xLSTM-UNet can be an Effective 2D & 3D Medical Image Segmentation Backbone with Vision-LSTM (ViL) better than its Mamba Counterpart

Tianrun Chen[1,2+], Chaotao Ding[2,3+], Lanyun Zhu[4+], Tao Xu[3+], Deyi Ji[6], Ying Zang[3*], Zejian Li[5], Lingyun Sun[1]

[1] College of Computer Science and Technology, Zhejiang University.
[2] KOKONI, Moxin (Huzhou) Tech. Co., LTD.
[3] School of Information Engineering, Huzhou University.
[4] Information Systems Technology and Design Pillar, Singapore University of Technology and Design.
[5] School of Software Technology, Zhejiang University.
[6] School of Information Science and Technology, University of Science and Technology of China.
+ Equal Contribution * Corresponding Author
tianrun.chen@zju.edu.cn; 02750@zjhu.edu.cn
Project Page: http://tianrun-chen.github.io/xLSTM-Unet/

TL;DR: We replace Mamba with xLSTM in UMamba, and it works better.

Abstract. Convolutional Neural Networks (CNNs) and Vision Transformers (ViT) have been pivotal in biomedical image segmentation, yet their ability to manage long-range dependencies remains constrained by inherent locality and computational overhead. To overcome these challenges, in this technical report, we first propose xLSTM-UNet, a UNet structured deep learning neural network that leverages Vision-LSTM (xLSTM) as its backbone for medical image segmentation. xLSTM is a recently proposed as the successor of Long Short-Term Memory (LSTM) networks and have demonstrated superior performance compared to Transformers and State Space Models (SSMs) like Mamba in Neural Language Processing (NLP) and image classification (as demonstrated in Vision-LSTM, or ViL implementation). Here, xLSTM-UNet we designed extend the success in biomedical image segmentation domain. By integrating the local feature extraction strengths of convolutional layers with the long-range dependency capturing abilities of xLSTM, xLSTM-UNet offers a robust solution for comprehensive image analysis. We validate the efficacy of xLSTM-UNet through experiments. Our findings demonstrate that xLSTM-UNet consistently surpasses the performance of leading CNN-based, Transformer-based, and Mamba-based segmentation networks in multiple datasets in biomedical segmentation including organs in abdomen MRI, instruments in endoscopic images, and cells in microscopic images. With comprehensive experiments performed, this technical report highlights the potential of xLSTM-based architectures in advancing biomedical image analysis in both 2D and 3D.

The code, models, and datasets are publicly available at http://tianrun-chen.github.io/xLSTM-Unet/.

1 Introduction

Biomedical image segmentation is a critical task in medical imaging, enabling precise delineation of anatomical structures and anomalies essential for diagnosis, treatment planning, and research [1,2]. In recent years, deep learning methods have achieved remarkable success in tumor segmentation [3,4] and organ segmentation in 3D Computed Tomography (CT) scans [5], as well as in cell segmentation in microscopy images [6,7,8,9]. These advancements underscore the transformative impact of deep learning on the landscape of biomedical image segmentation, paving the way for more accurate and efficient diagnostic and treatment planning tools. Traditionally, Convolutional Neural Networks (CNNs) have been the backbone of this domain in deep learning enabled method, leveraging their powerful local feature extraction capabilities [10,11,12,13,14]. More recently, Vision Transformers (ViTs) have gained popularity by offering a robust alternative, capable of capturing global context through self-attention mechanisms [15,16,17,18,19,20,21,22,23,24]. Despite their successes, both CNNs and ViTs face inherent limitations. CNNs struggle with long-range dependencies due to their localized receptive fields, while ViTs encounter substantial computational overhead [25,26,27,28], especially with high-resolution images or high-dimensional imaging modalities like 3D images or hyper-spectral imaging like stimulated Raman scattering (SRS) imaging [29,30,31,32] or mid-infrared (IR) spectroscopic imaging [33,34,35].

To address these challenges, recent work have proposed to integrate computation modules that have the long-range dependencies and also exhibits linear computational and memory complexity w.r.t. sequence length. Among these computation modules, State Space Models (SSMs) [36,37,38], like Mamba [39], has demonstrated its huge success. SSMs excel in handling long-range dependencies and have been successfully integrated into conventional UNet architectures. Variants like UMamba [40], VM-Unet[41,42,43], Mamba-Unet [44], Swin-UMamba [45], and SegMamba [46], have demonstrated their considerable success.

Meanwhile, Extended Long-Short Term Memory (xLSTM) has recently emerged as a powerful successor to Long Short-Term Memory (LSTM) networks, challenging Transformers in sequence modeling [47]. Like SSMs, xLSTM can handle long-range dependencies and maintain linear computational and memory complexity. However, xLSTM has demonstrated superior performance in neural language processing (NLP) and image classification (in its Vision-LSTM (ViL) implementation [48]). This success naturally raises the question: **Can xLSTM, or ViL, also excel in image segmentation, specifically in the field of medical image segmentation?**

The answer is **Yes!** In this technical report, we introduce xLSTM-UNet, the first xLSTM-enabled U-Net image segmentation network that can perform both 2D and 3D medical image segmentation task and achieves state-of-the-art (SOTA) results. We conducted comprehensive experiments in various 2D and 3D medical segmentation scenarios, including organs in abdominal MRI, instruments in endoscopy, cells in microscopy, and cancer segmentation in 3D brain MRI volumes. The results show that xLSTM-UNet outperforms existing CNN-based and Transformer-based segmentation methods, as well as its Mamba-based counterparts. These findings highlight the potential of xLSTM-based architecture to set new benchmarks in the field of medical image segmentation, offering improved accuracy and efficiency across a wide range of applications. To further advance research in this area, we will release the model and code in http://tianrun-chen.github.io/xLSTM-Unet/, enabling future explorations in various fields such as automated pathology detection, camouflaged image segmentation, precision agriculture, environmental monitoring, satellite imagery analysis, and industrial inspection, and so on.

2   Method

TL;DR: We just replaced Mamba with xLSTM blocks in Mamba-based UNet network [40,46] and it works. Fig. 1 showcases the com-
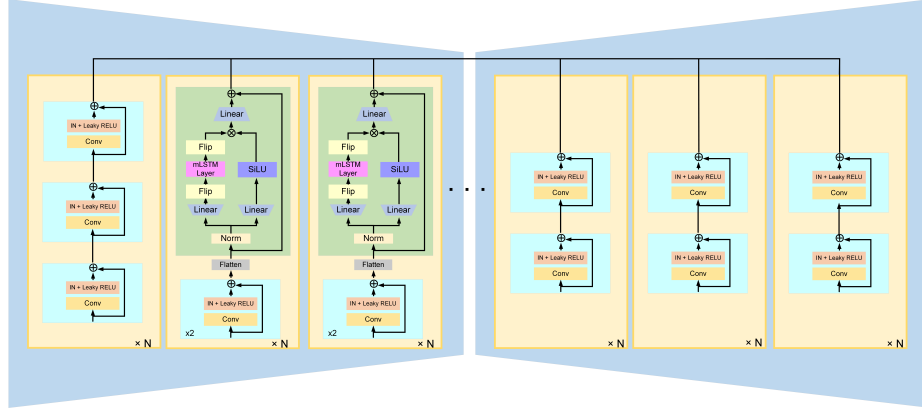

Fig. 1. The framework of the proposed method.

plete xLSTM-Unet network architecture, which following conventional UNet-like structure. The input information first pass a convolution layer for initial down-sampling. Then several subsequent layers that is constructed using the aforementioned xLSTM building blocks to capture both local features and long-range dependencies forms the main part of the encoder. Note that the xLSTM-Unet is designed with the goal to harness the best aspects of both U-Net and

xLSTM for improved global comprehension in medical image understanding. Therefore, instead of only applying xLSTM in the compressed latent space after the down-sampling has finished, we hereby use the xLSTM in multiple layers in the encoder, in which each layer contains two successive Residual blocks with one plain convolution and an Instance Normalization (IN) and followed by a xLSTM block as in [48]. Specifically, the image feature passing the residual blocks has a shape of (B, C, H, W, D), which is first flattened and transposed to $(B, H \times W \times D, C)$, followed by a layer normalization, and then feed to the ViL block. Such the practice of involving xLSTM in multiple layers helps the feature extractions in multiple resolutions / perception fields, and these information extracted by xLSTM blocks are reshaped to (B, C, H, W, D) and concatenate to the layers in the decoding steps to facilitate the segmentation mask generation.

After encoding, the decoder, comprising Residual blocks and transposed convolutions, concentrates on the meticulous recovery of detailed local information. Additionally, we inherit the skip connection from the U-Net architecture to interconnect the hierarchical features from the encoder to the decoder. The final decoder feature is fed into a 1 x 1 convolutional layer, coupled with a Softmax layer, to generate the ultimate segmentation probability map. Furthermore, following [40], we also implemented a variant where the U-xLSTM block is exclusively utilized in the bottleneck, denoted as 'ours_bot', while 'ours_enc,' denotes the network that applies xLSTM block in all encoder blocks.

3   Experiments

3.1   Datasets

TL;DR: We use the same benchmarks as UMamba [40] for 2D medical image segmentation, and the same benchmarks as SegMamba for 3D medical image segmentation.[46].[7]

To validate the effectiveness of our proposed method (xLSTM-UNet), we utilized several representative medical image segmentation datasets. These datasets encompass organ segmentation, instrument segmentation, and cell segmentation, and they span various resolutions and image modalities. By employing these diverse datasets, we can comprehensively assess the performance and applicability of xLSTM-UNet in different scenarios, thereby demonstrating its efficacy and superiority in medical image segmentation.

Abdomen MRI: We utilized the Abdomen MRI dataset from the MICCAI 2022 AMOS Challenge [5], which is dedicated to the segmentation of abdominal organs. The dataset was meticulously annotated by radiologists using Med-SAM [2] and ITK-SNAP [49]. In accordance with the U-Mamba [40] settings, we employed 60 labeled MRI scans for training and 50 MRI scans for testing, comprising 5615 slices in the training set and 3357 slices in the testing set. This

---

[7] Here we only test the open-sourced datasets BraTS2023 because only that one is publicly available.

dataset encompasses 13 distinct class labels: liver, spleen, pancreas, right kidney, left kidney, stomach, gallbladder, esophagus, aorta, inferior vena cava, right adrenal gland, left adrenal gland, and duodenum. For training purposes, in 2D segmentation tasks, the images were cropped into patches with a resolution of 320 x 320 pixels. For 3D segmentation tasks, the patches were set to a size of 48 x 160 x 224 pixels.

Endoscopy images: The endoscopy image dataset was sourced from the MICCAI 2017 EndoVis Challenge[50], focusing on the segmentation of seven surgical instruments from endoscopic images. These instruments include the large needle driver, prograsp forceps, monopolar curved scissors, cadiere forceps, bipolar forceps, vessel sealer, and an additional drop-in ultrasound probe. We adhered to the official dataset split, which comprises 1800 image frames for the training set and 1200 image frames for the testing set. The training images were extracted from eight videos, while the testing set consisted of unseen images from two new videos. The images were cropped to a size of (384, 640) pixels to accommodate the nnU-Net framework for both training and testing.

Microscopy images: The microscopy image dataset was obtained from the NeurIPS 2022 Cell Segmentation Challenge[8], which focuses on cell segmentation in various microscopy images. We used 1000 images for training and 101 images for evaluation. The original task was an instance segmentation task. In our experiments, we converted instance segmentation into a semantic segmentation task following the data processing method described in [40].

BraTS2023: The BraTS2023 dataset[51,52,53] comprises 1,251 3D brain MRI volumes. Each volume features four imaging modalities (T1, T1Gd, T2, and T2-FLAIR) and three segmentation targets: Whole Tumor (WT), Enhancing Tumor (ET), and Tumor Core (TC).

3.2   Implemetation details

The network is implemented based on UMamba [40]. The loss function used is the (Dice loss + cross-entropy loss). We employ the AdamW optimizer with a weight decay of 0.05. The learning rates were set to 0.005 for the Abdomen MRI dataset, 0.01 for the Endoscopy dataset, and for the Microscopy dataset, 0.007 for training xLSTM-UNet_Bot and 0.0015 for training xLSTM-UNet_Enc. All networks were trained from scratch for 1000 epochs on a single NVIDIA A100 GPU. For more implementation details, please refer to our codebase.

3.3   Baselines

TL;DR: The baseline is the same as that in the UMamba and SegMamba models [40,46] (and we outperform them!)

In 2D medical segmentation, to ensure a fair comparison, we following the evaluation protocol in UMamba [40]. We selected two CNN-based segmentation networks (nnU-Net [5] and SegResNet [54]) and two Transformer-based networks (UNETR [18] and SwinUNETR [17]), as well as the UMamba itself, which has

two variations: U-Mamba_Bot and U-Mamba_Enc. Similar to our configuration, U-Mamba_Bot is applied only at the bottleneck, while U-Mamba_Enc is used in each encoder. We used the Dice Similarity Coefficient (DSC) and Normalized Surface Distance (NSD) as evaluation metrics for the semantic segmentation tasks on the Abdomen MRI and Endoscopy datasets [55]. For cell segmentation on the Microscopy dataset, we employed the F1 score.

In 2D medical segmentation, to ensure a fair comparison, we following the evaluation protocol in SegMamba [46]. We use the same baselines, including three CNN-based methods (SegresNet [54], UX-Net [56], MedNeXt [57]), three transformer-based methods (UNETR [18], SwinUNETR [17], and SwinUNETR V2[58]), and Mamba-based methods SegMamba itself. Dice and HD95 were used as evaluation metrics.

3.4   Quantitative and Qualitative Results for 2D Segmentation

Table 1 presents the segmentation performance of various methods on the Abdomen MRI, Endoscopy, and Microscopy datasets. Our proposed xLSTM-UNet outperform all baseline methods and achieves the state-of-the-art (SOTA).

Table 1. Performance Comparison of Different Methods

Methods	Organs in Abdomen MRI		Instruments in Endoscopy		Cells in Microscopy
	DSC ↑	NSD ↑	DSC ↑	NSD ↑	F1 ↑
nnU-Net	0.7450 ± 0.1117	0.8153 ± 0.1145	0.6264 ± 0.3024	0.6412 ± 0.3074	0.5383 ± 0.2657
SegResNet	0.7317 ± 0.1379	0.8034 ± 0.1386	0.5820 ± 0.3268	0.5968 ± 0.3303	0.5411 ± 0.2633
UNETR	0.5747 ± 0.1672	0.6309 ± 0.1858	0.5017 ± 0.3201	0.5168 ± 0.3235	0.4357 ± 0.2572
SwinUNETR	0.7028 ± 0.1348	0.7669 ± 0.1442	0.5528 ± 0.3089	0.5683 ± 0.3123	0.3967 ± 0.2621
U-Mamba_Bot	0.7588 ± 0.1051	0.8285 ± 0.1074	0.6540 ± 0.3008	0.6692 ± 0.3050	0.5389 ± 0.2817
U-Mamba_Enc	0.7625 ± 0.1082	0.8327 ± 0.1087	0.6303 ± 0.3067	0.6451 ± 0.3104	0.5607 ± 0.2784
Ours_bot	0.7636 ± 0.1006	0.8322 ± 0.1034	$\mathbf{0.6843 \pm 0.3005}$	$\mathbf{0.7001 \pm 0.3046}$	0.5818 ± 0.2386
Ours_enc	$\mathbf{0.7747 \pm 0.0950}$	$\mathbf{0.8374 \pm 0.0951}$	$\mathbf{0.6843 \pm 0.3024}$	$\mathbf{0.7001 \pm 0.3067}$	$\mathbf{0.6036 \pm 0.2435}$

Notably, both variations of xLSTM-UNet show superior performance across all datasets. Specifically, xLSTM-UNet_Enc demonstrates the highest performance with a DSC of 0.7747 and an NSD of 0.8374 on the Abdomen MRI dataset, outperforming the previous state-of-the-art (SOTA) model, U-Mamba by a significant margin. Additionally, xLSTM-UNet_Bot achieves DSC and NSD scores of 0.7636 and 0.8322, respectively, surpassing the similarly structured U-Mamba_Bot. Similarly, on the Endoscopy dataset, both xLSTM-UNet_Bot and xLSTM-UNet_Enc achieve the best DSC and NSD scores of 0.6843 and 0.7001, respectively. For the Microscopy dataset, xLSTM-UNet_Enc and xLSTM-UNet_Bot achieve F1 scores of 0.6036 and 0.5818, respectively, both surpassing the previous SOTA results, indicating their robustness in cell segmentation tasks.

The visualized segmentation examples of 2D medical images further illustrate the effectiveness of xLSTM-UNet. As shown in the Figure 2, xLSTM-UNet is more robust to heterogeneous appearances and exhibits fewer segmentation outliers compared to other models. This visual evidence underscores the quantitative

results, highlighting the superior performance and reliability of xLSTM-UNet in diverse medical image segmentation tasks.
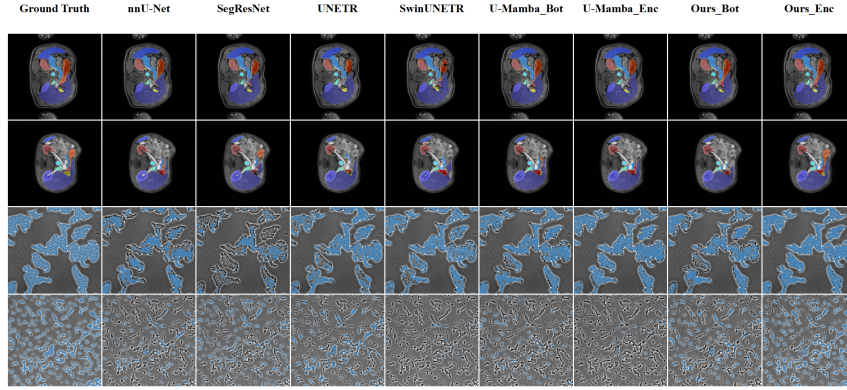

Fig. 2. Visualized examples of 2D medical segmentation. xLSTM-UNet demonstrates greater robustness to heterogeneous appearances and exhibits fewer segmentation errors.

3.5   Quantitative and Qualitative Results for 3D Segmentation

3D medical image segmentation is generally more challenging compared to its 2D counterpart, as it involves processing a larger amount of information. The increased dimensionality leads to a dramatic surge in computational complexity, with the resolution increase causing a cubic rise in the number of computations. Accurate spatial relationship modeling is also essential for achieving satisfactory segmentation results. These factors make the xLSTM-based building blocks, with their computational efficiency, well-suited for this task.

We conducted evaluations on the 3D segmentation dataset in BraTS2023 and Abdomen MRI 3D. Table 2 shows the performance comparison on the BraTS2023 dataset, including metrics for whole tumor (WT), tumor core (TC), and enhancing tumor (ET) regions. Our proposed method demonstrates superior performance across all evaluated metrics, including Dice and HD95, compared to other baseline methods such as SegresNet, UX-Net, MedNeXt, UNETR, SwinUNETR, SwinUNETR-V2, and SegFormer. Specifically, our method achieves the highest average Dice score of 91.80 and the lowest HD95 score of 4.03, highlighting its effectiveness in accurately segmenting brain tumor regions.

The ability of xLSTM to effectively model semantic information in complex spatial domains has been a key factor in our success. The superior experimental results also underscores the suitability of xLSTM-based building blocks for tackling the challenge of semantic segmentation in complex imaging applications.

Table 2. Performance Comparison on BraTS2023 Dataset

Methods	WT		TC		ET		Avg	
	Dice↑	HD95↓	Dice↑	HD95↓	Dice↑	HD95↓	Dice↑	HD95↓
SegresNet	92.02	4.07	89.10	4.08	83.66	3.88	88.26	4.01
UX-Net	93.13	4.56	90.03	5.68	85.91	4.19	89.69	4.81
MedNeXt	92.41	4.98	87.75	4.67	83.96	4.51	88.04	4.72
UNETR	92.19	6.17	86.39	5.29	84.48	5.03	87.68	5.49
SwinUNETR	92.71	5.22	87.79	4.42	84.21	4.48	88.23	4.70
SwinUNETR-V2	93.35	5.01	89.65	4.41	85.17	4.41	89.39	4.51
SegFormer	93.61	**3.37**	92.65	**3.85**	87.71	**3.48**	91.32	**3.56**
Ours	**93.84**	3.89	**92.77**	4.06	**88.79**	4.14	**91.80**	4.03

Table 3 details the performance comparison on the Organs in Abdomen MRI 3D dataset. Our proposed method, xLSTM-UNet_Bot, achieves the highest Dice score of 0.8483 and the best NSD score of 0.9153, surpassing other methods such as nnU-Net, SegResNet, UNETR, SwinUNETR, and U-Mamba_Bot. This demonstrates the robustness and accuracy of our approach in segmenting abdominal organs in MRI images.

Table 3. Performance Comparison on Organs in Abdomen MRI 3D Dataset

Methods	DSC ↑	NSD ↑
nnU-Net	0.8309 ± 0.0769	0.8996 ± 0.0729
SegResNet	0.8146 ± 0.0959	0.8841 ± 0.0917
UNETR	0.6867 ± 0.1488	0.7440 ± 0.1627
SwinUNETR	0.7565 ± 0.1394	0.8218 ± 0.1409
U-Mamba_Bot	0.8453 ± 0.0673	0.9121 ± 0.0634
xLSTM-UNet_Bot	**0.8483 ± 0.0774**	**0.9153 ± 0.0596**

4   Discussion

This study demonstrates that xLSTM, a model with linear computational complexity, can be an effective components in image segmentation networks. Our experimental results clearly show that xLSTM-UNet outperforms Mamba-based counterparts, underscoring the promising future of xLSTM. Given the recent huge interest in Mamba in academia, we believe that it is important to also

recognize and investigate the potential of xLSTM, which has shown remarkable efficacy in this domain.

Meanwhile, medical image segmentation is inherently challenging. General image segmentation foundation models, such as Segment Anything, often fail when applied to medical images [59,60,61]. This highlights the need for ongoing development of specialized medical image segmentation algorithms. Furthermore, medical imaging datasets are typically small. In this study, the datasets used were limited in size, which restricts our ability to explore the effects of different network scales and dataset sizes on segmentation outcomes. Investigating whether xLSTM-driven image algorithms adhere to scaling laws remains an open question for future research. Additionally, we need to examine the efficiency of xLSTM, including inference speed and memory usage relative to segmentation performance, to better understand its practical applications.

Currently, xLSTM lacks dedicated optimization for hardware such as NVIDIA GPUs, which presents an opportunity for the community to contribute. Collaborative efforts are essential to optimize xLSTM for various vision tasks, leveraging its full potential. By releasing our code, we aim to encourage and facilitate further research and development, enabling the community to build on our initial findings and drive progress in this area.

This research represents an initial exploration into the application of xLSTM in medical image segmentation. There are many peaks to climb and numerous scenarios to test in this field. We hope that our comprehensive experiments and tests will demonstrate the significant potential of xLSTM in practical applications, encouraging scholars to continue exploring this promising model. With further development and optimization, we envision xLSTM achieving success comparable to that of Mamba and even Transformers, becoming a cornerstone in image segmentation and beyond.

5   Conclusion

In this report, we have introduced xLSTM-UNet, the first U-Net architecture enhanced with Extended Long-Short Term Memory (xLSTM) for both 2D and 3D medical image segmentation tasks. Through extensive experiments across a variety of medical imaging scenarios—including abdominal MRI, endoscopy, microscopy, and brain MRI—we have demonstrated that xLSTM-UNet significantly outperforms existing CNN-based and Transformer-based methods, as well as its Mamba-based counterparts. These findings underscore the effectiveness of xLSTM in handling complex segmentation tasks, particularly in the challenging domain of 3D medical image segmentation.

Our results show that the xLSTM-based architecture can achieve state-of-the-art (SOTA) performance, offering enhanced accuracy and efficiency. This marks a significant advancement in the field of medical image segmentation, with potential applications extending beyond healthcare.

References

1. J. Ma and B. Wang, "Towards foundation models of biological image segmentation," *Nature Methods*, vol. 20, no. 7, pp. 953–955, 2023.
2. J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang, "Segment anything in medical images," *Nature Communications*, vol. 15, no. 1, p. 654, 2024.
3. P. Bilic, P. Christ, H. B. Li, E. Vorontsov, A. Ben-Cohen, G. Kaissis, A. Szeskin, C. Jacobs, G. E. H. Mamani, G. Chartrand *et al.*, "The liver tumor segmentation benchmark (lits)," *Medical Image Analysis*, vol. 84, p. 102680, 2023.
4. N. Heller, F. Isensee, K. H. Maier-Hein, X. Hou, C. Xie, F. Li, Y. Nan, G. Mu, Z. Lin, M. Han *et al.*, "The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge," *Medical image analysis*, vol. 67, p. 101821, 2021.
5. J. Ma, Y. Zhang, S. Gu, C. Ge, S. Ma, A. Young, C. Zhu, K. Meng, X. Yang, Z. Huang *et al.*, "Unleashing the strengths of unlabeled data in pan-cancer abdominal organ quantification: the flare22 challenge," *arXiv preprint arXiv:2308.05862*, 2023.
6. C. Stringer, T. Wang, M. Michaelos, and M. Pachitariu, "Cellpose: a generalist algorithm for cellular segmentation," *Nature methods*, vol. 18, no. 1, pp. 100–106, 2021.
7. E. Meijering, "Cell segmentation: 50 years down the road [life sciences]," *IEEE signal processing magazine*, vol. 29, no. 5, pp. 140–145, 2012.
8. J. Ma, R. Xie, S. Ayyadhury, C. Ge, A. Gupta, R. Gupta, S. Gu, Y. Zhang, G. Lee, J. Kim *et al.*, "The multimodality cell segmentation challenge: toward universal solutions," *Nature methods*, pp. 1–11, 2024.
9. A. Gupta, S. Gehlot, S. Goswami, S. Motwani, R. Gupta, A. G. Faura, D. Štepec, T. Martinčič, R. Azad, D. Merhof *et al.*, "Segpc-2021: A challenge & dataset on segmentation of multiple myeloma plasma cells from microscopic images," *Medical Image Analysis*, vol. 83, p. 102677, 2023.
10. O. Ronneberger, P. Fischer, and T. 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.* Springer, 2015, pp. 234–241.
11. F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV).* Ieee, 2016, pp. 565–571.
12. Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4.* Springer, 2018, pp. 3–11.
13. O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz *et al.*, "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.
14. H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y.-W. Chen, and J. Wu, "Unet 3+: A full-scale connected unet for medical image segmentation," in *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP).* IEEE, 2020, pp. 1055–1059.

15. J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

16. Y. Wang, Z. Li, J. Mei, Z. Wei, L. Liu, C. Wang, S. Sang, A. L. Yuille, C. Xie, and Y. Zhou, "Swinmm: masked multi-view with swin transformers for 3d medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*.   Springer, 2023, pp. 486–496.

17. A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu, "Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images," in *International MICCAI brainlesion workshop*.   Springer, 2021, pp. 272–284.

18. A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu, "Unetr: Transformers for 3d medical image segmentation," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2022, pp. 574–584.

19. A. Lin, B. Chen, J. Xu, Z. Zhang, G. Lu, and D. Zhang, "Ds-transunet: Dual swin transformer u-net for medical image segmentation," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–15, 2022.

20. A. M. Shaker, M. Maaz, H. Rasheed, S. Khan, M.-H. Yang, and F. S. Khan, "Unetr++: delving into efficient and accurate 3d medical image segmentation," *IEEE Transactions on Medical Imaging*, 2024.

21. F. Shamshad, S. Khan, S. W. Zamir, M. H. Khan, M. Hayat, F. S. Khan, and H. Fu, "Transformers in medical imaging: A survey," *Medical Image Analysis*, vol. 88, p. 102802, 2023.

22. R. Azad, A. Kazerouni, M. Heidari, E. K. Aghdam, A. Molaei, Y. Jia, A. Jose, R. Roy, and D. Merhof, "Advances in medical image analysis with vision transformers: a comprehensive review," *Medical Image Analysis*, p. 103000, 2023.

23. C. You, R. Zhao, F. Liu, S. Dong, S. Chinchali, U. Topcu, L. Staib, and J. Duncan, "Class-aware adversarial transformers for medical image segmentation," *Advances in Neural Information Processing Systems*, vol. 35, pp. 29 582–29 596, 2022.

24. X. Yu, Q. Yang, Y. Zhou, L. Y. Cai, R. Gao, H. H. Lee, T. Li, S. Bao, Z. Xu, T. A. Lasko *et al.*, "Unest: local spatial representation learning with hierarchical transformer for efficient medical segmentation," *Medical Image Analysis*, vol. 90, p. 102939, 2023.

25. Y. Pang, J. Liang, T. Huang, H. Chen, Y. Li, D. Li, L. Huang, and Q. Wang, "Slim unetr: Scale hybrid transformers to efficient 3d medical image segmentation under limited computational resources," *IEEE Transactions on Medical Imaging*, 2023.

26. F. Yuan, Z. Zhang, and Z. Fang, "An effective cnn and transformer complementary network for medical image segmentation," *Pattern Recognition*, vol. 136, p. 109228, 2023.

27. A. He, K. Wang, T. Li, C. Du, S. Xia, and H. Fu, "H2former: An efficient hierarchical hybrid transformer for medical image segmentation," *IEEE Transactions on Medical Imaging*, vol. 42, no. 9, pp. 2763–2775, 2023.

28. Y. Xie, J. Zhang, C. Shen, and Y. Xia, "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*.   Springer, 2021, pp. 171–180.

29. D. Zhang, M. N. Slipchenko, and J.-X. Cheng, "Highly sensitive vibrational imaging by femtosecond pulse stimulated raman loss," *The journal of physical chemistry letters*, vol. 2, no. 11, pp. 1248–1253, 2011.

30. D. Zhang, P. Wang, M. N. Slipchenko, and J.-X. Cheng, "Fast vibrational imaging of single cells and tissues by stimulated raman scattering microscopy," *Accounts of chemical research*, vol. 47, no. 8, pp. 2282–2290, 2014.

31. C. Zhang, D. Zhang, and J.-X. Cheng, "Coherent raman scattering microscopy in biology and medicine," *Annual review of biomedical engineering*, vol. 17, no. 1, pp. 415–445, 2015.

32. T. Chen, Y. Zang, J. Zhang, and D. Zhang, "High-quality label-free prediction of fluorescence images through stimulated raman scattering imaging and recurrent deep neural network," in *Medical Imaging 2023: Image Perception, Observer Performance, and Technology Assessment*, vol. 12467.   SPIE, 2023, pp. 348–352.

33. P. Fu, T. Chen, H. Zhu, C. Xu, M. Chen, and D. Zhang, "Pushing the resolution of infrared imaging by mid infrared photothermal microscopy," in *Advanced Optical Imaging Technologies III*, vol. 11549.   SPIE, 2020, pp. 68–74.

34. P. Fu, W. Cao, T. Chen, X. Huang, T. Le, S. Zhu, D.-W. Wang, H. J. Lee, and D. Zhang, "Super-resolution imaging of non-fluorescent molecules by photothermal relaxation localization microscopy," *Nature Photonics*, vol. 17, no. 4, pp. 330–337, 2023.

35. D. Zhang, C. Li, C. Zhang, M. N. Slipchenko, G. Eakins, and J.-X. Cheng, "Depth-resolved mid-infrared photothermal imaging of living cells and organisms with submicrometer spatial resolution," *Science advances*, vol. 2, no. 9, p. e1600521, 2016.

36. R. E. Kalman, "A new approach to linear filtering and prediction problems," 1960.

37. A. Gu, I. Johnson, K. Goel, K. Saab, T. Dao, A. Rudra, and C. Ré, "Combining recurrent, convolutional, and continuous-time models with linear state space layers," *Advances in neural information processing systems*, vol. 34, pp. 572–585, 2021.

38. A. Gu, *Modeling Sequences with Structured State Spaces*.   Stanford University, 2023.

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

40. J. Ma, F. Li, and B. Wang, "U-mamba: Enhancing long-range dependency for biomedical image segmentation," *arXiv preprint arXiv:2401.04722*, 2024.

41. J. Ruan and S. Xiang, "Vm-unet: Vision mamba unet for medical image segmentation," *arXiv preprint arXiv:2402.02491*, 2024.

42. M. Zhang, Y. Yu, L. Gu, T. Lin, and X. Tao, "Vm-unet-v2 rethinking vision mamba unet for medical image segmentation," *arXiv preprint arXiv:2403.09157*, 2024.

43. R. Wu, Y. Liu, P. Liang, and Q. Chang, "Ultralight vm-unet: Parallel vision mamba significantly reduces parameters for skin lesion segmentation," *arXiv preprint arXiv:2403.20035*, 2024.

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

45. J. Liu, H. Yang, H.-Y. Zhou, Y. Xi, L. Yu, Y. Yu, Y. Liang, G. Shi, S. Zhang, H. Zheng *et al.*, "Swin-umamba: Mamba-based unet with imagenet-based pretraining," *arXiv preprint arXiv:2402.03302*, 2024.

46. Z. Xing, T. Ye, Y. Yang, G. Liu, and L. Zhu, "Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation," *arXiv preprint arXiv:2401.13560*, 2024.

47. M. Beck, K. Pöppel, M. Spanring, A. Auer, O. Prudnikova, M. Kopp, G. Klambauer, J. Brandstetter, and S. Hochreiter, "xlstm: Extended long short-term memory," *arXiv preprint arXiv:2405.04517*, 2024.

48. B. Alkin, M. Beck, K. Pöppel, S. Hochreiter, and J. Brandstetter, "Vision-lstm: xlstm as generic vision backbone," *arXiv preprint arXiv:2406.04303*, 2024.

49. P. A. Yushkevich, Y. Gao, and G. Gerig, "Itk-snap: An interactive tool for semi-automatic segmentation of multi-modality biomedical images," in *2016 38th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*.   IEEE, 2016, pp. 3342–3345.

50. M. Allan, A. Shvets, T. Kurmann, Z. Zhang, R. Duggal, Y.-H. Su, N. Rieke, I. Laina, N. Kalavakonda, S. Bodenstedt *et al.*, "2017 robotic instrument segmentation challenge," *arXiv preprint arXiv:1902.06426*, 2019.

51. B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest *et al.*, "The multimodal brain tumor image segmentation benchmark (brats)," *IEEE transactions on medical imaging*, vol. 34, no. 10, pp. 1993–2024, 2014.

52. A. F. Kazerooni, N. Khalili, X. Liu, D. Haldar, Z. Jiang, S. M. Anwar, J. Albrecht, M. Adewole, U. Anazodo, H. Anderson *et al.*, "The brain tumor segmentation (brats) challenge 2023: Focus on pediatrics (cbtn-connect-dipgr-asnr-miccai brats-peds)," *arXiv preprint arXiv:2305.17033*, 2023.

53. S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. S. Kirby, J. B. Freymann, K. Farahani, and C. Davatzikos, "Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features," *Scientific data*, vol. 4, no. 1, pp. 1–13, 2017.

54. A. Myronenko, "3d mri brain tumor segmentation using autoencoder regularization," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II 4*.   Springer, 2019, pp. 311–320.

55. L. Maier-Hein, A. Reinke, P. Godau, M. D. Tizabi, F. Buettner, E. Christodoulou, B. Glocker, F. Isensee, J. Kleesiek, M. Kozubek *et al.*, "Metrics reloaded: recommendations for image analysis validation," *Nature methods*, vol. 21, no. 2, pp. 195–212, 2024.

56. H. H. Lee, S. Bao, Y. Huo, and B. A. Landman, "3d ux-net: A large kernel volumetric convnet modernizing hierarchical transformer for medical image segmentation," *arXiv preprint arXiv:2209.15076*, 2022.

57. S. Roy, G. Koehler, C. Ulrich, M. Baumgartner, J. Petersen, F. Isensee, P. F. Jaeger, and K. H. Maier-Hein, "Mednext: transformer-driven scaling of convnets for medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*.   Springer, 2023, pp. 405–415.

58. Y. He, V. Nath, D. Yang, Y. Tang, A. Myronenko, and D. Xu, "Swinunetr-v2: Stronger swin transformers with stagewise convolutions for 3d medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*.   Springer, 2023, pp. 416–426.

59. T. Chen, L. Zhu, C. Deng, R. Cao, Y. Wang, S. Zhang, Z. Li, L. Sun, Y. Zang, and P. Mao, "Sam-adapter: Adapting segment anything in underperformed scenes," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3367–3375.

60. T. Chen, L. Zhu, C. Ding, R. Cao, Y. Wang, Z. Li, L. Sun, P. Mao, and Y. Zang, "Sam fails to segment anything?–sam-adapter: Adapting sam in underperformed scenes: Camouflage, shadow, medical image segmentation, and more," *arXiv preprint arXiv:2304.09148*, 2023.

61. A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4015–4026.