Hetero-UNet: Heterogeneous Transformer with Mamba for Medical Image Segmentation

Zhiling Yan
Lehigh University
zhy423@lehigh.edu

Yixin Liu
Lehigh University
yila22@lehigh.edu

Xiang Li
Massachusetts General Hospital
and Harvard Medical School
xli60@mgh.harvard.edu

Lichao Sun*
Lehigh University
lis221@lehigh.edu

Abstract

Convolutional Neural Networks (CNNs) have significantly advanced medical image segmentation, offering unparalleled local feature extraction capabilities. However, CNNs face limitations in capturing long-range dependencies due to the local nature of convolutional operations. Recently, State-Space Models (SSMs), such as Mamba, have presented an efficient solution by incorporating gating, convolutions, and data-dependent filtering mechanisms for long-range interaction modeling. However, as an attention-free mechanism, SSMs are less efficient at handling variable distance token-to-token interactions compared to attention. In this paper, we introduce Hetero-UNet, a novel hybrid U-Net architecture that incorporates SSMs and attention mechanisms to map long-range dependencies. Featuring a hybrid Transformer-Mamba encoder within original U-Net architecture, it excels at extracting both local and global features. Our extensive experiments across diverse tasks—abdominal organ segmentation in CT and MR, instrument segmentation in endoscopy, and cell segmentation in microscopy—demonstrates Hetero-UNet's superior performance over previous state-of-the-art segmentation models, paving the way for hybrid long-range dependency modeling in medical imaging. The code is available at https://github.com/ZhilingYan/Hetero-UNet.

1 Introduction

Medical image segmentation is vital for helping healthcare professionals detect biological structures and assess their morphology, which aids in diagnosing various diseases. In recent years, convolutional neural networks (CNNs)-based models, such as U-Net [28] has become particularly popular due to its simple yet flexible architecture. Numerous variations and improvements have been proposed based on U-Net's distinctive U-shaped design [17, 3, 33, 34, 16, 15], most of which use a symmetric encoder-decoder structure to extract image features at different scales. These advancements have significantly impacted various medical imaging tasks, including abdominal multi-organ segmentation in MR images [19] and cell segmentation in microscopy images [23].

However, despite the success of CNN-based models, they face a key limitation: their inability to capture long-range dependencies in images. This issue arises from the fact that convolutional kernels are inherently local [5]. Although U-Net uses skip connections to combine low-level and high-level features, these connections primarily merge local features and do not significantly improve the

*Corresponding author.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

model's capacity to handle long-range dependencies. This shortcoming becomes more noticeable when dealing with significant inter-patient variations in size, shape, and other factors [5]. These variations make it challenging for CNNs to consistently and accurately capture information across broader spatial areas, underscoring the need for new methods to overcome this limitation.

Recognizing these limitations of CNNs, the research community has shifted interest towards State Space Models (SSMs) [25, 11, 30] due to their ability to establish long-distance dependencies and their increasingly competitive performance. These SSMs (e.g., Mamba [12]) integrate features like gating, convolutions, and input-dependent token selection to enhance performance. Their application extends across various domains, including language understanding [8] and general vision [35], demonstrating their versatility. Particularly in medical imaging, where datasets often exhibit large inter-patient variations, Mamba's capability to filter out irrelevant information [27] and then decrease the effect of inter-patient variations proves beneficial, allowing for a focused representation learning of medical images. Efforts to use pure Mamba [29, 31] or merge CNNs with SSMs [22, 32, 10] in medical image segmentation have shown promising results compared with state-of-the-art CNN-based and Transformer-based models, indicating SSMs' potential in this area.

While Mamba's performance in visual tasks, particularly medical image segmentation, has been impressive, applications have primarily focused on using Mamba alone or in conjunction with CNNs, overlooking the potential of integrating Transformers for global information capture. Transformers are known for their effectiveness in achieving input-dependent processing by computing all token-to-token interactions, a feature crucial for mixing information across sequences [2]. However, as an attention-free mechanism, SSMs are less efficient at handling variable distance token-to-token interactions compared to attention. In addition, the performance of SSMs is significantly influenced by the size of the hidden layer [27], and SSMs employ fixed filters defined by model weights for processing sequences, contrasting with input-adaptive functions of attention mechanisms [2]. Although SSMs allow for filtering of irrelevant information, excessive filtering in deeper models can lead to low-resolution features, results in challenges for the decoder in accurately reconstructing the input. Despite these challenges, merging multi-head attention (MHA) with SSMs has shown promise in processing long sequences [36, 25, 6], though the potential of interleaving Transformers and SSMs in vision domain, especially for enhancing medical image segmentation, is yet to be fully explored.

In this paper, we propose **Hetero-UNet**, a cutting-edge hybrid model that combines the strengths of SSMs with a U-shaped architecture, incorporating convolution encoder-decoder, attention mechanisms, and SSM blocks. In our work, we have preserved the original U-Net architecture, including the encoder, decoder, and skip connections, to ensure the model's ability to learn local features. We continue to employ CNN within both the encoder and decoder module. However, to enhance the model's long-range dependency learning capacity, we have incorporated a hybrid Transformer-Mamba (TransMamba) encoder following the CNN encoder. Specifically, the hybrid encoder consists of multiple layers of TransMamba blocks, each comprising a SSM block equipped with additional attention modules. Comprehensive experiments have been carried out across a variety of medical imaging tasks, including 3D abdominal multi-organ segmentation in CT and MR images, instrument segmentation in endoscopy images, and cell segmentation in microscopy images, to showcase the effectiveness of hybrid SSM-based model in medical image segmentation. Hetero-UNet achieves superior performance, outperforming networks solely based on Transformers, CNNs, and SSM techniques. This shed the light for future research in architectures that capture long-range dependencies in biomedical imaging. The primary contributions of this paper are highlighted as follows:

1) We propose a novel SSM-based model, marking a pioneering exploration for harnessing the capability of hybrid SSM-based models with attention mechanism for segmentation tasks in medical imaging domain.

2) We introduce the Hetero-UNet architecture, a sophisticated framework that utilizes hybrid Transformer-Mamba module to make strong U-Net encoder. This architecture is designed to adeptly capture both local details and global patterns within long-range data, ensuring comprehensive feature extraction.

3) Comprehensive experiments are conducted on four datasets, including both 3D and 2D images across a variety of imaging modalities (e.g. CT, MR, microscopy, and endoscopy), with results indicating that Hetero-UNet exhibits considerable competitiveness.

2 Related Work

2.1 Medical Image Segmentation

CNN-based and Transformer-based models have made significant strides in advancing medical image segmentation. U-Net [28], a key CNN-based model, uses a symmetric encoder-decoder architecture with skip connections to preserve detailed information. Several improvements [26], such as the self-adapting nnU-Net framework [18], have been developed based on this U-shaped structure, showing strong performance across diverse medical image segmentation tasks. On the Transformer side, TransUnet [4] integrates the Vision Transformer (ViT)[7] for feature extraction in the encoder while utilizing CNN for decoding, effectively capturing global information. Swin-UNETR[15] and UNETR [16] combine Transformer architectures with traditional U-Net designs to improve 3D image analysis. Furthermore, Swin-UNet [3] incorporates Swin Vision Transformer blocks [21] within the U-Net framework, further expanding the use of Transformer technology in medical imaging.

2.2 SSMs in Image Segmentation

State-space models (SSMs), such as Mamba, have recently emerged as a powerful tool for deep network development, delivering state-of-the-art results in analyzing long-sequence data [9, 8]. Early explorations of SSMs in the vision domain have shown encouraging results [35]. VMamba [20] introduces a Mamba-based vision backbone for hierarchical representations, incorporating a cross-scan module to tackle the dimensional mismatch between 1D sequences and 2D images. In medical image segmentation, U-Mamba [22] marks a groundbreaking hybrid approach that combines SSMs and CNNs, representing the first application of SSMs in this field. Further advancements include VM-UNet [29] and Mamba-UNet [31], which use a fully SSM-based encoder-decoder structure, and SegMamba [32] and nnMamba [10], which integrate SSMs in the encoder with CNNs in the decoder, showcasing the versatility and effectiveness of SSMs in advancing medical imaging analysis.

3 Method

The process of Hetero-UNet is shown in Figure 1. We implement a multi-layer encoder with CNN-based blocks to extract the local features of the input images. Then we introduce a novel hybrid Transformer-Mamba (TransMamba) encoder with multi-layer hybrid TransMamba blocks to learn the long-range dependencies of the input. For the decoder, we use the multi-layer CNN-based decoder to upsample and reconstruct the output segmentation label.

3.1 U-shape Convolutional Encoder and Decoder

The encoder, composed of residual blocks and transposed convolutions, focused on detailed local information and feature extraction. The decoder keeps the same as the encoder, while encoder is to decrease the input size and the decoder tries to increase the features size, focusing on resolution recovery. Moreover, we inherit the skip connection in U-Net to connect the hierarchical features from the encoder to the decoder. The final decoder feature is passed to a 3D convolutional layer with a softmax layer to produce the final segmentation probability map.

3.2 Hetero-UNet: Empowering SSM with Attention Mechanism

State Space Models and Mamba Block. Before introducing Hetero-UNet, we firstly introduce the technique details of the Mamba block, as a preliminary of the hybrid Hetero-UNet. State Space Models (SSMs) are traditionally identified as linear time-invariant frameworks that map an input signal $x(t) \in \mathbb{R}^N$ to an output signal $y(t) \in \mathbb{R}^N$ through the mediation of a hidden state $h(t) \in \mathbb{R}^N$. This process is captured by the equations:

$$h'(t) = \mathbf{A}h(t) + \mathbf{B}x(t), y(t) = \mathbf{C}h(t), \tag{1}$$

with $\mathbf{A} \in \mathbb{R}^{N\times N}$ representing the state transition matrix, and $\mathbf{B}, \mathbf{C} \in \mathbb{R}^N$ as the input and output matrices, respectively. Determining the output $y(t)$ at a specific time t often presents a significant challenge due to the difficulty in analytically solving for $h(t)$. Moreover, since real-world data frequently occurs in a discrete format, an alternative approach involves discretizing the continuous system described in Equation 1:

$$h_t = \overline{\mathbf{A}}h_{t-1} + \overline{\mathbf{B}}x_t,\ y_t = \mathbf{C}h_t, \tag{2}$$

where the discretized system parameters $\overline{\mathbf{A}} := \exp(\Delta \cdot \mathbf{A})$ and $\overline{\mathbf{B}} := (\Delta \cdot \mathbf{A})^{-1}(\exp(\Delta \cdot \mathbf{A}) - I) \cdot \Delta\mathbf{B}$, with Δ denoting the discretization interval.

Despite their profound theoretical underpinning, SSMs often encounter significant computational demands and potential for numerical instability. This has led to the emergence of Structured State

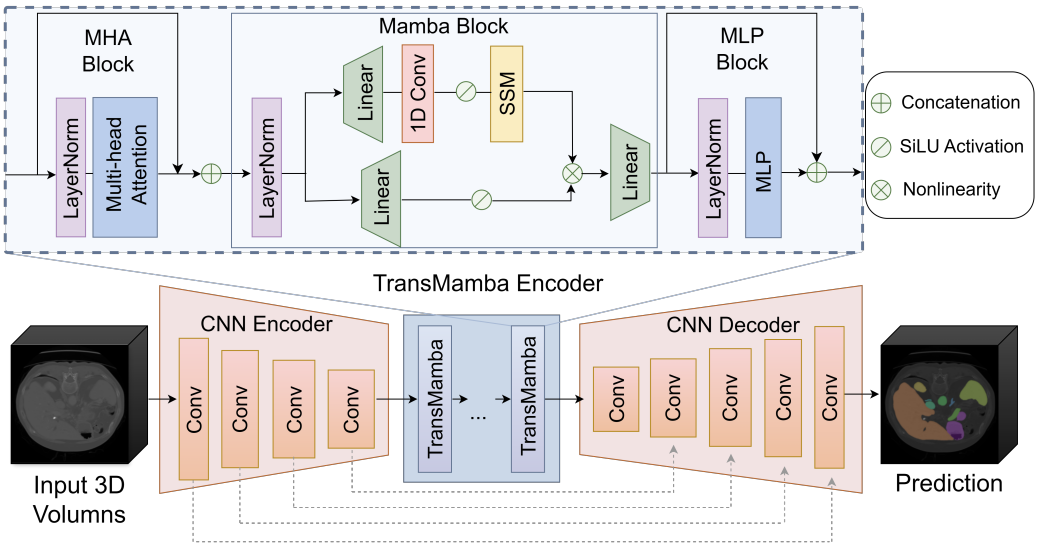


Figure 1: Overview of our proposed Hetero-UNet architecture for medical image segmentation. Hetero-UNet employs the U-shape encoder-decoder framework with a convolution encoder, a Transformer-Mamba (TransMamba) encoder, a convolution decoder and skip connections for enhanced feature fusion. The core of this architecture, the scalable TransMamba encoder, is composed of multiple TransMamba layers. Each hybrid layer within the encoder includes a MHA block and a MLP layer, with a SSM-based Mamba block positioned between them for enhanced long-range dependency modeling.

Space sequence models (S4) [14], which address these drawbacks by implementing a structured arrangement in the state transition matrix $\mathbf{A}$, notably through the use of HIPPO matrices [13]. Such structural modifications have yielded considerable improvements in both efficiency and performance metrics. Remarkably, S4 models have demonstrated superior performance compared to Transformers, especially in tasks requiring the modeling of long-range dependencies [14]. The more recent innovation, Mamba [12], enhances this model by incorporating an input-dependent selection mechanism alongside a faster, hardware-optimized algorithm.

Specifically, the selective scan mechanism (S6) [12] in Mamba block design the matrices $B \in \mathbb{R}^{B\times L\times N}$, $C \in \mathbb{R}^{B\times L\times N}$, and $\Delta \in \mathbb{R}^{B\times L\times C}$ are derived from the input data $x \in \mathbb{R}^{B\times L\times C}$. This implies that S6 is aware of the contextual information embedded in the input, ensuring the dynamism of weights within this mechanism.

Our Design: Hybrid TransMamba Block. Hybrid TransMamba block is designed to leverage the best of Transformer and Mamba blocks. In medical imaging, where datasets often exhibit large inter-patient variations, Mamba's capability to filter out irrelevant information [27] proves beneficial, while their performance being significantly influenced by the hidden layer size. Despite these obstacles, the combination of Transformer with SSMs has been found beneficial [36, 25] in long sequence processing. Considering the potential of interleaving Transformers and SSMs in the vision domain yet to be fully explored, we introduce a novel module, denoted as TransMamba Block, to leverage the best of Transformer and Mamba. It enables one to learn the global features within long-range data. To achieve this, we insert the Mamba block into the original Multi-Head Attention block, as shown in the blue box in Figure 1.

Given the input features $x \in \mathbf{R}^{B,L,C}$, they enter the hybrid TransMamba block. The first is the Transformer block. We keep the initial feature and concatenate it with the output feature from Multi-Head Attention block. The output of the Transformer block is: $x_{\text{output}} = \text{LayerNorm}(x + \text{MHA}(x))$. After the transformer block, the features are still in the shape of B, L, C. Then, with one more layer norm, the features are in the Mamba block that contains two parallel branches. In the first branch, the features are expanded to $(B, 2L, C)$ by a linear expansion layer $W_{\text{up}} \in \mathbf{R}^{C\times EC}$ with expansion ratio E followed by the SiLU activation function. In the second branch, the features firstly experience the same expansion to $(B, 2L, C)$ using the W_{up} of the same size. Then, it goes into a 1D convolutional layer, followed by a SiLU activation function and SSM layer. After that, the features from the two branches are merged together with the Hadamard product. Then, the features are projected back into

the original shape (B, L, C) using a linear down-projection layer $W_{\text{down}} \in \mathbf{R}^{EC \times C}$. The output can be calculated as:

$$x_{\text{output}} = W_{\text{down}}(\text{SSM}(\sigma(\text{Conv}(W_{\text{up}}x))) \odot \sigma(W_{\text{up}}x)). \tag{3}$$

Finally, the output features are put into the MLP block, with a layer norm and an MLP layer: $x_{\text{output}} = \text{LayerNorm}(x + \text{MLP}(x))$.

Implementation Note. The hybrid TransMamba encoder is composed of a stack of $N = 12$ identical layers. As shown in Figure 1, before entering the TransMamba block, the input $x \in \mathbf{R}^{B,C,H,W,D}$ are flattened and transposed to $x \in \mathbf{R}^{B,L,C}$ where $L = H \times W \times D$. After passing the layer norm, the features are fed into the hybrid TransMamba block. Go through $N = 12$ TransMamba layers, the features are fed into the final layer norm. The shape of x remains the same as (B, L, C). Finally, the features are reshaped and transposed into (B, C, H, W, D).

4 Experiments

In this section, we conduct comprehensive experiments on Hetero-UNet across various medical imaging tasks, including abdominal multi-organ segmentation in CT and MR images, instrument segmentation in endoscopy images, and cell segmentation in microscopy images. The results demonstrate that Hetero-UNet consistently outperforms existing methods, showcasing its superior capability in handling diverse segmentation challenges. This performance highlights Hetero-UNet's versatility and effectiveness in accurately segmenting medical images, affirming its potential as a leading solution in the field of medical image analysis.

4.1 Datasets and Implementation Details

To assess the performance and scalability of Hetero-UNet, we utilize four medical image datasets across a variety of segmentation tasks and imaging modalities, including Abdomen CT dataset [24], Abdomen MR dataset [19], Endoscopy dataset [1] and Microscopy dataset [23].

The setting of our experiments are the same as that in U-Mamba [22] and nnUNet [18] to ensure a fair comparison. Specifically, We adopt an unweighted combination of Dice loss and cross-entropy loss for all datasets and utilize the SGD optimizer with an initial learning rate of 1e-2. The training duration for each dataset is set to 1000 epochs, conducted on a single NVIDIA RTX A5000 GPU. Leveraging the self-configuring capabilities from nnUNet, the number of network blocks adjusts automatically according to the dataset. For evaluation metrics, we employ Dice Similarity Coefficient (DSC) and Normalized Surface Distance (NSD) to assess performance in abdominal multi-organ segmentation for CT and MR scans, and instrument segmentation in Endoscopy images. For the cell segmentation task, we utilize Precision, Recall, DSC, and F1 score to evaluate method performance.

4.2 Comparison Results

In our evaluation of Hetero-UNet, we compare against two prominent CNN-based segmentation methods: nnUNet [18] and SegResNet [26]. Additionally, we include a comparison with UNETR [16], a Transformer-based network that has gained popularity in medical image segmentation tasks. U-Mamba [22], a recent method based on the Mamba model, is also included in our comparison to provide a comprehensive overview of its performance. For each model, we implement their recommended optimizers to ensure consistency in training conditions. To maintain fairness across all comparisons, we apply the default image preprocessing in nnUNet [18].

Abdomen 3D Dataset. Table 1 and Table 2 present comparative results for the abdominal multi-organ segmentation task on CT and MRI images, showcasing Hetero-UNet's superior performance across CNN-based, Transformer-based, and SSM-based methods. Specifically, our method improves the DSC from 0.5% to 18.4% and NSD from 0.7% to 20.1% on CT 3D images. It also improves the DSC from 0.3% to 16.0% and NSD from 0.4% to 17.1% on MR 3D images. Despite nnUNet's strong performance, highlighting its robustness in segmentation tasks, SSM-based methods, including ours, excel in capturing long-range dependencies, critical for medical imaging analysis. The superior result demonstrates the great potential of the SSM-based network in medical 3D image segmentation. Hetero-UNet's advantage is particularly evident when compared to U-Mamba, an SSM-based method employing the same backbone as ours while lacking the integration of attention mechanisms. Our method consistently outperforms it, underscoring the advantage of the hybrid module that combines SSM with attention mechanisms for enhanced medical image segmentation. The visual comparisons in Figure 2 further affirm Hetero-UNet's precision in segmentation. It demonstrates fewer outliers and higher accuracy in recognizing organ shapes and types, notably producing more accurate segmentation masks for challenging organs. For instance, Hetero-UNet generates more accurate duodenum and

Table 1: Comparison of abdominal multi-organ segmentation results generated from our Hetero-UNet and other state-of-the-art methods on the Abdomen CT dataset.

Methods	LV	RK	Sp.	P.	Aor.	IVC	RA	LA	GB	Es.	St.	DD	LK	AVG
DSC [%] ↑														
nnUNet	96.9	**87.3**	**93.7**	85.0	95.8	87.8	**82.1**	79.9	72.0	86.2	88.0	77.0	85.8	86.0
SegResNet	95.3	81.2	86.8	74.6	93.4	84.0	75.4	70.7	66.9	80.7	82.1	61.3	79.5	79.4
UNETR	91.1	67.7	76.6	61.9	88.6	75.0	67.2	49.3	54.2	68.7	70.5	50.4	66.5	68.3
U-Mamba	**97.0**	85.8	90.8	85.2	95.6	**89.4**	82.0	**83.6**	72.5	87.0	**89.0**	**78.1**	84.7	86.2
Ours	96.6	86.4	91.1	**85.6**	**96.0**	88.1	80.4	83.5	**81.7**	**87.3**	88.7	73.6	**88.3**	**86.7**
NSD [%] ↑														
nnUNet	95.4	85.8	**92.8**	**93.0**	**97.4**	87.1	93.4	89.6	71.5	93.4	89.7	89.3	82.4	89.3
SegResNet	92.2	77.4	85.2	83.7	93.7	82.9	88.0	81.5	64.9	89.4	84.4	77.1	77.5	82.9
UNETR	82.1	64.0	72.7	69.0	86.7	72.4	80.7	60.4	49.3	78.6	67.4	69.9	62.3	70.4
U-Mamba	**96.0**	83.7	89.9	92.4	**97.4**	**89.3**	**94.0**	**93.5**	72.9	93.5	90.0	**89.4**	85.1	89.8
Ours	95.6	**86.1**	90.5	92.9	97.3	88.0	92.3	93.4	**82.4**	**93.7**	**90.1**	85.9	**88.9**	**90.5**

[1] LV: Liver, RK/LK: Right/Left kidney, Sp.: Spleen, P.: Pancreas, Aor.: Aorta, IVC: Inferior vena cava, RA/LA: Right/Left adrenal gland, GB: Gall bladder, Es.: Esophagus, St.: Stomach, DD: Duodenum

Table 2: Comparison of abdominal multi-organ segmentation results generated from our Hetero-UNet and other state-of-the-art methods on the Abdomen MR dataset.

Methods	LV	RK	Sp.	P.	Aor.	IVC	RA	LA	GB	Es.	St.	DD	LK	AVG
DSC [%] ↑														
nnUNet	**97.4**	**96.3**	92.5	85.6	94.4	83.0	61.0	69.8	80.0	77.7	82.2	70.1	**96.7**	83.6
SegResNet	96.5	93.9	89.8	83.8	92.1	81.6	61.0	62.6	80.0	71.8	74.3	67.2	96.0	80.8
UNETR	93.4	82.4	85.1	72.7	85.1	71.7	42.0	50.0	49.0	55.1	67.9	50.3	90.2	68.8
U-Mamba	**97.4**	96.1	**94.9**	87.1	**95.3**	**83.4**	63.2	71.5	80.9	77.4	83.3	**71.3**	96.6	84.5
Ours	97.3	96.1	92.8	**87.2**	94.7	83.1	**64.4**	**72.3**	**84.5**	**78.9**	**84.5**	70.2	96.5	**84.8**
NSD [%] ↑														
nnUNet	**97.6**	97.6	93.2	95.3	96.7	88.4	79.3	84.4	79.2	92.7	85.7	89.7	**98.7**	90.7
SegResNet	96.0	95.6	90.0	93.6	94.2	86.4	77.0	76.3	77.1	86.9	77.8	88.0	97.9	87.4
UNETR	89.6	82.7	82.2	82.8	86.4	75.8	58.3	63.6	41.9	72.9	70.6	75.4	88.3	74.7
U-Mamba	97.5	97.5	**95.4**	96.1	**97.9**	**88.6**	80.7	85.0	79.7	93.3	87.0	**90.2**	**98.7**	91.4
Ours	97.4	**97.7**	93.7	**96.2**	97.4	**88.6**	**82.2**	**86.2**	**83.9**	**93.4**	**88.1**	89.8	98.5	**91.8**

[1] Symbols are the same as those in Table 1.

stomach segmentation masks while the others have mis-segmentation and even mis-classification errors for duodenum masks as well as missing regions for stomach masks in Abdomen CT scans. Similarly, for pancreas segmentation in MR scans, Hetero-UNet successfully delineates its boundary while other methods generate various segmentation errors. This evidence emphasizes Hetero-UNet's potential as a leading solution for 3D medical image segmentation, combining the strengths of SSMs and attention mechanisms for superior performance.

Endoscopy Dataset. In our comparative analysis, Hetero-UNet is evaluated against a spectrum of models, including those based on CNNs, Transformer architectures, and SSMs. To visually differentiate each distinct method, we employ a color legend, where each model type is assigned a specific color. Figure 3a presents the segmentation performance on the Endoscopy dataset, with Hetero-UNet's performance denoted within the purple bar. Hetero-UNet achieves the highest scores with 70.0% in DSC and 71.6% in NSD among CNN-based, Transformer-based, and SSM-based methods, surpassing all baselines from 2.5% to 20.9% in DSC and from 2.6% to 20.9% in NSD. The superior performance of SSM-based methods including ours and U-Mamba emphasizes the importance of learning long-range dependencies in feature extraction and image segmentation, even in 2D images which are generally smaller than 3D images. The hybrid design of Hetero-UNet is also important for robust image segmentation, as evidenced by its performance exceeding that of U-Mamba and other baseline models. The qualitative visualization in the first two rows of Figure 4 further illustrates the effectiveness of the proposed Hetero-UNet. It most closely mirrors the ground truth, accurately delineating the boundaries of all instruments, while other methods often result in overlapping or over-segmentation errors, especially in the first case. The majority of these methods struggle with correctly identifying instruments such as the bipolar forceps and monopolar curved scissors, leading to erroneous segmentation masks. Moreover, a common issue among these methods is the incorrect identification of background areas as the ultrasound probe. This mis-classification

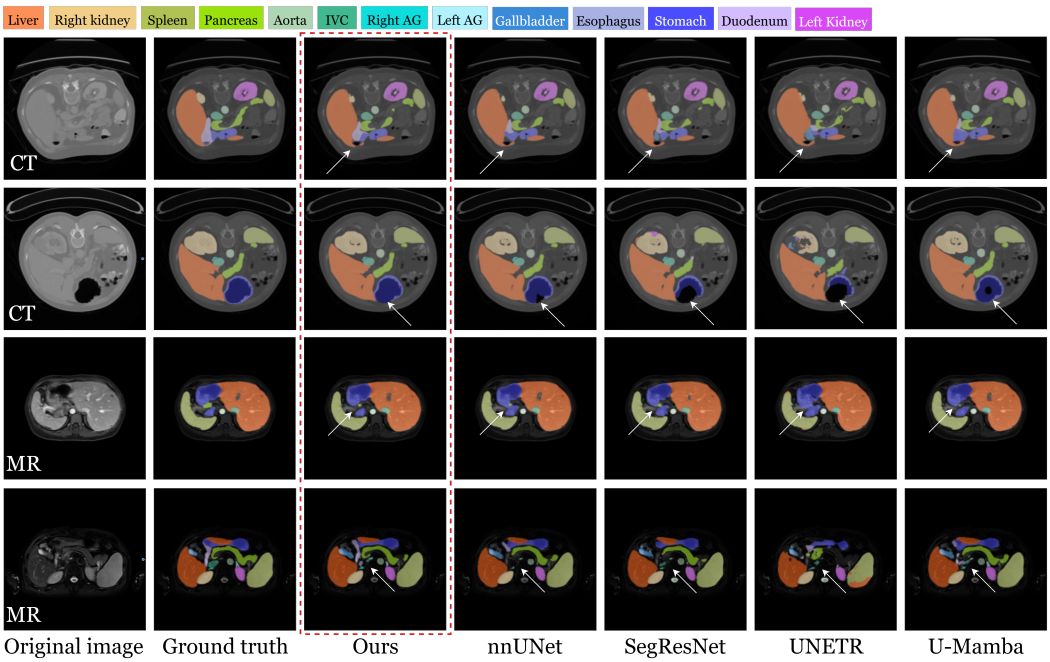
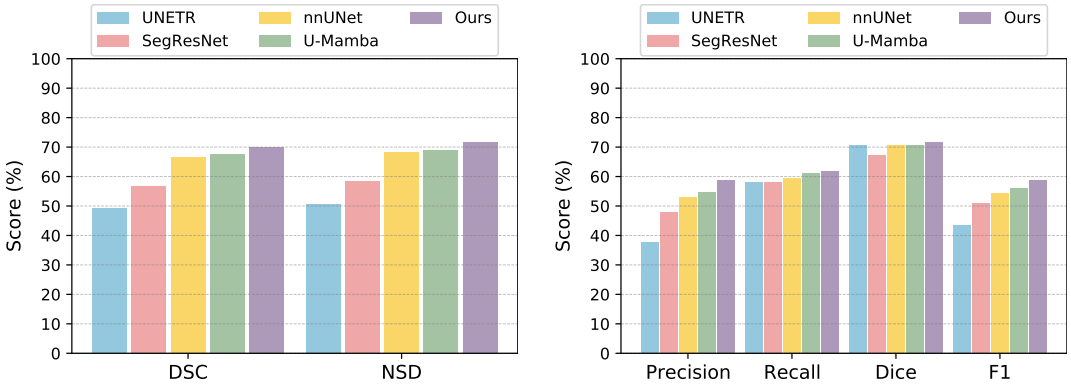


Figure 2: Qualitative visualization of abdominal segmentation results in CT (1st and 2nd rows) and MR scans (3rd and 4th rows). Distinct abdominal organs are represented by various colors, as indicated in the accompanying color bar.



(a) Segmentation results on Endoscopy dataset.

(b) Segmentation results on Microscopy dataset.

Figure 3: The comparison of our Hetero-UNet method and other state-of-the-art methods on the Endoscopy dataset (3a) and the Microscopy dataset (3b). Hetero-UNet continuously outperforms other methods on various metrics.

highlights challenges in distinguishing between the target instruments and background textures, a crucial aspect of medical image segmentation accuracy.

Microscopy Dataset. Figure 3b presents the segmentation performance on the Microscopy dataset. Hetero-UNet continues to outperform all baseline models by margins ranging from 2.9% to 15.4% on F1 score. It also achieves the highest scores, at 58.8% in Precision, 61.9% in Recall and 71.6% in DSC for cell segmentation, underscoring its comprehensive capability across various evaluation metrics. In addition, we still find SSM-based models significantly perform better than their CNN-based and Transformer-based baselines. It shows the great potential for SSM-based models in medical 2D images. In addition, the success of Hetero-UNet in the microscopy dataset, characterized by large image sizes, highlights the necessity for architectures adept at learning long-range dependencies like hybrid SSM-based models in pathology. The qualitative visualization results on the Microscopy dataset is illustrated in the last two rows in Figure 4. Models like nnUNet, SegResNet, UNETR and U-Mamba face challenges in fully delineating cell boundaries. These models often partially recognize

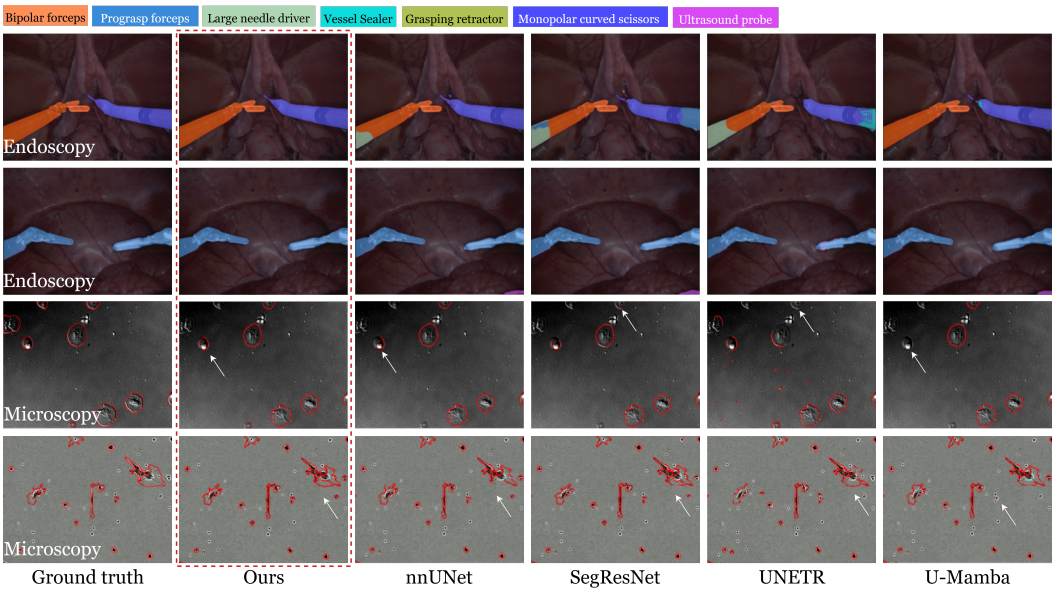


Figure 4: Our qualitative visualization on the Endoscopy and Microscopy datasets. In the Endoscopy images (1st and 2nd rows), different colors, as noted in the corresponding color bar, denote various instruments. For the Microscopy dataset (3rd and 4th rows), the cell boundary is uniformly delineated in red across all methods. Notably, Hetero-UNet exhibits fewer segmentation outliers compared to other methods, showcasing its enhanced accuracy and reliability in complex segmentation tasks.

Table 3: Ablation experiment on each key component in our method. The marker ✓ denotes that a specific component is used.

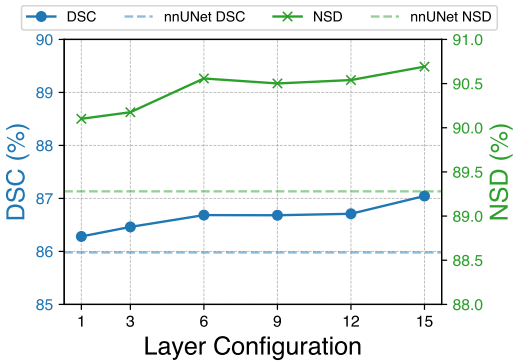
Methods	MLP	MHA	Mamba	Organs in Abdomen CT		Organs in Abdomen MR	
				DSC [%]	NSD [%]	DSC [%]	NSD [%]
nnUNet				86.0	89.3	83.6	90.7
Mamba			✓	86.6	90.2	84.4	91.2
MambaMLP	✓		✓	86.5	89.9	84.6	91.5
MambaFormer		✓	✓	86.4	90.1	84.3	91.3
Hetero-UNet	✓	✓	✓	**86.7**	**90.5**	**84.8**	**91.8**

cells or fail to detect them entirely, leading to incomplete or inaccurate segmentation. In contrast, Hetero-UNet demonstrates clear advantages, accurately capturing the entire boundary of cells and indicating its enhanced ability to comprehend global contexts.

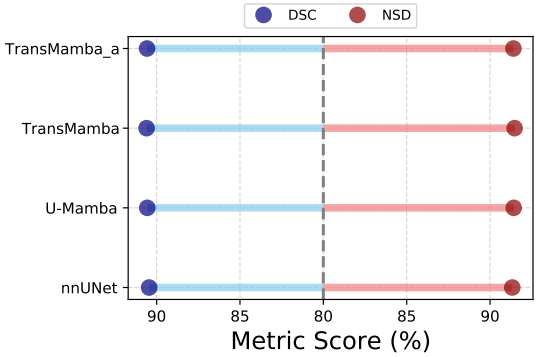
4.3 Ablation Study

In our ablation studies on the Abdomen dataset, we explore the Hetero-UNet method's design by examining: 1) the individual contribution of key components, including MHA blocks, SSM-based Mamba blocks, and MLP layers, to understand their impact on performance, 2) the effect of varying the number of layers in the hybrid TransMamba encoder to find the optimal balance between computational efficiency and segmentation accuracy, and 3) implementing the TransMamba block for each block in the U-Net encoder to further explore the performance of our method.

Effectiveness of Each Component. we conduct ablation studies to assess the significance of the core elements within our Hetero-UNet model—specifically, the Mamba block, MLP block, and MHA block. In Table 3, the presence of a component in the hybrid Mamba encoder is indicated by "MLP", "MHA", and "Mamba" with a ✓ marker. "Hetero-UNet" refers to our fully proposed architecture, integrating all three components. A comparison of the last four rows against the first row, which features nnUNet as the baseline, reveals that Hetero-UNet variants consistently surpass the baseline across both datasets, Abdomen CT and Abdomen MR, using two evaluation metrics, DSC and NSD. This highlights the Mamba's capability to capture long-range dependencies effectively through SSMs. An interesting observation from our experiments is that employing solely MLP layers or a single MHA block with Mamba results in a performance decrease compared to using the pure "Mamba" configuration. However, Hetero-UNet surpasses all other variations, showcasing the



(a) Results of encoder variants with different number of layers.



(b) Results of implementing TransMamba for all encoder blocks.

Figure 5: Figure 5a is the ablation experiment on varying number of layers in hybrid TransMamba encoder. We explore the performance of hybrid encoder with the number of TransMamba layers varying from 1 to 15 on the Abdomen CT dataset. Figure 5b shows results of implementing TransMamba for all encoder blocks on the Abdomen MR dataset, denoted as "TransMamba_a". The default setting of Hetero-UNet is denoted as "TransMamba".

superiority of its hybrid architecture. This architecture integrates attention mechanisms and SSMs, augmented with additional CNN modules, affirming its efficacy in medical image segmentation. In our experimental results, the performance of "MambaFormer" is not as strong as that of "MambaMLP." This suggests that directly and simply combining MHA with Mamba, as discussed in [12] and [27], may not be the optimal approach for integrating attention mechanisms with SSMs. A straightforward combination can sometimes even diminish the capacity of the attention technique. This might be due to the discrepancy between MLP modules and attention mechanisms. MLPs are focused on local, spatially specific features, enhancing detail-oriented processing, whereas attention layers capture global dependencies, prioritizing broad contextual understanding over local specifics. This contrast leads to overly abstract feature learning when SSM and attention modules are continuously combined together without the grounding effect of MLPs. Such abstraction may challenge the decoder's ability to accurately reconstruct specific features, thereby impacting segmentation accuracy.

Encoder Variants with Different Number of Layers. In our study, we delve into how each TransMamba block within the hybrid encoder influences performance by varying the layer count in ablation experiments, focusing solely on altering the number of TransMamba blocks while maintaining the convolutional encoder-decoder and skip connections unchanged. Our exploration on the Abdomen CT dataset reveals that all configurations of our model outperform the baseline nnUNet, which records a DSC of 86.0% and an NSD of 89.3%, as shown in Figure 5a. Notably, even the variant equipped with just one TransMamba block exceeds the baseline, underscoring the effectiveness of our module in facilitating comprehensive information integration across the image, irrespective of the encoder's depth. Furthermore, we observe a positive trend in performance on CT scans as the number of TransMamba blocks increases. This trend aligns with our understanding that our TransMamba block contributes to the model learning process. To strike an optimal balance between computational efficiency and model performance, we settle on a configuration of 12 blocks in the TransMamba encoder as our standard setup.

Implementing TransMamba for All Encoder Blocks. In the default setting of hybrid Hetero-UNet architecture, we use convolution encoder-decoder, with the hybrid TransMamba encoder to replace the bottleneck block in the U-Net encoder. To further explore the performance of the proposed TransMamba block, we introduce a new variant of Hetero-UNet, as each block of U-Net encoder is the TransMamba block, keeping the convolution decoder and skip connections in the default architecture. Then we conduct the experiments on the Abdomen MR dataset and compare the performance of this model with the original Hetero-UNet, the baseline nnUNet and U-Mamba, shown in Figure 5b. "TransMamba_a" segmentes abdominal organs on MR scans with 84.6% in DSC and 91.2% in NSD, outperforming both nnUNet and U-Mamba. This indicates the efficacy of the hybrid encoder over purely convolutional encoder and SSM-based convolutional encoder. However, its performance trails behind the default version of Hetero-UNet, possibly due to the implementation of only one layer of the TransMamba block within each encoder block, suggesting that optimizing the number of hybrid layers could further enhance segmentation accuracy.

5 Conclusion

In this study, we introduce Hetero-UNet, a hybrid segmentation architecture that blends novel SSM-based Transformer encoder to tackle medical image analysis. This architecture is designed to harness both local and global information, and enhance its capacity for learning long-range dependencies. Our results across a variety of tasks and imaging modalities—ranging from abdominal organ segmentation in CT and MR scans to instrument in endoscopy images and cell segmentation in microscopy images—demonstrate Hetero-UNet's potential for superior segmentation performance.

Acknowledgements

In this work, Prof. Lichao Sun was partially supported by the National Science Foundation Grants CRII-2246067, ATD-2427915, NSF POSE-2346158, and Lehigh Grant FRGS00011497.

References

[1] Allan, M., Shvets, A., Kurmann, T., Zhang, Z., Duggal, R., Su, Y.H., Rieke, N., Laina, I., Kalavakonda, N., Bodenstedt, S., et al.: 2017 robotic instrument segmentation challenge. arXiv preprint arXiv:1902.06426 (2019)

[2] Arora, S., Eyuboglu, S., Timalsina, A., Johnson, I., Poli, M., Zou, J., Rudra, A., Ré, C.: Zoology: Measuring and improving recall in efficient language models. arXiv preprint arXiv:2312.04927 (2023)

[3] Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation. In: European conference on computer vision. pp. 205–218. Springer (2022)

[4] Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)

[5] Chen, J., Mei, J., Li, X., Lu, Y., Yu, Q., Wei, Q., Luo, X., Xie, Y., Adeli, E., Wang, Y., et al.: 3d transunet: Advancing medical image segmentation through vision transformers. arXiv preprint arXiv:2310.07781 (2023)

[6] De, S., Smith, S.L., Fernando, A., Botev, A., Cristian-Muraru, G., Gu, A., Haroun, R., Berrada, L., Chen, Y., Srinivasan, S., et al.: Griffin: Mixing gated linear recurrences with local attention for efficient language models. arXiv preprint arXiv:2402.19427 (2024)

[7] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)

[8] Fu, D.Y., Dao, T., Saab, K.K., Thomas, A.W., Rudra, A., Ré, C.: Hungry hungry hippos: Towards language modeling with state space models. arXiv preprint arXiv:2212.14052 (2022)

[9] Goel, K., Gu, A., Donahue, C., Ré, C.: It's raw! audio generation with state-space models. In: International Conference on Machine Learning. pp. 7616–7633. PMLR (2022)

[10] Gong, H., Kang, L., Wang, Y., Wan, X., Li, H.: nnmamba: 3d biomedical image segmentation, classification and landmark detection with state space model. arXiv preprint arXiv:2402.03526 (2024)

[11] Gu, A.: Modeling Sequences with Structured State Spaces. Ph.D. thesis, Stanford University (2023)

[12] Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)

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

[14] Gu, A., Goel, K., Ré, C.: Efficiently modeling long sequences with structured state spaces. arXiv preprint arXiv:2111.00396 (2021)

[15] Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: International MICCAI Brainlesion Workshop. pp. 272–284. Springer (2021)

[16] Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: Unetr: Transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 574–584 (2022)

[17] Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., Han, X., Chen, Y.W., Wu, J.: Unet 3+: A full-scale connected unet for medical image segmentation. In: ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP). pp. 1055–1059. IEEE (2020)

[18] Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)

[19] Ji, Y., Bai, H., Ge, C., Yang, J., Zhu, Y., Zhang, R., Li, Z., Zhanng, L., Ma, W., Wan, X., et al.: Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. Advances in Neural Information Processing Systems **35**, 36722–36732 (2022)

[20] Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., Liu, Y.: Vmamba: Visual state space model. arXiv preprint arXiv:2401.10166 (2024)

[21] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)

[22] Ma, J., Li, F., Wang, B.: U-mamba: Enhancing long-range dependency for biomedical image segmentation. arXiv preprint arXiv:2401.04722 (2024)

[23] Ma, J., Xie, R., Ayyadhury, S., Ge, C., Gupta, A., Gupta, R., Gu, S., Zhang, Y., Lee, G., Kim, J., et al.: The multi-modality cell segmentation challenge: Towards universal solutions. arXiv preprint arXiv:2308.05864 (2023)

[24] Ma, J., Zhang, Y., Gu, S., Ge, C., Ma, S., Young, A., Zhu, C., Meng, K., Yang, X., Huang, Z., et al.: Unleashing the strengths of unlabeled data in pan-cancer abdominal organ quantification: the flare22 challenge. arXiv preprint arXiv:2308.05862 (2023)

[25] Mehta, H., Gupta, A., Cutkosky, A., Neyshabur, B.: Long range language modeling via gated state spaces. arXiv preprint arXiv:2206.13947 (2022)

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

[27] Park, J., Park, J., Xiong, Z., Lee, N., Cho, J., Oymak, S., Lee, K., Papailiopoulos, D.: Can mamba learn how to learn? a comparative study on in-context learning tasks. arXiv preprint arXiv:2402.04248 (2024)

[28] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015)

[29] Ruan, J., Xiang, S.: Vm-unet: Vision mamba unet for medical image segmentation. arXiv preprint arXiv:2402.02491 (2024)

[30] Wang, J., Zhu, W., Wang, P., Yu, X., Liu, L., Omar, M., Hamid, R.: Selective structured state-spaces for long-form video understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6387–6397 (2023)

[31] Wang, Z., Zheng, J.Q., Zhang, Y., Cui, G., Li, L.: Mamba-unet: Unet-like pure visual mamba for medical image segmentation. arXiv preprint arXiv:2402.05079 (2024)

[32] Xing, Z., Ye, T., Yang, Y., Liu, G., Zhu, L.: Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation. arXiv preprint arXiv:2401.13560 (2024)

[33] Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: Unet++: A nested u-net architecture for medical image segmentation. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4. pp. 3–11. Springer (2018)

[34] Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. IEEE transactions on medical imaging **39**(6), 1856–1867 (2019)

[35] Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., Wang, X.: Vision mamba: Efficient visual representation learning with bidirectional state space model. arXiv preprint arXiv:2401.09417 (2024)

[36] Zuo, S., Liu, X., Jiao, J., Charles, D., Manavoglu, E., Zhao, T., Gao, J.: Efficient long sequence modeling via state space augmented transformer. arXiv preprint arXiv:2212.08136 (2022)