U-ViL: A U-shaped Vision-LSTM Framework for Cardiac Image Segmentation

Ehsan Khodapanah Aghdam
Department of Computer Science
Vanderbilt University
ehsan.khodapanah.aghdam@vanderbilt.edu

Nick DiSanto
Department of Computer Science
Vanderbilt University
nicolas.c.disanto@vanderbilt.edu

Abstract

Deep learning techniques have demonstrated remarkable success in medical image segmentation, but challenges remain in simultaneously capturing both global contextual dependencies and local structural details in the presence of structural variability. To address these challenges, we propose **U-ViL** (U-Net-like Vision-LSTM), a novel U-shaped architecture fused with Vision Long Short-Term Memory units. U-ViL incorporates Vision-LSTM blocks as the backbone of an encoder-decoder framework, aiming to model both low-level features and long-range dependencies. We evaluate the proposed model on the Automated Cardiac Diagnosis Challenge dataset and benchmark it against widely used segmentation architectures, including the conventional U-Net and the transformer-based Swin-Unet. Although the current implementation of U-ViL does not yield overall superior segmentation accuracy, it reveals distinct qualitative feature representations and segmentation patterns compared to prevailing architectures and could benefit from further hierarchical refinement. These findings highlight the potential for implementing a unified framework combining recurrent mechanisms that capture long-range spatial dependencies with localized spatial precision for medical image analysis. All source code is publicly available at GitHub.

1 Introduction

In recent years, medical image segmentation has become a benchmark task in deep learning, serving both as a critical tool for clinical diagnosis and as a quantitative benchmark for evaluating novel deep learning architectures. Deep learning has become particularly valuable in high-stakes clinical contexts, such as cardiology and neuroimaging, where accurate and reproducible segmentation directly correlates with patient outcomes. Traditionally, Convolutional Neural Networks (CNNs) and their variants have been the dominant architectures in this field, with models such as U-Net [15] achieving state-of-the-art performance across diverse medical applications like ultrasounds [2] and MRIs [1]. These models are particularly effective in medical contexts due to their abilities to retain high-level semantic structure through their encoder-decoder structures and capture localized spatial information through skip connections. However, despite their empirical success, the focus of conventional CNNs and U-Nets on localized patterns can lead them to struggle to capture long-range dependencies and global contextual information [5]. This distinction is vital to maintain accurate

Preprint. Under review.

results in the presence of structural variability, occlusions, and low-contrast boundaries that are frequently inherent in medical imaging.

Recent advances in Vision Transformers (ViTs) [17, 18] have brought renewed interest to modeling long-range dependencies and global context within empirical applications. While this approach is promising in medical image analysis because they do not rely on locality-biased convolutional filters, ViTs often require extensive computational resources and large-scale datasets to train effectively. Additionally, ViTs typically operate over non-overlapping image patches, which can lead to a loss of granular spatial details necessary for precise segmentation. Given the need for both detailed localization and global representation learning, we propose U-ViL (U-Net-like Vision-LSTM), a novel segmentation architecture that integrates Vision-LSTM blocks as the backbone of an encoder-decoder framework inspired by U-Net. Our architecture seeks to leverage the long-range modeling capacity of Vision-LSTM modules to capture global dependencies while preserving the structural precision offered by U-Net.

In order to validate our proposed approach, we use the Automated Cardiac Diagnosis Challenge (ACDC) dataset [3], a benchmark consisting of cine-MRI sequences across multiple cardiac pathologies. The ACDC dataset presents a diverse range of cardiac pathologies and maintains consistent spatiotemporal resolution, making it a fitting benchmark for performance evaluation. Accordingly, to assess the effectiveness of U-ViL relative to both U-Net and ViT baselines, we employ standard segmentation metrics, such as Dice score and Hausdorff Distance, to assess region overlap and boundary accuracy. To this end, we seek to demonstrate whether U-ViL, a lightweight hybrid architecture of U-Net and extended LSTM units, can combine to perform high-performance medical image segmentation. In doing so, we evaluate whether combining global dependency modeling and spatial encoding should continue to be explored across future domains.

2 Related Work

2.1 U-Net and its Variants

U-Net, originally introduced by Ronneberger et al. [15], quickly emerged as a foundational architecture for a broad variety of medical image segmentation tasks. U-Net follows a symmetric encoder-decoder design that downsamples to extract high-level semantic representations and then upsamples with skip connections to reintegrate low-level spatial information. Since U-Net's inception, a variety of derivatives have been heuristically proposed to enhance feature extraction, improve generalization, and address domain-specific challenges. For example, Residual U-Net [19] incorporates residual blocks to mitigate vanishing gradients, while Attention U-Net [14] introduces attention gates within skip connections to selectively emphasize salient features. Despite these applications continuing to tackle unique tasks in a wide variety of domains, U-Net remains inherently limited by its lack of explicit mechanisms to model spatial recurrence. This motivates the integration of recurrent modules to enable the model to maintain coherence across feature hierarchies and improve its representation of complex anatomical structures.

2.2 Vision-LSTM Networks

Extended Long Short-Term Memory (xLSTM) and Vision-LSTM networks were introduced as an enhancement over traditional LSTM architectures to enable more effective information flow via parallelization [1]. In vision tasks, xLSTM's ability to capture fine-grained temporal relationships has made them attractive in applications like video sequencing [7] and stock price prediction [10]. Recent studies have also attempted to deploy Vision-LSTM Networks on semantic segmentation tasks, such as on remotely sensed images [21]. Additionally, Vision-LSTM networks have shown relevance in the medical field, with self-supervised xLSTM-UNet architectures demonstrating high performance on a variety of tumor segmentation tasks [13]. While segmentation pipelines are currently dominated by convolutional or transformer backbones, recurrent architectures like LSTM and GRU can offer advantages in both modeling continuity and structure—characteristics especially relevant for semantic segmentation.

2.3 Hybrid Architectures

Given the applicability of both architectures, recent research has explored hybrid approaches that integrate U-Net with Vision-LSTM modules to overcome locality constraints. For instance, Chen et al. [6] replaced Mamba (a state-space model) with xLSTM and outperformed CNN-based and Transformer-based baselines over multiple medical domains. Additionally, models like U-VixLSTM [9] have embedded Vision-LSTM blocks into the bottlenecks of a U-Net-shaped structure to attempt to retain spatial inductive biases while capturing temporal relationships. Additionally, Swin-Unet [4], a variant of the Swin Transformer [12] in a U-shaped structure, was established specifically for proficiency in medical image segmentation. The growing interest in implementing these hybrid architectures in the medical domain motivates our proposed U-ViL architecture, which seeks to address the contextual limitations inherent in cardiac image segmentation pipelines.

3 U-ViL: U-shaped Vision-LSTM

3.1 Architecture Overview

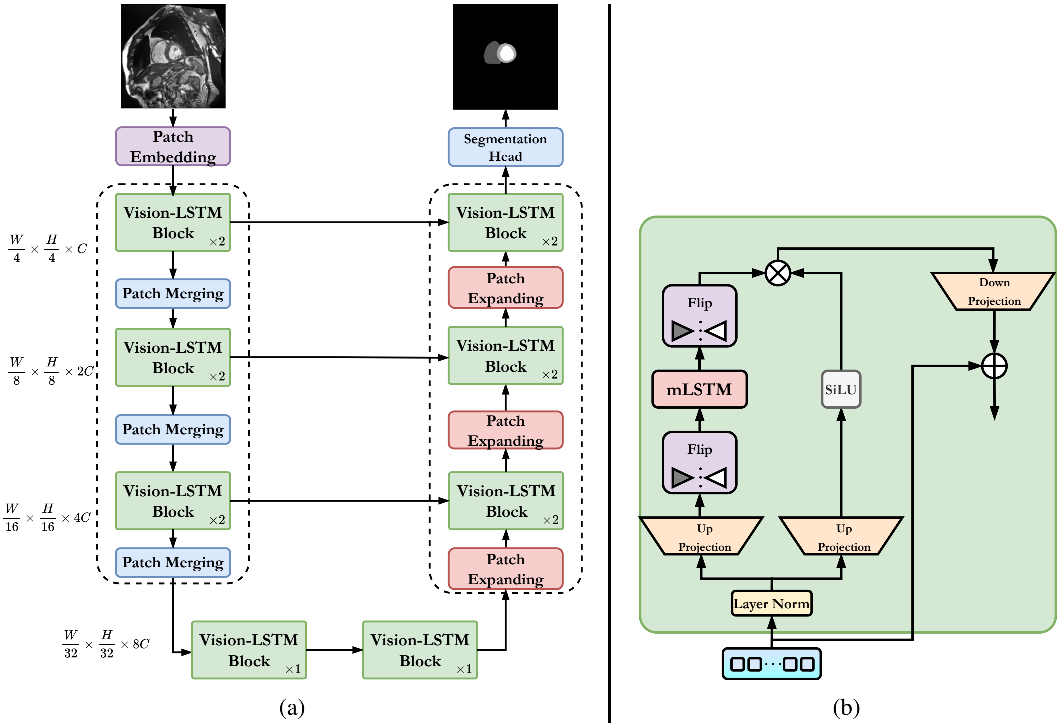
Patch Embedding
Vision-LSTM Block ×2
Patch Merging
$\frac{W}{4} \times \frac{H}{4} \times C$
$\frac{W}{8} \times \frac{H}{8} \times 2C$
$\frac{W}{16} \times \frac{H}{16} \times 4C$
$\frac{W}{32} \times \frac{H}{32} \times 8C$
Vision-LSTM Block ×1
Segmentation Head
Patch Expanding
(a)
Flip
mLSTM
SiLU
Up Projection
Down Projection
Layer Norm
(b)

Figure 1: **(a) The overview of the proposed U-ViL.** U-ViL consists of ViL blocks as the main backbone in a U-shaped structure, with contracting and expanding paths. **(b) The overview of the Vision-LSTM (ViL) block.** Each ViL block is built around the mLSTM, which parallelizes LSTM operations by extending them to matrix form.

The overall architecture of the proposed U-ViL is presented in Figure 1a. U-ViL consists of an encoder, decoder, bottleneck, and skip connections, with the foundational block based on **Vision-LSTM (ViL)**. At its core, the model leverages the ViL block as its fundamental unit, as seen in Figure 1b. In the encoder, images are initially divided into non-overlapping patches of size 4×4 and converted into sequential embeddings. A linear embedding layer then projects these into an arbitrary feature space, denoted as C. These patch tokens are passed through multiple ViL blocks and patch merging layers to build broad feature representations. The patch merging layer performs downsampling and increases the feature dimension, while the ViL block handles feature extraction and representation learning. Inspired by U-Net, a symmetric transformer-based decoder is constructed, consisting of ViL blocks and patch expanding layers. It integrates the extracted deep features with multi-scale encoder features through skip connections, mitigating spatial information loss from downsampling. In contrast to the merging layer, the patch expanding layer is used for $(2\times)$ upsampling, as it rearranges neighboring features into a larger resolution space. Finally, a last patch expanding layer executes $(4\times)$ upsampling to recover the original image resolution $(W \times H)$, and a linear projection layer is applied to generate pixel-level segmentation outputs.

3.2 Vision-LSTM (ViL) Block

The Vision-LSTM block introduces xLSTM to vision tasks, an exponential gating mechanism that dynamically filters information to enhance its capability to capture complex patterns. xLSTM introduces two distinct memory cells: scalar LSTM (sLSTM) and matrix LSTM (mLSTM). The sLSTM employs a scalar update mechanism to boost robustness in extensive sequence scenarios, while the mLSTM extends vector operations into matrix form to improve efficiency. xLSTM can be applied to images using patch tokenization, as in vision transformers, along with directional scanning strategies–such as top-to-bottom and bottom-to-top passes—adapted from Vision Mamba [20] and VMamba [11]. **The broad previous success of Vision-LSTM (ViL) blocks raises the question of whether it can serve as the main backbone for high-resolution vision tasks, such as image segmentation.** Our model's ViL blocks (see Figure 1b) alternate mLSTM modules in a bidirectional strategy to process patch tokenizations. Additionally, the $d \times d$ matrix memory in the mLSTM and associated scanning strategy enable the ViL blocks to operate within a single time step. The mLSTM utilizes a modified version of the self-attention mechanism introduced in the Vision Transformer [8], facilitating parallelization and high efficiency.

3.3 Encoder

The encoder is designed to progressively extract feature representations from the input image. The image is first partitioned into non-overlapping 4×4 patches, which are linearly projected to an embedding dimension C, which is 96. This produces patch tokens with initial resolutions of $\frac{H}{4} \times \frac{W}{4}$. At each encoder stage, the patch tokens are processed by two consecutive ViL blocks, preserving their spatial resolution while enriching feature representations. Following these blocks, a patch merging layer reduces the spatial resolution by a factor of 2 and doubles the feature dimension by concatenating spatially adjacent patches along the channel dimension, followed by a linear projection. This process is repeated across three encoder stages, resulting in a multi-scale feature hierarchy at progressively coarser resolutions.

3.4 Bottleneck

The bottleneck serves as the deepest layer of the network. Similarly to Swin-Unet [4], we employ two consecutive ViL blocks within the bottleneck to capture deep feature representations effectively. These blocks maintain both the feature dimension and spatial resolution, allowing the model to focus on refining representations without further downsampling.

3.5 Decoder

The decoder mirrors the structure of the encoder but operates in reverse to progressively recover spatial resolution and reconstruct the segmentation output. At each stage, the decoder employs a ViL block to process feature maps, followed by a patch expanding layer that performs 2×2 upsampling. The patch expanding operation begins by doubling the feature dimension via a linear projection and rearranging features from the channel dimension into the spatial domain. Finally, it doubles the resolution while reducing feature dimensionality by a factor of 2. After the final decoder stage, an additional patch expanding layer performs 4×4 upsampling to restore the original input resolution $H \times W$. A linear projection layer is applied to the final upsampled features to produce pixel-level segmentation outputs.

3.6 Skip Connections

Skip connections are integrated at each corresponding encoder-decoder stage to fuse low-level spatial features with high-level semantic features. Specifically, the features from the encoder are concatenated with the upsampled decoder features along the channel dimension. A linear projection layer is then applied to the concatenated features, aligning them with the dimensionality expected by subsequent ViL blocks. This strategy preserves fine-grained spatial details lost during downsampling while enhancing the decoder's capacity to maintain high precision.

4 Experiments

4.1 Dataset

We performed our experiments using the publicly accessible **Automated Cardiac Diagnosis Challenge (ACDC)** MRI cardiac segmentation dataset from the MICCAI 2017 Challenge [3]. This dataset includes MRI scans from 100 patients, each annotated with various cardiac structures. For each patient, the dataset provides short-axis cine-MRI images acquired over the cardiac cycle, along with manual annotations for key cardiac structures. The MR images are labeled to identify the left ventricle (LV), right ventricle (RV), and myocardium (MYO). The dataset encompasses a wide spectrum of pathological conditions grouped into five diagnostic categories: normal, myocardial infarction, dilated cardiomyopathy, hypertrophic cardiomyopathy, and abnormal right ventricle. Additionally, the ACDC dataset provides ground-truth segmentations for four primary regions of interest (ROIs): the LV cavity, RV cavity, myocardium, and background. This diversity facilitates a comprehensive feature space for training and evaluating segmentation algorithms, and supports reproducible comparisons across different segmentation methods in cardiac image analysis research.

4.2 Implementation

The implementation was conducted using PyTorch 2.5.1 and CUDA 12.4 and utilized both GPU and high-performance CPU capabilities. The full ACDC data processing pipeline, including data transfer, model training, and inference, took around 4 hours. The datasets were tailored for 2D image segmentation tasks, and U-ViL was trained for 10,000 iterations with a batch size of 4. Training employed the Stochastic Gradient Descent optimizer with a learning rate of 0.01 and a momentum of 0.9, and model performance was assessed on the validation set every 200 iterations.

4.3 Evaluation Metrics and Loss Functions

To evaluate U-ViL's performance, we employed three widely used metrics: Dice Similarity Coefficient (Dice), 95th percentile Hausdorff Distance (HD 95%), and Average Surface Distance (ASD). The Dice coefficient measures the overlap between the predicted and ground truth segmentations, with values ranging from 0 (no overlap) to 1 (perfect overlap). It is a widely adopted metric for evaluating segmentation accuracy, particularly in medical imaging. Given true positives (TP), false positives (FP), and false negatives (FN), the Dice coefficient is defined as:

$$\text{Dice} = \frac{2 \times TP}{2 \times TP + FP + FN}. \tag{1}$$

HD 95% quantifies the spatial discrepancy between the boundaries of predicted and ground truth segmentations by computing the 95th percentile of the Hausdorff Distance, reducing the sensitivity to outliers. It is formulated as:

$$\text{Hausdorff Distance (HD) } 95\% = \max\left(\max_{a\in A}\min_{b\in B} d(a,b),\ \max_{b\in B}\min_{a\in A} d(a,b)\right)_{95\%}. \tag{2}$$

The Average Surface Distance (ASD) computes the mean distance between the surfaces of the predicted and ground truth segmentations, offering a robust measure of boundary accuracy. It is defined as:

$$\text{Average Surface Distance (ASD)} = \frac{1}{|A|+|B|}\left(\sum_{a\in A}\min_{b\in B} d(a,b) + \sum_{b\in B}\min_{a\in A} d(a,b)\right), \tag{3}$$

where A and B are the sets of surface points from the predicted and ground truth segmentations, respectively, and $d(a, b)$ denotes the Euclidean distance between points $a \in A$ and $b \in B$.

Cross-entropy (CE) loss is widely used in classification and segmentation tasks, particularly when dealing with probabilistic outputs. It is derived from the Kullback–Leibler (KL) divergence, which measures the dissimilarity between two probability distributions. In semantic segmentation, CE loss quantifies the difference between the predicted class probabilities and the ground truth labels for each voxel. It is defined as:

$$L_{CE} = -\frac{1}{N} \sum_{i=1}^{C} \sum_{j=1}^{N} p_i^j \log q_i^j, \tag{4}$$

where C is the number of classes, N is the number of voxels, p_i^j is the ground truth indicator for class i at voxel j, and q_i^j represents the predicted probability for that class. This formulation encourages the model to assign high probabilities to the correct classes while penalizing incorrect predictions.

The Dice loss is derived from the Sørensen–Dice Coefficient (DSC), a statistical measure of set similarity. It is particularly effective for segmentation tasks involving class imbalance, as it directly optimizes for the overlap between predicted and ground truth masks. Introduced by Sudre et al. [16], Dice loss is defined as:

$$L_{\text{Dice}}(y, \hat{y}) = 1 - \frac{2y\hat{y} + \alpha}{y + \hat{y} + \alpha}, \tag{5}$$

where y and $\hat{y}$ represent the ground truth and predicted segmentation outputs and α is a small constant added to prevent division by zero. Owing to its robustness to class imbalance, Dice loss is well-suited for medical image segmentation. To balance voxel-wise accuracy and overlap, we optimize a weighted sum of cross-entropy and Dice loss as:

$$\mathcal{L}_{\text{tot}} = \lambda_1 \mathcal{L}_{ce} + \lambda_2 \mathcal{L}_{dice}, \tag{6}$$

with empirically chosen weights $\lambda_1 = \lambda_2 = \frac{1}{2}$.

4.4 Results

Table 1 compares the segmentation performance of U-ViL against U-Net and Swin-Unet on the ACDC cardiac MRI dataset using key metrics: Dice score, HD 95%, and Average Surface Distance (ASD). U-Net achieves the highest accuracy, with a Dice score of 0.8985, HD 95% of 1.12 mm, and ASD of 0.32 mm, indicating precise segmentation boundaries and minimal deviations. Swin-Unet, while requiring significantly greater model capacity (27.26M parameters) compared to U-Net (1.81M parameters), shows marginally lower segmentation performance, reflected by its DSC of 0.8844 and increased HD 95% and ASD values. The proposed U-ViL architecture, leveraging Vision-LSTM components, achieves lower computational cost (12.61 GFLOPS) compared to Swin-Unet (21.39 GFLOPS), but underperforms in segmentation accuracy across all metrics. Specifically, U-ViL records a Dice score of 0.7555, HD 95% of 3.97 mm, and ASD of 1.23 mm, suggesting the need for further refinement to balance efficiency with segmentation accuracy. These results highlight U-Net's efficiency and effectiveness, whereas the trade-offs in U-ViL emphasize the importance of optimizing complexity and performance in segmentation tasks.

Table 1: Comparison of segmentation performance on the ACDC dataset. RV (Right Ventricle), MYO (Myocardium), and LV (Left Ventricle) show class-wise Dice scores across cardiac regions.

Method	**Params (M)**	**FLOPS (G)**	**DSC** ↑	**HD 95%** ↓	**ASD** ↓	**RV**	**MYO**	**LV**
U-Net	1.81	4.51	0.8985	1.12	0.32	0.8917	0.8687	0.9351
Swin-Unet	27.26	21.39	0.8844	3.18	1.03	0.8864	0.8452	0.9216
U-ViL (Ours)	28.33	12.61	0.7555	3.97	1.23	0.7247	0.7042	0.8378

Table 2: Ablation study evaluating U-ViL skip connections and input resolution.

Setting	**DSC**	**HD 95%**
Using 1 skip connection	0.7330	4.52
Using 2 skip connections	0.7498	3.99
Using 3 skip connections	0.7555	3.97
Input image size 384×384	0.7578	4.07

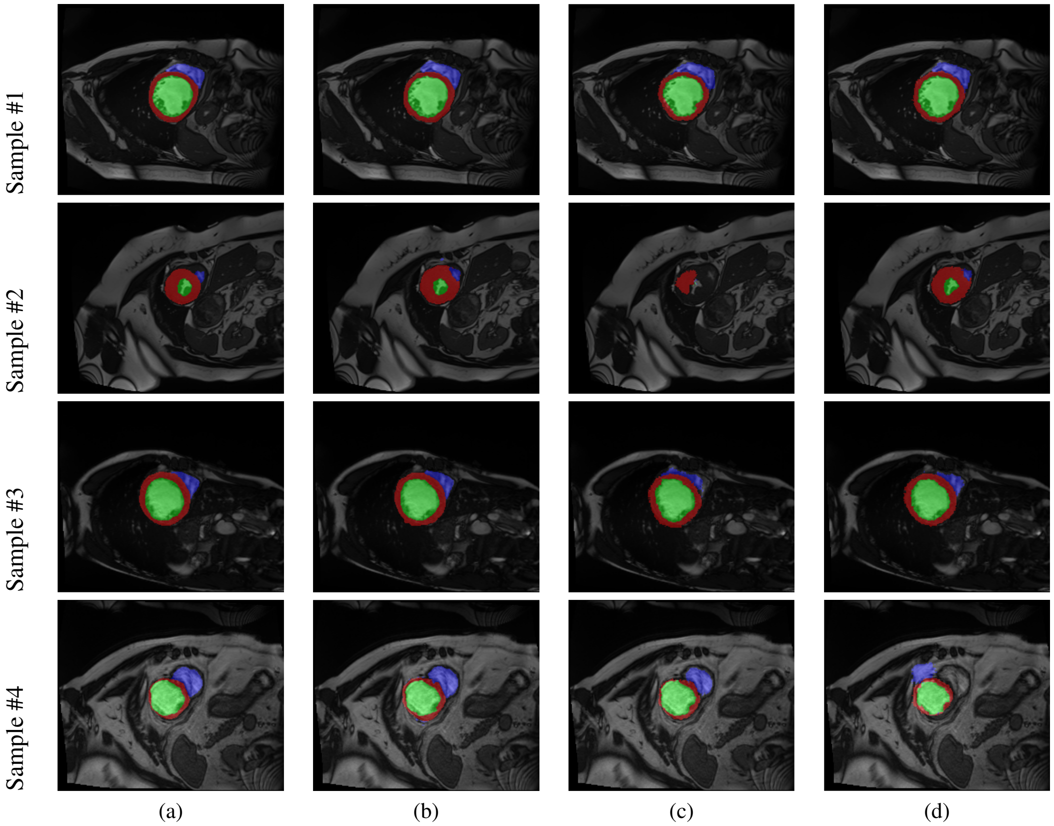

Figure 2: Visual comparisons of each method's performance on four samples from the ACDC dataset. Columns: (a) Ground Truth. (b) U-Net, (c) Swin-Unet, (d) Proposed method (U-ViL). The 🟦 Right Ventricle (RV), 🟩 Myocardium (MYO), and 🟥 Left Ventricle (LV) are highlighted.

To explore the influence of different architectural choices on U-ViL's performance, we conducted ablation studies focusing on two factors: the number of skip connections employing the proposed attention strategy and the input resolution. Specifically, we investigated the effect of incorporating our proposed attention at different resolution scales–$\frac{1}{4}$, $\frac{1}{8}$, and $\frac{1}{16}$–by varying their inclusion in the skip connections for $i \in 1, 2, 3$. This allowed us to assess how the depth of attention integration influences segmentation quality. Table 2 demonstrates that the segmentation performance improves when the number of our proposed modules in the skip connections increases, supporting the effectiveness of the proposed module in enhancing feature representation. We also analyzed the effect of increasing the input resolution to 384×384, and show slightly improved segmentation results with a high computational cost. These findings suggest that both deeper attention integration and higher input resolution can enhance the generalization capacity of U-ViL, as reflected by improved Dice scores.

5 Discussion

5.1 Summary of Findings

Although U-ViL did not outperform U-Net and Swin-Unet on standard segmentation metrics such as Dice score, HD 95%, or ASD, its unique results include several noteworthy observations. Quantitatively, U-ViL recorded lower Dice scores across all three classes, including reductions on the Myocardium (–0.167) and Right Ventricle (–0.0973) compared to U-Net. U-ViL also underperformed relative to Swin-Unet but offered a 42.5% reduction in floating point operations (FLOPs), a notable improvement in computational efficiency. These results show that while U-ViL does not yet outperform in segmentation quality, its efficiency and design offer a strong basis for future refinements.

Importantly, qualitative analysis reveals that U-ViL captures structural representations distinct from those of its baseline counterparts. For example, in Sample 2 of Figure 2, U-ViL segments the LV and MYO with far stronger alignment to the ground truth than Swin-Unet does. These results indicate that U-ViL introduces a novel inductive bias capable of emphasizing anatomical coherence in ways not reflected by general metrics alone.

5.2 Interpretation and Implications

Our findings point to both architectural strengths, as well as areas of improvement in U-ViL. Both the quantitative and qualitative findings suggest that the integration of Vision-LSTM blocks introduces a distinct form of context-aware spatial encoding that differs from that of CNNs or Transformers. While traditional CNNs emphasize locality and transformer models global attention, ViL encoders propagate sequential context, potentially offering improved modeling of structural consistency.

However, despite this theoretical benefit, U-ViL underperformed in traditional segmentation metrics. One plausible explanation is that xLSTM, while well-suited to modeling global continuity, may lack the fine-grained spatial precision necessary to capture detailed localized features unless explicitly guided by spatial priors or attention mechanisms. Additionally, while the network's skip connection strategy and decoder design were inspired by U-Net, it may not have been fully sufficient to fully recover spatial detail following the encoding stages. This may have contributed to boundary degradation, as evidenced by U-ViL's higher HD 95% and ASD values.

Nevertheless, U-ViL's competitive FLOPs and qualitative results make it a strong candidate as a backbone for future hybrid segmentation designs, suggesting potential for deployment in resource-constrained environments. The ability to integrate Vision-LSTM modules and long-range dependency modeling into a U-shaped segmentation framework opens many potential doors for an efficient and anatomically aware framework.

5.3 Limitations and Future Work

Several limitations in our implementation would show clear improvements in U-ViL's performance upon addressing. First, as previously mentioned, our ViL blocks may not explicitly capture spatial locality with full effectiveness. Without sufficient regularization, these modules may underutilize spatial features that are essential for high precision, so future work should integrate spatial attention gates to improve boundary precision.

Additionally, U-ViL was trained and evaluated only on 2D slices rather than full 3D volumes. This design limits its ability to capture inter-slice anatomical continuity–a critical aspect of volumetric medical imaging. Extending the model to process 3D volumes would allow it to learn consistent representations across adjacent slices, better leveraging the architecture's capacity for modeling long-range spatial dependencies within the volumetric context.

Finally, the size and diversity of the training dataset present an inherent limitation. While the ACDC dataset is a well-accepted benchmark for standard segmentation tasks, it is still relatively small for training deep architectures with data-hungry ViL modules. For optimal performance of an architecture like U-ViL, it may require a longer training schedule on a larger, more diverse dataset. Addressing these points will be a helpful direction for future exploration of this architecture.

In summary, future work may benefit from: (1) adding spatially aware attention layers, (2) extending to 3D volumetric modeling, and (3) scaling training to larger datasets. These directions will help determine whether the representational benefits of xLSTM can be more fully realized in practical segmentation tasks.

6 Conclusion

In this work, we introduced U-ViL, a novel U-shaped segmentation architecture that integrates Vision-LSTM blocks into an encoder-decoder framework. Our motivation stemmed from an interest in jointly modeling global contextual dependencies and localized anatomical structures for cardiac MRI segmentation. Although U-ViL did not surpass existing baselines such as U-Net or Swin-Unet in quantitative performance, it demonstrated both unique qualitative segmentation characteristics and superior computational efficiency relative to transformer-based models.

U-ViL's modularity and versatility make it a promising step toward future hybrid designs that optimize performance and efficiency. Additional research should continue to explore the integration of spatial priors, 3D volumetric modeling, and larger-scale training regimes to fully leverage the potential of recurrence-based spatial modeling in medical segmentation.

References

[1] Benedikt Alkin, Maximilian Beck, Korbinian Pöppel, Sepp Hochreiter, and Johannes Brandstetter. Vision-lstm: xLSTM as generic vision backbone. In *The Thirteenth International Conference on Learning Representations*, 2025.

[2] Maximilian Beck, Korbinian Pöppel, Markus Spanring, Andreas Auer, Oleksandra Prudnikova, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. xlstm: Extended long short-term memory. *Advances in Neural Information Processing Systems*, 37:107547–107603, 2025.

[3] Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE transactions on medical imaging*, 37(11):2514–2525, 2018.

[4] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *Computer Vision – ECCV 2022 Workshops*, pages 205–218, Cham, 2023. Springer Nature Switzerland.

[5] Jieneng Chen, Jieru Mei, Xianhang Li, Yongyi Lu, Qihang Yu, Qingyue Wei, Xiangde Luo, Yutong Xie, Ehsan Adeli, Yan Wang, Matthew P Lungren, Shaoting Zhang, Lei Xing, Le Lu, Alan Yuille, and Yuyin Zhou. TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers. *Med. Image Anal.*, 97:103280, 2024.

[6] Tianrun Chen, Chaotao Ding, Lanyun Zhu, Tao Xu, Deyi Ji, Yan Wang, Ying Zang, and Zejian Li. xlstm-unet can be an effective 2d & 3d medical image segmentation backbone with vision-lstm (vil) better than its mamba counterpart, 2024.

[7] Lihui Ding, Hongliang Wang, and Lijun Fu. Stse-xlstm: A deep learning framework for automated seizure detection in long video sequences using spatio-temporal and attention mechanisms. In *2024 10th International Conference on Computer and Communications (ICCC)*, pages 781–785, 2024.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[9] Pallabi Dutta, Soham Bose, Swalpa Kumar Roy, and Sushmita Mitra. Are vision xlstm embedded unet more reliable in medical 3d image segmentation?, 2024.

[10] Xiaojing Fan, Chunliang Tao, and Jianyu Zhao. Advanced stock price prediction with xlstm-based models: Improving long-term forecasting. In *2024 11th International Conference on Soft Computing & Machine Intelligence (ISCMI)*, pages 117–123, 2024.

[11] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. Vmamba: Visual state space model. *Advances in neural information processing systems*, 37:103031–103063, 2024.

[12] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002, 2021.

[13] Abdul Qayyum Moona Mazher and Steven A Niederer. Assessing self-supervised xlstm-unet. *Head and Neck Tumor Segmentation for MR-Guided Applications: First MICCAI Challenge, HNTS-MRG 2024, Held in Conjunction with MICCAI 2024, Marrakesh, Morocco, October 17, 2024, Proceedings*, page 166, 2024.

[14] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention u-net: Learning where to look for the pancreas, 2018.

[15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[16] Carole H Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3*, pages 240–248. Springer, 2017.

[17] Chunlong Xia, Xinliang Wang, Feng Lv, Xin Hao, and Yifeng Shi. Vit-comer: Vision transformer with convolutional multi-scale feature interaction for dense predictions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5493–5502, June 2024.

[18] Seokju Yun and Youngmin Ro. Shvit: Single-head vision transformer with memory efficient macro design. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5756–5767, June 2024.

[19] Zhengxin Zhang, Qingjie Liu, and Yunhong Wang. Road extraction by deep residual u-net. *IEEE Geoscience and Remote Sensing Letters*, 15(5):749–753, 2018.

[20] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: efficient visual representation learning with bidirectional state space model. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

[21] Qinfeng Zhu, Yuanzhi Cai, and Lei Fan. Seg-lstm: Performance of xlstm for semantic segmentation of remotely sensed images, 2024.