EccoMamba: Enhanced Cross-hierarchical Continuity Orthogonal Mamba for Medical Image Segmentation

Junlin Xu[1], Jincan Li[2], Feifei Cui[3], Zhuang Zhang[4], Jialiang Yang[5], Shuting Jin[1*], Qiangguo Jin[6*], Yajie Meng[4*]

[1]School of Computer Science and Technology, Wuhan University of Science and Technology, Hubei, China
[2]School of Mathematics and Statistics, Hainan Normal University, Hainan, China
[3]School of Computer Science and Technology, Hainan University, Hainan, China
[4]School of Computer Science and Artificial Intelligence, Wuhan Textile University, Hubei, China
[5]Geneis Beijing Co., Ltd, Beijing, China
[6]School of Software, Northwestern Polytechnical University, Shanxi, China
shutingjin@wust.edu.cn, qgking@nwpu.edu.cn, myj@hnu.edu.cn

Abstract

Medical image segmentation plays a crucial role in clinical diagnosis, lesion quantification, and preoperative planning. However, existing Mamba-based architectures, which rely on fixed-direction sequence modeling and flatten images into one-dimensional (1D) sequences, struggle to capture hierarchical anatomical features and spatial dependencies, thereby limiting their representational capacity for complex medical structures. To address these limitations, we propose EccoMamba (Enhanced Cross-hierarchical Continuity Orthogonal Mamba), a U-shaped encoder–decoder framework designed for medical image segmentation. In the encoder's downsampling path, we introduce a Hierarchical Aggregation Enhancement (HAE) module that integrates multi-scale convolutions with hierarchical attention mechanisms. The attention branch further incorporates cross-channel interactions, allowing the model to selectively enhance semantically relevant features while suppressing irrelevant background responses. For skip connections, we design a Structural Continuity Orthogonal (SCO) module to preserve spatial continuity by modeling cross-dimensional dependencies via orthogonal Axial Shifts (AS), thereby mitigating directional bias and improving anatomical consistency. Extensive experiments on four benchmark datasets—ISIC 2018, ISIC 2017, Synapse, and ACDC—show that EccoMamba consistently outperforms state-of-the-art methods in both segmentation accuracy and structural fidelity.

Code — https://github.com/Biowust/EccoMamba

Introduction

Medical image segmentation plays a crucial role in clinical diagnosis, lesion detection, preoperative planning, and treatment evaluation (Ronneberger, Fischer, and Brox 2015). It is essential for identifying complex organ structures and detecting subtle pathological regions. However, medical images often exhibit substantial modality heterogeneity (e.g., CT, MRI, WSI), ambiguous structural boundaries, high anatomical variability, and a scarcity of labeled data (Chen et al. 2024; Zhou et al. 2018). These challenges significantly hinder the generalization ability of traditional segmentation approaches.

Recent advances in medical image segmentation primarily fall into two categories. The first category focuses on CNN-based architectures. Models such as U-Net and its variants (e.g., UNet++ (Zhou et al. 2018), Attention-UNet (Zhu et al. 2022), UCTransNet (Wang et al. 2022), and UNetR (Hatamizadeh et al. 2022)) leverage encoder–decoder structures and skip connections to capture local spatial context and achieve strong boundary localization. Nevertheless, CNNs inherently suffer from limited receptive fields, restricting their ability to model global interactions and often resulting in fragmented anatomical predictions. The second category involves Transformer-based models (Dosovitskiy et al. 2020), including TransUNet (Chen et al. 2024), Swin-UNet (Cao et al. 2022), and MISSFormer (Huang et al. 2022). These models utilize self-attention to capture long-range dependencies but face two critical limitations: substantial computational overhead and insufficient spatial priors, both of which constrain boundary precision and positional consistency.

More recently, the Mamba architecture (Gu and Dao 2023), a selective state-space model, has emerged as a promising alternative to CNN- and Transformer-based designs for medical image segmentation, as demonstrated in VM-UNet (Ruan, Li, and Xiang 2024) and Mamba-UNet (Wang et al. 2024). Thanks to its linear computational complexity, Mamba efficiently models long-range dependencies with a relatively compact parameter budget. However, Mamba processes images as flattened one-dimensional (1D) sequences, which disrupts the intrinsic 2D or 3D spatial structures essential for medical imaging. This leads to difficulties in capturing fine anatomical details, such as cardiac subregions and tumor boundaries, that rely heavily on continuous spatial cues. Moreover, Mamba's unidirectional sequence processing introduces directional bias, reducing its ability to symmetrically perceive horizontal and vertical structures. In addition, the absence of explicit multi-scale mechanisms limits its capacity to model hierarchical anatomical features, resulting in imbalanced perceptual rep-

*Corresponding authors

resentations and suboptimal segmentation performance, particularly in anatomically complex regions.

To address these limitations, we propose the **E**nhanced **C**ross-hierarchical **C**ontinuity **O**rthogonal **Mamba** (EccoMamba), a U-shaped encoder–decoder architecture tailored for medical image segmentation. EccoMamba enhances the Mamba backbone by improving its ability to model local spatial continuity, multi-scale anatomical structures, and orthogonal directional dependencies. Specifically, we design two key components: the Hierarchical Aggregation Enhancement (HAE) module and the Structural Continuity Orthogonal (SCO) module. The HAE module enriches multi-scale feature representation through parallel multi-scale convolutions combined with a lightweight hierarchical attention mechanism. This attention mechanism integrates channel-wise modulation with spatial emphasis, enabling precise boundary localization and effective suppression of irrelevant background regions. The SCO module mitigates directional bias by performing orthogonal Axial Shifts (AS) along horizontal and vertical axes, thereby capturing cross-dimensional spatial dependencies. It further incorporates lightweight convolutional fusion to enhance continuity across symmetric structures and preserve fine-grained anatomical details.

The main contributions of this work are summarized as follows:

- We propose a novel HAE module to address Mamba's limitations in modeling hierarchical anatomical features. HAE integrates parallel multi-scale convolutions with hierarchical attention to capture features across varying spatial resolutions while enhancing anatomically relevant regions and preserving structural boundaries.
- We introduce the SCO module to overcome Mamba's directional bias and spatial discontinuity. SCO applies orthogonal AS in both horizontal and vertical directions, followed by feature fusion and context aggregation, yielding more consistent spatial representations and improved structural fidelity.
- Extensive experiments on four benchmark datasets (ISIC 2018, ISIC 2017, Synapse, and ACDC) demonstrate the effectiveness and strong generalization capability of EccoMamba across diverse evaluation metrics.

Related Work

Deep Learning in Medical Image Segmentation

Deep learning has become the dominant paradigm for medical image segmentation, with convolution-based and attention-based architectures achieving strong performance across diverse imaging modalities and anatomical targets. U-Net and its variants, such as U-Net++ (Zhou et al. 2018) and Attention-UNet (Zhu et al. 2022), effectively capture fine-grained spatial details through hierarchical feature extraction and skip connections. TransUNet (Chen et al. 2024) integrates Transformer blocks into the U-Net encoder to enhance global context modeling, while Swin-UNet (Cao et al. 2022) leverages the hierarchical design of Swin Transformers for improved multi-scale representation. UNetR (Hatamizadeh et al. 2022) directly applies ViT-style encoders to 3D volumes, enabling end-to-end learning of volumetric spatial dependencies. TransBTS (Wang et al. 2021) further explores modality-specific attention strategies for multi-modal fusion. SegFormer (Xie et al. 2021) balances accuracy and efficiency by combining a lightweight Transformer backbone with compact fusion heads, and DeiT-UNet (Touvron et al. 2021) employs knowledge distillation to improve training stability in data-scarce scenarios. Despite their effectiveness in modeling long-range dependencies, these hybrid CNN–Transformer models often exhibit high memory consumption and potentially unstable long-range attention, which limits their scalability to high-resolution inputs in practical applications.

State Space Models

As a selective state-space model, Mamba has recently attracted considerable attention in vision applications, including medical image segmentation. A number of recent works have explored its potential in various architectures. VM-UNet (Ruan, Li, and Xiang 2024) integrates enhanced attention mechanisms with multi-scale fusion, improving both segmentation accuracy and efficiency. LightM-UNet (Liao et al. 2024) introduces a lightweight Mamba module, improving suitability for deployment in resource-constrained environments. Swin-UMamba (Liu et al. 2024) combines pre-trained Swin Transformers with Mamba to strengthen feature extraction and generalization. Mamba-UNet (Wang et al. 2024) employs purely visual Mamba blocks to better handle complex imaging patterns. Ultra-Light VM-UNet (Wu et al. 2025b) adopts parallel Mamba branches, reducing parameter overhead while maintaining competitive performance in dermoscopic lesion segmentation. MTMamba (Lin et al. 2024) designs a Mamba-based decoder for dense multi-task segmentation, yielding consistent improvements across multiple subtasks. MMR-Mamba (Zou et al. 2025) incorporates Mamba with spatial frequency-aware fusion, significantly enhancing detail preservation and contrast consistency in multi-contrast MRI reconstruction.

Despite these advances, Mamba still exhibits inherent limitations. Its unidirectional sequential structure struggles to capture direction-invariant and symmetric spatial structures that are prevalent in medical images, often resulting in perceptual asymmetries between horizontal and vertical axes. Moreover, the absence of an explicit multi-scale mechanism restricts its ability to model semantic variations across organs of different sizes, which is essential for achieving accurate and robust medical image segmentation.

Method

Overview of EccoMamba

As shown in Fig. 1, EccoMamba is built upon a U-shaped encoder–decoder architecture inspired by VM-UNet (Ruan, Li, and Xiang 2024). It integrates selective state-space modeling with structural enhancement modules to effectively capture multi-scale features and spatial dependencies in

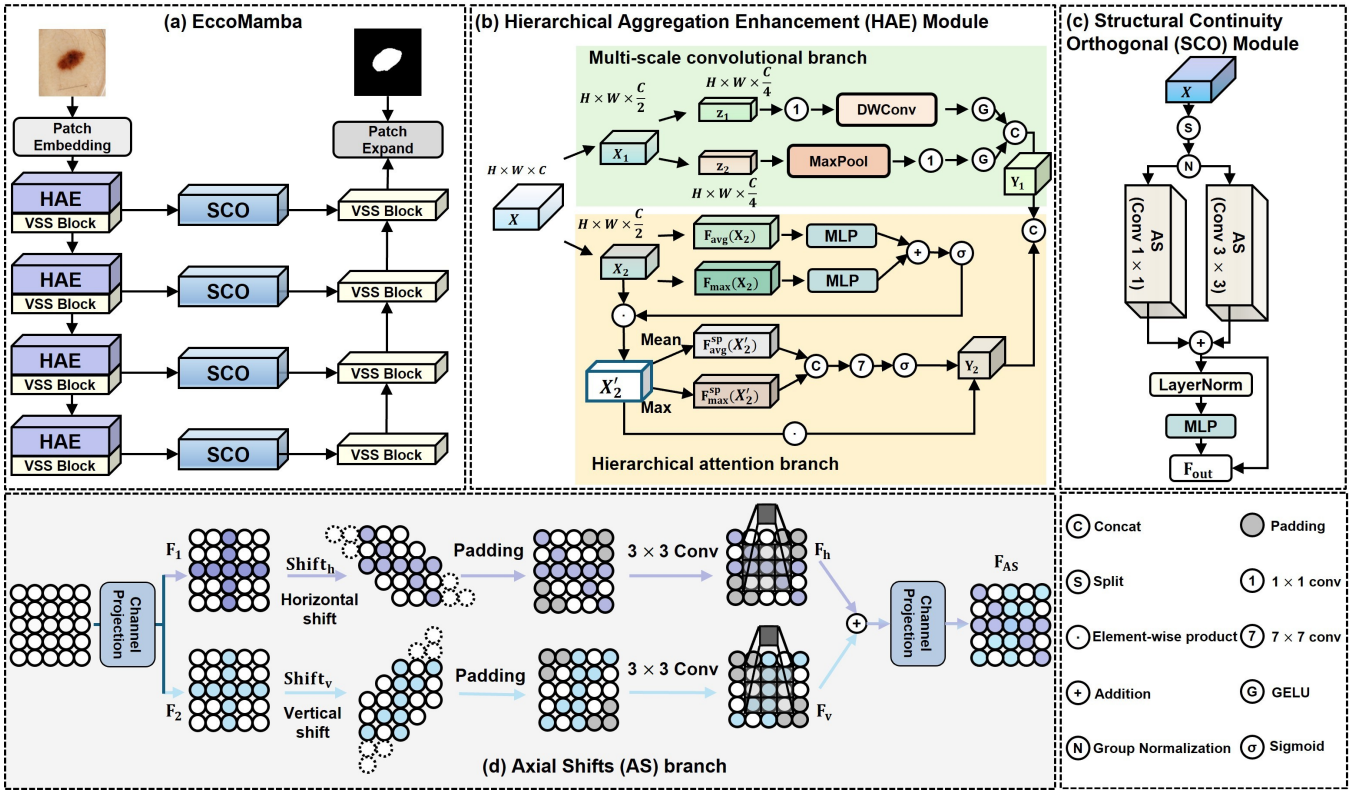

Figure 1: (a) Overview of the proposed EccoMamba architecture. (b) Hierarchical Aggregation Enhancement (HAE) module. (c) Structural Continuity Orthogonal (SCO) module. (d) Axial Shifts (AS) applies orthogonal shifts (horizontal and vertical) and dual convolutional pathways to model fine-grained spatial dependencies.

medical images. The input image is first projected into feature representations via a patch embedding layer, and is then processed by a four-stage encoder that performs hierarchical feature extraction and progressive downsampling. Each encoder stage incorporates a Hierarchical Aggregation Enhancement (HAE) module and a vision state space (VSS) block. The VSS block leverages the selective state-space mechanism to model long-range dependencies with linear computational complexity, enabling efficient capture of global contextual information. The HAE module extracts multi-scale features in parallel and integrates both channel-wise and spatial attention mechanisms to enhance feature representation. To mitigate the directional bias inherent in Mamba, the skip connections are augmented with a Structural Continuity Orthogonal (SCO) module, which performs Axial Shifts (AS) along horizontal and vertical directions. In the decoder, feature maps are progressively upsampled through VSS blocks and fused with the SCO-refined encoder features to restore spatial resolution and generate the final segmentation output.

Hierarchical Aggregation Enhancement

Medical images often exhibit highly variable spatial structures, ranging from fine cellular patterns to complex organ boundaries. This variability necessitates capturing hierarchical anatomical features across different scales while maintaining spatial coherence. To address this challenge, we introduce the HAE module.

The HAE module begins by splitting the input feature map $\boldsymbol{X} \in \mathbb{R}^{B\times H\times W\times C}$ into two equal parts along the channel dimension: $\boldsymbol{X}_1, \boldsymbol{X}_2 \in \mathbb{R}^{B\times H\times W\times (C/2)}$, where B is the batch size, H is the height, W is the width, and C is the number of channels. As shown in Fig. 1(b), HAE consists of two primary branches: a multi-scale convolutional branch and a hierarchical attention branch. In the convolutional branch, $\boldsymbol{X}_1$ is further divided into two pathways, $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$. $\boldsymbol{z}_1$ undergoes a 1×1 pointwise convolution, followed by a depth-wise separable 3×3 convolution (DWConv) and GELU activation to extract fine-grained local features. $\boldsymbol{z}_2$ passes through a 3×3 max-pooling operation, followed by a 1×1 convolution and GELU activation to aggregate coarse-scale context. The outputs of the two convolutional paths are fused to form $\boldsymbol{Y}_1$.

The second part of the input, $\boldsymbol{X}_2$, is processed through channel attention and spatial attention. Channel attention first applies global average pooling and max pooling to obtain $\boldsymbol{F}_{\text{avg}}(\boldsymbol{X}_2)$ and $\boldsymbol{F}_{\text{max}}(\boldsymbol{X}_2)$, which are passed through a two-layer multilayer perceptron (MLP) to generate channel-wise weights:

$$\begin{aligned} \boldsymbol{W}_{\text{ca}} &= \sigma(\text{MLP}(\boldsymbol{F}_{\text{avg}}(\boldsymbol{X}_2)) + \text{MLP}(\boldsymbol{F}_{\text{max}}(\boldsymbol{X}_2))), \\ \boldsymbol{X}_2' &= \boldsymbol{X}_2 \odot \boldsymbol{W}_{\text{ca}}, \end{aligned} \tag{1}$$

where $\sigma(\cdot)$ denotes the sigmoid function.

The spatial attention module applies average pooling and max pooling along the channel dimension of $\boldsymbol{X}'_2$ to generate $\boldsymbol{F}^{\text{sp}}_{\text{avg}}(\boldsymbol{X}'_2)$ and $\boldsymbol{F}^{\text{sp}}_{\text{max}}(\boldsymbol{X}'_2)$. These maps are concatenated and passed through a 7×7 convolution to obtain:

$$\boldsymbol{W}_{\text{sa}} = \sigma\Big(\text{Conv}_{7\times7}\Big(\text{Concat}\left[\boldsymbol{F}^{\text{sp}}_{\text{avg}}(\boldsymbol{X}'_2),\, \boldsymbol{F}^{\text{sp}}_{\text{max}}(\boldsymbol{X}'_2)\right]_c\Big)\Big). \tag{2}$$

The enhanced feature is given by:

$$\boldsymbol{Y}_2 = \boldsymbol{X}'_2 \odot \boldsymbol{W}_{\text{sa}}. \tag{3}$$

The final output is obtained by concatenating $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ and applying a residual connection with the original input.

Structural Continuity Orthogonal Module

Medical images inherently exhibit strong spatial continuity along anatomical structures. However, traditional Mamba architectures flatten 2D feature maps into 1D sequences and apply unidirectional processing, which disrupts spatial coherence and leads to structural discontinuities. The SCO module addresses this issue by explicitly modeling orthogonal dependencies across spatial axes.

As shown in Fig. 1(c), the input $\boldsymbol{X}$ is split into two parts along the channel dimension. Each part is processed by a distinct AS branch: one uses a 1×1 convolution for efficient channel-wise interaction, while the other uses a 3×3 convolution to capture richer local patterns. Within each AS branch (Fig. 1(d)), the features are projected into two components: one is shifted horizontally and the other vertically. Zero-padding preserves the original spatial size. The shifted features are convolved and fused through element-wise addition, followed by a second projection:

$$\begin{aligned} \boldsymbol{F}_{\text{h}} &= \text{Conv}(\text{Shift}_h(\boldsymbol{F}_1)), \\ \boldsymbol{F}_{\text{v}} &= \text{Conv}(\text{Shift}_v(\boldsymbol{F}_2)), \\ \boldsymbol{F}_{\text{AS}} &= \text{Conv}_{1\times1}(\boldsymbol{F}_{\text{h}} + \boldsymbol{F}_{\text{v}}). \end{aligned} \tag{4}$$

The outputs of the two AS branches are combined and refined:

$$\begin{aligned} \boldsymbol{F}_{\text{fused}} &= \boldsymbol{F}_{AS_{1\times1}} + \boldsymbol{F}_{AS_{3\times3}}, \\ \boldsymbol{F}_{\text{out}} &= \text{MLP}(\text{LayerNorm}(\boldsymbol{F}_{\text{fused}})) + \boldsymbol{F}_{\text{fused}}. \end{aligned} \tag{5}$$

This orthogonal design allows SCO to better capture structural continuity while preserving anatomical details. By combining cross-directional shifts with dual-scale convolutions and channel-wise interactions, SCO mitigates directional bias and strengthens boundary delineation.

Loss Function

To improve performance in multi-class medical image segmentation, we adopt a hybrid loss consisting of Dice loss and cross-entropy loss. Cross-entropy enforces accurate pixel-wise classification, while Dice loss directly optimizes the overlap between predictions and ground truth. The total loss is:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{ce}}\mathcal{L}_{\text{ce}} + \lambda_{\text{dice}}\mathcal{L}_{\text{dice}}, \tag{6}$$

where λ_{ce} and λ_{dice} balance the contributions of the two components.

Experiments

Datasets

Extensive experiments are conducted on four publicly available datasets to evaluate the effectiveness and robustness of the proposed method. The **ISIC 2017 and ISIC 2018 datasets** (Milton 2019), used for skin lesion segmentation, consist of 2,150 and 2,694 dermoscopic images with pixel-level annotations, respectively. For **ISIC 2017**, the data are split into 1,500 training and 650 testing images following the official protocol. For **ISIC 2018**, we follow a standard 7:3 train–test split, resulting in 1,886 training and 808 testing images. The **ACDC dataset** (Bernard et al. 2018), developed for automatic cardiac diagnosis, contains MRI scans from 100 patients with manual annotations for the right ventricle (RV), left ventricle (LV), and myocardium (MYO), partitioned into 70 training, 10 validation, and 20 testing samples. The **Synapse dataset** (Landman et al. 2015), designed for multi-organ CT segmentation, includes annotations for eight organs from 30 subjects, split into 18 training and 12 testing cases.

Evaluation Metrics

The performance of all methods on the Synapse and ACDC datasets is evaluated using the Dice similarity coefficient (DSC) and the 95th percentile Hausdorff distance (HD95). For the ISIC 2018 and ISIC 2017 datasets, we use sensitivity (Sens), specificity (Spec), accuracy (Acc), and DSC as evaluation metrics.

Implementation Details

All models are implemented in PyTorch and trained on a single NVIDIA RTX 4090 GPU. We adopt the AdamW optimizer with an initial learning rate of 1×10^{-3} and a weight decay of 1×10^{-2}, together with a cosine learning rate decay schedule. The batch size is uniformly set to 24 for all datasets. Each model is trained for 300 epochs, and the best checkpoint is selected based on the validation DSC. During inference on the Synapse dataset, a sliding-window strategy is employed to mitigate boundary artifacts. Mixed-precision training (AMP) is used to accelerate computation and reduce memory usage.

Comparisons with State-of-the-Art Methods

We compare EccoMamba with state-of-the-art (SOTA) methods from three categories. **U-Net-based methods** include MALUNet (Ruan et al. 2022), Attention Swin U-Net (Aghdam et al. 2023), ASP-VM-UNet (Bao et al. 2025), MSCD-VM-UNet (Huang et al. 2025), LeViT-UNet-384 (Xu et al. 2023), UNetR (Hatamizadeh et al. 2022), and TransClaw U-Net (Yao et al. 2022). **Transformer-based methods** include TransUNet (Chen et al. 2024), Swin-UNet (Cao et al. 2022), MISSFormer (Huang et al. 2022), SynergyNet (Gorade et al. 2024), MixFormer (Liu et al. 2025a), HiFormer-L (Heidari et al. 2023), CSWin-UNet (Liu et al. 2025b), and GLoG-CSUNet (Zarch et al. 2025). **Mamba-based methods** include VM-UNet (Ruan, Li, and Xiang 2024), HC-Mamba (Xu 2024), H-VmUNet (Wu et al. 2025a), and CC-ViM (Zhu et al.

Method	ISIC 2018				ISIC 2017			
	Sens	Spec	Acc	DSC	Sens	Spec	Acc	DSC
MALUNet	88.90	97.25	95.48	89.31	88.24	97.62	95.83	88.96
Attention Swin U-Net	80.57	**98.26**	94.80	85.40	84.92	**98.47**	95.91	88.59
UltraLight VM-UNet	88.32	95.96	93.86	88.76	85.02	98.11	96.14	86.89
LightM-UNet	88.29	97.65	**95.55**	88.98	87.04	97.74	95.95	87.80
HC-Mamba	88.90	97.08	94.84	89.26	86.99	97.47	95.17	88.18
VM-UNet	88.82	96.49	94.39	89.67	**91.87**	96.30	95.63	86.35
H-VmUNet	89.96	95.37	93.89	88.97	84.94	98.20	96.21	87.09
CC-ViM	88.74	97.32	95.23	90.06	88.70	97.19	95.60	88.74
MSCD-VM-UNet	89.71	97.05	95.26	90.21	90.44	97.23	95.92	88.78
ASP-VM-UNet	89.97	95.33	93.83	89.09	83.25	96.64	93.48	85.77
EccoMamba	**90.84**	96.82	95.17	**91.18**	88.60	98.29	**96.84**	**89.39**

Table 1: Comparison of segmentation performance on the ISIC 2018 and ISIC 2017 datasets.

2025). All methods are reproduced and evaluated under our experimental setting.

Skin Lesion Segmentation on the ISIC Datasets. As shown in Table 1, EccoMamba achieves the best overall performance on both ISIC 2018 and ISIC 2017. On ISIC 2018, it attains a DSC of 91.18%, outperforming the previous best method (88.98%) by 2.20 percentage points. On ISIC 2017, EccoMamba also obtains the highest DSC (89.39%) and the highest classification accuracy (96.84%), while maintaining competitive sensitivity and specificity. These results indicate that the proposed architecture is particularly effective for delineating challenging lesion boundaries under diverse imaging conditions.

Cardiac Segmentation on the ACDC Dataset. As reported in Table 2, EccoMamba achieves the highest performance on the ACDC dataset, with a DSC of 91.83%, outperforming VM-UNet (Ruan, Li, and Xiang 2024), which achieves 91.04%. These results highlight EccoMamba's strong capability in maintaining segmentation consistency and accuracy, particularly in anatomically complex regions.

Multi-Organ Segmentation on the Synapse Dataset. As summarized in Table 3, EccoMamba achieves the highest average DSC (81.51%) on the multi-organ segmentation task, together with a competitive HD95 of 22.02. It attains either the best or second-best performance on most organs, demonstrating strong generalization across both large and small anatomical structures. In particular, the high DSC scores on large organs (e.g., liver) and anatomically complex regions (e.g., stomach) highlight the effectiveness of the proposed modules in modeling multi-scale context and spatial continuity.

We further conduct paired t-tests on the DSC scores across all test subjects. EccoMamba achieves statistically significant improvements over all baseline models ($p < 0.01$), indicating that the observed gains are not due to random variation.

Ablation Study

The ablation study in Table 4 highlights the individual and combined contributions of the HAE and SCO modules on the ISIC 2018 dataset. Introducing the HAE module

Method	RV	MYO	LV	DSC	HD95
UNetR	85.29	86.52	94.02	86.61 ± 1.63	–
LeViT-UNet-384	89.55	87.64	93.76	90.32 ± 1.24	–
Swin-UNet	87.94	86.78	94.71	89.81 ± 1.11	1.32 ± 0.98
MISSFormer	86.36	85.75	91.59	87.90 ± 1.79	6.85 ± 0.21
SynergyNet	87.68	86.60	95.06	89.78 ± 1.28	1.50 ± 0.07
TransUNet	88.86	84.54	95.73	89.71 ± 1.15	1.82 ± 0.06
VM-UNet	88.44	89.22	95.45	91.04 ± 1.12	**1.30 ± 0.95**
MixFormer	89.02	88.46	95.55	91.04 ± 1.07	1.42 ± 0.21
GLoG-CSUNet	86.63	88.36	94.87	89.95 ± 1.26	4.14 ± 0.18
EccoMamba	**89.91**	**89.71**	**95.87**	**91.83 ± 1.06**	1.37 ± 0.11

Table 2: Comparison of segmentation performance on the ACDC dataset.

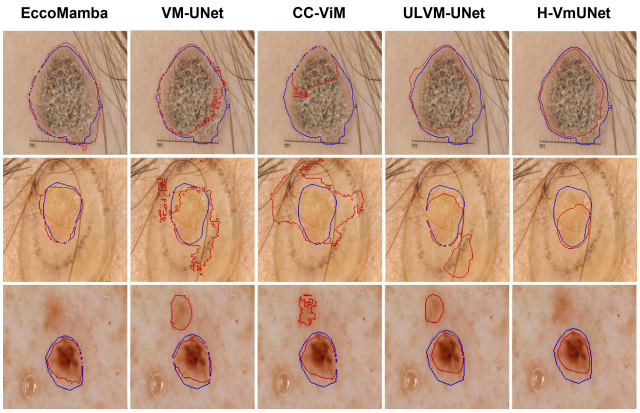

Figure 2: Visual comparison of EccoMamba and other models on the ISIC 2017 dataset. The ground truth is shown in blue contours, while the segmentation results are shown in red.

(Base+HAE) improves the DSC from 88.52% to 90.42%, demonstrating its effectiveness in capturing multi-scale contextual information and refining indistinct lesion boundaries. Adding the SCO module (Base+SCO) increases the DSC to 90.22%, indicating its ability to enhance spatial continuity and localize lesions more accurately in anatomically structured regions. When both modules are integrated in the full EccoMamba model, the DSC reaches 91.18%, confirming the complementary strengths of HAE and SCO in producing more precise and structurally coherent segmentations.

Visualization of Segmentation Results

To provide a more intuitive comparison of segmentation performance, we visualize the results of EccoMamba and several representative models on the ISIC 2017 dataset, as shown in Fig. 2. The visualizations demonstrate that EccoMamba produces more accurate and smoother segmentation contours, with higher overlap with the ground truth, particularly in challenging cases with blurry or ambiguous lesion boundaries. Compared to other methods, EccoMamba exhibits stronger contour alignment and better delineation of fine-grained structures, reflecting its superior ability to preserve anatomical consistency across diverse lesion types.

Method	Aorta	Gallbladder	Kidney (L)	Kidney (R)	Liver	Pancreas	Spleen	Stomach	DSC (%)	HD95
TransDeepLab	86.02	68.17	81.66	78.72	93.41	60.67	85.12	75.53	78.66 ± 1.71	29.38 ± 1.92
UNetR	**89.80**	56.30	85.60	84.52	94.57	60.47	85.00	70.46	78.34 ± 1.68	18.59 ± 1.74
TransClaw U-Net	85.87	61.38	84.83	79.36	94.28	57.65	87.74	73.55	78.08 ± 1.59	26.38 ± 1.83
LeViT-UNet-384	87.33	62.23	84.61	80.25	93.11	59.07	88.86	72.76	78.53 ± 1.43	**16.84 ± 1.64**
Swin-UNet	85.71	66.08	82.73	78.46	93.90	58.90	88.83	76.23	78.86 ± 1.39	24.80 ± 1.72
HiFormer-L	85.84	65.20	83.07	78.87	93.79	59.08	**91.00**	80.76	79.70 ± 1.41	20.82 ± 1.66
TransUNet	87.22	60.04	81.92	76.39	94.41	54.54	84.72	76.19	76.93 ± 1.48	30.18 ± 1.94
VM-UNet	88.17	69.51	84.35	80.95	93.57	62.17	88.45	76.21	80.42 ± 1.40	24.84 ± 1.78
CC-ViM	85.90	67.44	85.86	80.85	94.13	62.06	89.10	79.58	80.62 ± 1.44	29.39 ± 1.83
SWMA-UNet	86.04	68.84	**86.97**	**82.60**	94.19	57.33	89.82	78.05	80.48 ± 1.42	18.30 ± 1.73
CSWin-UNet	87.13	63.89	82.54	78.94	94.64	**64.05**	88.38	80.66	80.03 ± 1.39	26.74 ± 1.76
GLoG-CSUNet	88.21	70.44	81.67	80.55	93.33	61.45	88.24	80.37	80.53 ± 1.41	24.12 ± 1.74
EccoMamba	87.43	**72.01**	83.82	81.01	**95.29**	60.64	90.18	**81.67**	**81.51 ± 1.32**	22.02 ± 1.71

Table 3: Comparison of segmentation performance on the Synapse dataset. Results are averaged over three independent runs (mean ± std.).

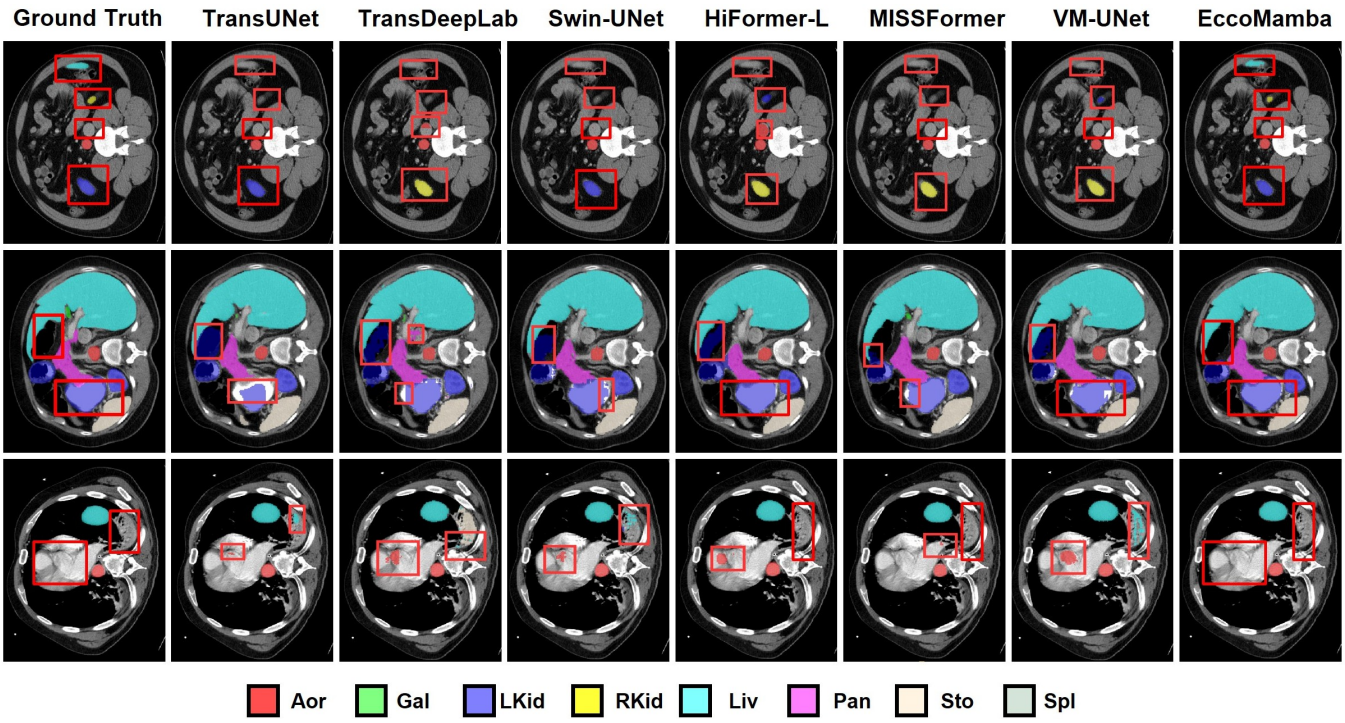

Figure 3: Visual comparison of EccoMamba and other models on the Synapse dataset. Red boxes highlight regions where EccoMamba yields more accurate segmentations.

Fig. 3 presents visual comparisons between EccoMamba and other competing models on the Synapse dataset. Red boxes highlight regions containing anatomically challenging organs, such as the kidneys, pancreas, and gallbladder. EccoMamba consistently generates smoother boundaries and more accurate shapes, particularly in areas with low contrast or structural ambiguity. For example, in the first row, it correctly distinguishes the left and right kidneys and provides precise liver delineation. In contrast, other models show boundary shifts and segmentation artifacts, validating the effectiveness of EccoMamba in modeling spatial continuity and capturing anatomical structures.

Fig. 4 shows visual comparisons on the ACDC dataset. Even in regions with unclear boundaries, complex tissue morphology, or tightly packed structures, EccoMamba accurately delineates cardiac structures such as the LV, RV, and MYO. In contrast, other models exhibit issues such as discontinuities or misaligned contours. EccoMamba, however, delivers better structural continuity and spatial alignment, further validating the practical efficacy and robustness of the proposed spatial enhancement modules.

	Sens	Spec	Acc	DSC
Base	86.63	96.56	93.84	88.52
Base+HAE	90.51	96.34	94.74	90.42
Base+SCO	90.57	96.15	94.62	90.22
EccoMamba	90.84	96.82	95.17	91.18

Table 4: Ablation study of EccoMamba on the ISIC 2018 dataset.

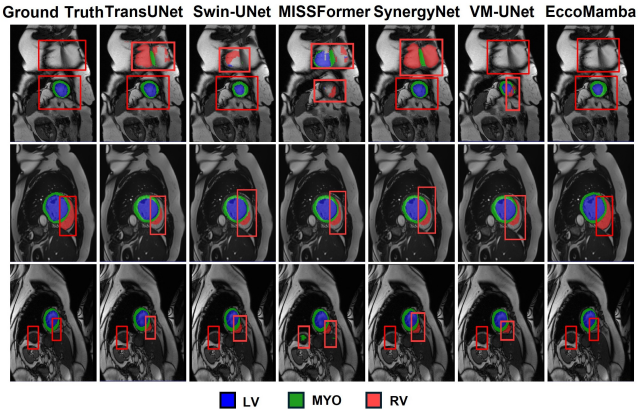

Figure 4: Segmentation results on the ACDC dataset. Red boxes highlight regions where EccoMamba yields more accurate predictions.

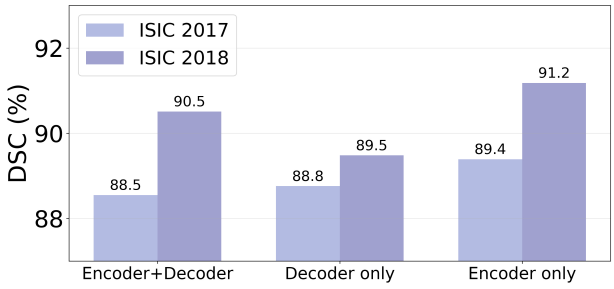

Figure 5: Stage-wise placement analysis of the HAE module.

Stage-wise Placement of the HAE Module

To investigate the optimal placement of the Hierarchical Aggregation Enhancement (HAE) module, we conduct experiments incorporating HAE at different positions: (1) both encoder and decoder stages, (2) decoder-only, and (3) encoder-only. Experimental results show that the encoder-only configuration achieves the best performance, with DSC scores of 91.18% on ISIC 2018 and 89.39% on ISIC 2017, outperforming both the encoder–decoder (90.51% and 88.55%) and decoder-only (89.48% and 88.76%) configurations. This demonstrates that early-stage hierarchical feature enhancement during encoding is most critical for optimal segmentation performance.

Method	Parameters	FLOPs	Memory
MISSFormer (2023)	42.46M	7.28G	0.56 GB
LeViT-UNet-384 (2023)	39.13M	2.25G	0.26 GB
Swin-UNet (2023)	27.17M	5.95G	0.31 GB
nnFormer (2023)	150.05M	704.23G	7.09 GB
HiFormer-L (2023)	25.51M	17.81G	0.42 GB
TransUNet (2024)	105.28M	593.00G	7.55 GB
VM-UNet (2024)	27.43M	3.15G	0.42 GB
H-VmUNet (2024)	8.97M	0.57G	0.35 GB
ULVM-UNet (2024)	0.18M	0.06G	0.15 GB
CSWin-UNet (2025)	23.57M	4.72G	0.31 GB
GLoG-CSUNet (2025)	40.62M	8.43G	0.58 GB
EccoMamba	38.30M	6.22G	0.63 GB

Table 5: Model parameters and computational complexity.

Memory and Computational Efficiency

We assess the efficiency of EccoMamba in terms of parameter count, FLOPs, and memory consumption. As shown in Table 5, EccoMamba requires 38.30M parameters, 6.22 GFLOPs, and 0.63 GB of memory, offering a favorable trade-off between accuracy and complexity. Compared with heavy models such as nnFormer (Zhou et al. 2023) and TransUNet (Chen et al. 2024), it is substantially more efficient, while still outperforming them in segmentation accuracy. Although ultra-light designs (e.g., ULVM-UNet (Wu et al. 2025b)) have fewer parameters and FLOPs, they achieve lower accuracy than EccoMamba, indicating that our design strikes a more balanced compromise between performance and computational cost.

Conclusion

In this paper, we have presented EccoMamba, a medical image segmentation framework that incorporates two key components: the Hierarchical Aggregation Enhancement (HAE) module and the Structural Continuity Orthogonal (SCO) module. The HAE module captures anatomical features across multiple scales through parallel multi-scale convolutions and a lightweight attention mechanism, enabling fine-to-coarse hierarchical representations. The SCO module enhances spatial continuity by applying orthogonal Axial Shifts (AS) that model cross-dimensional dependencies and reinforce structural consistency.

Extensive experiments across four benchmark datasets validate the effectiveness of EccoMamba. Specifically, evaluations on ISIC 2018, ISIC 2017, Synapse, and ACDC show that our method consistently achieves superior segmentation accuracy and robustness compared with existing state-of-the-art approaches. These results highlight the practical value of the proposed architectural design and its potential for broader applications in medical image analysis.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant Nos. 62302156, 62402351, 62402349, 62572401, and 62201460), the Natural Science

Foundation of Hubei Province (Grant Nos. 2024AFB275 and 2024AFB127), the Natural Science Foundation of Hunan Province (Grant No. 2023JJ40180), the Key Research and Development Program of Shaanxi (No. 2025SF-YBXM-424), and the Wuhan Textile University Foundation (Grant No. 2024309).

References

Aghdam, E. K.; Azad, R.; Zarvani, M.; and Merhof, D. 2023. Attention Swin U-Net: Cross-contextual attention mechanism for skin lesion segmentation. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, 1–5. IEEE.

Bao, M.; Lyu, S.; Xu, Z.; Zhao, Q.; Zeng, C.; Bai, W.; and Cheng, G. 2025. ASP-VMUNet: Atrous Shifted Parallel Vision Mamba U-Net for Skin Lesion Segmentation. *arXiv preprint arXiv:2503.19427*.

Bernard, O.; Lalande, A.; Zotti, C.; Cervenansky, F.; Yang, X.; Heng, P.-A.; Cetin, I.; Lekadir, K.; Camara, O.; Ballester, M. A. G.; et al. 2018. 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.

Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; and Wang, M. 2022. Swin-UNet: UNet-like pure transformer for medical image segmentation. In *European Conference on Computer Vision*, 205–218. Springer.

Chen, J.; Mei, J.; Li, X.; Lu, Y.; Yu, Q.; Wei, Q.; Luo, X.; Xie, Y.; Adeli, E.; Wang, Y.; et al. 2024. TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers. *Medical Image Analysis*, 97: 103280.

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

Gorade, V.; Mittal, S.; Jha, D.; and Bagci, U. 2024. SynergiNet: Bridging the gap between discrete and continuous representations for precise medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 7768–7777.

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

Hatamizadeh, A.; Tang, Y.; Nath, V.; Yang, D.; Myronenko, A.; Landman, B.; Roth, H. R.; and Xu, D. 2022. UNETR: Transformers for 3D medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 574–584.

Heidari, M.; Kazerouni, A.; Soltany, M.; Azad, R.; Aghdam, E. K.; Cohen-Adad, J.; and Merhof, D. 2023. HiFormer: Hierarchical multi-scale representations using transformers for medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 6202–6212.

Huang, X.; Deng, Z.; Li, D.; Yuan, X.; and Fu, Y. 2022. MissFormer: An effective transformer for 2D medical image segmentation. *IEEE Transactions on Medical Imaging*, 42(5): 1484–1494.

Huang, Z.; Wang, S.; Hou, M.; Yu, Z.; Wang, S.; Li, X.; Yan, Y.; Liu, Y.; and Gregersen, H. 2025. MSCD-VM-UNet: A Vision Mamba Combining Multi-Scale Global and Local Feature Extraction with Cross-Domain Feature Fusion for Medical Image Segmentation. *IEEE Journal of Biomedical and Health Informatics*.

Landman, B.; Xu, Z.; Iglesias, J.; Styner, M.; Langerak, T.; and Klein, A. 2015. MICCAI Multi-Atlas Labeling Beyond the Cranial Vault—Workshop and Challenge. In *Proceedings of MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, volume 5, 12.

Liao, W.; Zhu, Y.; Wang, X.; Pan, C.; Wang, Y.; and Ma, L. 2024. LightM-UNet: Mamba assists in lightweight UNet for medical image segmentation. *arXiv preprint arXiv:2403.05246*.

Lin, B.; Jiang, W.; Chen, P.; Zhang, Y.; Liu, S.; and Chen, Y.-C. 2024. MTMamba: Enhancing multi-task dense scene understanding by mamba-based decoders. In *European Conference on Computer Vision*, 314–330. Springer.

Liu, J.; Li, K.; Huang, C.; Dong, H.; Song, Y.; and Li, R. 2025a. MixFormer: A Mixed CNN-Transformer Backbone for Medical Image Segmentation. *IEEE Transactions on Instrumentation and Measurement*, 74: 1–20.

Liu, J.; Yang, H.; Zhou, H.-Y.; Yu, L.; Liang, Y.; Yu, Y.; Zhang, S.; Zheng, H.; and Wang, S. 2024. Swin-UMamba: Adapting Mamba-based vision foundation models for medical image segmentation. *IEEE Transactions on Medical Imaging*. Early access.

Liu, X.; Gao, P.; Yu, T.; Wang, F.; and Yuan, R.-Y. 2025b. CSWin-UNet: Transformer UNet with cross-shaped windows for medical image segmentation. *Information Fusion*, 113: 102634.

Milton, M. A. A. 2019. Automated skin lesion classification using ensemble of deep neural networks in ISIC 2018: Skin lesion analysis towards melanoma detection challenge. *arXiv preprint arXiv:1901.10802*.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241. Springer.

Ruan, J.; Li, J.; and Xiang, S. 2024. VM-UNet: Vision Mamba UNet for medical image segmentation. *arXiv preprint arXiv:2402.02491*.

Ruan, J.; Xiang, S.; Xie, M.; Liu, T.; and Fu, Y. 2022. MALUNet: A multi-attention and light-weight UNet for skin lesion segmentation. In *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 1150–1156. IEEE.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jégou, H. 2021. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, 10347–10357. PMLR.

Wang, H.; Cao, P.; Wang, J.; and Zaiane, O. R. 2022. UCTRANSNET: Rethinking the skip connections in U-Net from a channel-wise perspective with transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 2441–2449.

Wang, W.; Chen, C.; Ding, M.; Yu, H.; Zha, S.; and Li, J. 2021. Transbts: Multimodal brain tumor segmentation using transformer. In *International conference on medical image computing and computer-assisted intervention*, 109–119. Springer.

Wang, Z.; Zheng, J.-Q.; Zhang, Y.; Cui, G.; and Li, L. 2024. Mamba-UNet: UNet-like pure visual mamba for medical image segmentation. *arXiv preprint arXiv:2402.05079*.

Wu, R.; Liu, Y.; Liang, P.; and Chang, Q. 2025a. H-VMUNet: High-order vision mamba UNet for medical image segmentation. *Neurocomputing*, 624: 129447.

Wu, R.; Liu, Y.; Ning, G.; Liang, P.; and Chang, Q. 2025b. Ultralight VM-UNet: Parallel vision mamba significantly reduces parameters for skin lesion segmentation. *Patterns*, 101298.

Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J. M.; and Luo, P. 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34: 12077–12090.

Xu, G.; Zhang, X.; He, X.; and Wu, X. 2023. Levit-unet: Make faster encoders with transformer for medical image segmentation. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, 42–53. Springer.

Xu, J. 2024. HC-Mamba: Vision Mamba with hybrid convolutional techniques for medical image segmentation. *arXiv preprint arXiv:2405.05007*.

Yao, C.; Hu, M.; Li, Q.; Zhai, G.; and Zhang, X.-P. 2022. TransClaw U-Net: Claw U-Net with transformers for medical image segmentation. In *2022 5th International Conference on Information Communication and Signal Processing (ICICSP)*, 280–284. IEEE.

Zarch, N. E.; et al. 2025. GLoG-CSUNet: Enhancing Vision Transformers with Adaptable Radiomic Features for Medical Image Segmentation. In *Proceedings of the 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Zhou, H.-Y.; Guo, J.; Zhang, Y.; Han, X.; Yu, L.; Wang, L.; and Yu, Y. 2023. nnFormer: Volumetric medical image segmentation via a 3D transformer. *IEEE Transactions on Image Processing*, 32: 4036–4045.

Zhou, Z.; Rahman Siddiquee, M. M.; Tajbakhsh, N.; and Liang, J. 2018. 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*, 3–11. Springer.

Zhu, Y.; Zhang, D.; Lin, Y.; Feng, Y.; and Tang, J. 2025. Merging Context Clustering with Visual State Space Models for Medical Image Segmentation. *IEEE Transactions on Medical Imaging*, 44(5): 2131–2142.

Zhu, Z.; Yan, Y.; Xu, R.; Zi, Y.; and Wang, J. 2022. Attention-UNet: A deep learning approach for fast and accurate segmentation in medical imaging. *Journal of Computer Science and Software Applications*, 2(4): 24–31.

Zou, J.; Liu, L.; Chen, Q.; Wang, S.; Hu, Z.; Xing, X.; and Qin, J. 2025. MMR-Mamba: Multi-modal MRI reconstruction with Mamba and spatial-frequency information fusion. *Medical Image Analysis*, 102: 103549.