ViL-UNet: Beyond Transformers for 3D Medical Image Segmentation with Vision xLSTM

Yong Wu[1], Wenjun Huang[2], Ziyu Hu[3], Yiqi Ma[4],
Pi Fang[1], Yuechen Yin[1], Qinghua Zhong[1,*]
[1]South China Normal University, China [2]Sun Yat-Sen University, China
[3]Aalto University, Finland [4]Shanghai Jiao Tong University, China
Email: {2024025400, 2024023790, 2024025405}@m.scnu.edu.cn, huangwj98@mail2.sysu.edu.cn,
ziyu.hu@aalto.fi, yiqi_17@sjtu.edu.cn, zhongqinghua@m.scnu.edu.cn
*Corresponding author

***Abstract*—Efficient segmentation of medical images has evolved from relying solely on Convolutional Neural Networks (CNNs) to exploring hybrid models that integrate CNNs with Vision Transformers (ViTs). Despite the advantages of transformers in capturing global dependencies, their high computational and storage demands pose challenges. This research introduces the novel *ViL-UNet*, which combines CNNs with Vision Extended Long Short-Term Memory (*Vision-xLSTM*), aiming to balance performance and computational efficiency. The *Vision-xLSTM* blocks capture long-range spatial and global relationships, both at the network's bottleneck and within the skip connections to refine multi-scale feature fusion. Our primary goal is to demonstrate that *Vision-xLSTM* provides a suitable backbone for medical image segmentation, offering superior performance with reduced computational costs. The *ViL-UNet* achieves state-of-the-art results on the publicly available *Synapse* and *ACDC* datasets.**

***Index Terms*—Medical Image Segmentation, xLSTM, 3D Imaging.**

I. Introduction

Deep learning models [1]–[13], particularly Convolutional Neural Networks (CNNs) like U-Net [14], are pivotal for automated medical image segmentation [15]. However, the limited receptive field of CNNs hinders their ability to capture global context. Vision Transformers (ViTs) and hybrid models like TransUNet [16] overcome this by modeling long-range dependencies using self-attention [17]. Despite their success, the quadratic computational complexity of self-attention imposes significant memory and computational burdens, limiting their clinical applicability [18].

To address this trade-off, we explore eXtended Long-Short Term Memory (xLSTM) [19], a recent sequence modeling architecture with linear complexity. Its vision-centric adaptation, Vision-xLSTM (ViL) [20], presents a promising, efficient alternative to Transformers. This paper introduces *ViL-UNet*, a U-shaped network that integrates a CNN encoder with a ViL bottleneck to balance local feature extraction and global context modeling.

The need for computationally efficient yet accurate segmentation algorithms, combined with xLSTM's advantages, motivated the development of the *ViL-UNet* model. Our approach uniquely integrates CNNs with ViL within a U-shaped framework [21]–[23]. Our contributions are threefold: we design a hybrid encoder where CNNs capture local patterns and a ViL block encodes global context; we use a decoder with skip connections to reconstruct high-resolution segmentation; and we demonstrate through experiments on the *Synapse* and *ACDC* datasets that our model achieves state-of-the-art performance with high efficiency.

II. Related Work

A. CNN-based Medical Image Segmentation

Deep learning has been applied across various domain [24]–[37]. And Convolutional Neural Networks have long been the foundation of medical image segmentation. The seminal U-Net introduced an encoder-decoder structure with skip connections for multi-scale feature propagation. Representative variants include Attention U-Net [38] and UNet++ [39]. Despite their success, the locality of convolution limits their ability to capture long-range spatial dependencies in volumetric medical images.

B. Transformer-based Medical Image Segmentation

To address the limitations of CNNs, transformer-based architectures have been introduced to medical image segmentation [40]–[52]. Trans UNet pioneered the integration of transformers with U-Net, employing self-attention mechanisms to capture global context. Swin-Unet [53] adopted hierarchical Swin Transformers with shifted window attention, achieving improved computational efficiency. Despite their success in modeling long-range dependencies, transformer-based methods suffer from quadratic computational complexity, resulting in substantial memory consumption and computational overhead, particularly for high-resolution 3D medical volumes [54]–[57]. This limitation has motivated research into more efficient alternatives.

C. State Space Models for Medical Imaging

Recent advances in sequence modeling have introduced State Space Models (SSMs) as efficient alternatives to transformers. VM-UNet [58]incorporated visual state space blocks with multi-scale feature fusion for medical image

segmentation. Swin-UMamba [59]combined hierarchical designs with Mamba's efficient sequence modeling, while HC-Mamba [60]proposed hybrid convolutional-Mamba architectures. MSVM-UNet [61]and MixFormer [62]further explored multi-scale feature fusion strategies. While these approaches have shown improved efficiency over transformers, they primarily focus on unidirectional or bidirectional scanning strategies.

D. Vision-xLSTM for Medical Image Segmentation

The recently proposed eXtended Long Short-Term Memory (xLSTM) represents a significant advancement in recurrent architectures, introducing exponential gating and modified memory structures. Vision-xLSTM (ViL) adapted xLSTM for computer vision tasks, demonstrating competitive performance with linear computational complexity. However, the application of Vision-xLSTM to medical image segmentation remains largely unexplored. Our work addresses this gap by systematically integrating ViL blocks into a U-shaped architecture, introducing a quad-directional scanning strategy specifically designed for 3D medical volumes, and demonstrating that Vision-xLSTM can serve as an effective and efficient backbone for medical image segmentation tasks.

III. Methodology

The proposed *ViL-UNet* adopts a U-shaped architecture with a feature extraction (encoder) path and a feature reconstruction (decoder) path, as illustrated in Fig. 1. The encoder employs CNN layers to capture hierarchical features and a Vision-xLSTM (ViL) block at the bottleneck to model long-range dependencies. The decoder progressively upsamples the features and integrates multi-level encoder information via skip connections to produce the final segmentation map.

A. Feature Extraction Path

This path consists of two key stages: hierarchical feature learning using CNNs and global context modeling using ViL blocks.

1) Hierarchical Feature Learning: The volumetric input image $I \in \mathbb{R}^{H\times W\times D\times 1}$ is passed through a series of convolution layers to hierarchically construct an intermediate abstract and high-level representation. This process downsamples the input by a factor of 8, yielding a high-level feature representation of size $\mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times\frac{D}{8}\times 8C}$. Here, H, W, D represent the height, width, and depth of the feature volume, respectively, with C corresponding to the number of channels.

This feature volume is then tokenized. It is divided into non-overlapping $P \times P \times P$ patches. These 3D patches are flattened into 1D vectors to yield a tokenized representation $\mathbf{t} \in \mathbb{R}^{N\times(P^3\frac{C}{4})}$. The flattened tokens $(t^1, t^2, \ldots, t^N)$ are subsequently projected into a Z-dimensional embedding space. To retain spatial information, learnable positional embeddings are added to these token embeddings. This process is mathematically expressed as:

$$\mathbf{p} = [t^1\mathbf{K}; t^2\mathbf{K}; \ldots; t^N\mathbf{K}] + \mathbf{K_{pos}}, \tag{1}$$

where $\mathbf{K} \in \mathbb{R}^{(P^3\frac{C}{4}\times Z)}$ is the trainable projection matrix and $\mathbf{K_{pos}} \in \mathbb{R}^{N\times Z}$ is the position embedding matrix. $\mathbf{p} \in \mathbb{R}^{N\times Z}$ represents the final sequence of embedded patches fed into the ViL blocks.

2) Global Dependencies: The sequence of embedded patches $\mathbf{p}$ is processed by a stack of ViL blocks at the model's bottleneck. To comprehensively capture 3D spatial dependencies, we introduce a quad-directional scanning strategy. This strategy operates on the flattened sequence of 3D feature patches by cyclically applying four distinct traversal methods in consecutive ViL blocks, as illustrated in Fig. 2: **(a) Forward Traversal:** processing in the default flattened order. **(b) Axis-Permuted Forward Traversal:** permuting the spatial axes before forward processing. **(c)Backward Traversal:** processing the reversed sequence. **(d) Axis-Permuted Backward Traversal:** axis-permuting and then reversing the sequence.

This alternating scanning mechanism enables the model to effectively capture long-range dependencies along all principal axes and in both directions, thereby building a more robust representation of 3D anatomical structures.

Internally, each ViL block first projects the input tokens to a higher-dimensional space and then splits them into two parallel paths. One path is processed by the core mLSTM layer to capture long-range sequential dependencies, while the other acts as a gating branch to modulate the propagated information. The two outputs are then fused and projected back to the original embedding dimension, and the representations from the four scanning directions are merged to form the final feature representation.

3) Computational Complexity Analysis and Design Motivation: A key motivation behind integrating Vision-xLSTM (ViL) into the proposed architecture is its favorable computational complexity compared to transformer-based self-attention mechanisms. For a sequence of N tokens, self-attention incurs a quadratic complexity of $\mathcal{O}(N^2)$ in both computation and memory, which becomes prohibitive for high-resolution 3D medical volumes where N grows cubically with spatial resolution. In contrast, the recurrent formulation of xLSTM enables linear complexity $\mathcal{O}(N)$, making it significantly more scalable for volumetric data.

In the context of 3D medical image segmentation, this efficiency is particularly critical. Volumetric scans such as CT and MRI inherently contain dense spatial information along three axes, and modeling long-range dependencies across slices is essential for accurate anatomical understanding. By employing ViL blocks at the bottleneck, our model captures global contextual information across the entire 3D volume while maintaining manageable memory consumption and computational cost.

Furthermore, the proposed quad-directional scanning strategy enhances the expressive power of ViL without increasing asymptotic complexity. By alternately traversing the token sequence along different spatial permutations and directions, the model effectively aggregates contextual information from all principal axes. This design enables ViL-UNet to model complex anatomical structures with long-range spatial correlations

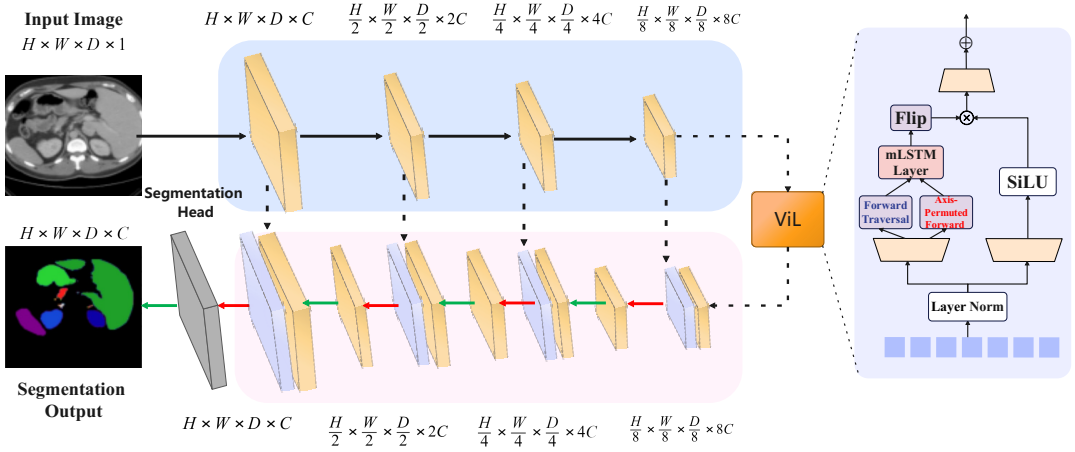

Fig. 1. Architectural framework of *ViL-UNet* depicting the input being processed by stacked layers of CNNs and ViL block. The feature representation from the ViL block is upsampled through the feature reconstruction path to obtain the final segmentation output.

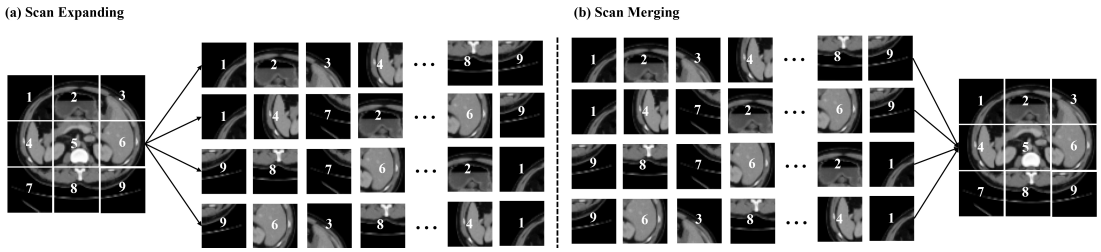

Fig. 2. The ViL module employs a quad-directional scanning strategy to capture 3D spatial dependencies. This strategy alternately traverses the image patch sequence using four methods, including forward, backward, and axis-permuted traversals. Finally, the results from all scanning directions are merged to reconstruct the final feature representation.

while remaining computationally efficient, thereby offering a practical alternative to transformer-based architectures for 3D medical image segmentation.

B. Feature Reconstruction Path

The feature reconstruction path symmetrically mirrors the encoder. At each level l, a trilinear upsampling operation τ is first applied to increase the spatial dimensions of the feature map $\mathbf{F}_{l+1}$ from the previous, deeper level.

$$\mathbf{U}_l = \tau(\mathbf{F}_{l+1}), \tag{2}$$

where $\mathbf{U}_l$ is the upsampled feature map.

Next, to bridge the semantic gap and refine multi-scale feature fusion, we embed ViL blocks within the skip connections. Instead of directly concatenating the feature map $\mathbf{E}_l$ from the encoder path, it is first processed by a Vision-xLSTM (ViL) block to model its internal spatial dependencies and enhance its feature representation, yielding a refined feature map $\mathbf{E'}_l$.

$$\mathbf{E'}_l = \text{ViL}(\mathbf{E}_l) \tag{3}$$

This refined, context-aware feature map $\mathbf{E'}_l$ is then concatenated with the upsampled feature map $\mathbf{U}_l$. This step re-introduces high-resolution spatial details that have been contextually enhanced before fusion.

$$\mathbf{C}_l = \text{Concat}(\mathbf{U}_l, \mathbf{E'}_l). \tag{4}$$

Finally, the concatenated feature volume $\mathbf{C}_l$ is refined through a series of $3 \times 3 \times 3$ convolutions to produce the output volume $\mathbf{R}_l$ for that level.

$$\mathbf{R}_l = \text{Conv}_{3\times3\times3}(\mathbf{C}_l). \tag{5}$$

This process is repeated until the feature map is restored to the original input image resolution, at which point a final convolution layer produces the pixel-wise segmentation map.

1) Rationale for ViL-Enhanced Skip Connections: While traditional U-shaped architectures directly concatenate encoder and decoder features through skip connections, such a strategy may introduce a semantic gap between low-level spatial details and high-level contextual representations. This issue is particularly pronounced in 3D medical image segmentation, where fine-grained boundary information and global anatomical context must be jointly preserved.

To address this challenge, we embed Vision-xLSTM (ViL) blocks within the skip connections to refine encoder features before fusion. By modeling long-range spatial dependencies within each encoder feature map, the ViL-enhanced skip connections provide context-aware representations that are more semantically aligned with the decoder features. This design allows the network to preserve high-resolution spatial details while incorporating global contextual cues, which is especially beneficial for segmenting small organs and thin anatomical structures such as the myocardium.

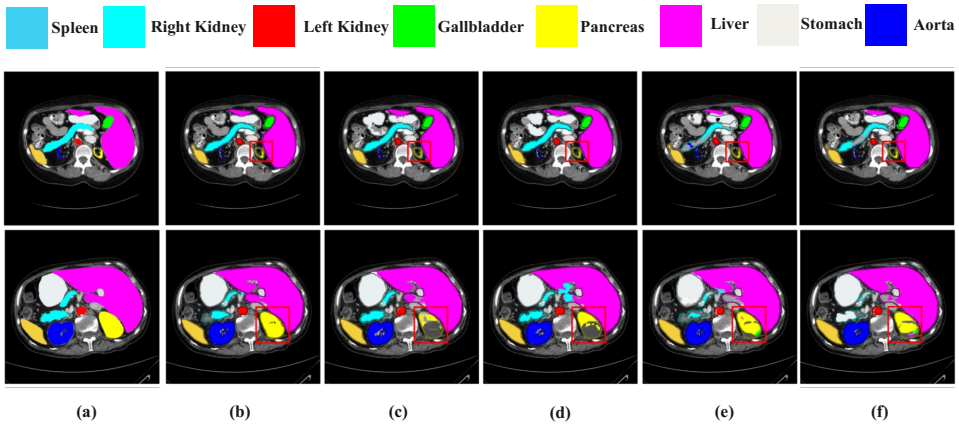

Fig. 3. Visual comparison of different methods on the Synapse multi-organ dataset: (a) Ground Truth, (b) ViL-UNet, (c) MSVM-UNet, (d) MixFormer, (e) Swin-UMamba, and (f) Swin-Unet.

Experimental results on both the Synapse and ACDC datasets demonstrate that this strategy contributes to improved boundary delineation and segmentation consistency, highlighting the importance of contextual refinement within multi-scale feature fusion.

IV. Experimental Details

A. Datasets

To comprehensively evaluate the performance of our proposed ViL-UNet model, we conducted extensive experiments on two widely-used public benchmarks for medical image segmentation: the Synapse abdominal multi-organ segmentation dataset and the Automated Cardiac Diagnosis Challenge (ACDC) dataset, which include CT and MRI images.

1) Synapse Multi-Organ Segmentation Dataset: The Synapse dataset originates from the MICCAI 2015 Multi-Atlas Abdomen Labelling Challenge, which comprises clinical CT images from 30 patients, providing a total of 3779 axial contrast-enhanced images for abdominal multi-organ segmentation. The images have a consistent resolution of 512 × 512 pixels and provide precise annotations of organ contours. Following the approach of TransUNet, we randomly divided the dataset into 18 cases for training and 12 cases for testing, focusing on the segmentation of 8 abdominal organs: the aorta, gallbladder, left kidney, right kidney, liver, pancreas, spleen, and stomach. This dataset presents significant challenges due to the high variability in organ shapes, sizes, and positions across different patients, as well as the presence of complex anatomical structures and potential pathological variations.

2) ACDC Cardiac Segmentation Dataset: The ACDC dataset originates from the Automated Cardiac Diagnosis Challenge, which gathers cardiac MRI scan images from various patients with different cardiac pathologies. Each patient scan is manually annotated with ground truth for three cardiac sub-structures: the left ventricle (LV), right ventricle (RV), and myocardium (Myo). The dataset includes images acquired at both end-diastolic (ED) and end-systolic (ES) phases of the cardiac cycle, capturing the full range of cardiac motion. In line with TransUNet, we selected 70 cases for training, 10 for validation, and 20 for testing. This dataset poses unique challenges including varying image quality, different cardiac pathologies (such as dilated cardiomyopathy, hypertrophic cardiomyopathy, and abnormal right ventricle), and the need to accurately delineate thin myocardial walls and complex ventricular geometries.

B. Implementation Details

We implemented *ViL-UNet* using PyTorch and MONAI, and the model was trained on a single NVIDIA RTX A100 GPU. For optimization, we employed the AdamW optimizer with an initial learning rate of 1×10^{-4} and a weight decay of 1×10^{-5}. The model performance was comprehensively evaluated using the Dice Score Coefficient (DSC) and the 95% Hausdorff Distance ($HD95$).

Architecture Configuration: The backbone follows a U-shaped design. The encoder path hierarchically extracts features, which are then processed by a bottleneck consisting of 6 Vision-xLSTM (ViL) blocks to capture long-range spatial dependencies. In our implementation, the feature volume is tokenized into non-overlapping $P \times P \times P$ patches and projected into a Z-dimensional embedding space. The decoder integrates multi-scale features via skip connections enhanced by ViL blocks to refine the final segmentation.

Dataset Setup: For the Synapse dataset, we evaluated the segmentation of 8 abdominal organs from 30 CT scans. For the ACDC dataset, 150 cardiac MRI scans were used to segment the Right Ventricle (RV), Left Ventricle (LV), and Myocardium (Myo).

C. Results and Discussion

The performance of our *ViL-UNet* was next compared with that of the state-of-the-art (SOTA) algorithms in the literature in terms of metrics DSC, $HD95$ on *Synapse* and *ACDC* datasets. Table I presents a comprehensive breakdown of performance for different organs (in *Synapse* data) in the context of *DSC*. The average results for $HD95$ are provided for the nine abdominal organs. *ViL-UNet* is found to outperform other SOTA, with respect to the mean *DSC* and $HD95$,

TABLE I

COMPARISON WITH STATE-OF-THE-ART MODELS ON MULTI-ORGAN SEGMENTATION (SYNAPSE) DATASET, WITH BEST RESULTS MARKED IN **BOLD**.

Methods	Year	DSC(%)↑	HD95 (mm)↓	Aorta	Gallbladder	Kidney(L)	Kidney(R)	Liver	Pancreas	Spleen	Stomach
Swin-Unet [53]	2021	79.13	21.55	85.47	66.53	83.28	79.61	94.29	56.58	90.66	76.60
VM-UNet [58]	2024	82.38	16.22	87.00	69.37	85.52	82.25	94.10	65.77	91.54	83.51
HC-Mamba [60]	2024	79.58	26.34	89.93	67.65	84.57	78.27	95.38	52.08	89.49	79.84
Swin-UMamba [63]	2024	82.26	19.51	86.32	70.77	83.66	81.60	95.23	69.36	89.95	81.14
MSVM-UNet [61]	2024	85.00	14.75	88.73	74.90	85.62	84.47	95.74	71.53	92.52	86.51
MixFormer [62]	2025	82.64	12.67	87.36	71.53	86.22	83.19	95.23	66.82	89.98	80.77
ViL-UNet(ours)	2025	**85.12**	**12.49**	**90.04**	**74.93**	**89.17**	**84.72**	**95.83**	**71.91**	**92.70**	**86.70**

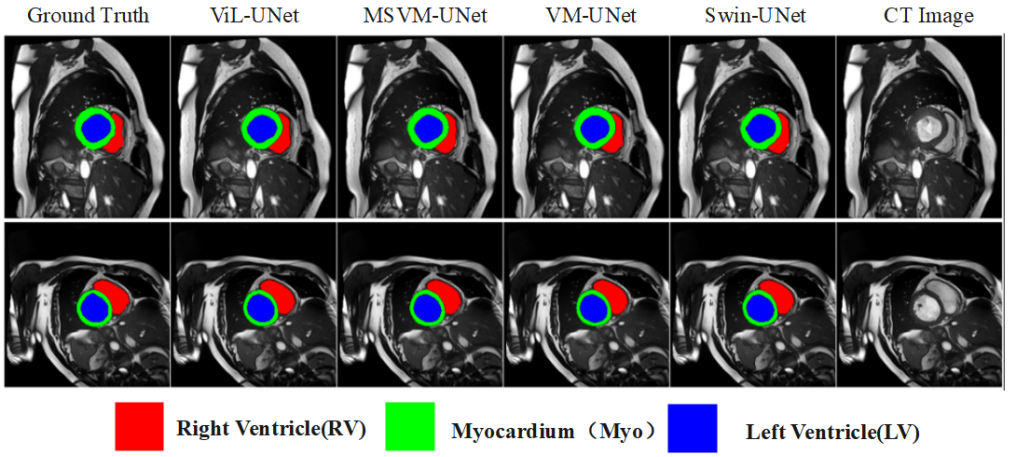

Fig. 4. Qualitative comparison of cardiac segmentation results on the ACDC dataset. From left to right: Ground Truth, ViL-UNet (ours), MSVM-UNet, VM-UNet, Swin-Unet, and the original MRI image. The color-coded overlays represent: Right Ventricle (RV) in red, Myocardium (Myo) in green, and Left Ventricle (blue). Our ViL-UNet produces segmentation masks that closely match the ground truth, demonstrating superior boundary delineation for all three cardiac structures, particularly for the challenging myocardium region.

TABLE II

COMPARATIVE EXPERIMENTAL RESULTS OF OUR VIL-UNET AND SOTA MODELS ON THE ACDC DATASET. HERE, BOLD BLACK DATA INDICATES THE BEST RESULT, AND UNDERLINED BLACK DATA DENOTES THE SECOND-BEST RESULT.

Methods	Year	DSC(%)↑	RV	Myo	LV
Swin-Unet [53]	2021	90.00	88.55	85.62	95.83
VM-Unet [58]	2024	92.24	90.74	89.93	96.03
Swin-UMamba [63]	2024	92.14	90.90	89.80	95.72
MSVM-UNet [61]	2024	92.58	91.00	90.35	96.39
MixFormer [62]	2025	91.01	89.02	88.46	95.55
ViL-UNet(ours)	2025	**92.79**	**91.16**	**90.54**	**96.56**

TABLE III

COMPARISON OF DIFFERENT VARIANTS OF *ViL-UNet* WITH INCREASING NUMBER OF VIL BLOCKS ($\times L$) AND CONVOLUTION LAYERS, ON *Synapse* DATA.

Ablations	Model Variants	mDSC (%)	mHD95 (mm)
# ViL blocks	x 3	82.18	20.21
	x 6	**85.12**	**12.49**
	x 12	84.89	12.57
	x 18	84.01	13.34
	x 24	84.99	16.89
# Convolution Layers	3	84.43	13.48
	4	**85.12**	**12.49**
	5	83.14	12.85

with scores of 85.12% and 12.49, respectively. The proposed *ViL-UNet* achieves higher segmentation accuracy compared to existing methods in generating the highest DSC values for the segmentation of larger organs (such as the spleen, liver) and smaller organs (such as the kidneys, pancreas and gall bladder). Performance remained stable, despite the reduction in organ size, as indicated by the consistently high DSC scores observed in smaller and larger organs. This illustrates the robustness of our model in utilizing acquired knowledge on anatomical structures with varying shapes and sizes. It is found to precisely identify and define the target areas possessing unique structures.

Table II presents a detailed performance comparison of our proposed ViL-UNet model against several state-of-the-art (SOTA) methods on the ACDC dataset. The evaluation is primarily based on the Dice Similarity Coefficient (DSC) metric, reporting the results for three cardiac structures—Right Ventricle (RV), Myocardium (Myo), and Left Ventricle (LV)—as well as the average DSC score. The results demonstrate that our proposed ViL-UNet model achieves the best performance across all evaluated metrics. Its average DSC of 92.79% significantly outperforms all other baseline models. Fig. 4 provides a qualitative comparison, where our ViL-UNet demonstrates superior segmentation quality with more accurate boundary delineation, especially for the thin myocardial walls and irregular ventricular shapes.

Ablation Study Analysis: Table III shows our ablation study on the impact of ViL block depth and encoder con-

TABLE IV
COMPARISON OF MODEL COMPLEXITY AND EFFICIENCY BETWEEN VIL-UNET AND OTHER SOTA MODELS. HERE, BOLD BLACK DATA INDICATES THE BEST RESULT, AND UNDERLINED BLACK DATA DENOTES IS THE SECOND-BEST RESULT.

Methods	Year	#Params(M)	#FLOPs(G)
Swin-Unet [53]	2021	<u>27.17</u>	<u>16.16</u>
VM-UNet [58]	2024	44.27	17.46
Swin-UMamba [63]	2024	60	163.6
MSVM-UNet [61]	2024	35.93	29.53
MixFormer [62]	2025	123.35	108.32
ViL-UNet(ours)	2025	**16.48**	17.93

volution layers.

Impact of ViL Block Depth: Increasing ViL blocks from 3 to 6 improves mDSC from 82.18% to 85.12% and reduces $HD95$ from 20.21mm to 12.49mm, demonstrating that deeper ViL stacks better capture long-range spatial dependencies. However, further increasing to 12 blocks yields only marginal improvement (84.89% mDSC, 12.57mm $HD95$), while 18 and 24 blocks show performance degradation (84.01% and 84.99% mDSC, 13.34mm and 16.89mm $HD95$), indicating overfitting with excessive depth. The optimal 6-block configuration achieves the best balance between model capacity and generalization.

Impact of Encoder Depth: Comparing convolution layer configurations, 4 layers achieve optimal performance (85.12% mDSC, 12.49mm $HD95$), outperforming both 3 layers (84.43% mDSC, 13.48mm $HD95$) and 5 layers (83.14% mDSC, 12.85mm $HD95$). The 3-layer configuration provides insufficient feature abstraction, while 5 layers cause excessive spatial downsampling that loses critical boundary information despite slightly better $HD95$.

Per-Organ Analysis on Synapse: ViL-UNet excels across diverse organ types. For large organs (liver: 95.83%, spleen: 92.70%), the model leverages abundant contextual information. The aorta (90.04%) benefits from our quad-directional scanning capturing tubular structures. Critically, small organs show substantial improvements: gallbladder (74.93% vs. 74.90% second-best) and pancreas (71.91% vs. 71.53%), where long-range context is crucial for disambiguating low-contrast boundaries. Kidneys achieve 89.17% (left) and 84.72% (right), with asymmetry reflecting anatomical variability.

Cardiac Structure Analysis on ACDC: The left ventricle (96.56%) achieves highest accuracy due to clear boundaries. Right ventricle (91.16%) and myocardium (90.54%) present greater challenges from thin walls and irregular shapes, yet ViL-UNet outperforms all baselines by 0.16% and 0.19% respectively. The ViL blocks in skip connections prove essential for preserving fine-grained myocardial details during upsampling.

Efficiency Analysis: Table IV highlights our model's efficiency. With only 16.48M parameters, ViL-UNet is 39% lighter than Swin-Unet (27.17M), 63% lighter than VM-UNet (44.27M), and 87% lighter than MixFormer (123.35M). FLOPs (17.93G) remain competitive with Swin-Unet (16.16G) while being substantially lower than Swin-UMamba (163.6G) and MixFormer (108.32G), making it suitable for resource-constrained clinical environments.

Qualitative Analysis: Fig. 3 displays sample segmentation maps from ViL-UNet and several other frameworks on the Synapse dataset. ViL-UNet outperforms the baselines, providing enhanced depiction of organs such as the pancreas, spleen, and stomach. This indicates the proposed model's ability to accurately identify relevant anatomical features.

V. CONCLUSIONS

In this paper, we introduced ViL-UNet, an efficient and reliable model for 3D medical image segmentation. The proposed framework combines CNN-based local detail extraction with Vision-xLSTM (ViL)-based long-range context modeling. By placing ViL blocks at the bottleneck and within skip connections, the model effectively integrates global context with high-resolution spatial features. Its gating mechanism and linear complexity contribute to strong performance with improved parameter efficiency compared with transformer-based methods. Results on the Synapse and ACDC datasets indicate that ViL-UNet is a competitive and practical framework for medical image segmentation, while further evaluation on more diverse datasets and deployment settings remains an important direction for future work.

REFERENCES

[1] J. Jiang, P. Yang, R. Zhang, and F. Liu, "Towards efficient large language model serving: A survey on system-aware kv cache optimization," *Authorea Preprints*, 2025.

[2] P. Yang, N. Akhtar, Z. Wen, M. Shah, and A. S. Mian, "Re-calibrating feature attributions for model interpretation," in *ICLR*, 2023.

[3] P. Yang, N. Akhtar, Z. Wen, and A. Mian, "Local path integration for attribution," in *AAAI*, 2023.

[4] P. Yang, N. Akhtar, M. Shah, and A. Mian, "Regulating model reliance on non-robust features by smoothing input marginal density," in *ECCV*, 2024.

[5] P. Yang, N. Akhtar, J. Jiang, and A. Mian, "Backdoor-based explainable ai benchmark for high fidelity evaluation of attribution methods," *arXiv preprint arXiv:2405.02344*, 2024.

[6] J. Wei, S. Guan, D. Li, Z. Hou, M. Su, Y. Guo, X. Jin, J. Guo, and X. Cheng, "A survey of link prediction in n-ary knowledge graphs," in *EMNLP*, 2025.

[7] J. Wei, S. Guan, X. Jin, J. Guo, and X. Cheng, "Inductive link prediction in n-ary knowledge graphs," in *Coling*, 2025.

[8] C. Xiao, J. Dou, Z. Lin, Z. Ke, and L. Hou, "From points to coalitions: Hierarchical contrastive shapley values for prioritizing data samples," in *AAAI*, 2026.

[9] C. Xiao, T. Xu, Y. Jiang, H. Gao, Y. Wu *et al.*, "Reversible primitive–composition alignment for continual vision–language learning," in *ICLR*, 2026.

[10] J. Zhang, C. Zhao, C. Xiao, R. Duan, W. Mo, H. Gao, and W. Wang, "Pi-cca: Prompt-invariant cca certificates for replay-free continual multimodal learning," in *ICLR*, 2026.

[11] X. Chen, C. Xiao, and Y. Liu, "Confusion-resistant federated learning via diffusion-based data harmonization on non-iid data," in *NeurIPS*, 2024.

[12] L. Li, S. Jia, J. Wang, Z. Jiang, F. Zhou, J. Dai, T. Zhang, Z. Wu, and J.-N. Hwang, "Human motion instruction tuning," in *CVPR*, 2025.

[13] R. Gu, S. Jia, Y. Ma, J. Zhong, J.-N. Hwang, and L. Li, "Mocount: Motion-based repetitive action counting," in *ACM MM*, 2025.

[14] O. Ronneberger, P. Fischer, and *et al.*, "*U*-Net: Convolutional networks for biomedical image segmentation," in *MICCAI*. Springer, 2015.

[15] X. Chen, X. Wang, and *et al.*, "Recent advances and clinical applications of deep learning in medical image analysis," *Medical Image Analysis*, 2022.

[16] J. Chen, Y. Lu, and *et al.*, "TransUNet: Transformers make strong encoders for medical image segmentation," 2021.

[17] A. Vaswani, N. Shazeer, and *et al.*, "Attention is all you need," in *NeurIPS*, 2017.

[18] Z. Liu, Y. Lin, and *et al.*, "Swin Transformer: Hierarchical vision transformer using shifted windows," in *ICCV*, 2021.

[19] M. Beck, K. Pöppel, and *et al.*, "xLSTM: Extended Long Short-Term Memory," *arXiv preprint arXiv:2405.04517*, 2024.

[20] B. Alkin, M. Beck, and *et al.*, "Vision-LSTM: xLSTM as generic vision backbone," *arXiv preprint arXiv:2406.04303*, 2024.

[21] M. M. Rahman, S. Shokouhmand, and *et al.*, "MIST: Medical Image Segmentation Transformer with Convolutional Attention Mixing (CAM) decoder," in *WACV*, 2024.

[22] X. Yan, H. Tang, and *et al.*, "After-UNet: Axial fusion transformer U-Net for medical image segmentation," in *WACV*, 2022.

[23] P. Dutta and S. Mitra, "Efficient global-context driven volumetric segmentation of abdominal images," in *BIBM*, 2023.

[24] L. Li, "Cpseg: Finer-grained image semantic segmentation via chain-of-thought language prompting," in *WACV*, 2024.

[25] J. Shi, Q. Ma, H. Ma, and L. Li, "Scaling law for time series forecasting," *NeurIPS*, 2024.

[26] Y. Tang, T. Wang, Y. Chen, B. Zhang, Q. Guan, and R. Tang, "Shifting uncertainty to critical moments: Towards reliable uncertainty quantification for vla model," 2026.

[27] P. Zhang, X. Gao, Y. Wu, K. Liu, D. Wang, Z. Wang, B. Zhao, Y. Ding, and X. Li, "Moma-kitchen: A 100k+ benchmark for affordance-grounded last-mile navigation in mobile manipulation," in *ICCV*, 2025.

[28] P. Zhang, Y. Su, P. Wu, D. An, L. Zhang, Z. Wang, D. Wang, Y. Ding, B. Zhao, and X. Li, "Cross from left to right brain: Adaptive text dreamer for vision-and-language navigation," *arXiv preprint arXiv:2505.20897*, 2025.

[29] S. Jiang, T. Zheng, Y. Zhang, Y. Jin, L. Yuan, and Z. Liu, "Med-moe: Mixture of domain-specific experts for lightweight medical vision-language models," in *Findings of EMNLP*, 2024.

[30] S. Jiang, Y. Wang, S. Song, T. Hu, C. Zhou, B. Pu, Y. Zhang, Z. Yang, Y. Feng, J. T. Zhou *et al.*, "Hulu-med: A transparent generalist model towards holistic medical vision-language understanding," *arXiv preprint arXiv:2510.08668*, 2025.

[31] S. Lian, Y. Wu, J. Ma, Y. Ding, Z. Song, B. Chen, X. Zheng, and H. Li, "Ui-agile: Advancing gui agents with effective reinforcement learning and precise inference-time grounding," *arXiv preprint arXiv:2507.22025*, 2025.

[32] Y. Liu, J. Zhu, Y. Mo, G. Li, X. Cao, J. Jin, Y. Shen, Z. Li, T. Yu, W. Yuan *et al.*, "Palm: Progress-aware policy learning via affordance reasoning for long-horizon robotic manipulation," *arXiv preprint arXiv:2601.07060*, 2026.

[33] F. Xu, G. Zhai, X. Kong, T. Fu, D. F. Gordon, X. An, and B. Busam, "Stare-vla: Progressive stage-aware reinforcement for fine-tuning vision-language-action models," *arXiv preprint arXiv:2512.05107*, 2025.

[34] Y. Shen, Y. Liu, J. Zhu, X. Cao, X. Zhang, Y. He, W. Ye, J. M. Rehg, and I. Lourentzou, "Fine-grained preference optimization improves spatial reasoning in vlms," *arXiv preprint arXiv:2506.21656*, 2025.

[35] B. Li, Y. Shen, Y. Liu, Y. Xu, J. Liu, X. Li, Z. Li, J. Zhu, Y. Zhong, F. Lan *et al.*, "Toward cognitive supersensing in multimodal large language model," *arXiv preprint arXiv:2602.01541*, 2026.

[36] D. Zhao, W. Li, Z. Shen, Y. Qiu, B. Xu, H. Chen, and Y. Chen, "Bias is a subspace, not a coordinate: A geometric rethinking of post-hoc debiasing in vision-language models," *arXiv preprint arXiv:2511.18123*, 2025.

[37] S. Du, Y. Zou, Z. Wang, X. Li, Y. Li, C. Shang, and Q. Shen, "Unsupervised hyperspectral image super-resolution via self-supervised modality decoupling," *International Journal of Computer Vision*, 2026.

[38] O. Oktay, J. Schlemper, and *et al.*, "Attention U-Net: Learning where to look for the pancreas," in *Proceedings of the Medical Imaging with Deep Learning*, 2018.

[39] Z. Zhou, R. Siddiquee, and *et al.*, "A nested U-Net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer, 2018, pp. 3-11.

[40] S. Zeng, D. Qi, X. Chang, F. Xiong, S. Xie, X. Wu, S. Liang, M. Xu, and X. Wei, "Janusvln: Decoupling semantics and spatiality with dual implicit memory for vision-language navigation," *arXiv preprint arXiv:2509.22548*, 2025.

[41] X. Hou, S. Xu, M. Biyani, M. Li, J. Liu, T. C. Hollon, and B. Wang, "Codev: Code with images for faithful visual reasoning via tool-aware policy optimization," *arXiv preprint arXiv:2511.19661*, 2025.

[42] C. Zhao, Y. Lyu, A. Chowdury, E. Harake, A. Kondepudi, A. Rao, X. Hou, H. Lee, and T. Hollon, "Towards scalable language-image pre-training for 3d medical imaging," *arXiv preprint arXiv:2505.21862*, 2025.

[43] Y. Li, C. Yang, H. Zeng, Z. Dong, Z. An, Y. Xu, Y. Tian, and H. Wu, "Frequency-aligned knowledge distillation for lightweight spatiotemporal forecasting," in *ICCV*, 2025.

[44] Y. Li, S. Meng, C. Yang, W. Feng, J. Liu, Z. An, Y. Wang, and Y. Tian, "A comprehensive survey of interaction techniques in 3d scene generation," *Authorea Preprints*, 2026.

[45] L. Li, S. Jia, and J.-N. Hwang, "Multiple human motion understanding," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 40, no. 8, 2026, pp. 6297–6305.

[46] L. Liu, S. Chen, S. Jia, J. Shi, C. Jin, Z. Wu, J.-N. Hwang, and L. Li, "Graph canvas for controllable 3d scene generation," in *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025, pp. 2536–2545.

[47] Y. Li, K. Li, X. Yin, Z. Yang, Z. Dong, Z. Yao, H. Xu, Y. Tian, and Y. Lu, "Sepprune: Structured pruning for efficient deep speech separation," in *AAAI*, 2026.

[48] C. Zhu, Y. Lin, S. Chen, Y. Wang, and J. Lin, "Medeyes: Learning dynamic visual focus for medical progressive diagnosis," in *AAAI*, 2026.

[49] C. Zhu, Y. Lin, J. Shao, J. Lin, and Y. Wang, "Pathology-aware prototype evolution via llm-driven semantic disambiguation for multicenter diabetic retinopathy diagnosis," in *ACM Multimedia*, 2025.

[50] Y. Li, Z. Zhou, Z. Peng, J. Dong, H. You, R. Yan, S. Wen, Y. Tian, and T. Huang, "A preference-driven methodology for efficient code generation," *IEEE Transactions on Artificial Intelligence*, 2025.

[51] R. Zhang, Y. Lou, Y. Huang, Y. Xin, Y. Cao, and H. Wang, "Beyond conservation: Flexible molecular assembly with unbalanced diffusion bridge," in *AAAI*, 2026.

[52] Y. Li, K. Ding, C. Yang, H. Wang, H. Wang, H. Duan, J. Liu, and Y. Tian, "Ddtime: Dataset distillation with spectral alignment and information bottleneck for time-series forecasting," *arXiv preprint arXiv:2511.16715*, 2025.

[53] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-unet: Unet-like pure transformer for medical image segmentation," in *ECCV*, 2022.

[54] Y. Li, H. Zeng, F. Zhang, C. Yang, Y. Li, and W. Ding, "Efficient Medical Image Segmentation via Reinforcement Learning-Driven K-Space Sampling," *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2025.

[55] Y. Chu, G. Luo, L. Zhou, S. Cao, G. Ma, X. Meng, J. Zhou, C. Yang, D. Xie, D. Mu *et al.*, "Deep learning-driven pulmonary artery and vein segmentation reveals demography-associated vasculature anatomical differences," *Nature Communications*, 2025.

[56] Y. Chu, Y. Zhang, Z. Han, C. Yang, L. Zhou, G. Luo, C. Huang, and X. Gao, "Improving representation of high-frequency components for medical visual foundation models," *IEEE Transactions on Medical Imaging*, 2025.

[57] Y. Chu, L. Zhou, G. Luo, Z. Qiu, and X. Gao, "Topology-preserving computed tomography super-resolution based on dual-stream diffusion model," in *MICCAI*. Springer, 2023.

[58] J. Ruan and S. Xiang, "Vm-unet: Vision mamba unet for medical image segmentation," *ArXiv*, vol. abs/2402.02491, 2024.

[59] J. Liu, H. Yang, and *et al.*, "Swin-UMamba: Mamba-based *U*-Net with ImageNet-based pretraining," in *MICCAI*, 2024.

[60] J. Xu, "Hc-mamba: Vision mamba with hybrid convolutional techniques for medical image segmentation," *arXiv preprint arXiv:2405.05007*, 2024.

[61] C. Chen, L. Yu, S. Min, and S. Wang, "Msvm-unet: Multi-scale vision mamba unet for medical image segmentation," *BIBM*, 2024.

[62] J. Liu, K. Li, C. Huang, H. Dong, Y. Song, and R. Li, "Mixformer: A mixed cnn–transformer backbone for medical image segmentation," *IEEE Transactions on Instrumentation and Measurement*, 2025.

[63] J. Liu, H. Yang, H.-Y. Zhou, Y. Xi, L. Yu, C. Li, Y. Liang, G. Shi, Y. Yu, S. Zhang *et al.*, "Swin-umamba: Mamba-based unet with imagenet-based pretraining," in *MICCAI*, 2024.