2024 IEEE International Symposium on Biomedical Imaging (ISBI) | 979-8-3503-1333-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/ISBI56570.2024.10635844

SAM3D: SEGMENT ANYTHING MODEL IN VOLUMETRIC MEDICAL IMAGES

Nhat-Tan Bui[1*], *Dinh-Hieu Hoang*[2,3*], *Minh-Triet Tran*[2,3], *Gianfranco Doretto*[4]
Donald Adjeroh[4], *Brijesh Patel*[4], *Arabinda Choudhary*[5], *Ngan Le*[1]

[1]AICV Lab, University of Arkansas, Arkansas, USA
[2]University of Science, Vietnam National University, Ho Chi Minh City, Vietnam
[3]John von Neumann Institute, Vietnam National University, Ho Chi Minh City, Vietnam
[4]West Virginia University, West Virginia, USA
[5]University of Arkansas for Medical Sciences, Arkansas, USA

ABSTRACT

Image segmentation remains a pivotal component in medical image analysis, aiding in the extraction of critical information for precise diagnostic practices. With the advent of deep learning, automated image segmentation methods have risen to prominence, showcasing exceptional proficiency in processing medical imagery. Motivated by the Segment Anything Model (SAM)—a foundational model renowned for its remarkable precision and robust generalization capabilities in segmenting 2D natural images—we introduce SAM3D, an innovative adaptation tailored for 3D volumetric medical image analysis. Unlike current SAM-based methods that segment volumetric data by converting the volume into separate 2D slices for individual analysis, our SAM3D model processes the entire 3D volume image in a unified approach. Extensive experiments are conducted on multiple medical image datasets to demonstrate that our network attains competitive results compared with other state-of-the-art methods in 3D medical segmentation tasks while being significantly efficient in terms of parameters. Code and checkpoints are available at https://github.com/UARK-AICV/SAM3D.

Index Terms— 3D Medical Segmentation, Foundation Model, Transfer Learning, Segment Anything Model

1. INTRODUCTION

Volumetric segmentation is crucial in medical image analysis, finding applications in pathology diagnosis, surgical planning, and computer-aided diagnosis. Volumetric medical images like CT, MRI, OCT, and DBT offer a 3D view of anatomical structures. Segmentation identifies regions of interest for better interpretation.

Deep learning, particularly UNet [1] and variants [2, 3, 4], made strides in 3D medical segmentation but faced limitations. Transformer-based models like Vision Transformer (ViT) [5] and Swin-UNet [4] showed promise in capturing long-range relationships. Combining CNNs and Transformers in models like TransUNet [3], UNETR [6], and HiFormer [7], yielded promising results. However, these models prioritize precision, leading to increased complexity and training time. Leveraging pretrained models offers an alternative. SAM, a transformer-based model pretrained on large-scale datasets, has shown generalizability in segmentation tasks. SAM-based models for medical images piqued interest.

This work introduces **SAM3D**, an architecture for volumetric medical segmentation, combining the SAM encoder and a lightweight 3D CNN decoder. Unlike traditional slice-by-slice processing, SAM3D extracts features across the entire volume, improving segmentation while maintaining simplicity and computational efficiency. Contributions include applying the SAM encoder to process 3D volumes, designing SAM3D for effective 3D medical segmentation, and validating its performance on various datasets, such as ACDC [8], Synapse [9], MSD BraTS [10], and MSD Lung [10]. SAM3D demonstrates competitive results, marking a novel approach to 3D volumetric imaging.

2. METHODOLOGY

In this section, we introduce our model, SAM3D, and explain the rationale behind its simple design. Our goal is to leverage SAM without the need for extensive parameter retraining or complex task-specific modules.

Overall Architecture. SAM was trained on an extensive dataset comprising 1 million images and 1.1 billion masks, and it features a robust image encoder tailored for natural images. However, applying SAM directly to 3D medical images poses challenges due to inherent domain differences. We posit that the SAM image encoder retains valuable low-level features, e.g. edges and boundaries, which have relevance across various image domains.

In contrast to SAMed [11] and MedSAM [12], where all three components of SAM are fine-tuned, our approach involves freezing SAM's image encoder and training a new lightweight 3D decoder. SAM3D leverages SAM by initially

[*]Equal contribution

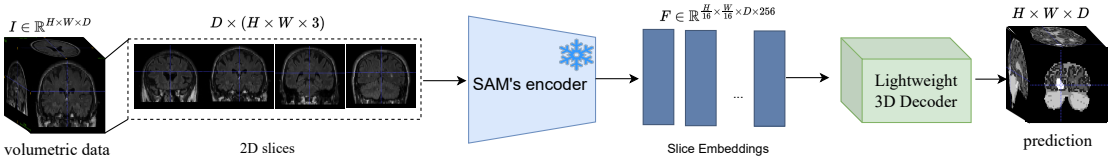

Fig. 1: Overall architecture of the proposed SAM3D. Given a volumetric image $I \in \mathbb{R}^{H\times W\times D}$, SAM3D initially applies SAM to process each of the D slices individually, producing slice embeddings denoted as $F \in \mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times D\times 256}$. These embeddings are then decoded by a lightweight 3D decoder, ultimately yielding the segmentation prediction.

processing images slice by slice and then incorporating a lightweight 3D decoder to capture depth-wise relationships between slices. The overall architecture of SAM3D is depicted in Figure 1 and can be summarized as follows: a volumetric input $I \in \mathbb{R}^{H\times W\times D}$ is divided into D 2D slices, each of dimension $H \times W$. We duplicate each channel three times to generate the slices that have dimension of $H\times W\times 3$. The pretrained SAM encoder processes these slices, generating 3D slice embeddings denoted as F. The depth-wise relationships among these slice embeddings are effectively captured by our proposed 3D decoder. Additionally, we remove the prompt encoder from SAM to ensure that feature extraction remains uninhibited across different modalities.

Encoder. SAM's image encoder extracts robust low-level information. Thus, it is plausible to tackle the notorious weak boundary in the medical image domain by using features extracted by SAM's image encoder. Formally, let $I \in \mathbb{R}^{H\times W\times D}$ be the input, and Enc represent the slice encoder. We split I into D slices I_i along the depth dimension, each slice is in $3 \times H \times W$, and feed them into Enc. The output slice embeddings are stacked and transposed to obtain the final 3D slice embeddings $F = [f_i]_{i=1}^{D}$.

$$f_i = Enc(I_i), \text{where } f_i \in \mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times 256} \quad (1)$$

We stack these slice embeddings and transpose the result to obtain the final 3D slice embedding, $F = [f_i]_{i=1}^{D}, F \in \mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times D\times 256}$.

Decoder. Because our decoder must handle 3D volumetric data, we cannot utilize SAM's mask decoder, which is specifically designed for 2D natural images. Instead, we propose the development of an appropriate 3D decoder. However, creating a 3D network with the Vision Transformer [5] and its variants can be resource-intensive, requiring significant computational power and increasing inference time, especially when dealing with a large value of D. Therefore, we suggest the design of a lightweight 3D decoder comprising four 3D convolutional blocks with skip connections [13] and a segmentation head, as elaborated in Figure 2.

Objective Function. We train our SAM3D network with a combination loss of both the dice loss and cross-entropy loss. The formulation is as follows:

$$\mathcal{L}(Y,\hat{Y}) = -\sum_{n=1}^{N}\sum_{k=1}^{K}\left(\frac{2\times Y_{k,n}\hat{Y}_{k,n}}{Y_{k,n}^2+\hat{Y}_{k,n}^2} + Y_{k,n}log\hat{Y}_{k,n}\right) \quad (2)$$

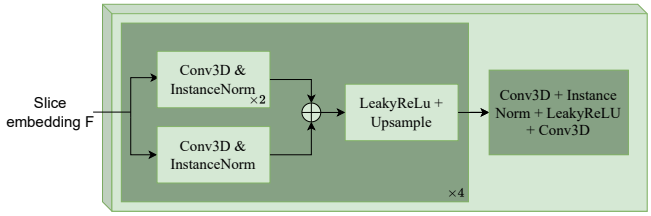

Fig. 2: Architecture of the proposed lightweight 3D decoder.

here, Y is the predicted segmenting result from SAM3D, and $\hat{Y}$ is the ground truth. N represents the number of classes, K denotes the number of voxels, and $Y_{k,n}$ and $\hat{Y}_{k,n}$ refer to the predictions and the ground truths at voxel j for class i, respectively.

Additionally, we employ the deep supervision technique for multiple decoding stages. Specifically, the output features of each decoding stage pass through a segmentation block, consisting of one 3 x 3 x 3 and one 1 x 1 x 1 convolution layer, to generate predictions for one typical stage. To calculate the loss value for one typical stage, we down-sample the ground truth to match the prediction resolution. Consequently, the final loss can be defined as $\mathcal{L}_{total} = \sum_{l=1}^{L}\alpha_l \times \mathcal{L}_l$, where L is set to 3, representing the number of decoder layers. α_l signifies the hyperparameter controlling the contribution of different resolutions to the final loss function. In practice, we set $\alpha_2 = \frac{\alpha_1}{2}$ and $\alpha_3 = \frac{\alpha_1}{4}$ with all α is normalized to 1.

3. EXPERIMENTS

A. Datasets: We conduct the experiments on four datasets: Multi-organ CT Segmentation (Synapse) [9], Automated Cardiac Diagnosis (ACDC) [8], Brain Tumor Segmentation (BraTS) [10], and Lung Tumor Segmentation (Lung) [10]. BraTS and Lung come from the Medical Segmentation Decathlon challenge (MSD) [10]. For a fair comparison, we follow the data splitting of previous works, e.g. nnFormer [16] and UNETR++ [17].

B. Implementation Details. Our model is implemented based on Python 3.8.10 with PyTorch library and trained on a single NVIDIA RTX 2080 Ti GPU with 11GB memory. We use **ViT-B** version as our backbone for the SAM's image encoder due to the limited resources. Instead of exhaustively finding an overfitting training procedure, we trained our model with the general training strategy of nnFormer [16] and UNETR++ [17], the stochastic gradient descent (SGD) with a momentum of 0.99 and a weight

Table 1: Quantitative results on Synapse dataset.

SAM	Networks	Methods	Params ↓	Average		DSC on individual abdominal organs							
				HD ↓	DSC ↑	RKid	LKid	Spl	Gal	Sto	Pan	Aor	Liv
✗	2D	TransUNet [3]	96.07M	31.69	77.49	77.02	81.87	85.08	63.16	75.62	55.86	87.23	94.08
		Swin-Unet [4]	27.17M	21.55	79.13	79.61	83.28	90.66	66.53	76.60	56.58	85.47	94.29
		TransDeepLab [14]	21.14M	21.25	80.16	79.88	84.08	89.00	69.16	78.40	61.19	86.04	93.53
		HiFormer-S [7]	23.25M	18.85	80.29	64.84	82.39	91.03	73.29	78.07	60.84	85.63	94.22
		HiFormer-B [7]	25.51M	14.70	80.39	79.77	85.23	90.99	65.69	81.08	59.52	86.21	94.61
		HiFormer-L [7]	29.52M	19.14	80.69	78.37	84.23	90.44	68.61	82.03	60.77	87.03	94.07
	3D	MISSFormer [15]	-	18.20	81.96	82.00	85.21	91.92	68.65	80.81	65.67	86.99	94.41
		nnFormer [16]	150.50M	10.63	86.57	86.25	86.57	90.51	70.17	86.83	83.35	92.04	96.84
		UNETR [6]	92.49M	18.59	78.35	84.52	85.60	85.00	56.30	70.46	60.47	89.80	94.57
		UNETR++ [17]	42.95M	7.53	87.22	87.18	87.54	95.77	71.25	86.01	81.10	92.52	96.42
✓	2D	SAMed [11]	18.81M	20.64	81.88	79.95	80.45	88.72	69.11	82.06	72.17	87.77	94.80
		SAMed_s [11]	6.32M	31.72	77.78	78.92	79.63	85.81	57.11	77.49	65.66	83.62	93.98
	3D	**SAM3D (Ours)**	1.88M	17.87	79.56	85.64	86.31	84.29	49.81	76.11	69.32	89.57	95.42

Table 2: Quantitative results on ACDC dataset.

Methods	Params↓	Average	DSC on individual regions		
		DSC ↑	RV	LV	MYO
TransUNet [3]	96.07M	89.71	88.86	84.54	95.73
Swin-Unet [4]	27.17M	90.00	88.55	85.62	95.83
UNETR [6]	92.49M	86.61	85.29	86.52	94.02
MISSFormer [15]	–	87.90	86.36	85.75	91.59
nnFormer [16]	150.5M	92.06	90.94	89.58	95.65
UNETR++ [17]	66.80M	92.83	91.89	90.61	96.00
SAM3D (Ours)	1.88M	90.41	89.44	87.12	94.67

Table 3: Quantitative results on Lung dataset.

Methods	Params ↓	Average DSC ↑
nnUNet [2]	–	74.31
Swin UNETR [18]	62.83M	75.55
nnFormer [16]	150.5M	77.95
UNETR [6]	92.49M	73.29
UNETR++ [17]	121.17M	80.68
SAM3D (Ours)	1.88M	71.42

Table 4: Quantitative results on BraTS dataset.

Methods	Params↓	Average		WT		ET		TC	
		HD ↓	DSC ↑	HD ↓	DSC ↑	HD ↓	DSC ↑	HD ↓	DSC ↑
TransUNet [3]	96.07M	12.98	64.4	14.03	70.6	10.42	54.2	14.50	68.4
UNETR [6]	92.49M	8.82	71.1	8.27	78.9	9.35	58.5	8.85	76.1
nnFormer [16]	150.5M	4.05	86.4	3.80	91.3	3.87	81.8	4.49	86.0
UNETR++ [17]	42.65M	5.85	77.7	4.79	91.2	4.22	78.5	6.78	78.4
SAM3D (Ours)	4.63M	8.72	72.9	6.03	88.0	10.05	69.6	9.79	76.6

decay of 3e-5. The learning rate scheduler is defined as $lr = init\_lr \times (1 - \frac{epoch}{max\_epoch})^{power}$, where $init\_lr$ = 1e-2, $power$ = 0.9, and max_epoch = 1000. One epoch consists of 250 iterations. For ACDC, Synapse, BraTS, and Lung datasets, SAM3D is trained with the 3D volume sizes of 160 x 160 x 14, 176 x 176 x 64, 64 x 64 x 64 and 192 x 192 x 34, respectively. We also utilize the same data augmentation techniques including rotation, scaling, brightness adjustment, gamma augmentation, and mirroring. The batch size is set to 4 for ACDC and 2 for Synapse, BraTS, and Lung.

C. Performance Comparisons.

We compared our SAM3D with recent SOTA methods on both CNNs-based networks, e.g. nnFormer [16] and Transformer-based networks, e.g. TransUNet [3], Swin-Unet [4], TransDeepLab [14], HiFormer [7], MISSFormer [15], UNETR [6] and SAM-based models SAMed and SAMed_s [11]. The performance comparisons are reported in Tables 1, 2, 3, and 4 including both accuracy (i.e. HD95 and DSC metrics) and network complexity (#params).

Synapse comprises eight abdominal organs in a large dataset and the performance comparison is shown in Table 1. Among the models evaluated, UNETR++ (a Transformer-based model) achieved the best results with 42.9M parameters, while nnFormer ranked second with 150.5M parameters. Notably, SAMed_s distinguishes itself by achieving impressive results with a modest 6.32M parameters and a DSC of 77.78%. SAMed_s shares a similar architecture with our SAM3D, fine-tuned from SAM, but differs in processing methods. SAMed_s employs a straightforward slice-by-slice approach, while SAM3D considers depth-wise information. Despite this difference, both models are efficient in parameter usage. SAMed_s requires 6.32M parameters, whereas SAM3D excels with just 1.88M parameters. Furthermore, SAM3D achieves a DSC score exceeding 1.78%, demonstrating superior performance compared to SAM-based methods with lightweight models.

While SAMed is exclusive to the Synapse dataset, our SAM3D can be evaluated on a variety of other datasets, including Cardiac, Brain Tumor, and Lung. In Table 2 and 3, it is evident that SAM3D competes favorably with SOTA CNNs/Transformer-based networks on the Cardiac ACDC and Lung datasets. For instance, SAM3D surpasses TransUnet's performance on the ACDC dataset with a 0.41% increase in DSC while utilizing less than 50× the number of parameters. Table 4 further illustrates SAM3D's competitiveness with other leading models on the Brain Tumor Brats dataset, despite its significantly lower parameter count. For example, SAM3D achieves a 1.8% DSC improvement compared to UNETR, while requiring less than 20× the number of params. It is worth noting that the MRI scans in Brats contain four modalities, which explains SAM3D's parameter count being four times that of other single-modality models.

D. Ablation Study.

To assess the impact of skip connections in our proposed lightweight 3D decoder, we conducted an ablation study on ACDC and Synapse datasets as depicted in Table 5. The results clearly indicate that these skip connections contribute positively to the model's performance, resulting in improvement. We believe that these skip connections play a crucial role in preserving information related to edges and boundaries from lower-level features, enhancing the precision of the segmentation process.

Table 5: Ablation study of the skip connection in our lightweight 3D decoder on ACDC and Synapse datasets.

(a) ACDC dataset.

Settings	Average	DSC on individual regions		
	DSC ↑	RV	LV	MYO
w/o skip connection	89.73	88.46	94.41	86.32
w skip connection	90.41	89.44	94.67	87.12

(b) Synapse dataset.

Settings	Average		DSC on individual abdominal organs							
	HD↓	DSC↑	RKid	LKid	Spl	Gal	Sto	Pan	Aor	Liv
w/o skip connection	25.87	79.33	84.68	85.20	85.26	50.55	75.07	68.83	90.10	94.98
w skip connection	17.87	79.56	85.64	86.31	84.29	49.81	76.11	69.32	89.57	95.42

4. CONCLUSION

In this study, we introduce SAM3D, an efficient and simple SAM-based model tailored for volumetric medical image segmentation. Our approach harnesses the capabilities of a SAM pre-trained encoder coupled with a lightweight 3D decoder. Through extensive experimentation, we have established that SAM3D competes effectively with current SOTA 3D neural networks and Transformer-based models while demanding significantly fewer parameters (50× fewer). Furthermore, SAM3D outperforms other lightweight networks in the context of volumetric segmentation. As SAM has already made a substantial impact on natural image segmentation, our research extends its potential to the domain of medical image segmentation. We anticipate that this work will serve as an inspiration for future researchers, fostering advancements in the field of medical segmentation

Discussion. In our experiments, we employed the smallest SAM variant, which utilizes ViT-B backbone, primarily due to resource and time constraints. We hypothesize that ViT-L and ViT-H pre-trained models may yield even more remarkable results. Consequently, we encourage researchers to explore these options for our segmentation task.

Additionally, our simple decoder leaves room for developing a more complex architecture, which could potentially enhance the model's performance. This presents a promising avenue for further research and development.

Acknowledgement: Nhat-Tan Bui, Thinh Phan, and Ngan Le are supported by the National Science Foundation (NSF) under Award No OIA-1946391 RII Track-1, NSF 1920920 RII Track 2 FEC, NSF 2223793 EFRI BRAID, NSF 2119691 AI SUSTEIN, NSF 2236302. Minh-Triet Tran is sponsored by VNU-HCM (DS2020-42-01). Dinh-Hieu Hoang is funded by VINIF, VINBIGDATA (code VINIF.2022.ThS.JVN.04).

5. REFERENCES

[1] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *MICCAI*, 2015. 1

[2] Fabian Isensee, Paul Jaeger, et al., "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, 2021. 1, 3

[3] Jieneng Chen, Yongyi Lu, et al., "TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation," *arXiv preprint arXiv:2102.04306*, 2021. 1, 3

[4] Hu Cao, Yueyue Wang, et al., "Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation," in *ECCVW*, 2022. 1, 3

[5] Alexey Dosovitskiy, Lucas Beyer, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020. 1, 2

[6] 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 *WACV*, 2022. 1, 3

[7] Moein Heidari, Amirhossein Kazerouni, et al., "Hiformer: Hierarchical multi-scale representations using transformers for medical image segmentation," in *WACV*, 2023. 1, 3

[8] Olivier Bernard, Alain Lalande, et al., "Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: Is the problem solved?," *IEEE TMI*, 2018. 1, 2

[9] Bennett Landman, Zhoubing Xu, et al., "Multi-Atlas Labeling Beyond the Cranial Vault - Workshop and Challenge," in *MICCAI-W*, 2015. 1, 2

[10] Amber L. Simpson, Michela Antonelli, et al., "A large annotated medical image dataset for the development and evaluation of segmentation algorithms," *arXiv preprint arXiv:1902.09063*, 2019. 1, 2

[11] Kaidong Zhang and Dong Liu, "Customized Segment Anything Model for Medical Image Segmentation," *arXiv preprint arXiv:2304.13785*, 2023. 1, 3

[12] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang, "Segment anything in medical images," *arXiv preprint arXiv:2304.12306*, 2023. 1

[13] Kaiming He, Xiangyu Zhang, et al., "Deep Residual Learning for Image Recognition," in *CVPR*, 2016. 2

[14] Reza Azad, Moein Heidari, et al., "TransDeepLab: Convolution-Free Transformer-based DeepLab v3+ for Medical Image Segmentation," in *PRIME*, 2022. 3

[15] Xiaohong Huang, Zhifang Deng, et al., "MISSFormer: An Effective Medical Image Segmentation Transformer," *arXiv preprint arXiv:2109.07162*, 2021. 3

[16] Hong-Yu Zhou, Jiansen Guo, Zhang Yinghao, Lequan Yu, Liansheng Wang, and Yizhou Yu, "nnFormer: Interleaved Transformer for Volumetric Segmentation," *arXiv preprint arXiv:2109.03201*, 2021. 2, 3

[17] Abdelrahman Shaker, Muhammad Maaz, et al., "UNETR++: Delving into Efficient and Accurate 3D Medical Image Segmentation," *arXiv:2212.04497*, 2022. 2, 3

[18] Ali Hatamizadeh, Vishwesh Nath, et al., "Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images," in *MICCAI-W*, 2021, pp. 272–284. 3