

# SegMaFormer: A Hybrid State-Space and Transformer Model for Efficient Segmentation

Duy D. Nguyen<sup>1,2</sup>, Phat T. Tran-Truong<sup>1,2</sup> (✉) 

<sup>1</sup> Department of Software Engineering, Faculty of Computer Science and Engineering, Ho Chi Minh City University of Technology (HCMUT), 268 Ly Thuong Kiet Street, Dien Hong Ward, Ho Chi Minh City, Vietnam

<sup>2</sup> Vietnam National University Ho Chi Minh City, Linh Trung Ward, Ho Chi Minh City, Vietnam

(✉) phatttt@hcmut.edu.vn

**Abstract.** The advent of Transformer and Mamba-based architectures has significantly advanced 3D medical image segmentation by enabling global contextual modeling, a capability traditionally limited in Convolutional Neural Networks (CNNs). However, state-of-the-art Transformer models often entail substantial computational complexity and parameter counts, which is particularly prohibitive for volumetric data and further exacerbated by the limited availability of annotated medical imaging datasets. To address these limitations, this work introduces SegMaFormer, a lightweight hybrid architecture that synergizes Mamba and Transformer modules within a hierarchical volumetric encoder for efficient long-range dependency modeling. The model strategically employs Mamba-based layers in early, high-resolution stages to reduce computational overhead while capturing essential spatial context, and reserves self-attention mechanisms for later, lower-resolution stages to refine feature representation. This design is augmented with generalized rotary position embeddings to enhance spatial awareness. Despite its compact structure, SegMaFormer achieves competitive performance on three public benchmarks (Synapse, BraTS, and ACDC), matching the Dice coefficient of significantly larger models. Empirically, our approach reduces parameters by up to 75× and substantially decreases FLOPs compared to current state-of-the-art models, establishing an efficient and high-performing solution for 3D medical image segmentation.

**Keywords:** State-space Model · Transformer Model · Supervised Learning · Feature Extraction · Medical and Public Health Application

## 1 Introduction

The proliferation of deep learning technologies is revolutionizing healthcare systems by providing robust capabilities for learning and analyzing complex patterns within medical data. A critical component of this advancement is semantic 3D volumetric segmentation, a foundational task in medical image analysis. This

capability is indispensable for various clinical applications, including tumor detection and multi-organ identification, which are vital for accurate diagnosis and treatment planning.

At the dawn of the deep learning era, convolutional-neural-network-driven encoder–decoder architectures [19,23] became the dominant paradigm for medical image segmentation. Nonetheless, because their receptive fields are inherently limited due to the characteristics of CNN, these models cannot capture sufficient global context. Transformer models, utilizing the powerful attention mechanism [25], have significantly revolutionized research in numerous fields, including computer vision and natural language processing. By adopting attention, vision models from ViT [6] can effectively capture long-range global dependencies through their attention layers [12,32], a capability that fundamentally distinguishes them from convolutional architectures, whose operations are constrained by local inductive biases.

To address the locality limitations of CNNs, many works integrate convolutional layers with Transformer blocks to capture global context. TransUNet [5] combines a ViT encoder with a CNN decoder, while UNETR [12] applies Transformer layers directly to volumetric inputs. Subsequent models such as SwinUNETR [11] and nnFormer [32] introduce hierarchical and window-based attention for 3D data. Lighter hybrids focus attention on selected stages, as in TransBTS [28] and CoTr [30], which apply attention only at coarse scales. Despite their theoretical advantages, Transformer-based models often struggle to match the generalization performance achieved by their CNN counterparts [15,26]. Indeed, recent large-scale analyses [26] show that many CNN-Transformer hybrids derive most of their performance from convolutional components, and attention-heavy designs often underperform strong CNN baselines such as nnU-Net. These findings highlight architectural limitations and emphasize the need for improved positional embeddings, as demonstrated by PRIMUS [26].

Although Transformers face challenges compared to CNN-based SOTA models [15], their strong sequence-modeling capability makes them suited for medical image segmentation. Medical data is often multi-modal—combining scans, clinical notes, and representing images as tokens enables Transformers to integrate these heterogeneous sources within a unified model. Moreover, Transformer architectures can be computationally efficient, often requiring fewer parameters and FLOPs than CNN-based models [15,16], making them attractive for resource-constrained medical settings.

State-space models (SSMs) have emerged as efficient alternatives for long-range sequence modeling, offering *linear* complexity. Structured SSMs [9] and the selective Mamba architecture [8] enable hardware-efficient sequence mixing, inspiring vision variants such as VMamba [18]. In medical imaging, U-Mamba [21] integrates Mamba blocks into a U-Net architecture to combine CNN-based local features with SSM-driven global dependencies. Nevertheless, Isensee et al. [15] report that Mamba layers alone contribute minimal gains without careful design. Swin-UMamba [17] addresses this by replacing Swin attention with

Mamba blocks and leveraging ImageNet pretraining, achieving improvements over CNNs, ViTs, and prior Mamba variants across multiple datasets.

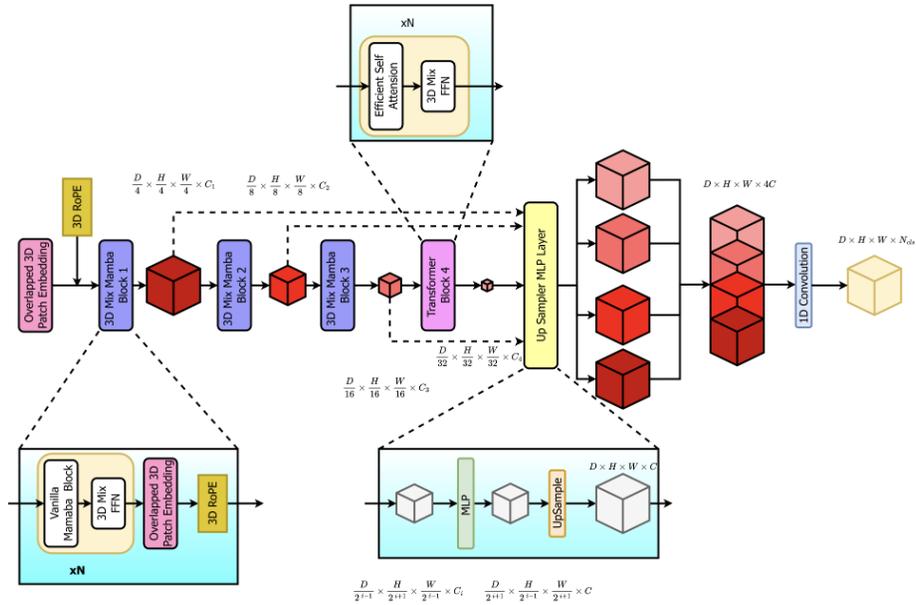
In many public hospitals, especially in developing regions, limited computational capacity makes large segmentation models impractical. This underscores the need for lightweight architectures that maintain high accuracy while reducing parameter counts and FLOPs, enabling reliable 3D segmentation on modest hardware. To resolve these challenges, we propose SegMaFormer, a lightweight hybrid architecture that combines the sequence-modeling efficiency of Mamba[8] with the global context modeling of Transformers[25] for 3D medical image segmentation. Building upon advances in SegFormer-style tokenization [29,22], SegMaFormer preserves the benefits of patch-based volumetric representation while substantially reducing complexity.

In this work, we present **SegMaFormer**, a highly efficient hybrid state-space and transformer architecture for 3D medical image segmentation. Our contributions are summarized as follows:

- We introduce a Hybrid Transformer-Mamba encoder that combines Mamba-based state-space layers for early-stage sequence mixing with self-attention operating on compact, low-resolution tokens. This structure captures global anatomical dependencies while significantly reducing the computational burden of 3D attention, especially given that volumetric medical images expand into very long token sequences once converted into patches.
- We enhance the baseline 3D overlapped patch embedding by integrating 3D Rotary Positional embedding (3D-RoPE), which supplies rotation-consistent positional cues while maintaining local voxel structure, boosting the performance of both Mamba and Transformer. This improves the model’s ability to capture spatial relationships in complex anatomical regions.
- With only 2M parameters and 15 GFLOPs, SegMaFormer achieves accuracy comparable to exceeding much larger CNN and Transformer architectures, offering up to  $75\times$  fewer parameters and substantially lower computational complexity. Notably, SegMaFormer consistently surpasses the SegFormer3D[22] baseline across all three benchmark datasets without pretraining progress, demonstrating that strong segmentation performance can be achieved without heavy model capacity.

## 2 Methodology

While the attention mechanism and transformer architecture have shown a significant impact on the performance of semantic segmentation in medical images, transformer-based models have also demonstrated better computational efficiency compared to conventional CNN-based models. As well as Segformer[22], the base model provides an effective and lightweight framework that achieves strong segmentation performance with considerably reduced model parameters and computational complexity. However, the original Segformer baseline still suffers from quadratic complexity from the self-attention mechanism, which constrains its scalability for high-resolution or volumetric medical data and poses



**Fig. 1.** Architecture Overview: The model input is a 3D volume  $\mathbb{R}^{C \times D \times H \times W}$ . A four-stage hierarchical Mamba-Transformer Block is adopted to derive multiscale volumetric representations. These features are then upsampled and fused by an all-MLP decoder, integrating local and global attention cues to generate the final segmentation mask.

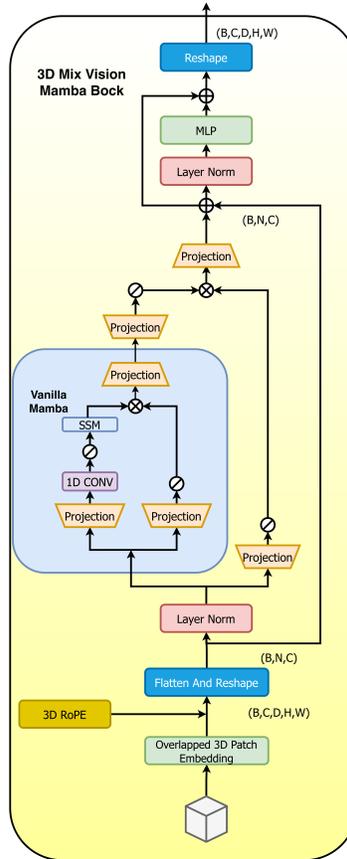
inference-related challenges on resource-limited or low-end computational devices.

**Table 1.** Relative comparison of model complexity in terms of parameters and GFLOPs.

Architecture	Params(M)	GFLOPs
nnFormer[32]	150.5	213.4
TransUNet[5]	96.07	88.91
UNETR[12]	92.49	75.76
SwinUNETR[11]	62.83	384.2
Segformer3D[22]	4.51	17.5
<b>SegMaFormer (ours)</b>	<b>2.02</b>	<b>15.2</b>

With the aim of alleviating this limitation, a Mamba block [8] is incorporated into the original Segformer encoder to create a hybrid transformer-Mamba model, which in turn enhances computational efficiency while maintaining SOTA performance.

**Embedding:** This paper continues to use the overlapped 3D patch embedding from the baseline model. While rotary position embedding has shown clear benefits for models requiring robust sequence-dependency modeling such as Mamba and Transformers, we adopt 3D-RoPE after the embedding, following the implement idea from [24,7].



**Fig. 2.** 3D Mix Vision Mamba Block with minimal reliance on convolutional layers.

**Encoder:** Processing volumetric medical data within attention-based architectures is inherently expensive because the tokenization of 3D voxel grids leads to an exponential growth in sequence length relative to spatial resolution. When a volumetric scan of resolution  $D \times H \times W$  is decomposed into tokens, the total sequence length becomes  $N = D \cdot H \cdot W$ . Standard self-attention mechanisms require the computation of pairwise correlations among all tokens, resulting in

a computational complexity of  $\mathcal{O}(N^2 \cdot d)$ , where  $d$  denotes the feature embedding dimension [25]. In 3D imaging, a moderate input resolutions (e.g.,  $128^3$  yield  $N = 32,768$  tokens), and thus the quadratic attention term  $\mathcal{O}(N^2 \cdot d)$  increases prohibitively large, leading to excessive memory consumption and computational latency. To overcome the quadratic complexity of full self-attention, The proposed Transformer-Mamba hybrid encoder modified the representation hierarchy by integrating a state-space model in the early stages and self-attention blocks in the last stages.

The Mamba-based state-space layers model perform sequence mixing with a computational cost of  $\mathcal{O}(N/r \cdot d^2)$  [8], where  $N$  is the token count,  $d$  the embedding dimension, and  $r$  the spatial-reduction ratio. Otherwise, in hierarchical encoder architectures, the later stages operate on feature representations with a smaller spatial resolution but a larger embedding dimension due to progressive downsampling. Under these hierarchical conditions, self-attention becomes particularly effective, as the reduced sequence length  $N$  significantly lowers its quadratic computational burden, while the expanded embedding dimension  $d$  enhances representational capacity for rich global context modeling. In the case of volumetric medical images, the spatial resolution after progressive downsampling is often considerably smaller than the embedding dimension ( $N \ll d$ ). Consequently, applying self-attention at these deeper stages is computationally more efficient than in early layers and enables the model to capture global anatomical dependencies without incurring the prohibitive  $\mathcal{O}(N^2 \cdot d)$  cost associated with full-resolution attention.

This paper proposes a 3D Mamba-based block that integrates long-range sequence modeling with efficient channel mixing. Given a volumetric input feature map  $X$ , the structure of the block is illustrated in Fig. 2. Given a volumetric

input feature map  $X$ , the block is formulated as

$$\begin{aligned}
\mathbf{X}_{\text{seq}} &= \text{Flatten}(\mathbf{X}), \\
\hat{\mathbf{X}} &= \text{LayerNorm}(\mathbf{X}_{\text{seq}}), \\
\mathbf{G} &= \text{Mamba}(\hat{\mathbf{X}}), \\
\hat{\mathbf{G}} &= \text{SiLU}(\mathbf{W}_a \mathbf{G}), \\
\mathbf{H} &= \text{SiLU}(\mathbf{W}_b \hat{\mathbf{X}}), \\
\mathbf{F} &= \mathbf{W}_p \left( \hat{\mathbf{G}} \odot \mathbf{H} \right), \\
\mathbf{X}^{(1)} &= \mathbf{X}_{\text{seq}} + \mathbf{F}, \\
\mathbf{Z} &= \text{LayerNorm}(\mathbf{X}^{(1)}), \\
\mathbf{X}^{(2)} &= \mathbf{X}^{(1)} + \text{MLP}(\mathbf{Z}), \\
\mathbf{X}_{\text{out}} &= \text{Reshape}(\mathbf{X}^{(2)}).
\end{aligned} \tag{1}$$

This block employs a GRU-inspired gating mechanism to dynamically regulate the information flowing through the Mamba layer, thereby preserving both data-dependent local biases and the long-range dependencies modeled by Mamba, while simultaneously accelerating convergence during training.

Traditional self-attention incurs a quadratic cost  $\mathcal{O}(N^2)$ , which becomes prohibitive for long 3D sequences. Efficient attention [27,29] reduces this cost by reshaping and projecting the keys,

$$\begin{aligned}
\hat{\mathbf{K}} &= \text{Reshape}\left(\frac{N}{r}, C \cdot r\right) \mathbf{K}, \\
\mathbf{K} &= \text{Linear}(C \cdot r, C) \hat{\mathbf{K}}.
\end{aligned}$$

lowering the complexity to  $\mathcal{O}(N^2/r)$  in Transformer and  $\mathcal{O}(N/r)$  in Mamba Block. This paper retain this formulation but apply a fixed reduction ratio  $r = 1$  across all encoder stages to evaluate the full potential of Mamba without relying on Convolution components, which can diminish the effectiveness of both Mamba and Transformer models.

**Decoder:** The decoding stage is in encoder-decoder architectures widely used for medical image segmentation, such as UNet and its variants [23,14]. However, instead of the typical successive 3D convolutions, a lightweight decoder based on linear layers is efficient, avoids over-parameterization, and effectively reconstructs volumetric features.

The proposed architecture follows the baseline paradigm [22,29] of fusing multi-scale features. As shown below, this process consists of four steps:

$$\tilde{\mathbf{X}}_i = \text{Linear}(C_i, C)(\mathbf{X}_i), \quad \forall i \in \{1, \dots, 4\}, \quad (2)$$

$$\mathbf{X}_i^{\text{up}} = \text{Upsample}_{s_i}(\tilde{\mathbf{X}}_i), \quad \forall i \in \{1, \dots, 4\}, \quad (3)$$

$$\mathbf{Z} = \text{Linear}(4C, C)(\text{Concat}(\mathbf{X}_1^{\text{up}}, \dots, \mathbf{X}_4^{\text{up}})), \quad (4)$$

$$\mathbf{Y} = \text{Linear}(C, N_{\text{cls}})(\mathbf{Z}). \quad (5)$$

Features from each encoder stage are aggregated and projected to a unified dimensionality, similar to Unet and its variants. After standardization, the feature maps are upsampled, concatenated, and fused through a linear transformation. The resulting representation is finally passed through a linear prediction head (equivalent to a 3D  $1 \times 1 \times 1$  convolution) to generate the segmentation masks. Furthermore, this work implements optional Deep Supervision (DS) auxiliary heads, similar to SOTA models. However, for tasks involving small anatomical structures, it can be observed that such auxiliary supervision does not improve performance and may even negatively impact fine-grained feature learning. Moreover, the weight of the deep supervision loss requires careful tuning, as improper balancing can lead to degraded training stability or suboptimal feature learning.

### 3 Experimental Result

Following the standard baseline model[22], this work adopts identical datasets and evaluation strategy to enable fair comparison between networks architectures. The proposed model is trained and assessed on three commonly used dataset benchmarks without relying on any external pretraining data. The Brain Tumor Segmentation (BraTS) [1], Synapse Multi-Organ Segmentation (Synapse) [10], and Automatic Cardiac Diagnosis (ACDC) [2] datasets are evaluated in sequence.

All experiments are conducted on a dual NVIDIA RTX 4060Ti GPU using PyTorch version 2.8. We adopt the nnUnet[14] framework setup for training, validation, and prediction. A weighted combination of Dice and Cross-Entropy losses is utilized to stabilize optimization and improve convergence. The learning rate is linearly warmed up from min learning rate to the initial learning rate and then decayed using cosine annealing. Training uses the AdamW optimizer[20] with a base learning rate of  $3e-4$ .

#### 3.1 Brain Tumor Segmentation (BraTs)

This research utilizes the standard BraTS dataset from [1], which contains 484 MRI scans of brain tumors in BraTS 2016 and 2017 Challenges from 19 hospitals, across four modals (FLAIR, T1w, T1gd, T2w). The original annotation masks contain three tumor subregions, which are edema (ED), enhancing tumor (ET), and non-enhancing tumor (NET). Following standard benchmarking practices,

these labels are reorganized into whole tumor (WT), enhancing tumor (ET), and tumor core (TC) for comparison with Transformer-based methods. In this benchmark, the proposed model completely outperforms the baseline model. This demonstrates the representation-learning capability of the efficient self-attention module and the Mamba component, both of which effectively analyze the full sequence of patches.

**Table 2.** Comparison of segmentation performance across methods on the BraTS benchmark. Our model achieves competitive accuracy while using significantly fewer parameters and small FLOPS.

Methods	Params(M)	Avg(%)	Whole Tumor	Enhancing Tumor	Tumor Core
nnFormer [32]	150.5	86.4	91.3	81.8	86.0
<b>Ours</b>	<b>2.0</b>	<b>83.79</b>	<b>91.0</b>	<b>76.2</b>	<b>84.16</b>
SegFormer3D [22]	4.5	82.1	89.9	74.2	82.2
UNETR [12]	92.49	71.1	78.9	58.5	76.1
TransBTS [28]	–	69.6	77.9	57.4	73.5
CoTr [30]	41.9	68.3	74.6	55.7	74.8
CoTr w/o CNN Enc. [30]	–	64.4	71.2	52.3	69.8
TransUNet [5]	96.07	64.4	70.6	54.2	68.4

### 3.2 Multi-Organ CT Segmentation (Synapse)

The Synapse dataset consists of 30 abdominal CT scans with annotations for multiple organs. Following the data split used in prior work, 18 scans are used for training and 12 for testing. Model performance is evaluated using the Dice score across eight organs, including the aorta, gallbladder, spleen, kidneys, liver, pancreas, and stomach. Table 3 presents the quantitative results, demonstrating how our method compares against previous architectures in terms of accuracy, parameter efficiency. The result indices that our model achieves third-best performance, outperforming the baseline model and many larger models. High-capacity architectures such as U-Mamba [21] and nnFormer [32] achieve slightly higher averages, respectively, but require over  $75\times$  more parameters. Moreover, Mamba-based models, such as U-Mamba and our approach, demonstrate enhanced long-range dependency modeling, which is particularly beneficial for large-scale organs. Likewise, the Mamba-based architectures have witnessed a notable performance drop in small organs.

### 3.3 Automated Cardiac Diagnosis (ACDC)

The ACDC dataset [2] contains imaging data from 100 patients and is widely used to evaluate 3D segmentation methods for the left ventricle (LV), right ventricle (RV), and myocardium (Myo). Although resampling to a  $1\times 1\times 1$  mm isotropic resolution has been shown to be an effective preprocessing strategy validated by Wald et al. and Isensee et al. [26,15], we refrain from using this

**Table 3.** Within the Synapse benchmark, models are ranked by their average Dice scores across all organ classes. The SegMaFormer performs exceptionally well, outperforming numerous established baselines and ranking just behind nnFormer and U-Mamba, even though these models use more than  $75\times$  the parameters and require substantially higher FLOPs.

Methods	Params(M)	Avg(%)	AOR	LIV	LKID	RKID	GAL	PAN	SPL	STO
U-Mamba [21]	172.63	87.98	90.8	96.90	94.6	94.5	73.80	79.3	95.80	81.70
nnFormer [32]	150.5	86.57	92.04	96.84	86.57	86.25	70.17	83.35	90.51	86.83
<b>Ours</b>	<b>2.0</b>	<b>83.33</b>	<b>89.98</b>	<b>96.47</b>	<b>90.49</b>	<b>90.53</b>	<b>57.29</b>	<b>70.57</b>	<b>93.03</b>	<b>78.70</b>
SegFormer3D[22]	4.5	82.15	90.43	95.68	86.53	86.13	55.26	73.06	89.02	81.12
MISSFormer [13]	-	81.96	86.99	94.41	85.21	82.00	68.65	65.67	91.92	80.81
UNETR [12]	92.49	79.56	89.99	94.46	85.66	84.80	60.56	59.25	87.81	73.99
SwinUNet [3]	-	79.13	85.47	94.29	83.28	79.61	66.53	56.58	90.66	76.60
LeViT-UNet-384 [31]	52.17	78.53	87.33	93.11	84.61	80.25	62.23	59.07	88.86	72.76
TransClaw U-Net [4]	-	78.09	85.87	94.28	84.83	79.36	61.38	57.65	87.74	73.55
TransUNet [5]	96.07	77.48	87.23	94.08	81.87	77.02	63.16	55.86	85.08	75.62

technique in order to ensure a fair comparison with prior benchmark studies, specially the baseline SegFormer3D [22] model. As shown in Table 4, illustrating quantitative benchmarks of models, Mamba-based architectures typically experience a performance drop on this dataset due to their weaker inherent local inductive and spatial bias, a phenomenon also observed in recent re-evaluations of segmentation backbones [15], where U-Mamba underperforms its non-Mamba counterpart. Adopting 3DRoPE into our design effectively strengthens local spatial awareness, enabling our model to recover accuracy and achieve competitive performance while keeping 2% margin of the SOTA performance with models on average  $10\times$  higher in parameter count in computational complexity. These results further indicate that ViT-inspired tokenization strategies, which transform images into patches, when combined with appropriately designed Mamba and Transformer components, can fully replace CNNs and provide both higher efficiency and competitive accuracy.

**Table 4.** Comparison on the ACDC benchmark. Models are ranked by their average Dice scores across RV, Myo, and LV. Our method achieves competitive performance with significantly fewer parameters.

Methods	Params(M)	Avg(%)	RV	Myo	LV
Primus-S [26]	23.9	92.46	-	-	-
nnFormer [32]	150.5	92.06	90.94	89.58	95.65
<b>Ours</b>	<b>2.0</b>	<b>91.11</b>	<b>90.06</b>	<b>89.1</b>	<b>94.14</b>
SegFormer3D [22]	4.5	90.96	88.5	88.86	95.53
LeViT-UNet-384 [31]	52.17	90.32	89.55	87.64	93.76
SwinUNet [3]	-	90.00	88.55	85.62	95.83
TransUNet [5]	96.07	89.71	88.86	84.54	95.73
UNETR [12]	92.49	88.61	85.29	86.52	94.02
R50-VIT-CUP [5]	86.00	87.57	86.07	81.88	94.75
VIT-CUP [5]	86.00	81.45	81.46	70.71	92.18

## 4 Limitation and Conclusion

In this work, an efficient lightweight model was proposed to address the growing computational and generalization challenges of 3D medical image segmentation. By integrating a 3D-RoPE in patch merging embedding, Mamba-based state-space layers in the early high-resolution stages, and applying self-attention only at deeper, low-resolution scales, the proposed framework achieves an effective balance between accuracy and complexity. This design enables substantial reductions in parameters and FLOPs while maintaining competitive performance across multiple benchmarks, demonstrating that compact architectures can remain highly effective even in data-limited medical imaging scenarios. Furthermore, these findings suggest that the parameter number requires careful consideration, as over-parameterization does not lead to significant performance gains. Despite its promising results, the model still presents opportunities for improvement. The performance on small organs and sharp anatomical boundaries indicates room for further refinement, particularly in capturing fine-grained spatial details under limited training data. The proposed approach leverages lightweight hybrid architectures and provides a practical foundation for scalable, resource-efficient, and clinically deployable 3D medical image segmentation.

## Acknowledgement

We acknowledge the support of time and facilities from Ho Chi Minh City University of Technology (HCMUT), VNU-HCM for this study.

## References

1. Antonelli, Michela, e.a.: The Medical Segmentation Decathlon. *Nature Communications* **13**(1), 4128 (Jul 2022)
2. Bernard, Olivier, e.a.: Deep Learning Techniques for Automatic MRI Cardiac Multi-Structures Segmentation and Diagnosis: Is the Problem Solved? *IEEE Transactions on Medical Imaging* **37**(11), 2514–2525 (Nov 2018)
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 (May 2021), arXiv:2105.05537 [eess]
4. Chang, Y., Menghan, H., Guangtao, Z., Xiao-Ping, Z.: TransClaw U-Net: Claw U-Net with Transformers for Medical Image Segmentation (Jul 2021), arXiv:2107.05188 [cs]
5. 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 (Feb 2021), arXiv:2102.04306 [cs]
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale (Jun 2021), arXiv:2010.11929 [cs]

7. Gervet, T., Xian, Z., Gkanatsios, N., Fragkiadaki, K.: Act3D: 3D Feature Field Transformers for Multi-Task Robotic Manipulation (Oct 2023), arXiv:2306.17817 [cs]
8. Gu, A., Dao, T.: Mamba: Linear-Time Sequence Modeling with Selective State Spaces (May 2024), arXiv:2312.00752 [cs]
9. Gu, A., Goel, K., Ré, C.: Efficiently Modeling Long Sequences with Structured State Spaces (Aug 2022), arXiv:2111.00396 [cs]
10. Harrigr: Segmentation Outside the Cranial Vault Challenge (2015)
11. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H., Xu, D.: Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images (Jan 2022), arXiv:2201.01266 [eess]
12. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H., Xu, D.: UNETR: Transformers for 3D Medical Image Segmentation (Oct 2021), arXiv:2103.10504 [eess]
13. Huang, X., Deng, Z., Li, D., Yuan, X.: MISSFormer: An Effective Medical Image Segmentation Transformer (Dec 2021), arXiv:2109.07162 [cs]
14. Isensee, F., Ulrich, C., Wald, T., Maier-Hein, K.H.: Extending nnU-Net is all you need (Aug 2022), arXiv:2208.10791 [eess]
15. Isensee, F., Wald, T., Ulrich, C., Baumgartner, M., Roy, S., Maier-Hein, K., Jaeger, P.F.: nnU-Net Revisited: A Call for Rigorous Validation in 3D Medical Image Segmentation (Jul 2024), arXiv:2404.09556 [cs]
16. Kuang, H., Wang, Y., Tan, X., Yang, J., Sun, J., Liu, J., Qiu, W., Zhang, J., Zhang, J., Yang, C., Wang, J., Chen, Y.: LW-CTrans: A lightweight hybrid network of CNN and Transformer for 3D medical image segmentation. *Medical Image Analysis* **102**, 103545 (2025)
17. Liu, J., Yang, H., Zhou, H.Y., Xi, Y., Yu, L., Yu, Y., Liang, Y., Shi, G., Zhang, S., Zheng, H., Wang, S.: Swin-UMamba: Mamba-based UNet with ImageNet-based pretraining (Mar 2024), arXiv:2402.03302 [eess]
18. Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., Jiao, J., Liu, Y.: VMamba: Visual State Space Model (Dec 2024), arXiv:2401.10166 [cs]
19. Long, J., Shelhamer, E., Darrell, T.: Fully Convolutional Networks for Semantic Segmentation (Mar 2015), arXiv:1411.4038 [cs]
20. Loshchilov, I., Hutter, F.: Decoupled Weight Decay Regularization (Jan 2019), arXiv:1711.05101 [cs]
21. Ma, J., Li, F., Wang, B.: U-Mamba: Enhancing Long-range Dependency for Biomedical Image Segmentation (Jan 2024), arXiv:2401.04722 [eess]
22. Perera, S., Navard, P., Yilmaz, A.: SegFormer3D: an Efficient Transformer for 3D Medical Image Segmentation (Apr 2024), arXiv:2404.10156 [cs]
23. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation (May 2015), arXiv:1505.04597 [cs]
24. Su, J., Lu, Y., Pan, S., Murtadha, A., Wen, B., Liu, Y.: RoFormer: Enhanced Transformer with Rotary Position Embedding (Nov 2023), arXiv:2104.09864 [cs]
25. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention Is All You Need (Aug 2023), arXiv:1706.03762 [cs]
26. Wald, T., Roy, S., Isensee, F., Ulrich, C., Ziegler, S., Trofimova, D., Stock, R., Baumgartner, M., Köhler, G., Maier-Hein, K.: Primus: Enforcing Attention Usage for 3D Medical Image Segmentation (Mar 2025), arXiv:2503.01835 [cs]
27. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions (Aug 2021), arXiv:2102.12122 [cs]

28. Wang, W., Chen, C., Ding, M., Li, J., Yu, H., Zha, S.: TransBTS: Multimodal Brain Tumor Segmentation Using Transformer (Jun 2021), arXiv:2103.04430 [cs]
29. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers (Oct 2021), arXiv:2105.15203 [cs]
30. Xie, Y., Zhang, J., Shen, C., Xia, Y.: CoTr: Efficiently Bridging CNN and Transformer for 3D Medical Image Segmentation (Mar 2021), arXiv:2103.03024 [cs]
31. Xu, G., Wu, X., Zhang, X., He, X.: LeViT-UNet: Make Faster Encoders with Transformer for Medical Image Segmentation (Jul 2021), arXiv:2107.08623 [cs]
32. Zhou, H.Y., Guo, J., Zhang, Y., Yu, L., Wang, L., Yu, Y.: nnFormer: Interleaved Transformer for Volumetric Segmentation (Feb 2022)