

SegMaFormer: A Hybrid State-Space and Transformer Model for Efficient 3D Medical Image Segmentation

Duy D. Nguyen Phat T. Tran-Truong Xuan-Bach Le
Ho Chi Minh City University of Technology (HCMUT), VNU-HCM, Vietnam

Abstract

The advent of Transformer- and Mamba-based architectures has significantly advanced 3D medical image segmentation by enabling long-range contextual modeling, a capability traditionally limited in convolutional neural networks (CNNs). However, state-of-the-art Transformer models often incur substantial computational complexity and parameter overhead, which becomes particularly prohibitive for volumetric inputs and is further exacerbated by the limited availability of annotated medical imaging data. To address these limitations, this work introduces SegMaFormer, a lightweight stage-aware hybrid architecture that integrates Mamba, Gated Hybrid Conv-SSM (GHConv-SSM), and efficient self-attention within a hierarchical volumetric encoder. Specifically, SegMaFormer employs a Mamba block at the first high-resolution stage to capture dense spatial dependencies with reduced overhead, uses GHConv-SSM blocks at intermediate stages to adaptively fuse state-space contextual modeling with local convolutional priors, and reserves self-attention for the final low-resolution stage to refine compact semantic representations. This design is further enhanced with generalized rotary position embeddings to improve spatial awareness in 3D feature learning. Despite its compact structure, SegMaFormer achieves competitive performance on three public benchmarks, including BTCV, BraTS, and ACDC, while using substantially fewer parameters and lower computational cost than larger Transformer- and Mamba-based counterparts. These results demonstrate that stage-aware allocation of Mamba, gated Conv-SSM fusion, and attention provides an effective accuracy-efficiency trade-off for 3D medical image segmentation. Code: <https://anonymous.4open.science/r/nnUNet-Custom-4B43/>

1. Introduction

Deep learning has become a central tool for medical image analysis, enabling models to learn complex visual patterns from volumetric data. Among these tasks, 3D semantic segmentation is fundamental because it supports applications

such as tumor delineation, organ identification, diagnosis, and treatment planning. However, accurate 3D segmentation remains challenging due to the high computational cost of volumetric inputs and the limited availability of annotated medical images.

Convolutional encoder-decoder architectures [21, 24] have long been the dominant paradigm for medical image segmentation because they provide strong local inductive bias and efficient feature extraction. Purely convolutional models, however, may require deep hierarchies to capture long-range context. Transformer models address this limitation through self-attention [30], enabling direct interaction among distant tokens, and Vision Transformer variants [6, 10, 38] have therefore been explored for 3D medical segmentation, where global anatomical relationships can matter.

To address the locality of CNNs, many works combine convolutional layers with Transformer blocks. TransUNet [5] pairs a ViT encoder with a CNN decoder, while UNETR [10] applies Transformer layers directly to volumetric inputs. SwinUNETR [11] and nnFormer [38] introduce hierarchical and window-based attention for 3D data, whereas lighter hybrids such as TransBTS [32] and CoTr [35] restrict attention to coarse scales. Despite these advances, Transformer-heavy designs often fail to match the generalization of strong CNN baselines such as nnU-Net in limited-data settings [14, 31]. A recent large-scale analysis attributes much of these hybrids’ performance to their convolutional components rather than to attention [31], pointing to positional encoding and operator placement, rather than attention capacity alone, as the levers that matter, as PRIMUS [31] demonstrates. Lightweight convolutional models such as LightMedSeg [29] pursue the same efficiency goal from a different direction, reducing redundancy through Ghost and depthwise convolutions.

State-space models (SSMs) are well suited to this role, offering long-range sequence modeling at *linear* complexity. Structured SSMs [8] and the selective Mamba architecture [7] enable hardware-efficient sequence mixing, inspiring vision variants such as VMamba [20].

Building on this principle, we propose SegMaFormer, a

compact hybrid architecture for 3D medical image segmentation aimed at resource-constrained clinical settings where large models are impractical. SegMaFormer combines the linear-complexity sequence modeling of Mamba [7] with the global context modeling of Transformers [30], and builds on SegFormer-style volumetric tokenization [23, 34] to retain patch-based representation at substantially reduced complexity. SegMaFormer is designed to reduce memory and computational requirements, making training feasible on low- to mid-tier GPUs while preserving competitive segmentation performance.

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

- We introduce SegMaFormer, a lightweight hybrid Conv-SSM–Transformer encoder that assigns a global-mixing operator to each stage by resolution: a pure Mamba block at the first, highest-resolution stage, where token sequences are longest and attention is most expensive; learned-gated GHConv-SSM fusion at the two intermediate stages; and self-attention at the final, most downsampled stage for semantic refinement.
- We apply parameter-free 3D rotary positional embeddings (3D-RoPE) to the token sequence immediately after overlapped patch embedding, so every stage whether mixed by Mamba, GHConv-SSM, or self-attention, operates on position-aware tokens rather than only the attention stage, adding explicit depth/height/width cues without increasing model size.
- We conduct controlled ablations on stage placement, Conv-SSM design, 3D-RoPE, pure-Mamba stages. The results support the principle that state-space mixing is most effective at the longest-token stage, whereas attention remains useful after spatial downsampling.
- With approximately 1.9-M parameters and 13.94 GFLOPs, SegMaFormer achieves competitive performance on BraTS, BTCV, and ACDC while remaining substantially smaller than large Transformer- and Mamba-based baselines, positioning it as an accuracy–efficiency-oriented alternative for resource-constrained 3D medical image segmentation.

2. Related Work

CNN-based volumetric segmentation. Fully convolutional networks established dense per-pixel prediction without fully connected layers [21], and the U-Net encoder-decoder with skip connections became the dominant template for medical image segmentation [24]. Volumetric extensions such as 3D U-Net [40] and nested-skip variants such as UNet++ [39] improved volumetric prediction and feature reuse, while broad surveys document the rapid adoption of these designs across modalities [18]. The

self-configuring nnU-Net framework remains a strong and reproducible baseline, and its residual-encoder extensions continue to rival or exceed more elaborate architectures on standard benchmarks [13, 14]. Complementary training strategies such as deep supervision further stabilize these networks under limited data [17]. Collectively, this line of work shows that strong local inductive bias and carefully tuned training pipelines are difficult to surpass, which motivates adding global context without discarding convolutional priors.

Transformer and hybrid CNN–Transformer models.

The Transformer [30] and the Vision Transformer [6] introduced global token interaction through self-attention, prompting many medical adaptations. In 2D, TransUNet pairs a ViT encoder with a convolutional decoder [5], SwinUNet builds a fully transformer U-shape [3], MISSFormer designs an efficient medical-segmentation transformer [12], TransClaw U-Net augments a claw-shaped network with attention [4], and LeViT-UNet targets faster encoders [37]. Efficient backbones such as the Pyramid Vision Transformer [33] and SegFormer, which combines hierarchical attention with a lightweight all-MLP decoder [34], reduce the cost of dense prediction. In 3D, UNETR applies a transformer encoder to volumetric patches [10], SwinUNETR and nnFormer adopt windowed or hierarchical attention [11, 38], and UNETR++ introduces efficient paired attention for accuracy at reduced cost [25]. Lighter hybrids restrict attention to coarse scales, as in TransBTS and CoTr [32, 35], and SegFormer3D ports the efficient-attention and MLP-decoder recipe to volumes [23]. Despite these advances, transformer-heavy models frequently fail to match strong CNN baselines in limited-data regimes, and a recent large-scale study attributes much of their measured performance to convolutional components and positional encoding rather than to attention capacity [31]. SegMaFormer responds to this evidence by applying global mixing selectively rather than uniformly, choosing each operator by the token length of the stage on which it acts.

State-space models in vision and medical imaging.

Structured state-space models [8] and the selective Mamba architecture [7] model long sequences in linear time through input-dependent recurrent state updates, and vision adaptations such as VMamba extend this mechanism to spatial feature maps [20]. In medical imaging, U-Mamba integrates Mamba blocks into a U-Net to couple convolutional locality with state-space context [22], while Swin-UMamba replaces windowed attention with Mamba and leverages ImageNet pretraining for consistent gains [19]. More recently, Hymavi combines Mamba and attention in a parallel hybrid architecture and further exploits multi-view volumetric representations for medical image segmentation [28].

Continued refinement of the underlying state-space formulation [16] indicates further headroom for these mixers. Isensee et al. nonetheless caution that state-space layers add little without careful placement [14]. SegMaFormer takes this caution literally: rather than inserting Mamba uniformly, it reserves a pure state-space pathway for the highest-resolution stage, where linear-time mixing is most valuable, and combines it with convolution and attention at later stages.

Positional encoding for flattened volumetric tokens. Flattening volumetric features removes explicit spatial coordinates. RoPE restores positional information through parameter-free channel rotations [27]. PRIMUS [31] highlights the importance of positional encoding for 3D segmentation. SegMaFormer therefore applies parameter-free 3D-RoPE directly to volumetric tokens, encoding depth, height, and width before every mixer stage rather than only within attention.

Lightweight volumetric segmentation. Efficient 3D segmentation matters for resource-limited clinical and research settings, where full-resolution volumetric patches are costly. Existing approaches reduce cost through depth-wise and Ghost convolutions, as in LightMedSeg [29], or through compact CNN–Transformer hybrids such as LW-CTrans [15], alongside more general tactics such as reduced attention, smaller channel widths, and compact decoders. SegMaFormer shares this objective but obtains its savings structurally: high-resolution attention is replaced by a linear-time Mamba scan, intermediate stages use learned-gated GHConv-SSM fusion, and attention is retained only where the token count is already small. This stage-aware allocation, rather than block-level compression alone, is the source of its accuracy–efficiency trade-off.

3. Method

3.1. Architecture Overview

SegMaFormer follows a hierarchical encoder-decoder design for dense 3D prediction. Given an input volume $\mathbf{X} \in \mathbb{R}^{C \times D \times H \times W}$, overlapped 3D patch embeddings generate four multi-scale feature maps. Spatial resolution is gradually reduced while channel capacity is increased. The encoder uses the selected stage pattern $M/H/H/A$, where M denotes a pure Mamba block, H denotes a Gated Hybrid Conv-SSM (GHConv-SSM) block, and A denotes efficient self-attention. The decoder follows the SegFormer-style all-MLP design: each encoder feature is projected to a shared dimension, upsampled to a common resolution, concatenated, fused, and mapped to segmentation logits.

3.2. Volumetric Patch Embedding and 3D-RoPE

Overlapped patch embedding converts each 3D feature map into a token sequence while preserving local neighborhood continuity. For stage i , let the spatial size be $D_i \times H_i \times W_i$ and the channel dimension be C_i . The token count is $N_i = D_i H_i W_i$. After flattening, however, the explicit depth-height-width coordinate structure is no longer directly visible to the sequence mixer. This is undesirable in medical volumes, where similar local appearance can correspond to different anatomical locations.

To retain volumetric position information, SegMaFormer applies parameter-free 3D rotary position embedding (3D-RoPE) to the token sequence after patch embedding. For token coordinate (d, h, w) , the channel vector is split into depth, height, and width subspaces:

$$\mathbf{z}_{d,h,w} = [\mathbf{z}_{d,h,w}^D; \mathbf{z}_{d,h,w}^H; \mathbf{z}_{d,h,w}^W], \quad (1)$$

and each subspace is rotated according to the corresponding coordinate:

$$\text{RoPE}_{3D}(\mathbf{z}_{d,h,w}) = \begin{bmatrix} R_D(d)\mathbf{z}_{d,h,w}^D; \\ R_H(h)\mathbf{z}_{d,h,w}^H; \\ R_W(w)\mathbf{z}_{d,h,w}^W \end{bmatrix}, \quad (2)$$

where R_D , R_H , and R_W are block-diagonal rotary matrices applied to paired feature dimensions. Because the rotation is applied to the tokens themselves, every stage—whether mixed by Mamba, GHConv-SSM, or attention—operates on position-aware features; at the state-space and convolutional stages this acts as an absolute rotary encoding, rather than the relative encoding RoPE induces inside the attention dot product. As 3D-RoPE is deterministic and parameter-free, it improves spatial awareness without adding learned parameters.

3.3. Gated Hybrid Conv-SSM Block

The Gated Hybrid Conv-SSM (GHConv-SSM) block combines two complementary inductive biases: sequence-level communication from state-space modeling and local spatial evidence from convolution. Given an input feature map $\mathbf{X}$, we flatten the spatial dimensions and normalize the token sequence:

$$\mathbf{X}_{\text{seq}} = \text{Flatten}(\mathbf{X}), \quad \hat{\mathbf{X}} = \text{LayerNorm}(\mathbf{X}_{\text{seq}}). \quad (3)$$

Two parallel branches are then applied:

$$\mathbf{Y}_{\text{SSM}} = \text{Mamba}(\hat{\mathbf{X}}), \quad \mathbf{Y}_{\text{Conv}} = \text{ConvBranch}(\hat{\mathbf{X}}). \quad (4)$$

The Mamba branch captures long-range dependencies with approximately linear sequence complexity, while the convolution branch preserves local boundary, shape, and texture

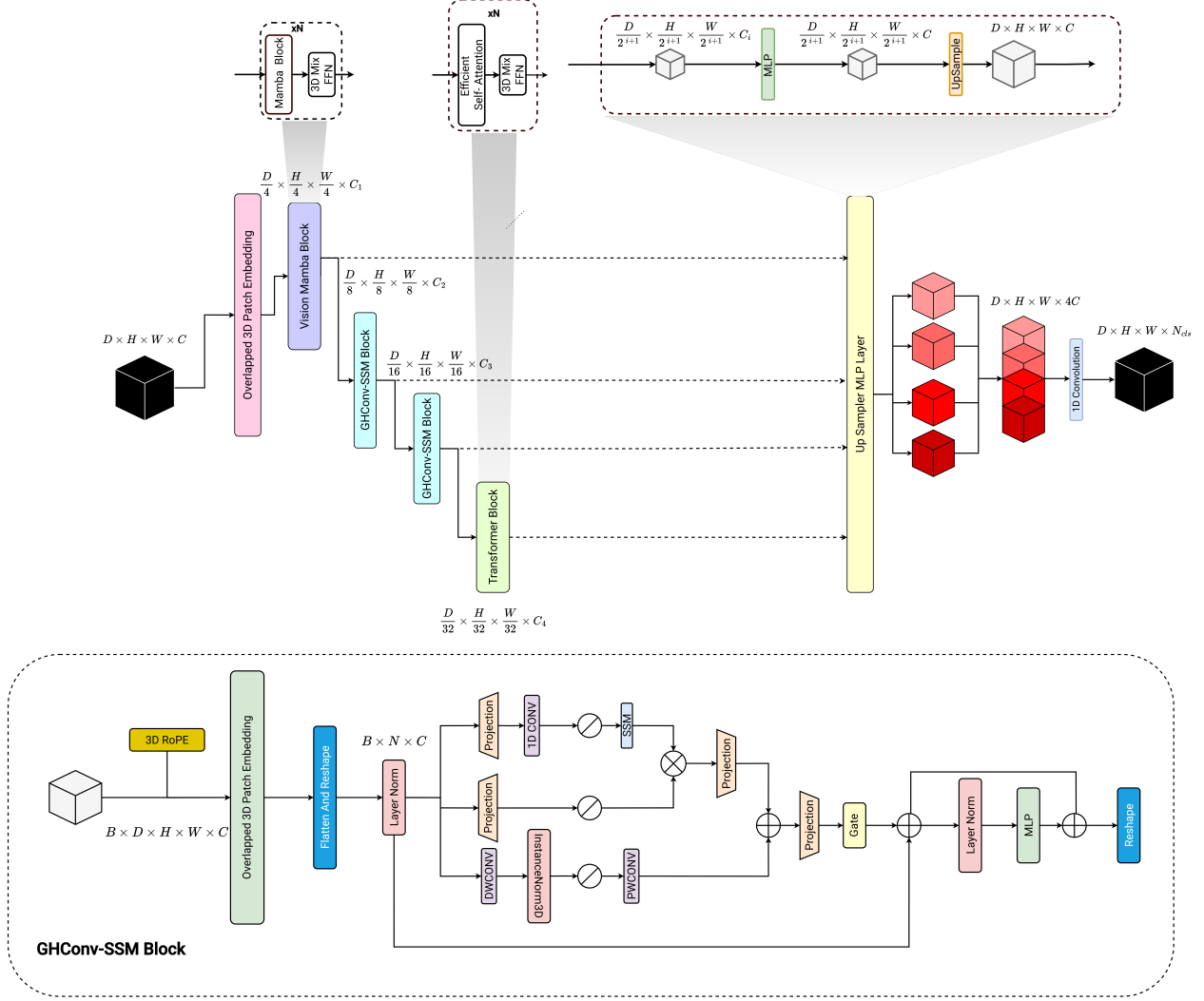


Figure 1. Overview of SegMaFormer. The encoder follows a stage-aware M/H/H/A design: a Mamba block is used at the first high-resolution stage, GHConv-SSM blocks are used at the intermediate stages to adaptively fuse state-space and convolutional features, and efficient attention is applied at the final low-resolution stage. Multi-scale features are fused by an MLP decoder for volumetric segmentation.

cues. Rather than using a fixed sum, GHConv-SSM learns token-wise fusion weights:

$$\begin{aligned} \mathbf{g} &= \sigma(\text{Gate}([\mathbf{Y}_{\text{SSM}}; \mathbf{Y}_{\text{Conv}}])), \\ \mathbf{Y} &= \mathbf{g} \odot \mathbf{Y}_{\text{SSM}} + (1 - \mathbf{g}) \odot \mathbf{Y}_{\text{Conv}}. \end{aligned} \quad (5)$$

The gate is trained end-to-end with the segmentation objective and provides content-adaptive fusion: instead of applying one fixed rule to every token, it can favor broader SSM context in homogeneous organ interiors and local convolutional evidence at boundaries and small structures. The fused representation is projected back to the model dimension

and followed by residual MLP refinement:

$$\begin{aligned} \mathbf{X}^{(1)} &= \mathbf{X}_{\text{seq}} + \text{Proj}(\mathbf{Y}), \\ \mathbf{X}^{(2)} &= \mathbf{X}^{(1)} + \text{MLP}(\text{LayerNorm}(\mathbf{X}^{(1)})), \\ \mathbf{X}_{\text{out}} &= \text{Reshape}(\mathbf{X}^{(2)}). \end{aligned} \quad (6)$$

In SegMaFormer, GHConv-SSM is used only at stages 2–3, keeping the first stage as a clean Mamba pathway and avoiding unnecessary hybridization at the final compact stage, where attention is already affordable.

3.4. Stage-Aware Mixer Placement

SegMaFormer assigns each stage the mixer whose inductive bias and cost best match its token count $N_i = D_i H_i W_i$,

where $N_1 > N_2 > N_3 > N_4$ under stagewise downsampling.

Stage 1 is the most constrained: it has the longest sequence and the lowest-level features, which excludes every simpler operator. Self-attention is too costly ($\mathcal{O}(N_i^2 C_i)$), with even efficient variants projecting all tokens); pure convolution cannot propagate long-range anatomical cues without a deep stack; and a hybrid GHConv-SSM block is redundant, since the convolutional patch embedding and the depthwise convolution inside the Mamba block already encode local detail, so an added convolutional branch over-emphasizes texture at the longest and most expensive sequence. A pure Mamba block is the only operator that provides global mixing at linear cost while keeping the high-resolution pathway uninterrupted. At the intermediate stages the sequence is shorter and the features more semantic, so gated GHConv-SSM fusion couples state-space context with a local convolutional prior; at stage 4 the token count is smallest, making self-attention tractable and restoring the explicit token–token interaction a Mamba-only encoder would forgo.

Although flattening could weaken locality, overlapped patch embedding, the Mamba block’s internal convolution, and 3D-RoPE together keep the stage position-aware. Our ablation in Table 1 study favors this allocation over early hybridization. The contribution is therefore not a uniform replacement of attention by Mamba, but a stage-specific allocation of sequence mixers by token regime and feature semantics.

3.5. All-MLP Decoder

SegMaFormer uses an all-MLP decoder to fuse multi-scale encoder features without introducing a heavy convolutional upsampling path. Let $\mathbf{X}_i$ be the output of encoder stage i with channel dimension C_i . Each feature is projected to a common channel dimension C :

$$\tilde{\mathbf{X}}_i = \text{Linear}(C_i, C)(\mathbf{X}_i), \quad i \in \{1, 2, 3, 4\}. \quad (7)$$

The projected features are upsampled to the same spatial resolution, concatenated, and fused:

$$\mathbf{Z} = \text{Linear}(4C, C) \left(\text{Concat}(\tilde{\mathbf{X}}_1^{up}, \tilde{\mathbf{X}}_2^{up}, \tilde{\mathbf{X}}_3^{up}, \tilde{\mathbf{X}}_4^{up}) \right). \quad (8)$$

A final linear prediction head produces class logits:

$$\mathbf{Y} = \text{Linear}(C, N_{cls})(\mathbf{Z}). \quad (9)$$

This keeps the decoder’s parameter count low, so that most of the model’s efficiency gain comes from the stage-aware encoder rather than from a heavy decoder.

4. Experiments

4.1. Datasets and Training Protocol

We evaluate SegMaFormer on three public 3D medical image segmentation benchmarks: Brain Tumor Segmentation (BraTS)[1], Beyond The Cranial Vault(BTCV)[9], and Automated Cardiac Diagnosis Segmentation (ACDC)[2]. BraTS contains multi-modal brain MRI scans with tumor subregion annotations, which are reported as whole tumor (WT), enhancing tumor (ET), and tumor core (TC). BTCV evaluates abdominal multi-organ CT segmentation over eight organs: aorta, liver, left kidney, right kidney, gallbladder, pancreas, spleen, and stomach. ACDC evaluates cardiac MRI segmentation over the right ventricle (RV), myocardium (Myo), and left ventricle (LV).

All internally reproduced experiments are conducted within the same nnU-Net framework [14], using identical data splits, preprocessing, and training protocols without external pretraining. Parameter counts vary slightly across datasets due only to dataset-specific config. When controlled reproduction is not feasible, we include results reported in the original publications for context only and explicitly distinguish them from our reproduced results. Full implementation details are provided in Appendix A.

4.2. Ablation and Inference Efficiency

Table 1 reports controlled BraTS ablations for stage placement and efficiency. A, H, and M denote efficient attention, GHConv-SSM, and pure Mamba stages, respectively. The selected SegMaFormer configuration is $M/H/H/A$, meaning that stage 1 uses pure Mamba, stages 2–3 use GHConv-SSM, and stage 4 uses attention. This configuration obtains the highest validation Dice among the tested variants while keeping the parameter count and FLOPs low.

The ablation supports three observations. First, 3D-RoPE improves Dice without increasing the number of learned parameters, suggesting that coordinate-aware phase information is beneficial for flattened volumetric tokens. Second, the selected M/H/H/A allocation outperforms both the attention-only encoder and the H/H/H/A hybrid-heavy variant. This indicates that the first high-resolution stage benefits from an uninterrupted Mamba pathway rather than from immediate hybrid fusion. Third, pure Mamba variants are efficient, but their Dice is lower than the selected stage-aware mixture. The best trade-off is therefore obtained by combining early Mamba, intermediate GHConv-SSM fusion, and final attention. This also explains why SegFormer3D can show lower measured latency despite higher FLOPs and its attention blocks benefit from spatial reduction and highly optimized GPU matrix-multiplication kernels, whereas selective state-space modules depend more heavily on scan-kernel efficiency and memory-access behavior.

Table 1. BraTS ablation and inference-efficiency analysis. A, H, and M denote attention, GHConv-SSM, and pure Mamba stages, respectively. Dice scores are reported as mean $\pm$ standard deviation over three random seeds, except for SegFormer3D, which is reported from a single run. Inference time is reported as mean $\pm$ standard deviation over 100 timed forward passes after 30 warm-up iterations.

Configuration	Stage mix	3D-RoPE	Mean Dice (%) $\uparrow$	Params (M) $\downarrow$	GFLOPs $\downarrow$	Time/patch (ms) $\downarrow$
SegFormer3D	A/A/A/A	$\times$	83.74	4.50	17.50	5.87 $\pm$ 0.28
Attention-only encoder	A/A/A/A	$\checkmark$	83.82 $\pm$ 0.13	2.04	15.95	7.58 $\pm$ 0.36
GHConv-SSM heavy	H/H/H/A	$\checkmark$	84.21 $\pm$ 0.06	1.92	14.27	8.19 $\pm$ 0.78
Stage-1 hybrid only	H/A/A/A	$\checkmark$	84.16 $\pm$ 0.31	1.90	14.55	7.23 $\pm$ 0.28
Stage-1 Mamba only	M/A/A/A	$\checkmark$	84.18 $\pm$ 0.34	1.89	14.22	6.87 $\pm$ 0.39
All early Mamba	M/M/M/A	$\checkmark$	84.38 $\pm$ 0.12	1.85	13.80	7.13 $\pm$ 0.23
SegMaFormer, RoPE off	M/H/H/A	$\times$	84.19 $\pm$ 0.27	1.91	13.94	5.75 $\pm$ 0.47
SegMaFormer (ours)	M/H/H/A	$\checkmark$	84.42 $\pm$ 0.15	1.91	13.94	7.65 $\pm$ 0.19

Table 2. Cross-dataset evaluation of representative stage-allocation strategies. BTCV-8 uses the standard eight-organ subset of BTCV, whereas BTCV-13 evaluates all 13 annotated organs.

Configuration	Stage mix	BTCV-8 Dice (%) $\uparrow$	BTCV-13 Dice (%) $\uparrow$	ACDC Dice (%) $\uparrow$
Stage-1 Mamba Only	M/A/A/A	80.53	76.34	91.02
Stage-1 Hybrid only	H/A/A/A	81.68	77.76	91.27
All early Mamba	M/M/M/A	82.68	78.05	90.96
SegMaFormer (ours)	M/H/H/A	83.47	79.02	91.25

Table 3. Complexity comparison on BraTS. Cross-paper latency should be interpreted cautiously because hardware and implementations differ.

Architecture	Params (M) $\downarrow$	FLOPs (G) $\downarrow$	Mem. (MB) $\downarrow$	Lat. (ms) $\downarrow$
nnUNet [13]	31.22	44	1135	32.4
nnFormer [38]	149.60	421.50	1269	149
UNETR [10]	92.49	75.76	3379	—
UNETR++ [25]	40.6	70.1	2400	62.4
SwinUNETR [11]	62.83	384.20	20173	—
SegFormer3D [23]	4.51	17.50	430	5.87
LightMedSeg [29]	0.48	14.64	1174	13.70
SegMaFormer (M/H/H/A)	1.91	13.94	423	7.65

4.3. Brain Tumor Segmentation

Table 4 compares SegMaFormer with representative Transformer and hybrid methods on BraTS. SegMaFormer obtains 84.42% average Dice with about 2M parameters. It improves over SegFormer3D and LightMedSeg in the reported average Dice while remaining substantially smaller than nnFormer. Whole-tumor segmentation is the strongest component, consistent with the need for broad contextual modeling. Enhancing tumor remains more challenging because ET regions are smaller and more variable. Finally, the qualitative results in Figure 2 illustrate competitive performance against well-established architectures.

4.4. Beyond the Cranial Vault

BTCV contains abdominal structures with substantial variation in size and appearance, ranging from large organs such as the liver and spleen to smaller structures such as the gallbladder and pancreas. Following the conventional

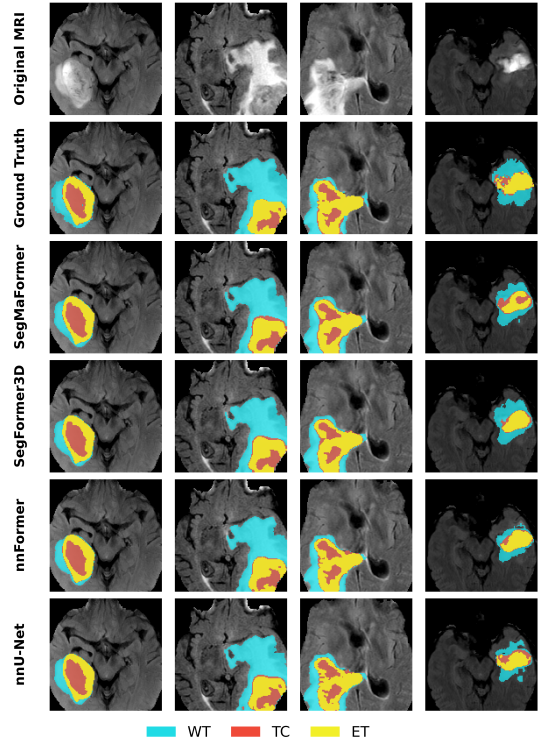


Figure 2. Qualitative results of the original MRI scans, ground-truth annotations, and predicted segmentation results for three representative cases from the BraTS dataset.

Table 4. BraTS benchmark comparison in Average Dice score (%). WT, ET, TC denote whole tumor, enhancing tumor, and tumor core.

Method	Params (M) ↓	Avg. ↑	WT ↑	ET ↑	TC ↑
<i>Controlled reproductions</i>					
nnU-Net [13]	31.22	84.88	91.21	78.86	84.57
nnFormer [38]	149.22	84.38	89.90	79.53	83.69
SegFormer3D [23]	4.5	83.74	90.21	77.57	83.66
<i>Original Reports</i>					
nnFormer [38]	150.50	86.40	91.30	81.80	86.00
SegFormer3D [23]	4.52	82.10	89.90	74.20	82.20
LightMedSeg [29]	0.48	83.40	89.80	77.00	83.50
UNETR [10]	92.49	71.10	78.90	58.50	76.10
TransBTS [32]	–	69.60	77.90	57.40	73.50
CoTr [35]	41.90	68.30	74.60	55.70	74.80
TransUNet [5]	96.07	64.40	70.60	54.20	68.40
SegMaFormer (ours)	1.91	84.42	91.02	78.22	84.01

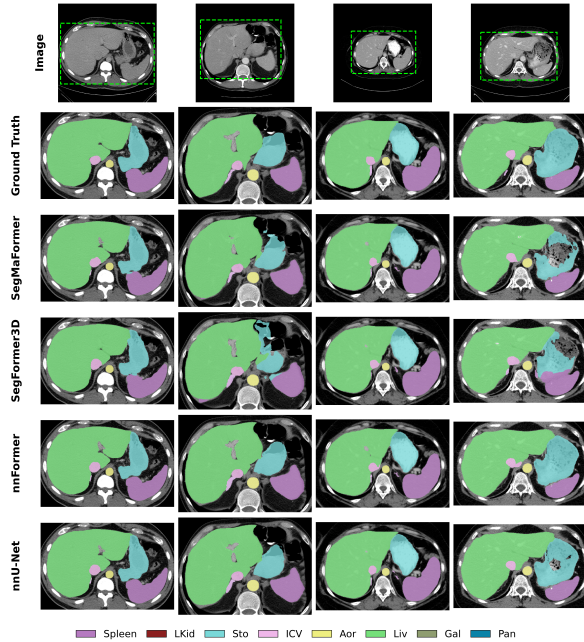


Figure 3. The qualitative evaluation on the BTCV dataset illustrates the visual segmentation performance of different methods on abdominal CT images. In the figure, each column corresponds to an individual CT slice.

eight-organ evaluation protocol, SegMaFormer achieves a Dice score of 83.47% (Table 5). It outperforms SegFormer3D while using fewer parameters and remains competitive with substantially larger architectures. The organ-wise results show strong performance on the aorta, liver, kidneys, and spleen, whereas the gallbladder and pancreas remain more challenging. These results indicate that small and anatomically variable structures may benefit from additional boundary-aware or small-organ supervision.

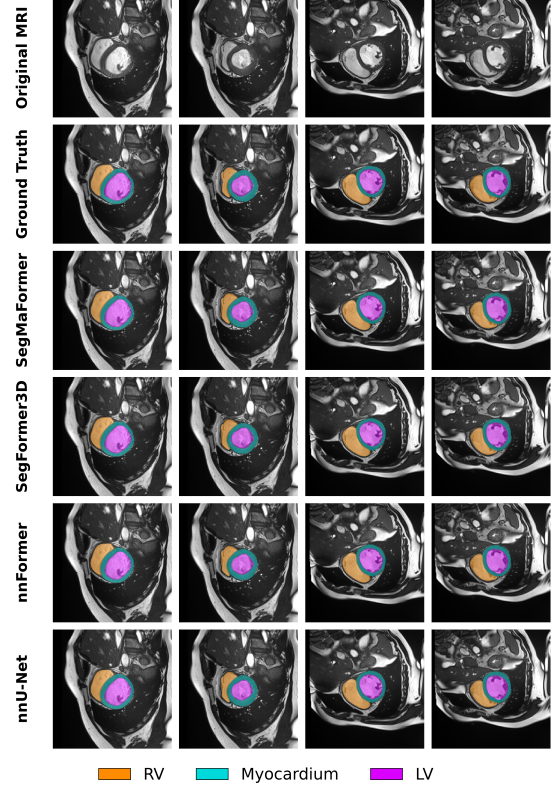


Figure 4. The qualitative evaluation on the ACDC dataset depicts the visual segmentation performance of different cases.

4.5. Automated Cardiac Diagnosis Segmentation

Table 6 and Figure 4 report results on ACDC. SegMaFormer obtains 91.25% average Dice, slightly above SegFormer3D while using fewer parameters. The gap to nnFormer and UNETR++ is small but still visible. This result supports the same conclusion as BTCV: compact state-space/attention hybridization can substantially improve efficiency while remaining competitive, although larger or pretraining-oriented models may still achieve the best absolute Dice on some datasets.

4.6. Accuracy–Efficiency Trade-off

The benchmark results reveal two operating regimes. High-capacity models such as nnFormer and nnU-Net can achieve higher absolute Dice on some datasets, but they require tens to hundreds of millions of parameters. Lightweight models such as SegFormer3D, LightMedSeg, and SegMaFormer target a different setting, where model size, computation, and memory usage are important. SegMaFormer improves this lightweight operating point by using fewer parameters while maintaining competitive segmentation performance.

Table 5. BTCV-8 comparison in Dice (%). AOR: aorta; LIV: liver; LKID/RKID: left/right kidney; GAL: gallbladder; PAN: pancreas; SPL: spleen; STO: stomach.

Method	Params (M)	Avg.	AOR	LIV	LKID	RKID	GAL	PAN	SPL	STO
<i>Controlled reproductions</i>										
nnU-Net [13]	31.20	86.86	91.85	96.89	95.10	92.46	59.89	80.18	93.61	84.91
nnFormer [38]	149.59	87.93	91.25	97.07	93.61	95.05	67.46	78.89	93.83	86.25
SegFormer3D [23]	4.50	81.61	88.66	95.48	92.48	89.51	54.82	68.99	88.69	74.24
<i>Original reports</i>										
U-Mamba [22]	172.63	87.98	90.80	96.90	94.60	94.50	73.80	79.30	95.80	81.70
nnFormer [38]	150.50	86.57	92.04	96.84	86.57	86.25	70.17	83.35	90.51	86.83
MISSFormer	–	81.96	86.99	94.41	85.21	82.00	68.65	65.67	91.92	80.81
UNETR [10]	92.49	79.56	89.99	94.46	85.66	84.80	60.56	59.25	87.81	73.99
SwinUNet	–	79.13	85.47	94.29	83.28	79.61	66.53	56.58	90.66	76.60
TransUNet [5]	96.07	77.48	87.23	94.08	81.87	77.02	63.16	55.86	85.08	75.62
SegMaFormer (ours)	1.83	83.47	90.11	96.57	91.25	88.97	59.61	72.01	90.94	78.29

Table 6. ACDC comparison in Dice (%). RV, Myo, and LV denote right ventricle, myocardium, and left ventricle.

Method	Params (M)	Avg.	RV	Myo	LV
<i>Controlled reproductions</i>					
nnU-Net [13]	30.37	91.01	88.70	89.30	95.03
nnFormer [38]	150.50	91.78	90.27	89.47	95.64
SegFormer3D [23]	4.44	90.97	89.00	88.49	95.41
<i>Original reports</i>					
Hymavi [28]	287	92.67	91.67	90.23	96.10
SwinUNETR-V2 [26]	–	92.24	90.86	90.05	95.80
SegMamba [36]	–	92.27	91.50	89.45	96.10
UNETR++ [25]	42.96	92.83	91.89	90.61	96.00
nnFormer [38]	150.50	92.06	90.94	89.58	95.65
LightMedSeg [29]	0.44	91.24	91.77	86.24	95.72
LeViT-UNet-384	52.17	90.32	89.55	87.64	93.76
SegMaFormer (ours)	1.83	91.25	88.90	89.18	95.67

Why lower FLOPs do not always imply lower latency.

Although SegMaFormer reduces both parameter count and FLOPs, its measured latency is not always lower than attention-based variants at small input sizes. FLOPs measure arithmetic complexity but do not fully capture GPU execution efficiency. Attention operations can benefit from highly optimized dense matrix-multiplication kernels, whereas selective state-space layers rely on scan-style computation, memory movement, and custom kernels whose realized efficiency depends strongly on hardware implementation. This distinction has also been observed in recent SSM work [7, 16], where theoretically favorable complexity does not necessarily translate into lower latency at every operating point. The resolution-scaling analysis in Appendix B provides additional context.

4.7. Stage Placement and Discussion

The ablation results in Table 1 show that *where* sequence-mixing operators are placed is more important than simply increasing the number of Mamba or hybrid blocks. The H/H/H/A variant does not outperform M/H/H/A, while the M/M/M/A variant is slightly less accurate despite its efficiency. These results support the proposed stage-aware

allocation: Mamba is used at the highest-resolution stage, where the sequence is longest and attention is most costly; GHConv-SSM is applied at intermediate stages to balance global context with local anatomical evidence; and attention is retained at the final compact stage, where global token interaction becomes affordable.

SegMaFormer should therefore be viewed as an efficiency-oriented architecture, enabling competitive segmentation accuracy with substantially fewer parameters than many high-capacity Transformer and Mamba baselines. This accuracy–efficiency trade-off is particularly relevant to 3D medical segmentation, where computational and memory resources are often constrained.

5. Conclusion and Limitations

We presented SegMaFormer, a compact stage-aware architecture for efficient 3D medical image segmentation. SegMaFormer uses Mamba at the highest-resolution stage, GHConv-SSM at intermediate stages, efficient attention at the final compact stage, parameter-free 3D-RoPE for volumetric position encoding, and an all-MLP decoder for lightweight multi-scale fusion. Experiments on BraTS, BTCV, and ACDC show that this architecture provides a favorable accuracy-efficiency trade-off, improving over baselines in the reported comparisons while using fewer parameters and remaining competitive with much larger Transformer and Mamba baselines.

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 a lightweight hybrid architecture and provides a practical foundation for scalable, resource-efficient 3D medical image segmentation.

References

- [1] Michela Antonelli et al. The Medical Segmentation Decathlon. *Nature Communications*, 13(1):4128, 2022. 5
- [2] Olivier Bernard et al. 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, 2018. 5
- [3] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation, 2021. arXiv:2105.05537 [eess]. 2
- [4] Yao Chang, Hu Menghan, Zhai Guangtao, and Zhang Xiaoping. TransClaw U-Net: Claw U-Net with Transformers for Medical Image Segmentation, 2021. arXiv:2107.05188 [cs]. 2
- [5] 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, 2021. arXiv:2102.04306 [cs]. 1, 2, 7, 8
- [6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, 2021. arXiv:2010.11929 [cs]. 1, 2
- [7] Albert Gu and Tri Dao. Mamba: Linear-Time Sequence Modeling with Selective State Spaces, 2024. arXiv:2312.00752 [cs]. 1, 2, 8
- [8] Albert Gu, Karan Goel, and Christopher Ré. Efficiently Modeling Long Sequences with Structured State Spaces, 2022. arXiv:2111.00396 [cs]. 1, 2
- [9] Robert Harrigan, Bennett Landman, and Zhoubing Xu. Multi-Atlas Labeling Beyond the Cranial Vault - Workshop and Challenge, 2015. Synapse ID: syn3193805. 5
- [10] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger Roth, and Daguang Xu. UNETR: Transformers for 3D Medical Image Segmentation, 2021. arXiv:2103.10504 [eess]. 1, 2, 6, 7, 8
- [11] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger Roth, and Daguang Xu. Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images, 2022. arXiv:2201.01266 [eess]. 1, 2, 6
- [12] Xiaohong Huang, Zhifang Deng, Dandan Li, and Xueguang Yuan. MISSFormer: An Effective Medical Image Segmentation Transformer, 2021. arXiv:2109.07162 [cs]. 2
- [13] Fabian Isensee, Constantin Ulrich, Tassilo Wald, and Klaus H. Maier-Hein. Extending nnU-Net is all you need, 2022. arXiv:2208.10791 [eess]. 2, 6, 7, 8
- [14] Fabian Isensee, Tassilo Wald, Constantin Ulrich, Michael Baumgartner, Saikat Roy, Klaus Maier-Hein, and Paul F. Jaeger. nnU-Net Revisited: A Call for Rigorous Validation in 3D Medical Image Segmentation, 2024. arXiv:2404.09556 [cs]. 1, 2, 3, 5
- [15] Hulin Kuang, Yahui Wang, Xianzhen Tan, Jialin Yang, Jiarui Sun, Jin Liu, Wu Qiu, Jingyang Zhang, Jiulou Zhang, Chunfeng Yang, Jianxin Wang, and Yang Chen. LW-CTrans: A lightweight hybrid network of CNN and Transformer for 3D medical image segmentation. *Medical Image Analysis*, 102: 103545, 2025. 3
- [16] Aakash Lahoti, Kevin Y. Li, Berlin Chen, Caitlin Wang, Aviv Bick, J. Zico Kolter, Tri Dao, and Albert Gu. Mamba-3: Improved sequence modeling using state space principles, 2026. 3, 8
- [17] Renjie Li, Xinyi Wang, Guan Huang, Wenli Yang, Kaining Zhang, Xiaotong Gu, Son N. Tran, Saurabh Garg, Jane Alty, and Quan Bai. A Comprehensive Review on Deep Supervision: Theories and Applications, 2022. arXiv:2207.02376 [cs]. 2
- [18] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A. W. M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017. 2
- [19] Jiarun Liu, Hao Yang, Hong-Yu Zhou, Yan Xi, Lequan Yu, Yizhou Yu, Yong Liang, Guangming Shi, Shaoting Zhang, Hairong Zheng, and Shanshan Wang. Swin-UMamba: Mamba-based UNet with ImageNet-based pretraining, 2024. arXiv:2402.03302 [eess]. 2
- [20] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. VMamba: Visual State Space Model, 2024. arXiv:2401.10166 [cs]. 1, 2
- [21] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully Convolutional Networks for Semantic Segmentation, 2015. arXiv:1411.4038 [cs]. 1, 2
- [22] Jun Ma, Feifei Li, and Bo Wang. U-Mamba: Enhancing Long-range Dependency for Biomedical Image Segmentation, 2024. arXiv:2401.04722 [eess]. 2, 8
- [23] Shehan Perera, Pouyan Navard, and Alper Yilmaz. SegFormer3D: an Efficient Transformer for 3D Medical Image Segmentation, 2024. arXiv:2404.10156 [cs]. 2, 6, 7, 8
- [24] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation, 2015. arXiv:1505.04597 [cs]. 1, 2
- [25] Abdelrahman Shaker, Muhammad Maaz, Hanoona Rasheed, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Unetr++: Delving into efficient and accurate 3d medical image segmentation, 2024. 2, 6, 8
- [26] Mahsa Shakeri, Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Can Zhao, Dong Yang, Holger R Roth, and Daguang Xu. Swinunetr-v2: Stagewise convolutions for 3d medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 574–584. Springer, 2023. 8
- [27] Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced Transformer with Rotary Position Embedding, 2023. arXiv:2104.09864 [cs]. 3
- [28] Sy Dat Tran and Jin Kyu Gahm. Hymavi: A hybrid mamba-attention network in multi-view framework for volumetric medical image segmentation. In *2026 IEEE/CVF Winter*

- Conference on Applications of Computer Vision (WACV)*, pages 6006–6015, 2026. [2](#), [8](#)
- [29] Kavyansh Tyagi, Vishwas Rathi, and Puneet Goyal. Lightmedseg: Lightweight 3d medical image segmentation with learned spatial anchors, 2026. [1](#), [3](#), [6](#), [7](#), [8](#)
 - [30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need, 2023. arXiv:1706.03762 [cs]. [1](#), [2](#)
 - [31] Tassilo Wald, Saikat Roy, Fabian Isensee, Constantin Ulrich, Sebastian Ziegler, Dasha Trofimova, Raphael Stock, Michael Baumgartner, Gregor Köhler, and Klaus Maier-Hein. Primus: Enforcing Attention Usage for 3D Medical Image Segmentation, 2025. arXiv:2503.01835 [cs]. [1](#), [2](#), [3](#)
 - [32] Wenxuan Wang, Chen Chen, Meng Ding, Jiangyun Li, Hong Yu, and Sen Zha. TransBTS: Multimodal Brain Tumor Segmentation Using Transformer, 2021. arXiv:2103.04430 [cs]. [1](#), [2](#), [7](#)
 - [33] 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, 2021. arXiv:2102.12122 [cs]. [2](#)
 - [34] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers, 2021. arXiv:2105.15203 [cs]. [2](#)
 - [35] Yutong Xie, Jianpeng Zhang, Chunhua Shen, and Yong Xia. CoTr: Efficiently Bridging CNN and Transformer for 3D Medical Image Segmentation, 2021. arXiv:2103.03024 [cs]. [1](#), [2](#), [7](#)
 - [36] Zhaohu Xing, Tian Ye, Yijun Yang, Guang Liu, and Lei Zhu. Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation, 2024. [8](#)
 - [37] Guoping Xu, Xingrong Wu, Xuan Zhang, and Xinwei He. LeViT-UNet: Make Faster Encoders with Transformer for Medical Image Segmentation, 2021. arXiv:2107.08623 [cs]. [2](#)
 - [38] Hong-Yu Zhou, Jiansen Guo, Yinghao Zhang, Lequan Yu, Liansheng Wang, and Yizhou Yu. nnFormer: Interleaved Transformer for Volumetric Segmentation, 2022. [1](#), [2](#), [6](#), [7](#), [8](#)
 - [39] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. UNet++: A Nested U-Net Architecture for Medical Image Segmentation, 2018. arXiv:1807.10165 [cs]. [2](#)
 - [40] Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation, 2016. arXiv:1606.06650 [cs]. [2](#)