mathbf{Z}) \tag{10}$$

3.Contract the hidden state $\mathbf{H}$ with $\mathbf{C}$ to produce the output $\mathbf{Y}$:

$$\mathbf{Y} = \text{contract}(\text{LN}, \text{ND} \rightarrow \text{LD})(\mathbf{C}, \mathbf{H}). \tag{11}$$

These steps replace the traditional recurrent computations with parallelizable operations, significantly enhancing training and inference speeds.

In summary, VSSD allows for more flexible processing of sequence data by removing the constraints of causality, leading to improved performance and efficiency in vision tasks.

3.3 Encoder

In the encoder of the VSSD-UNet, C-dimensional tokenized inputs pass through two sequential VSSD blocks to extract features without changing their size or dimension. The patch merging layer is utilized for downsampling in the encoder of VSSD-UNet, reduces the token count by $\frac{1}{2}$ while doubling feature dimensions by $2\times$, by segmenting inputs into quadrants by $\frac{1}{4}$, concatenating them, and then normalizing dimensions through a layernorm each time.

3.4 Decoder

The decoder also uses two VSSD blocks in succession to reconstruction the features. Instead of merging layers, it uses patch expansion layers to upscale deep features. This process effectively halves feature dimensions by $\frac{1}{2}$ while enhancing image resolution ($2\times$ upscaling). It works by an initial layer that doubles feature dimensions before reorganizing and reducing them for resolution enhancement.

Moreove, Mamba2 demonstrates that integrating SSD with standard Multi-head Self Attention (MSA) yields additional improvements. In a similar way, our model incorporates self-attention. However, unlike Mamba2, which uniformly intersperses self-attention throughout the network, we strategically replace the VSSD block with self-attention module exclusively in the last stage. This modification leverages the robust capabilities of self-attention in processing high-level features, as evidenced by prior works Lin et al. (2023); Ren et al. (2023); Fan et al. (2024) in vision tasks.

Table 1: Comparative experimental results on the ISIC2017 dataset. The best results are highlighted in bold fonts. " ↑ "and " ↓ " indicate that larger or smaller is better.

Model	Year	mIoU(%)↑	DSC(%)↑	Acc(%)↑	Spe(%)↑	Sen(%)↑
UNet	2015	75.97	86.34	95.53	97.75	84.47
R2UNet	2018	73.43	84.68	95.08	97.86	81.25
UNet++	2019	77.85	87.55	95.91	97.94	85.82
R2AttUNet	2021	75.07	85.76	95.24	97.17	85.63
SwinUnet	2022	67.93	80.90	93.75	96.69	79.11
MISSFormer	2022	75.84	86.26	95.62	**98.34**	82.09
MALUNet	2022	74.69	85.51	95.15	97.10	85.46
H2Former	2023	76.27	86.54	95.58	97.72	84.90
EGE-UNet	2023	76.50	86.68	95.65	97.88	84.55
MHorunet	2024	78.16	87.73	95.77	97.15	85.99
VMUNet	2024	77.24	87.16	95.78	97.82	85.62
VMUNet v2	2024	75.25	85.88	95.34	97.47	84.71
H-vmunet	2024	78.18	87.75	95.82	97.12	85.72
ULVM-UNet	2024	78.13	87.72	95.78	97.59	83.61
VSSD-UNet	-	**78.30**	**87.83**	**96.00**	97.99	**86.14**

3.5 Bottleneck & Skip Connetions

In the VSSD-UNet bottleneck, we use two VSSD blocks to process the features. At each stage of the encoder and decoder, skip connections are utilized to blend features from multiple scales with the upscaled image outputs. This process merges information from both shallow and deep layers, which enhances the spatial details in the segmentation results. After that, a linear layer is applied to keep the combined features' dimensions the same as the upsampled resolution, ensuring consistency with the upscaled resolution.

4 Experiments

4.1 Data Sets

In this section, we conducted extensive experiments using two prominent lesion segmentation datasets that are publicly available: the International Skin Imaging Collaboration's 2017 and 2018 challenge datasets (ISIC2017 and ISIC2018), to train and evaluate the proposed model. These datasets consist of a substantial collection of dermoscopic images, with ISIC2017 containing 2,150 images and ISIC2018 containing 2,694 images, all of which are accompanied by segmentation mask labels. Following the methods employed in prior research Ruan et al. (2022; 2023), we segmented these datasets into training and test subsets at a ratio of 7:3. To elaborate, the ISIC2017 dataset was divided into a training set of 1,500 images and a test set of 650 images. Similarly, the ISIC2018 dataset was split into a training set comprising 1,886 images and a test set comprising 808 images. This approach allowed us to train and assess the performance of our proposed model across a broad spectrum of lesion segmentation tasks.

4.2 Implementation Details

We implemented our VSSD-UNet using PyTorch 1.13 and trained it on an A100-PCIE-40G GPU with 24 GB of memory for 300 epochs with a batch size of 32. The input images are uniformly resized to 224 × 224. We employed data augmentation techniques such as random flipping and random rotation to prevent overfitting. We used the AdamW optimizer with an initial learning rate of 1×10^{-3}, β_1 of 0.9, β_2 of 0.999, and weight decay of 1×10^{-4}. Additionally, we applied a cosine annealing learning rate decay strategy and an early stopping mechanism. To ensure reproducibility, we set the random seed to 42.

Table 2: Comparative experimental results on the ISIC2018 dataset. The best results are highlighted in bold fonts. " ↑ "and " ↓ " indicate that larger or smaller is better.

Model	Year	mIoU(%)↑	DSC(%)↑	Acc(%)↑	Spe(%)↑	Sen(%)↑
UNet	2015	77.22	87.15	93.86	96.56	85.47
R2UNet	2018	71.74	83.55	92.36	96.41	79.74
UNet++	2019	79.14	88.36	94.40	96.69	87.28
R2AttUNet	2021	75.24	85.87	93.15	95.62	85.47
SwinUnet	2022	74.26	85.23	92.87	95.55	84.54
MISSFormer	2022	77.94	87.60	94.11	96.89	85.48
MALUNet	2022	78.09	87.70	94.07	96.41	86.80
H2Former	2023	77.33	87.21	93.89	96.57	85.56
EGE-UNet	2023	78.90	88.20	94.25	96.17	88.29
MHorunet	2024	79.40	88.52	94.47	96.70	87.55
VMUNet	2024	74.14	85.15	93.03	96.54	82.10
VMUNet v2	2024	78.25	87.80	94.09	96.25	87.38
H-vmunet	2024	79.41	88.52	94.37	96.03	89.20
ULVM-UNet	2024	78.74	88.10	94.29	96.68	86.85
VSSD-UNet	-	**80.65**	**89.29**	**94.73**	**97.24**	**90.18**

4.3 Evaluation Metrics

We used five metrics to assess the quality of the segmentations: Mean Intersection over Union (mIoU), Dice Similarity Score (DSC), Accuracy (Acc), Sensitivity (Sen), and Specificity (Spe). The mathematical formulations for these metrics are summarized as follows:

$$mIoU = \frac{TP}{TP + FP + FN} \tag{12}$$

$$DSC = \frac{2TP}{2TP + FP + FN} \tag{13}$$

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{14}$$

$$Sen = \frac{TP}{TP + FN} \tag{15}$$

$$Spe = \frac{TN}{TN + FP} \tag{16}$$

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

4.4 Comparison Results

To validate the effectiveness of our approach, we compared VSSD-UNet with other state-of-the-art methods. Specifically, this comparison includes UNet Ronneberger et al. (2015), R2UNet Alom et al. (2018), UNet++ Zhou et al. (2019), R2AttUNet Zuo et al. (2021), SwinUnet Aghdam et al. (2023), MISSFormer Huang et al. (2023), MALUNet Ruan et al. (2022), H2Former He et al. (2023), EGEUNet Ruan et al. (2023), MHorunet Wu et al. (2024a), VMUNet Ruan & Xiang (2024a), VMUNet v2 Zhang et al. (2024), H-vmunet Wu et al. (2024b), UltraLight-VM-UNet Wu et al. (2024c), and VSSD-NUet. 1 and 2 show the comparative results on the ISIC2017 and ISIC2018 datasets, respectively. Our proposed VSSD-UNet outperformed the other models in terms of mIoU, DSC, Acc, Spe, and Sen metrics.

4.5 Ablations

To validate the effectiveness of the proposed modules, we conducted detailed ablation experiments on the VSSD-UNet model. Using the SSD block as the token mixer and patchified downsamplers (e.g. convolution with 4×4 kernel and stride of 4 in stem) following Swin Liu et al. (2021) and vallina VMamaba Liu et al. (2024a), we established the baseline configuration, detailed in the first

row of Tab. 3. Our ablation study was conducted on an A100-PCIE-40G GPU with a batch size of 128 using FP16 precision.

Table 3: **Ablation study of VSSD-UNet.** Our VSSD consistently outperforms vallina SSD and Bi-SSD in terms of accuracy and efficiency.

Op. Type	Downsampler	Layers	Top-1 Acc(%)	#Params	FLOPs (G)	Thru. (imgs/sec)	Train Thru. (imgs/sec)
SSD	Patch	2, 4, 8, 4	81.0	14.8 M	2.1	1818	523
Bi-SSD	Patch	2, 4, 8, 4	81.4	15.2 M	2.2	1741	399
VSSD	Patch	2, 4, 8, 4	81.6	14.8 M	2.1	1843	606
Hybrid	Patch	2, 4, 8, 4	82.3	13.4 M	2.1	1890	622

Different SSD Mechanisms. In our ablation study for the token mixer, we explored different scanning routes for SSD. Specifically, we introduced Bi-SSD, which splits channels and reverses one part to create backward scanning sequences. These sequences are then concatenated post-SSD block. As shown in Tab. 3, our VSSD model outperforms both the vanilla SSD and Bi-SSD by 0.6% and 0.2% in top-1 accuracy, respectively. Moreover, both training and inference throughput are enhanced, with VSSD improving training throughput by nearly 50% compared to the Bi-SSD approach.

Hybrid Architecture and Overlapped Downsampler. The effectiveness of incorporating standard attention in the last stage is demonstrated in the last row of Tab. 3. Specifically, replacing VSSD with standard attention in the last stage results in a 0.7% improvement in accuracy while slightly reducing the parameters.

5 Conclusion

In this paper, we introduced VSSD-UNet, which is a mamba-based UNet style network for medical image segmentation. The performance demonstrates that VSSD-UNet superior performance against classical similar network such as UNet and Swin-UNet. In the future, we aim to conduct more in-depth explorations on more medical image segmentation tasks from different modalities and targets, with comparisons to more segmentation backbones. Besides, we aim to extend VSSD-UNet to 3D medical images to further enhance the developments in medical imaging.

REFERENCES

Ehsan Khodapanah Aghdam, Reza Azad, Maral Zarvani, and Dorit Merhof. Attention swin u-net: Cross-contextual attention mechanism for skin lesion segmentation. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5. IEEE, 2023. doi: 10.1109/ISBI53787.2023.10230337.

Md Zahangir Alom, Mahmudul Hasan, Chris Yakopcic, Tarek M Taha, and Vijayan K Asari. Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation. *arXiv preprint arXiv:1802.06955*, 2018. doi: 10.48550/arXiv.1802.06955.

Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(12): 2481–2495, 2017.

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. *arXiv preprint arXiv:2105.05537*, 2021.

Gongping Chen, Lei Li, Yu Dai, Jianxun Zhang, and Moi Hoon Yap. Aau-net: an adaptive attention u-net for breast lesions segmentation in ultrasound images. *IEEE Trans. Med. Imaging*, 2022.

Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

Keyan Chen, Bowen Chen, Chenyang Liu, Wenyuan Li, Zhengxia Zou, and Zhenwei Shi. Rsmamba: Remote sensing image classification with state space model. *IEEE Geoscience and Remote Sensing Letters*, 2024a.

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(4):834–848, 2017.

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Eur. Conf. Comput. Vis.*, pp. 801–818, 2018.

Tianxiang Chen, Zhentao Tan, Tao Gong, Qi Chu, Yue Wu, Bin Liu, Jieping Ye, and Nenghai Yu. Mim-istd: Mamba-in-mamba for efficient infrared small target detection. *arXiv preprint arXiv:2403.02148*, 2024b.

Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*, 2020.

Tri Dao and Albert Gu. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. *arXiv preprint arXiv:2405.21060*, 2024.

Bo Dong, Wenhai Wang, Deng-Ping Fan, Jinpeng Li, Huazhu Fu, and Ling Shao. Polyp-pvt: Polyp segmentation with pyramid vision transformers. *arXiv preprint arXiv:2108.06932*, 2021.

Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In *CVPR*, 2022.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Chieu Ho, Martin Chan, Xiang Yang, Vincent Vanhoucke, Anil Jaitly, and Lukasz Kaiser. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Chengbin Du, Yanxi Li, and Chang Xu. Understanding robustness of visual state space models for image classification. *arXiv preprint arXiv:2403.10935*, 2024.

Isensee et al. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.

Qihang Fan, Huaibo Huang, Mingrui Chen, Hongmin Liu, and Ran He. Rmt: Retentive networks meet vision transformers. In *CVPR*, 2024.

Daniel Y Fu, Tri Dao, Khaled K Saab, Armin W Thomas, Atri Rudra, and Christopher Ré. Hungry hungry hippos: Towards language modeling with state space models. *arXiv preprint arXiv:2212.14052*, 2022.

Albert Gu. *Modeling Sequences with Structured State Spaces*. Phd thesis, Stanford University, 2023. ProQuest Document ID: 2880853867.

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

Albert Gu, Tri Dao, Stefano Ermon, Atri Rudra, and Christopher Ré. Hippo: Recurrent memory with optimal polynomial projections. *NeurIPS*, 2020.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021a.

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, 2021b.

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. *NeurIPS*, 2021c.

Dongchen Han, Ziyi Wang, Zhuofan Xia, Yizeng Han, Yifan Pu, Chunjiang Ge, Jun Song, Shiji Song, Bo Zheng, and Gao Huang. Demystify mamba in vision: A linear attention perspective. *arXiv preprint arXiv:2405.16605*, 2024.

Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. Transformer in transformer. In *NeurIPS*, 2021.

Ali Hassani, Steven Walton, Jiachen Li, Shen Li, and Humphrey Shi. Neighborhood attention transformer. In *CVPR*, 2023.

Along He, Kai Wang, Tao Li, Chengkun Du, Shuang Xia, and Huazhu Fu. H2former: An efficient hierarchical hybrid transformer for medical image segmentation. *IEEE Transactions on Medical Imaging*, 42(9):2763–2775, 2023. doi: 10.1109/TMI.2023.3264513.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

Gao Huang, Zhuang Liu, Geoff Pleiss, Laurens Van Der Maaten, and Kilian Weinberger. Convolutional networks with dense connectivity. *IEEE TPAMI*, 2019.

Huimin Huang, Lanfen Lin, Ruofeng Tong, Hongjie Hu, Qiaowei Zhang, Yutaro Iwamoto, Xianhua Han, Yen-Wei Chen, and Jian Wu. Unet 3+: A full-scale connected unet for medical image segmentation. In *ICASSP*, pp. 1055–1059. IEEE, 2020.

Tao Huang, Xiaohuan Pei, Shan You, Fei Wang, Chen Qian, and Chang Xu. Localmamba: Visual state space model with windowed selective scan. *arXiv preprint arXiv:2403.09338*, 2024.

Xiaohong Huang, Zhifang Deng, Dandan Li, Xueguang Yuan, and Ying Fu. Missformer: An effective transformer for 2d medical image segmentation. *IEEE Transactions on Medical Imaging*, 42 (5):1484–1494, 2023. doi: 10.1109/TMI.2022.3230943.

Nabil Ibtehaz and Daisuke Kihara. Acc-unet: A completely convolutional unet model for the 2020s. In *Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, pp. 692–702. Springer, 2023.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *ICML*, 2020.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *NeurIPS*, 2012.

Yann LeCun and Yoshua Bengio. *Convolutional Networks for Images, Speech, and Time Series*, pp. 255–258. MIT Press, Cambridge, MA, USA, 1998. ISBN 0262511029.

Kunchang Li, Xinhao Li, Yi Wang, Yinan He, Yali Wang, Limin Wang, and Yu Qiao. Videomamba: State space model for efficient video understanding. *arXiv preprint arXiv:2403.06977*, 2024.

Dingkang Liang, Xin Zhou, Xinyu Wang, Xingkui Zhu, Wei Xu, Zhikang Zou, Xiaoqing Ye, and Xiang Bai. Pointmamba: A simple state space model for point cloud analysis. *arXiv preprint arXiv:2402.10739*, 2024.

Weifeng Lin, Ziheng Wu, Jiayu Chen, Jun Huang, and Lianwen Jin. Scale-aware modulation meet transformer. In *ICCV*, 2023.

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*, 2024a.

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*, 2024b.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Int. Conf. Comput. Vis.*, pp. 10012–10022, 2021.

Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *CVPR*, 2022a.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *CVPR*, 2022b.

Ange Lou, Shuyue Guan, and Murray Loew. Dc-unet: rethinking the u-net architecture with dual channel efficient cnn for medical image segmentation. In *Med. Imaging 2021: Image Process.*, volume 11596, pp. 758–768. SPIE, 2021.

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

Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

Xiaohuan Pei, Tao Huang, and Chang Xu. Efficientvmamba: Atrous selective scan for light weight visual mamba. *arXiv preprint arXiv:2403.09977*, 2024.

Zhen Qin, Weixuan Sun, Hui Deng, Dongxu Li, Yunshen Wei, Baohong Lv, Junjie Yan, Lingpeng Kong, and Yiran Zhong. cosformer: Rethinking softmax in attention. *arXiv preprint arXiv:2202.08791*, 2022.

Md Mostafijur Rahman and Radu Marculescu. Medical image segmentation via cascaded attention decoding. In *IEEE/CVF Winter Conf. Appl. Comput. Vis.*, pp. 6222–6231, January 2023a.

Md Mostafijur Rahman and Radu Marculescu. Multi-scale hierarchical vision transformer with cascaded attention decoding for medical image segmentation. In *Med. Imaging Deep Learn.*, 2023b.

Sucheng Ren, Xingyi Yang, Songhua Liu, and Xinchao Wang. Sg-former: Self-guided transformer with evolving token reallocation. In *ICCV*, 2023.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, pp. 234–241. Springer, 2015.

Jiacheng Ruan and Suncheng Xiang. Vm-unet: Vision mamba unet for medical image segmentation. *ArXiv*, abs/2402.02491, 2024a. doi: 10.48550/arXiv.2402.02491.

Jiacheng Ruan and Suncheng Xiang. Vm-unet: Vision mamba unet for medical image segmentation. *arXiv preprint arXiv:2402.02491*, 2024b.

Jiacheng Ruan, Suncheng Xiang, Mingye Xie, Ting Liu, and Yuzhuo Fu. Malunet: A multi-attention and light-weight unet for skin lesion segmentation. In *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 1150–1156. IEEE, 2022. doi: 10.1109/BIBM55620.2022.9995040.

Jiacheng Ruan, Mingye Xie, Jingsheng Gao, Ting Liu, and Yuzhuo Fu. Ege-unet: an efficient group enhanced unet for skin lesion segmentation. In *International conference on medical image computing and computer-assisted intervention*, pp. 481–490. Springer, 2023. doi: 10.1007/978-3-031-43901-8_46.

Yuheng Shi, Minjing Dong, and Chang Xu. Multi-scale vmamba: Hierarchy in hierarchy visual state space model. *arXiv preprint arXiv:2405.14174*, 2024.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Jimmy TH Smith, Andrew Warrington, and Scott W Linderman. Simplified state space layers for sequence modeling. *arXiv preprint arXiv:2208.04933*, 2022.

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, 2019.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *ICML*, 2021.

Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxvit: Multi-axis vision transformer. In *ECCV*, 2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

Haonan Wang, Peng Cao, Jiaqi Wang, and Osmar R Zaiane. Uctransnet: rethinking the skip connections in u-net from a channel-wise perspective with transformer. In *AAAI*, volume 36, pp. 2441–2449, 2022a.

Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *ICCV*, 2021.

Wenhai Wang, Jifeng Dai, Zhe Chen, Zhenhang Huang, Zhiqi Li, Xizhou Zhu, Xiaowei Hu, Tong Lu, Lewei Lu, Hongsheng Li, et al. Internimage: Exploring large-scale vision foundation models with deformable convolutions. *arXiv preprint arXiv:2211.05778*, 2022b.

Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 2022c.

Renkai Wu, Pengchen Liang, Xuan Huang, Liu Shi, Yuandong Gu, Haiqin Zhu, and Qing Chang. Mhorunet: High-order spatial interaction unet for skin lesion segmentation. *Biomedical Signal Processing and Control*, 88:105517, 2024a. doi: 10.1016/j.bspc.2023.105517.

Renkai Wu, Yinghao Liu, Pengchen Liang, and Qing Chang. H-vmunet: High-order vision mamba unet for medical image segmentation. *arXiv preprint arXiv:2403.13642*, 2024b. doi: 10.48550/arXiv.2403.13642.

Renkai Wu, Yinghao Liu, Pengchen Liang, and Qing Chang. Ultralight vm-unet: Parallel vision mamba significantly reduces parameters for skin lesion segmentation. *arXiv preprint arXiv:2403.20035*, 2024c. doi: 10.48550/arXiv.2403.20035.

Zhuofan Xia, Xuran Pan, Shiji Song, Li Erran Li, and Gao Huang. Dat++: Spatially dynamic vision transformer with deformable attention. *arXiv preprint arXiv:2309.01430*, 2023.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Adv. Neural Inform. Process. Syst.*, 34:12077–12090, 2021.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *CVPR*, 2017.

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 preprint arXiv:2403.17695*, 2024a.

Yuhuan Yang, Chaofan Ma, Jiangchao Yao, Zhun Zhong, Ya Zhang, and Yanfeng Wang. Remamber: Referring image segmentation with mamba twister. *arXiv preprint arXiv:2403.17839*, 2024b.

Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan. Metaformer is actually what you need for vision. In *CVPR*, 2022.

Mingya Zhang, Yue Yu, Sun Jin, Limei Gu, Tingsheng Ling, and Xianping Tao. Vm-unet-v2: rethinking vision mamba unet for medical image segmentation. In *International Symposium on Bioinformatics Research and Applications*, pp. 335–346. Springer, 2024. doi: 10.1007/978-981-97-5128-0_27.

Yundong Zhang, Huiye Liu, and Qiang Hu. Transfuse: Fusing transformers and cnns for medical image segmentation. In *Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, pp. 14–24. Springer, 2021.

Zhuoran Zheng and Chen Wu. U-shaped vision mamba for single image dehazing. *arXiv preprint arXiv:2402.04139*, 2024.

Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learn. Med. Image Anal. Multimodal Learn. Clin. Decis. Support*, pp. 3–11. Springer, 2018.

Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Transactions on Medical Imaging*, 2019. doi: 10.1109/TMI.2019.2959609.

Lei Zhu, Xinjiang Wang, Zhanghan Ke, Wayne Zhang, and Rynson Lau. Biformer: Vision transformer with bi-level routing attention. In *CVPR*, 2023.

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*, 2024a.

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*, 2024b.

Qiang Zuo, Songyu Chen, Zhifang Wang, and Liguo Zhang. R2au-net: Attention recurrent residual convolutional neural network for multimodal medical image segmentation. *Sec. and Commun. Netw.*, 2021, jan 2021. ISSN 1939-0114. doi: 10.1155/2021/6625688.