3DSAM-adapter: Holistic Adaptation of SAM from 2D to 3D for Promptable Medical Image Segmentation

Yu Jiening
Yu Jiening@umac.mo

um 澳大

- arXiv, 23 Jun 2023,

3DSAM-adapter: Holistic Adaptation of SAM from 2D to 3D for Promptable Medical Image Segmentation

Shizhan Gong, Yuan Zhong, Wenao Ma, Jinpeng Li, Zhao Wang, Jingyang Zhang, Pheng-Ann Heng, Qi Dou
Department of Computer Science and Engineering
The Chinese University of Hong Kong

Abstract

Despite that the segment anything model (SAM) achieved impressive results on general-purpose semantic segmentation with strong generalization ability on daily images, its demonstrated performance on medical image segmentation is less precise and not stable, especially when dealing with tumor segmentation tasks that involve objects of small sizes, irregular shapes, and low contrast. Notably, the original SAM architecture is designed for 2D natural images, therefore would not be able to extract the 3D spatial information from volumetric medical data effectively. In this paper, we propose a novel adaptation method for transferring SAM from 2D to 3D for promptable medical image segmentation. Through a holistically designed scheme for architecture modification, we transfer the SAM to support volumetric inputs while retaining the majority of its pre-trained parameters for reuse. The fine-tuning process is conducted in a parameter-efficient manner, wherein most of the pre-trained parameters remain frozen, and only a few lightweight spatial adapters are introduced and tuned. Regardless of the domain gap between natural and medical data and the disparity in the spatial arrangement between 2D and 3D, the transformer trained on natural images can effectively capture the spatial patterns present in volumetric medical images with only lightweight adaptations. We conduct experiments on four open-source tumor segmentation datasets, and with a single click prompt, our model can outperform domain state-of-the-art medical image segmentation models on 3 out of 4 tasks, specifically by 8.25%, 29.87%, and 10.11% for kidney tumor, pancreas tumor, colon cancer segmentation, and achieve similar performance for liver tumor segmentation. We also compare our adaptation method with existing popular adapters, and observed significant performance improvement on most datasets. Our code and models are available at: https://github.com/med-air/3DSAM-adapter.

1 Introduction

Foundation models trained on massive data have demonstrated impressive ability on various general tasks [1, 2, 3], and are envisaged to impact downstream domains, especially where data collection and labeling are expensive. Segment anything model (SAM) [4] is the one from the computer vision field, which has shown success for general-purpose promptable object segmentation. As is known, such powerful discrimination ability relies on the coverage of distributions as exhibited in training data. This is also the underlying reason for the reported suboptimal performance when applying SAM to

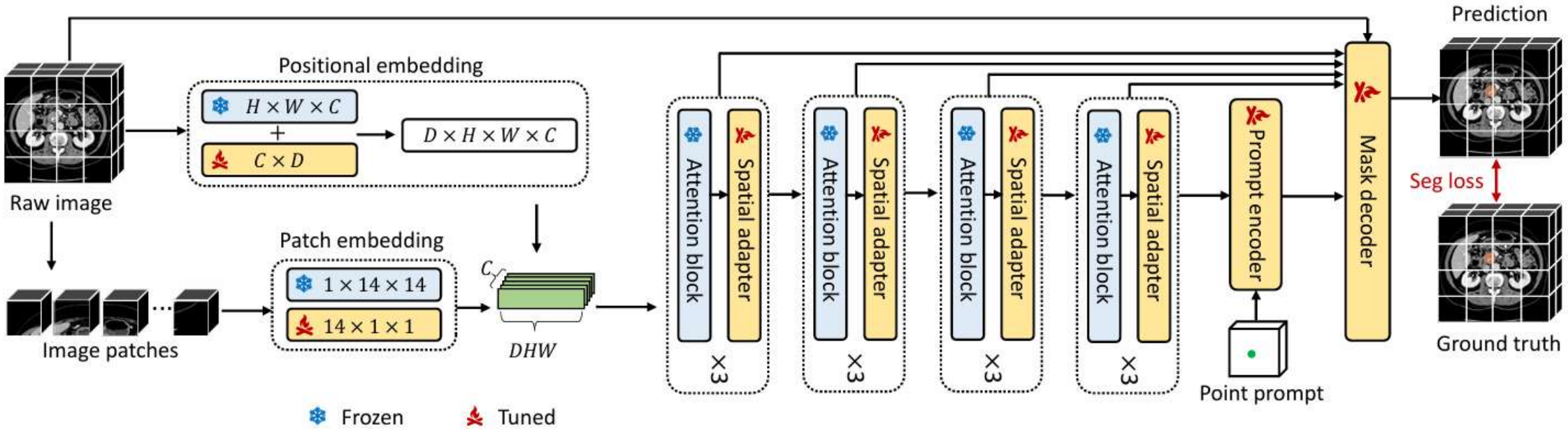
Catalogue

- Introduction
- Method
- Experimental
- Conclusion

Introduction

- The original SAM architecture is designed for 2D natural images, therefore would not be able to extract the 3D spatial information from volumetric medical data effectively.
- We propose a holistic 2D to 3D adaptation method via carefully designed modification of SAM architecture.
- We introduce a novel parameter-efficient fine-tuning method to effectively capitalize a large image model pre-trained on 2D images for 3D medical image segmentation

Method

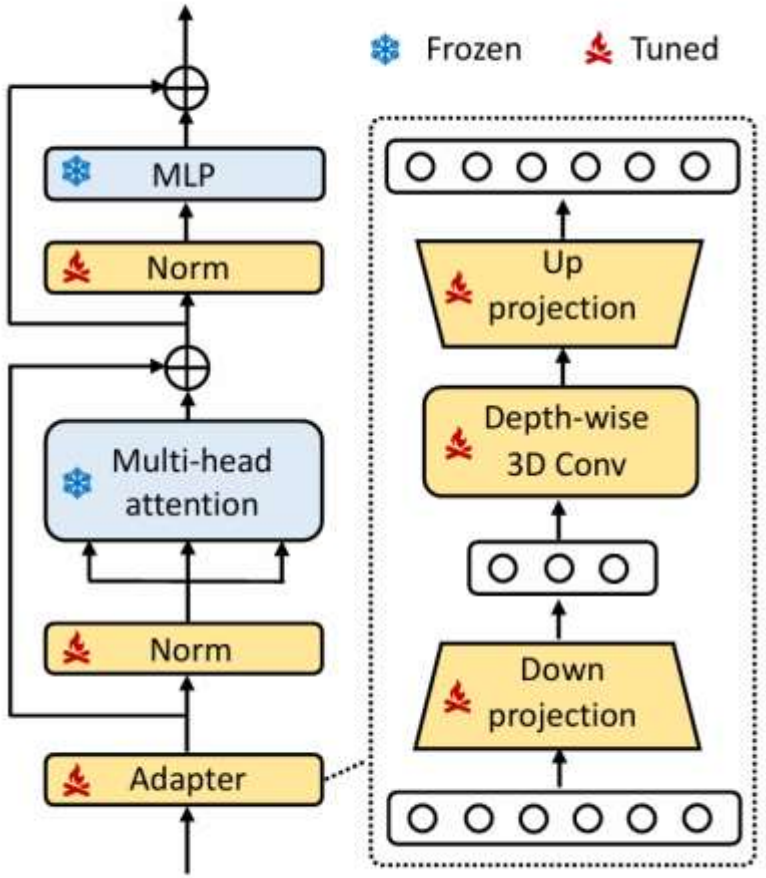


- The original ViT is modifed to support volumetric inputs. The prompt encoder is redesigned to support 3D point prompt, and the mask decoder is updated to 3D CNN with multi-layer aggregation to generate 3D segmentation.

Adapting Image Encoder for Volumetric Inputs

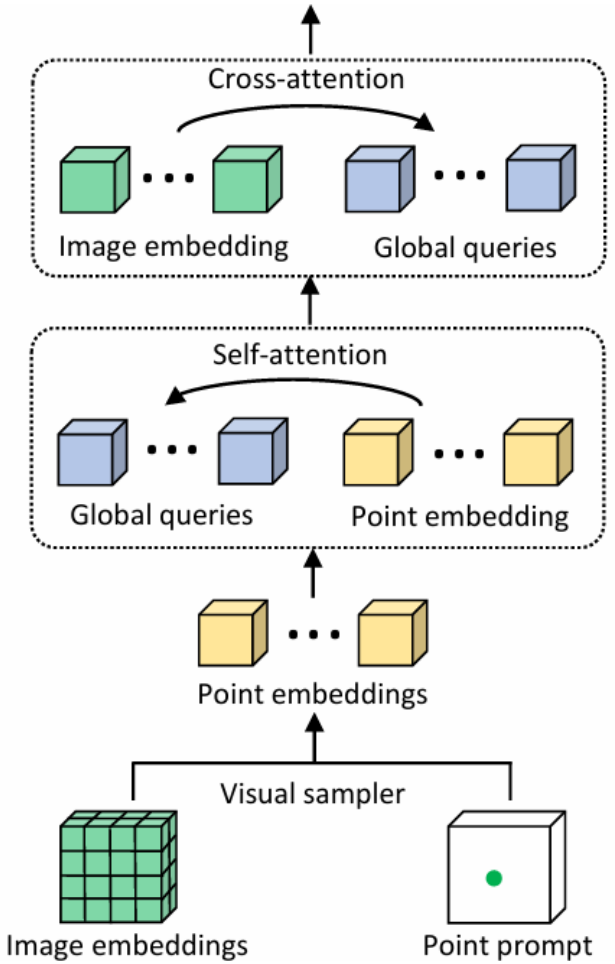
- During the training phase, we only tune the parameters of convolutions, spatial adapters, and normalization layers, while keeping all other parameters frozen.

$$\text{Apater}(\mathbf{X}) = \mathbf{X} + \sigma(\mathbf{X}W_{down})W_{up}$$

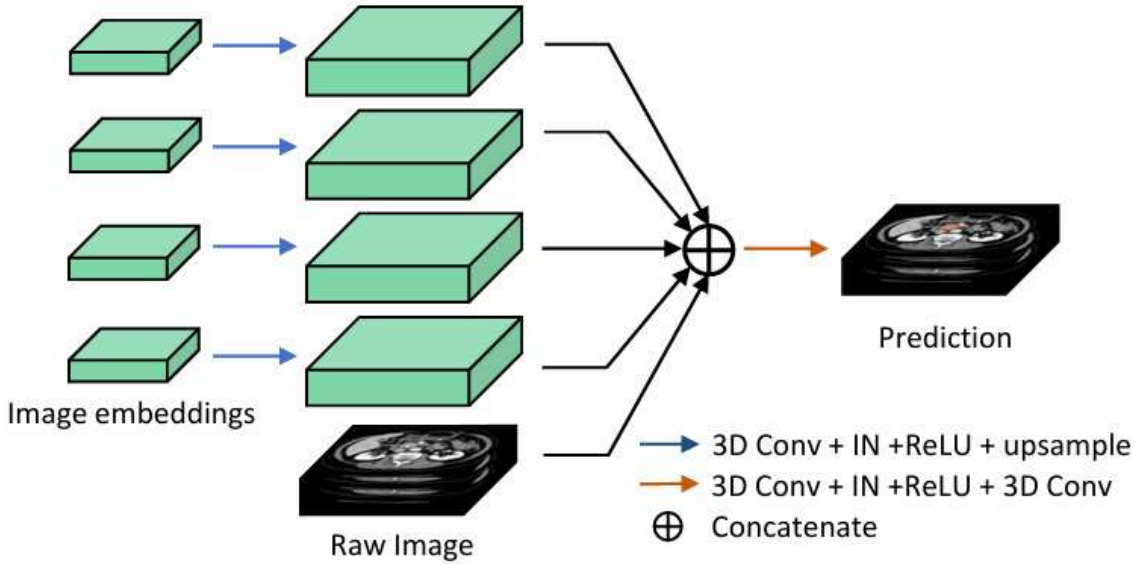


Prompt Encoding by Visual Sampler

- Structure of our prompt encoder based on visual sampler and global queries cross-attention.



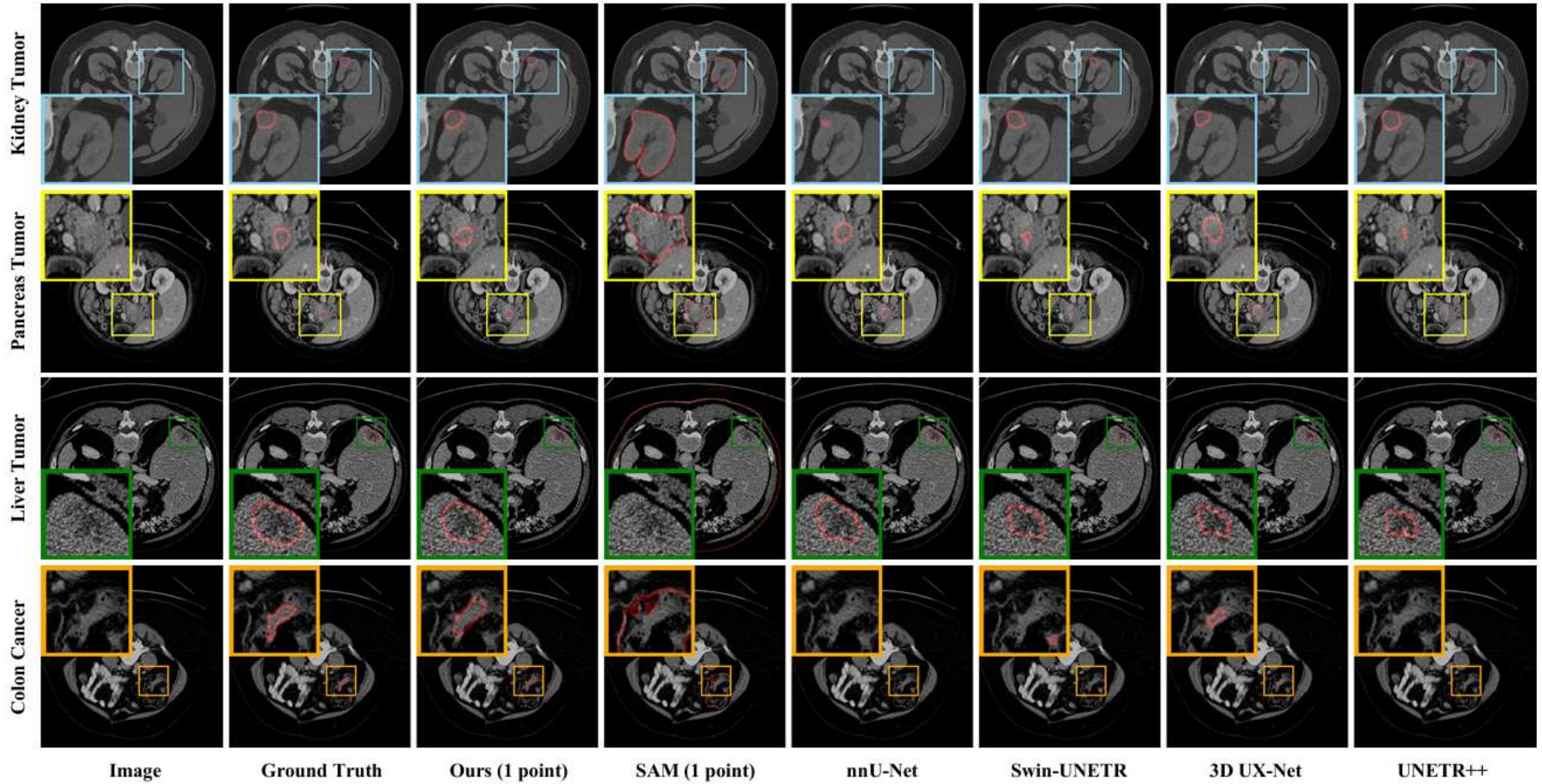
Lightweight Mask Decoder



- To alleviate this issue and meanwhile maintain the lightweight property, we utilize a multi-layer aggregation mechanism in our decoder, where the intermediate output of the encoder is concatenated together to produce a mask feature map while the whole structure remains lightweight.

Experiments

- KiTS21(肾肿瘤), 300 abdominal CT scans
- MSD-Pancreas(胰腺), 281 abdominal CT scans
- LiTS17(肝脏肿瘤), 118 abdominal CT scans
- MSD-Colon(结肠癌), 126 abdominal CT scans
- The datasets are randomly split into 70%, 10%, and 20% for training, validation, and testing



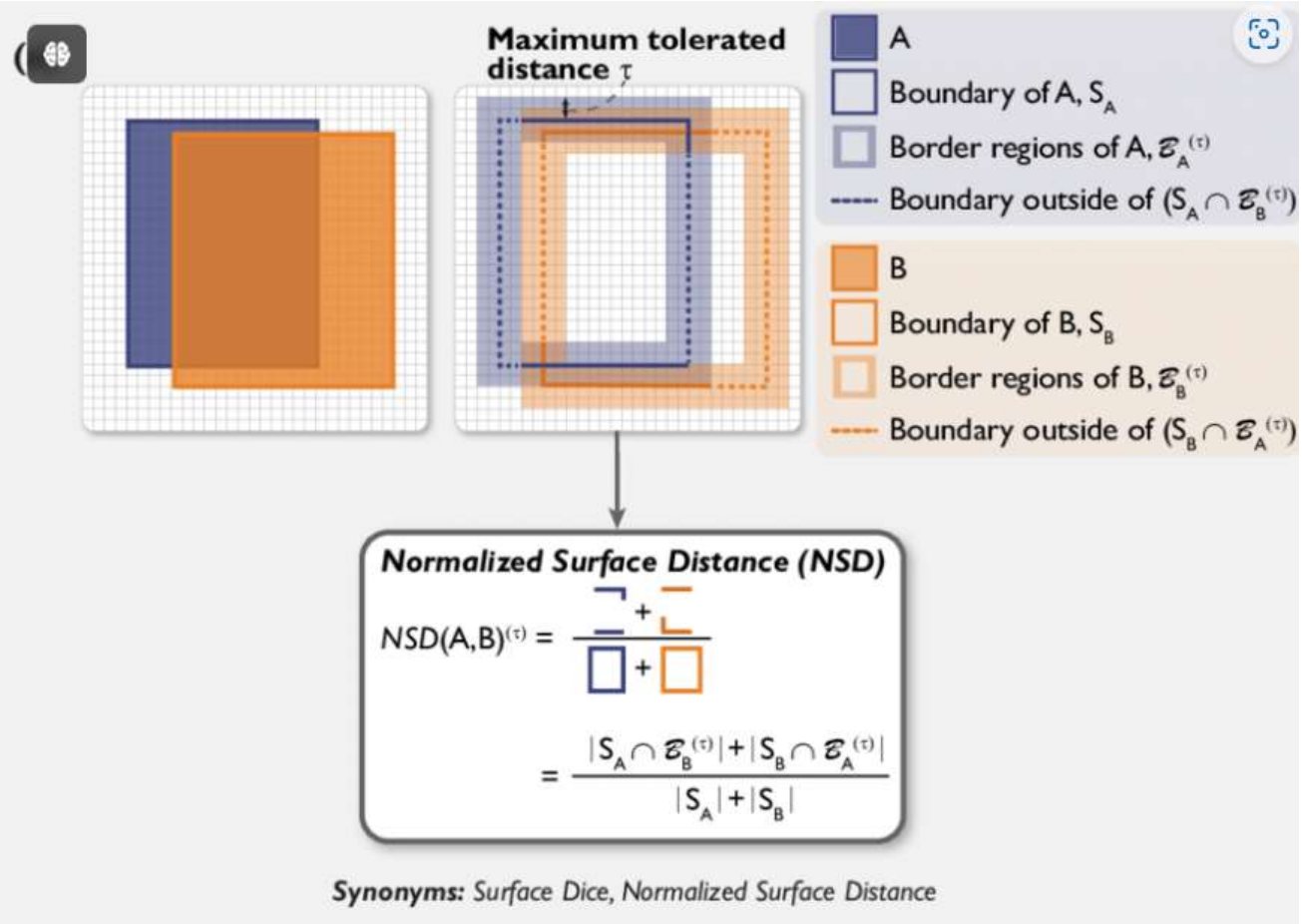
- Qualitative visualizations of the proposed method and baseline approaches on kidney tumor, pancreas tumor, liver tumor and colon cancer segmentation tasks.

Comparison with classical medical image segmentation methods on four tumor segmentation datasets.

Methods	Kidney Tumor		Pancreas Tumor		Liver Tumor		Colon Cancer		#Tuned Params
	Dice ↑	NSD ↑	Dice ↑	NSD ↑	Dice ↑	NSD ↑	Dice ↑	NSD ↑	
nnU-Net (Nat. Methods 2021) [22]	73.07	77.47	41.65	62.54	**60.10**	**75.41**	43.91	52.52	30.76M
TransBTS (MICCAI 2021) [52]	40.79	37.74	31.90	41.62	34.69	49.47	17.05	21.63	32.33M
nnFormer (arXiv 2021) [53]	45.14	42.28	36.53	53.97	45.54	60.67	24.28	32.19	149.49M
Swin-UNETR (CVPR 2022) [54]	65.54	72.04	40.57	60.05	50.26	64.32	35.21	42.94	62.19M
UNETR++ (arXiv 2022) [42]	56.49	60.04	37.25	53.59	37.13	51.99	25.36	30.68	55.70M
3D UX-Net (ICLR 2023) [43]	57.59	58.55	34.83	52.56	45.54	60.67	28.50	32.73	53.01M
SAM-B (1 pt/slice) [4]	36.30	29.86	24.01	26.74	6.71	7.63	28.83	33.63	–
Ours (1 pt/volume)	**73.78**	**83.86**	**54.09**	**76.27**	54.78	69.55	**48.35**	**63.65**	25.46M
SAM-B (3 pts/slice)) [4]	39.66	34.85	29.80	33.24	7.87	6.76	35.26	39.31	–
Ours (3 pts/volume)	**74.91**	**84.35**	**54.92**	**77.57**	56.30	70.02	**49.43**	**65.02**	25.46M
SAM-B (10 pts/slice)) [4]	40.07	34.96	30.55	32.91	8.56	5.97	39.14	42.70	–
Ours (10 pts/volume)	**75.95**	**84.92**	**57.47**	**79.62**	56.61	69.52	**49.99**	**65.67**	25.46M

- Distinct improvements can be specifically observed for pancreas tumors and colon cancers, of 12.44% and 10.11% in Dice against the prior state-of-the-art.

Normalized surface dice (NSD)

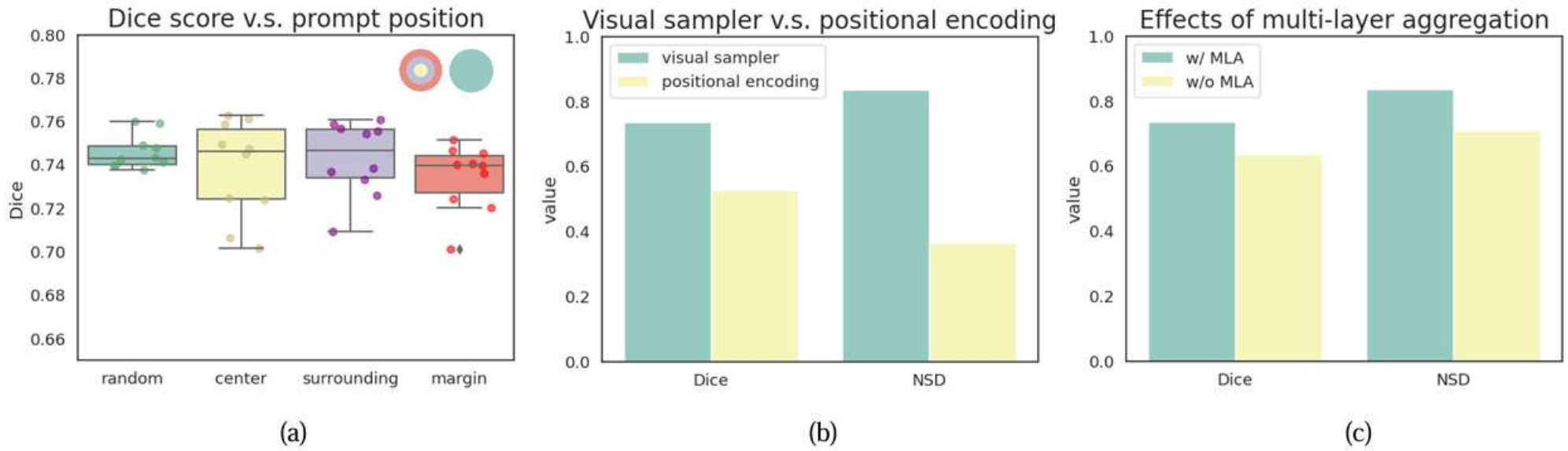


Comparison with existing parameter-efficient and full fine-tuning methods

Methods	Kidney Tumor		Pancreas Tumor		Liver Tumor		Colon Cancer		#Tuned Params (image encoder)
	Dice ↑	NSD ↑	Dice ↑	NSD ↑	Dice ↑	NSD ↑	Dice ↑	NSD ↑	
Full fine-tuning	52.31	50.35	26.49	33.28	45.59	52.53	24.63	**40.67**	89.67M
Adapter (ICML 2019) [16]	46.99	43.76	20.28	30.81	42.17	57.52	22.55	38.10	7.61M
Pro-tuning (arXiv 2022) [14]	50.73	50.81	18.93	30.45	47.33	55.61	21.24	25.10	7.17M
ST-Adapter (NeurIPS 2022) [19]	47.30	45.61	**30.27**	43.53	51.93	59.93	28.41	34.60	7.15M
Med-tuning (arXiv 2023) [20]	44.73	40.28	22.87	30.02	**52.06**	**68.44**	21.08	37.78	11.10M
Ours	**61.60**	**70.40**	30.20	**45.37**	49.26	59.48	**31.97**	**40.67**	15.21M

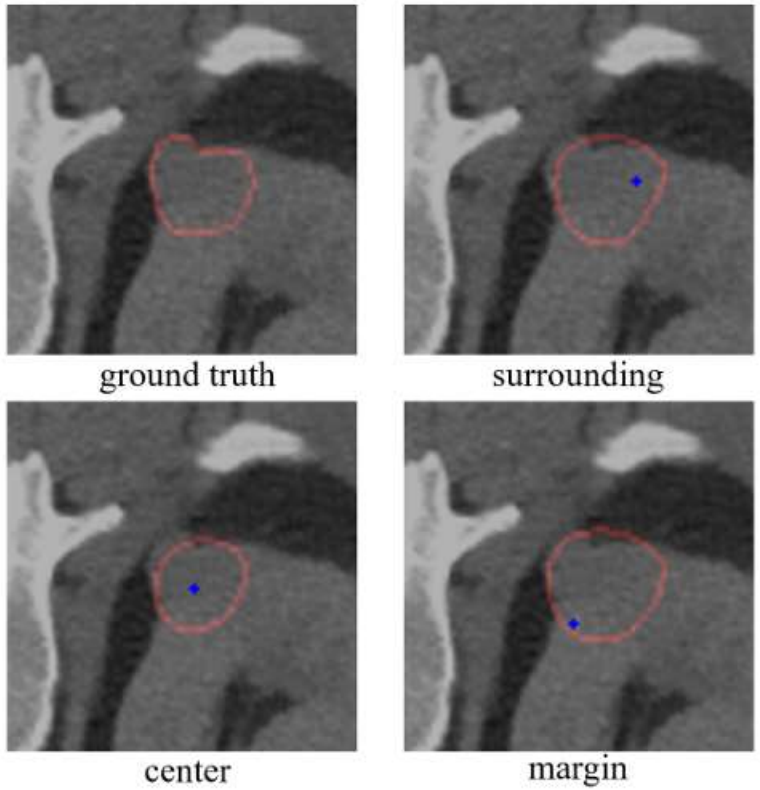
- We discard the prompt encoder and only tune the image encoder and mask decoder for fully automatic segmentation.

Ablation Studies



- The model is not very sensitive to the position of the point prompts. Giving prompt at different positions yield almost the same results.

- We analyze how the model performance is different when the point prompt is given at the center of the objects, at its adjacent regions, or at the margin.



Feature Level				Dice ↑	NSD ↑
1	2	3	4		
✓	✓	✓	✓	64.61	71.33
✓			✓	66.82	74.87
	✓		✓	69.17	77.49
		✓	✓	65.39	72.56
			✓	**73.78**	**83.86**

- We conduct experiments to plug in the prompt encoder to other feature levels besides the final bottleneck levels.
- We find incorporating deep prompts brings no gain or even degenerates the performance.

Conclusion

- We propose a comprehensive scheme to adapt SAM from a 2D natural image generalist to a volumetric medical imaging expert, especially for tumor segmentation.
- Through parameter-efficient fine-tuning, our method significantly improves SAM’s performance in the medical domain, making it outperform SOTA, with only a single coarse click as the prompt.
- Our proposed method can also beat existing adaptation methods for volumetric adaptation.

Thank You!