LHU-Net: a Lean Hybrid U-Net for Cost-efficient, High-performance Volumetric Segmentation

Yousef Sadegheih [†1], Afshin Bozorgpour [†1], Pratibha Kumari[1], Reza Azad[2], and Dorit Merhof[1,3]

[1] Faculty of Informatics and Data Science, University of Regensburg, Regensburg, 93053, Germany
[2] Faculty of Electrical Engineering and Information Technology, RWTH Aachen University, 52062 Aachen, Germany
[3] Fraunhofer Institute for Digital Medicine MEVIS, Bremen 28359, Germany
dorit.merhof@ur.de
† *Indicates equal contribution*

Abstract. The rise of Transformer architectures has advanced medical image segmentation, leading to hybrid models that combine Convolutional Neural Networks (CNNs) and Transformers. However, these models often suffer from excessive complexity and fail to effectively integrate spatial and channel features, crucial for precise segmentation. To address this, we propose LHU-Net, a Lean Hybrid U-Net for volumetric medical image segmentation. LHU-Net prioritizes spatial feature extraction before refining channel features, optimizing both efficiency and accuracy. Evaluated on four benchmark datasets (Synapse, Left Atrial, BraTS-Decathlon, and Lung-Decathlon), LHU-Net consistently outperforms existing models across diverse modalities (CT/MRI) and output configurations. It achieves state-of-the-art Dice scores while using four times fewer parameters and 20% fewer FLOPs than competing models, without the need for pre-training, additional data, or model ensembles. With an average of 11 million parameters, LHU-Net sets a new benchmark for computational efficiency and segmentation accuracy. Our implementation is available on github.com/xmindflow/LHUNet.

Keywords: Volumetric Medical Image Segmentation · Light Hybrid Architecture · Computational Efficiency

1 Introduction

Medical image acquisition technologies such as MRI, CT, and X-ray enable non-invasive imaging of anatomical structures, making image segmentation essential for diagnosis, intervention planning, and disease assessment. Manual segmentation is time-consuming and prone to inconsistencies, necessitating automated methods. While deep learning approaches, particularly Convolutional Neural Networks (CNNs), have advanced medical segmentation, their performance can be limited by a lack of global context [2,11]. Vision Transformers (ViTs) [6], which use self-attention to capture global context, have addressed this gap but often

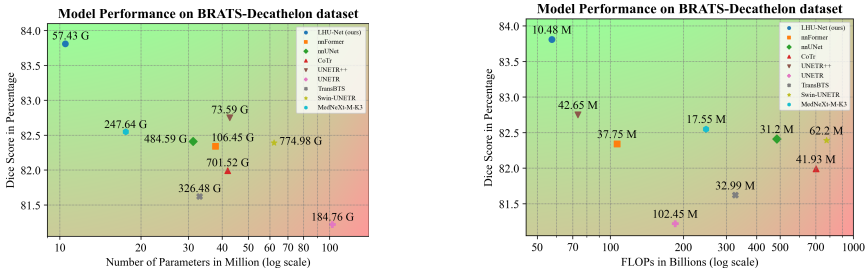

Fig. 1: Model performance on the BraTS-Decathlon dataset: (left) DSC vs. parameter count with FLOPs annotated next to each point and (right) DSC vs FLOPs with parameter count annotated next to each point.

fail to preserve fine-grained local details essential for accurate segmentation [3]. Segmentation becomes even more challenging in the 3D domain due to the increased data volume and complexity. However, 3D models have been shown to outperform 2D models by capturing better context and improving segmentation accuracy [11,4]. Despite this, 3D models typically require higher computational power and parameter counts [4,8]. A common trend is to use the same module across all layers to achieve state-of-the-art performance [4,27,5,20]. However, we argue that using tailored modules for different layers can make the model more efficient, achieving better segmentation with lower computational costs. Hybrid models, which combine CNNs for local feature extraction with ViTs for global context, have gained popularity in tackling these challenges. While promising, many existing hybrid models increase complexity without proportional improvements in performance, leading to excessive computational costs [28,9,7]. To address this, LHU-Net (**L**ean **H**ybrid **U-Net**) optimizes attention mechanisms by using spatial attention in early layers for local feature extraction and channel attention in deeper layers for broader contextual understanding. This approach balances model complexity with efficiency, significantly reducing computational cost while improving 3D medical image segmentation performance. In this paper, we present LHU-Net with the following contributions: ❶ **Efficient Hybrid Attention selection for Better Contextual Understanding:** LHU-Net utilizes two specialized attention mechanisms within ViTs to capture both local and global contexts effectively. It combines Large Kernel Attention with an extra deformable attention layer (LKAd) to manage long-range dependencies and maintain high-frequency details. In the early layers, spatial attention focuses on local features, while in deeper layers, channel attention captures global feature interactions, ensuring a comprehensive feature extraction process suited for medical image segmentation. ❷ **High Efficiency with Minimal Cost:** On BraTS, it reduces parameters by 75% and FLOPS by 21% while maintaining top DSC performance, averaging 11 M parameters across datasets (Fig. 1). ❸ **Robust**

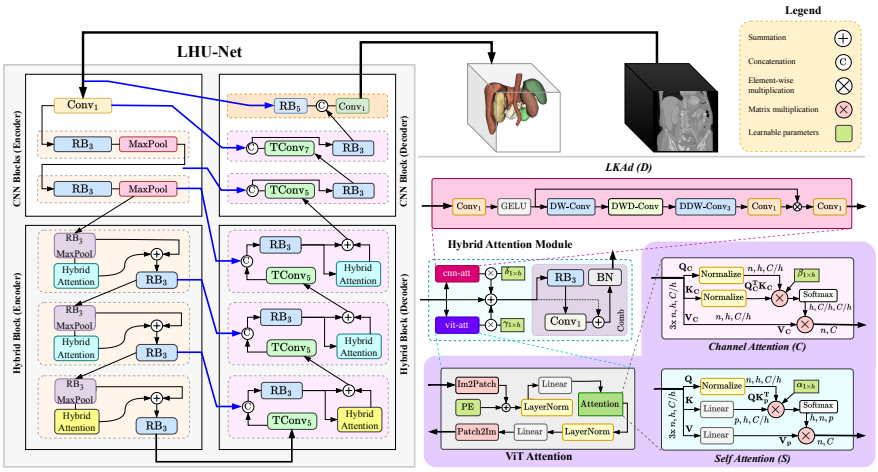

Fig. 2: Overview of the LHU-Net's encoder-decoder structure, featuring CNN and hybrid blocks. The hybrid block can switch between OmniFocus Attention (Channel Attention) and Self-Adaptive Contextual Fusion (Self Attention).

Across Modalities: Excelling in CT, MR, and multimodal datasets, LHU-Net handles both single- and multi-label segmentation tasks with high versatility.

2 Methodology

LHU-Net extracts features efficiently by combining convolutional blocks with hybrid attention mechanisms to capture both local and global contexts. As shown in Fig. 2, its U-Net encoder-decoder processes 3D patches, with the encoder refining multi-scale features and skip connections transferring key details for segmentation reconstruction. Initial convolutional blocks enhance local features and reduce spatial dimensions, while core hybrid blocks integrate large kernel convolutional attention followed by deformable convolution and spatial-channel ViT attention to capture global features and long-range dependencies. Detailed explanations of each stage follow.

2.1 CNN Blocks

Our architecture employs CNN blocks to capture local features at high and mid-frequency levels efficiently. Unlike models relying solely on ViTs, we use ResBlocks with MaxPool to downsample spatial dimensions while preserving key details for later hybrid processing. Before the encoder, an initial refinement is applied using a point-wise convolution (PW-Conv) followed by PReLU and batch normalization to maintain spatial resolution and sharpen boundaries. Each encoder layer employs a ResBlock that begins with a depth-wise convolution, batch

Table 1: LHU-Net training configuration along with parameter and FLOPs comparisons with SOTA models. Blue and red indicate the best and second-best results, respectively.

LHU-Net Details	**Synapse**		**Lung-Decathelon**		**BraTS-Decathelon**		**LA**	
Patch size	128 x 128 x 64		192 x 192 x 32		128 x 128 x 128		96 x 96 x 96	
Base learning rate	0.003		0.003		0.01		0.01	
Downsample[4]	[2, 2, 1], 2, 2, 2, 2		[2, 2, 1], [2, 2, 1], 2, 2, 2		2, 2, 2, 2, 2		2, 2, 2, 2, 2	
Methods	**Params.↓**	**FLOPs↓**	**Params.↓**	**FLOPs↓**	**Params.↓**	**FLOPs↓**	**Params.↓**	**FLOPs↓**
nnUNet [11]	30.71 M	476.58 G	30.79 M	561.82 G	31.20 M	484.59 G	30.79 M	539.34 G
MedNeXt-M-K3 [19]	17.55 M	123.47 G	17.55 M	137.48 G	17.55 M	247.64 G	17.55 M	103.84 G
MedNeXt-M-K5 [19]	18.26 M	153.52 G	18.26 M	171.29 G	18.26 M	307.74 G	18.26 M	129.20 G
CoTr [25]	41.87 M	334.23 G	41.86 M	375.11 G	41.93 M	701.52 G	41.86 M	281.33 G
nnFormer [28]	150.5 M	213.4 G	149.12 M	136.10 G	37.75 M	106.45 G	149.22 M	102.35 G
UNETR++ [20]	42.96 M	47.98 G	121.18 M	125.84 G	42.65 M	73.59 G	29.54 M	29.74 G
Swin-UNETR [7]	62.83 M	384.2 G	62.19 M	429.95 G	62.2 M	774.98 G	62.19 M	319.38 G
TransBTS [24]	32.79 M	324.13 G	32.79 M	323.73 G	32.99 M	326.48 G	31.58 M	119.81 G
UNETR [8]	92.49 M	75.76 G	121.18 M	98.40 G	102.45 M	184.76 G	92.78 M	73.51 G
LHU-Net	10.47 M	37.49 G	14.68 M	52.26 G	10.48 M	57.43 G	8.53 M	22.20 G

normalization, and leakyReLU, followed by another depth-wise convolution with batch normalization. A residual connection, processed through PW-Conv and batch normalization, is added before a final leakyReLU activation. MaxPool is then used to downsample the spatial dimensions, which reduces the feature map size and aggregates prominent features, making the subsequent processing more effective. This approach achieves two key objectives. It reduces the computational cost of processing large spatial dimensions with ViT blocks and preserves local features by delaying downsampling until after initial refinement, thereby optimizing parameter usage.

2.2 Hybrid Blocks (Hybrid Attention)

In the intermediate and deep layers, we employ hybrid blocks to enhance segmentation by balancing local detail with global context. By integrating the LKAd module as a CNN attention block, these blocks improve boundary delineation and object identification. Building on robust local features from earlier stages, our hybrid blocks use larger kernel sizes to capture broader spatial representations and deploy deformable attention to focus on relevant receptive fields. Spatial ViT attention is applied in the early part of this stage, while channel ViT attention in later layers aggregates high-level features. This design overcomes the limitations of repeated fixed modules [20,18,14] and avoids the redundancy seen in methods that repeat blocks across levels [4,22,10,13]. Our level-specific modules ensure that each stage effectively captures its unique features, resulting in improved segmentation performance with fewer parameters and lower computational cost.

Self-Adaptive Contextual Fusion Module: This module is integrated into the top hybrid blocks to enhance global structure capture by fusing spa-

[4] A single integer indicates uniform downsampling across all axes.

Table 2: Comparison of LHU-Net with SOTA methods on the Synapse dataset. Blue and red indicate the best and second-best results, respectively. Metrics include DSC for individual organs, average DSC, and HD95. Parameter counts (M) and FLOPs (G) are also reported.

Methods	Params↓	FLOPS↓	Spl	RKid	LKid	Gal	Liv	Sto	Aor	Pan	Average DSC ↑	Average HD95 ↓
UNETR [8]	92.49 M	75.76 G	85.00	84.52	85.60	56.30	94.57	70.46	89.80	60.47	78.35	18.59
Swin-UNETR [7]	62.83 M	384.2 G	95.37	86.26	86.99	66.54	95.72	77.01	91.12	68.80	83.48	10.55
nnFormer [28]	150.5 M	213.4 G	90.51	86.25	86.57	70.17	96.84	86.83	92.04	83.35	86.57	10.63
UNETR++ [20]	42.96 M	47.98 G	95.77	87.18	87.54	71.25	96.42	86.01	92.52	81.10	87.22	7.53
TC-CoNet [5]	313.67 M	699.58 G	91.77	87.92	88.16	62.00	96.35	79.40	92.45	72.78	83.86	9.64
D-LKA Net [4]	42.35 M	66.96 G	95.88	88.50	87.64	72.14	96.25	85.03	92.87	81.64	87.49	9.57
TransBTS [24]	32.79 M	324.13 G	91.65	86.99	87.46	62.52	96.42	77.39	91.71	72.12	83.28	12.34
CoTr [25]	41.87 M	334.23 G	94.93	86.80	87.67	62.90	96.37	80.46	92.43	78.84	85.05	9.04
nnUNet [11]	30.71 M	476.58 G	91.16	86.21	86.92	69.77	96.49	85.92	91.78	83.23	86.44	10.91
MedNeXt-M-K3 [19]	17.55 M	123.47 G	90.63	86.50	87.66	73.00	96.92	77.89	92.25	80.81	85.71	19.10
MedNeXt-M-K5 [19]	18.26 M	153.52 G	91.16	87.51	87.67	71.31	97.01	80.46	92.48	80.20	85.97	17.59
LHU-Net	10.47 M	37.49 G	96.02	87.46	87.75	74.30	96.83	85.73	92.53	82.04	87.83	6.26

tial attention maps from both the LKAd (D) and self-attention (S) mechanisms. This effectively preserves global context and minimizes information loss in deep encoder layers. Within the module, LKAd uses a deformable grid to adaptively capture local and global features, while S handles long-range dependencies and reduces texture bias to maintain high-frequency details. The final output is computed as:

$$F_S = Comb\left(F + \delta_s \mathbf{D}(F) + \gamma_s \mathbf{S}(F)\right) \tag{1}$$

where $F \in \mathbb{R}^{C\times H\times W\times D}$ is the input feature map, F_S is the output, and δ_s and γ_s are learnable weights that balance the contributions of the two attention mechanisms. The $Comb$ function, as shown in Fig. 2, processes the fused output through a ResBlock and a PW-Conv, followed by a residual connection and batch normalization. This design ensures effective feature fusion and robust contextual representation.

LKAd (D): This module computes attention following the method in [4]. The input tensor $x \in \mathbb{R}^{C\times H\times W\times D}$ undergoes these steps:

$$\begin{aligned} \hat{x} &= \text{GELU}(\text{PW-Conv}(x)), \\ x_{\text{LKAd}} &= \text{PW-Conv}(\text{DDW-Conv}_3(\text{DWD-Conv}(\text{DW-Conv}(\hat{x})))) \otimes \hat{x}, \end{aligned} \tag{2}$$

First, x is refined via a pointwise convolution and GELU activation, producing $\hat{x}$. Next, sequential depth-wise and dilated convolutions extract multi-scale local features, The key step is the deformable depth-wise convolution (DDW-Conv$_3$), which enables adaptive sampling, enhancing fine-detail and long-range dependency capture. Finally, element-wise multiplication with $\hat{x}$ yields x_{LKAd}, a rich, contextually enhanced representation of x.

Self-Attention (S): The self-attention module computes $S(x)$ from a normalized tensor x of shape $n \times C$, where $n = H \times W \times D$. First, three linear layers generate queries, keys, and values as $Q = W^Q x$, $K = W^K x$, $V = W^V x$, each

Table 3: Comparison of LHU-Net with SOTA models on BraTS, Lung, and LA datasets. Blue and red indicate the best and second-best results, respectively.

Method	BRATS (MRI)					Lung (CT)		LA (MRI)	
	WT	ET	TC	DSC ↑	HD95 ↓	DSC ↑	HD95 ↓	DSC ↑	HD95 ↓
UNETR [8]	90.35	76.30	77.02	81.22	6.61	73.29	23.84	91.25	9.23
TransBTS [24]	90.91	77.86	76.10	81.62	5.80	70.38	30.09	92.25	4.92
Swin-UNETR [7]	91.12	77.65	78.41	82.39	5.33	75.55	28.74	91.89	5.96
CoTr [25]	91.01	77.52	77.43	81.99	5.78	75.74	27.91	92.60	4.87
nnUNet [11]	91.21	77.96	78.05	82.41	5.58	74.31	28.52	92.75	4.35
nnFormer [28]	91.23	77.84	77.91	82.34	5.18	77.95	16.25	92.02	5.08
MedNeXt-M-K3 [19]	91.42	78.24	77.98	82.55	5.13	80.14	2.85	92.68	4.49
MedNeXt-M-K5 [19]	91.21	78.15	78.03	82.46	5.37	79.51	2.84	92.50	4.68
UNETR++ [20]	91.27	78.39	78.60	82.75	5.05	80.68	2.79	92.55	5.08
LHU-Net	91.56	80.03	79.83	83.81	4.83	81.27	3.23	92.91	3.99

reshaped to $n \times h \times \frac{C}{h}$ for h heads. Following [20], two additional projections transform K and V into learnable representations: $K_p = W_K^p K$, $V_p = W_V^p V$, reducing the spatial dimension from n to p (with $p \ll n$), which improves efficiency and focuses attention on the most representative features. Next, the normalized Q is multiplied by K_p^T (scaled by $\sqrt{d}$) and the resulting values are adjusted using the learnable parameter α and passed through a softmax (σ)to obtain similarity scores, which are then multiplied by V_p to yield the spatial attention where it is further refined by a linear normalization and a linear layer.

OmniFocus Attention Block: At the deepest network level, this block processes the richest feature layers by operating residually in the encoder and collectively in the decoder. It enhances feature representation by reducing noise and capturing essential details through integrated convolutional flows. The block leverages ViT channel attention (C) alongside the LKAd module to learn inter-channel relationships and long-range dependencies. Specifically, the block applies the LKAd module (see Equation 2) and a channel attention module that computes inter-channel interactions using dot-product attention on projections Q_C, K_C, and V_C:

$$x_C = V_C \cdot \sigma\left(\frac{Q_C^T K_C}{\sqrt{d}}\right). \tag{3}$$

After linear normalization and an additional linear layer, the final output is produced by a *Comb* function that fuses the LKAd and channel attention outputs with learnable weights (similar to the Self-Adaptive Contextual Fusion Module). This design avoids redundant module repetition and effectively extracts rich contextual information, thereby enhancing overall segmentation performance.

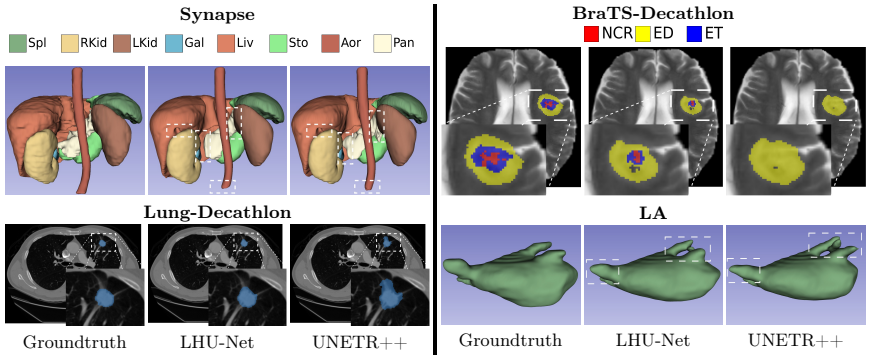

Fig. 3: Qualitative comparison of LHU-Net and UNETR++ across Synapse, BraTS-Decathlon, Lung-Decathlon, and LA datasets.

3 Experimental setup and results

3.1 Datasets, experimental setup and evaluation metrics

To assess model effectiveness, we used four datasets: two CT-based and two MRI-based. **Synapse** [12] consists of 30 abdominal CT scans, split as in [4,20] (18 for training, 12 for testing). Evaluated on eight organs: aorta, gallbladder, spleen, left/right kidneys, liver, pancreas, and stomach. **Lung-Decathlon** [1,21] consists of 63 CT scans for lung cancer segmentation, using an 80:20 training-validation split [20]. **BraTs-Decathlon** [1,16] which consists of 484 MRI scans (FLAIR, T2w, T1w, T1ce) for segmenting whole tumor (WT), enhancing tumor (ET), and tumor core (TC). Data split follows [20]. **Left Atrial (LA)** [26] consists of 100 MRI scans with varying sizes, preprocessed using Z-score normalization [15,23]. Five-fold cross-validation was used to segment the left atrium. Evaluation metrics include DSC and HD95, with final segmentation generated using 0.5 patch overlaps. All datasets used a batch size of 2. The model, implemented in PyTorch 2.1.0 [17] and for fair comparison integrated into the nnUNet framework [11], was trained on two NVIDIA A100 GPUs (80 GB VRAM). Training used a composite loss (DICE and cross-entropy, weighted 1:1) and the SGD optimizer with Nesterov momentum (0.99) and a weight decay of 3×10^{-5}. The learning rate followed a polynomial decay strategy (power 0.9), set to an initial learning rate of 0.01 for isometric patches and 0.003 otherwise. Data augmentation followed [28,20,11]. Training ran for 1000 epochs with 250 iterations per epoch (250,000 iterations). Table 1 details the network settings for each dataset.

3.2 Quantitative and qualitative results

Table 2 shows the performance of LHU-Net in comparison with SOTA models on the Synapse dataset. LHU-Net achieves the highest DSC while maintaining the lowest parameter count and FLOPs, offering a 4× reduction in parameters

Table 4: Ablation study on the impact of different attention mechanisms on parameters, FLOPs, and DSC for the BraTS dataset. Each repeated entry represents attention used in successive hybrid layers. The best model is in **bold**.

ViT Attn.	CNN Attn.	Params.↓	FLOPs↓	DSC↑	ViT Attn.	CNN Attn.	Params.↓	FLOPs↓	DSC↑
SSS	DDD	10.51 M	57.44 G	82.64	SSC	DDI	9.44 M	57.33 G	82.35
CCC	DDD	7.97 M	57.25 G	82.52	SSC	III	8.88 M	55.12 G	81.48
SSC	**DDD**	**10.48 M**	**57.43 G**	**83.81**	SSC	LLL	9.02 M	55.55 G	82.32
SCC	DDD	10.33 M	57.41 G	83.12	SC	DD	5.12 M	56.78 G	82.85

and 44% lower FLOPs compared to the second-best DSC model. Additionally, LHU-Net improves average HD95 by 16%, enhancing segmentation accuracy. Although MedNext [19] has a competitive parameter count, its DSC is 2% lower than LHU-Net, with significantly higher FLOPs. Table 3 presents the quantitative results for the three datasets. As observed, LHU-Net surpasses other methods by a large margin. However, the HD95 in the Lung dataset reveals some shortcomings, indicating room for improvement. The parameters and FLOPs of each SOTA model (Table 1) highlight the efficiency of LHU-Net while outperforming SOTA models in Average DSC. Fig. 3 illustrates qualitative comparisons with UNETR++, one of the leading SOTA models. Across different datasets, UNETR++ exhibits over-segmentation or label omission, whereas LHU-Net consistently achieves more precise segmentation. This highlights how selecting the right module for each layer enhances efficiency while setting new benchmarks across datasets.

4 Ablation studies

We conducted an ablation study on the BraTS dataset to assess the impact of different attention mechanisms on segmentation performance. Table 4 presents the results, where attention mechanisms were applied to successive hybrid layers, including large kernel attention (L), LKAd (D), Self-Attention (S), Channel-Attention (C), and a baseline without CNN attention (I). The SSC-DDD configuration, which applies Self-attention (S) and channel attention (C) on the ViT side and LKAd (D) on the CNN side, achieves the highest DSC (83.81) with efficient computation. This demonstrates the advantage of selective attention combinations over uniform mechanisms. Reducing hybrid layers from three to two (SC-DD) slightly lowers DSC (82.85) but significantly reduces parameters (5.12M), indicating a strong trade-off between accuracy and efficiency. The III configuration (no CNN attention) performs worst (81.48 DSC), confirming the necessity of CNN attention mechanisms. Additionally, SCC-DDD achieves 83.12 DSC, suggesting that transitioning from self-attention (S) to channel attention (C) in earlier layers further enhances segmentation. Overall, large kernel convolutional attention followed by deformable convolution is crucial as a CNN attention in the hybrid layers, while progressive ViT attention transitions im-

prove segmentation. Fewer hybrid layers can still offer competitive performance, making SC-DD a strong alternative for efficiency-focused applications.

The learnable weights (γ and δ) that balance the contributions of CNN and ViT attention are crucial, as they vary across datasets (e.g., 0.45 for Synapse and 0.11 for BraTS). This demonstrates that fixed weights are suboptimal; learnable parameters allow the model to adapt and find the most suitable weight for each segmentation task.

5 Conclusion

In this work, we introduced LHU-Net, a lean hybrid U-Net for volumetric medical image segmentation. By strategically using spatial attention in early layers and channel attention in deeper layers, LHU-Net efficiently handles diverse datasets and segmentation tasks with only about 11 million parameters. Our results demonstrate that high segmentation accuracy can be achieved with a simpler model, advancing the development of accessible and effective medical image analysis tools.

Acknowledgments. The authors gratefully acknowledge the computational and data resources provided by the Leibniz Supercomputing Centre. Also, the authors gratefully acknowledge the scientific support and HPC resources provided by the Erlangen National High-Performance Computing Center (NHR@FAU) of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) under the NHR project "b213da." NHR funding is provided by federal and Bavarian state authorities. NHR@FAU hardware is partially funded by the German Research Foundation (DFG) – 440719683. Also, this work was supported by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) under the grant no. 417063796.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., Ronneberger, O., Summers, R.M., et al.: The medical segmentation decathlon. Nature communications **13**(1), 4128 (2022)
2. Azad, R., Aghdam, E.K., Rauland, A., Jia, Y., Avval, A.H., Bozorgpour, A., Karimijafarbigloo, S., Cohen, J.P., Adeli, E., Merhof, D.: Medical image segmentation review: The success of u-net. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
3. Azad, R., Kazerouni, A., Heidari, M., Aghdam, E.K., Molaei, A., Jia, Y., Jose, A., Roy, R., Merhof, D.: Advances in medical image analysis with vision transformers: a comprehensive review. Medical Image Analysis **91**, 103000 (2024)
4. Azad, R., Niggemeier, L., Hüttemann, M., Kazerouni, A., Aghdam, E.K., Velichko, Y., Bagci, U., Merhof, D.: Beyond self-attention: Deformable large kernel attention for medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1287–1297 (2024)

5. Chen, Y., Lu, X., Xie, Q.: Collaborative networks of transformers and convolutional neural networks are powerful and versatile learners for accurate 3d medical image segmentation. Computers in Biology and Medicine **164**, 107228 (2023)
6. Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., Brendel, W.: Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. arXiv preprint arXiv:1811.12231 (2018)
7. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: International MICCAI Brainlesion Workshop. pp. 272–284. Springer (2021)
8. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: Unetr: Transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 574–584 (2022)
9. He, Y., Nath, V., Yang, D., Tang, Y., Myronenko, A., Xu, D.: Swinunetr-v2: Stronger swin transformers with stagewise convolutions for 3d medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 416–426. Springer (2023)
10. Hong, Z., Chen, M., Hu, W., Yan, S., Qu, A., Chen, L., Chen, J.: Dual encoder network with transformer-cnn for multi-organ segmentation. Medical & Biological Engineering & Computing **61**(3), 661–671 (2023)
11. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)
12. Landman, B., Xu, Z., Igelsias, J., Styner, M., Langerak, T., Klein, A.: Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In: Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge. vol. 5, p. 12 (2015)
13. Li, J., Ye, J., Deng, H., Shi, H.: Cpftransformer: transformer fusion context pyramid medical image segmentation network. Frontiers in Neuroscience **17**, 1288366 (2023)
14. Liu, Y., Zhang, Z., Yue, J., Guo, W.: Scanext: Enhancing 3d medical image segmentation with dual attention network and depth-wise convolution. Heliyon (2024)
15. Luo, X., Chen, J., Song, T., Wang, G.: Semi-supervised medical image segmentation through dual-task consistency. In: Proceedings of the AAAI conference on artificial intelligence. vol. 35, pp. 8801–8809 (2021)
16. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al.: The multimodal brain tumor image segmentation benchmark (brats). IEEE transactions on medical imaging **34**(10), 1993–2024 (2014)
17. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
18. Rahman, M.M., Shokouhmand, S., Bhatt, S., Faezipour, M.: Mist: Medical image segmentation transformer with convolutional attention mixing (cam) decoder. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 404–413 (2024)
19. Roy, S., Koehler, G., Ulrich, C., Baumgartner, M., Petersen, J., Isensee, F., Jaeger, P.F., Maier-Hein, K.H.: Mednext: transformer-driven scaling of convnets for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 405–415. Springer (2023)

20. Shaker, A.M., Maaz, M., Rasheed, H., Khan, S., Yang, M.H., Khan, F.S.: Unetr++: Delving into efficient and accurate 3d medical image segmentation. IEEE Transactions on Medical Imaging (2024). https://doi.org/10.1109/TMI.2024.3398728
21. Simpson, A.L., Antonelli, M., Bakas, S., Bilello, M., Farahani, K., Van Ginneken, B., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., et al.: A large annotated medical image dataset for the development and evaluation of segmentation algorithms. arXiv preprint arXiv:1902.09063 (2019)
22. Wang, J., Zhao, H., Liang, W., Wang, S., Zhang, Y.: Cross-convolutional transformer for automated multi-organs segmentation in a variety of medical images. Physics in Medicine & Biology **68**(3), 035008 (2023)
23. Wang, Y., Xiao, B., Bi, X., Li, W., Gao, X.: Mcf: Mutual correction framework for semi-supervised medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15651–15660 (2023)
24. Wenxuan, W., Chen, C., Meng, D., Hong, Y., Sen, Z., Jiangyun, L.: Transbts: Multimodal brain tumor segmentation using transformer. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 109–119 (2021)
25. Xie, Y., Zhang, J., Shen, C., Xia, Y.: Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24. pp. 171–180. Springer (2021)
26. Xiong, Z., Xia, Q., Hu, Z., Huang, N., Bian, C., Zheng, Y., Vesal, S., Ravikumar, N., Maier, A., Yang, X., et al.: A global benchmark of algorithms for segmenting the left atrium from late gadolinium-enhanced cardiac magnetic resonance imaging. Medical image analysis **67**, 101832 (2021)
27. Yang, F., Wang, F., Dong, P., Wang, B.: Hca-former: Hybrid convolution attention transformer for 3d medical image segmentation. Biomedical Signal Processing and Control **90**, 105834 (2024)
28. Zhou, H.Y., Guo, J., Zhang, Y., Han, X., Yu, L., Wang, L., Yu, Y.: nnformer: Volumetric medical image segmentation via a 3d transformer. IEEE Transactions on Image Processing (2023)