MA-SAM: Modality-agnostic SAM Adaptation for 3D Medical Image Segmentation

YU JIENING
YU JIENING@umac.mo

um 澳大

arXiv preprint
arXiv:2309.08842,
2023.9.16

Computer Science > Computer Vision and Pattern Recognition

[Submitted on 16 Sep 2023]

MA-SAM: Modality-agnostic SAM Adaptation for 3D Medical Image Segmentation

Cheng Chen, Juzheng Miao, Dufan Wu, Zhiling Yan, Sekeun Kim, Jiang Hu, Aoxiao Zhong, Zhengliang Liu, Lichao Sun, Xiang Li, Tianming Liu, Pheng-Ann Heng, Quanzheng Li

The Segment Anything Model (SAM), a foundation model for general image segmentation, has demonstrated impressive zero-shot performance across numerous natural image segmentation tasks. However, SAM's performance significantly declines when applied to medical images, primarily due to the substantial disparity between natural and medical image domains. To effectively adapt SAM to medical images, it is important to incorporate critical third-dimensional information, i.e., volumetric or temporal knowledge, during fine-tuning. Simultaneously, we aim to harness SAM's pre-trained weights within its original 2D backbone to the fullest extent. In this paper, we introduce a modality-agnostic SAM adaptation framework, named as MA-SAM, that is applicable to various volumetric and video medical data. Our method roots in the parameter-efficient fine-tuning strategy to update only a small portion of weight increments while preserving the majority of SAM's pre-trained weights. By injecting a series of 3D adapters into the transformer blocks of the image encoder, our method enables the pre-trained 2D backbone to extract third-dimensional information from input data. The effectiveness of our method has been comprehensively evaluated on four medical image segmentation tasks, by using 10 public datasets across CT, MRI, and surgical video data. Remarkably, without using any prompt, our method consistently outperforms various state-of-the-art 3D approaches, surpassing nnU-Net by 0.9%, 2.6%, and 9.9% in Dice for CT multi-organ segmentation, MRI prostate segmentation, and surgical scene segmentation respectively. Our model also demonstrates strong generalization, and excels in challenging tumor segmentation when prompts are used. Our code is available at: this https URL.

Catalogue

- Introduction
- Method
- Experiments
- Conclusion

1.Introduction

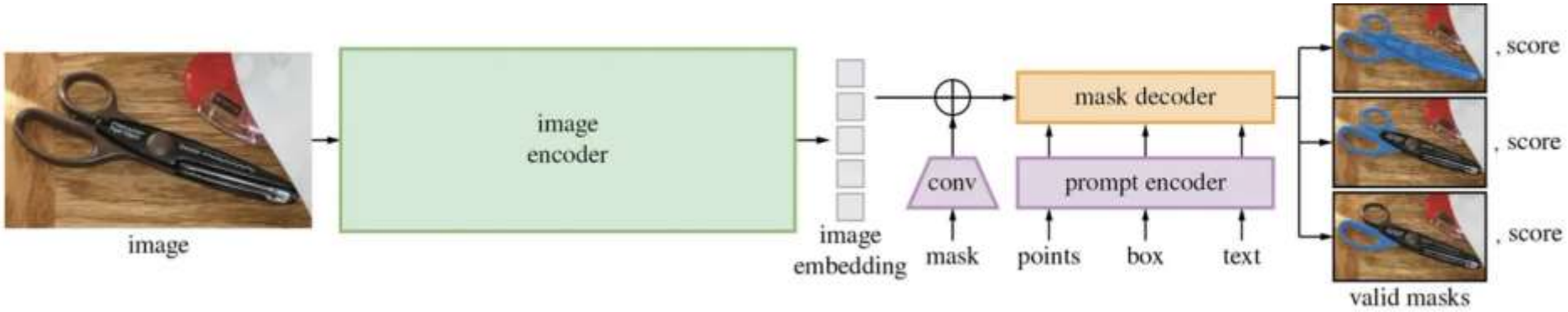
- SAM does not perform well when applied directly to medical images and needs to be **fine-tuned**.
- It's overlooked for the **crucial 3D spatial information in medical volumetric data** and the **temporal information in medical video data.**
- In this paper, we propose a modality-agnostic SAM adaptation method for medical image segmentation, named as **MA-SAM.**

Main contributions of this paper:

- Propose a **parameter-efficient fine-tuning method** to adapt SAM to **volumetric and video medical data**.
- Demonstrate that their SAM adaptation can be applied to **various medical imaging modalities**, including CT, MRI, and surgical video data, for anatomy, surgical scene, and tumor segmentation.
- Validate that after fine-tuning on medical images, the obtained models present outstanding **generalization capability**.

2.Method

2.1 Overview of SAM



- SAM consists of image encoder， prompt encoder and mask decoder.

2.2 Parameter-efficient fine-tuning of image encoder

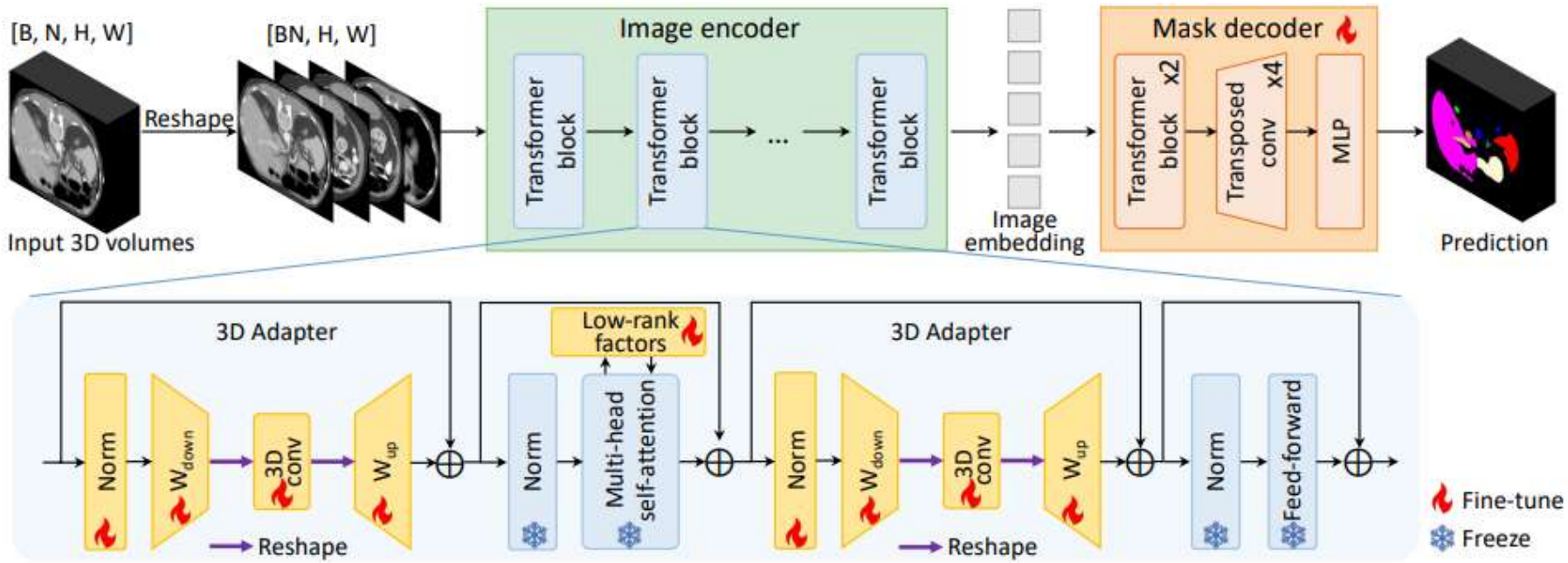
- Adopt **FacT** (Factor-Tuning), a SOTA PETL technique, adapt Transformer. The weight increments can then be calculated using the following equation:

$$\Delta \boldsymbol{W}_{j,k} = s \cdot \sum_{t_1=1}^{r} \sum_{t_2=1}^{r} \boldsymbol{\Sigma}_{t_1,t_2} \boldsymbol{U}_{j,t_1} \boldsymbol{V}_{k,t_2},$$

With the FacT weight increments, the query and value transformations become:

$$\boldsymbol{W}_{q/v} = \boldsymbol{W}_0 + s \cdot \boldsymbol{U}\boldsymbol{\Sigma}_{q/v}\boldsymbol{V}^T,$$

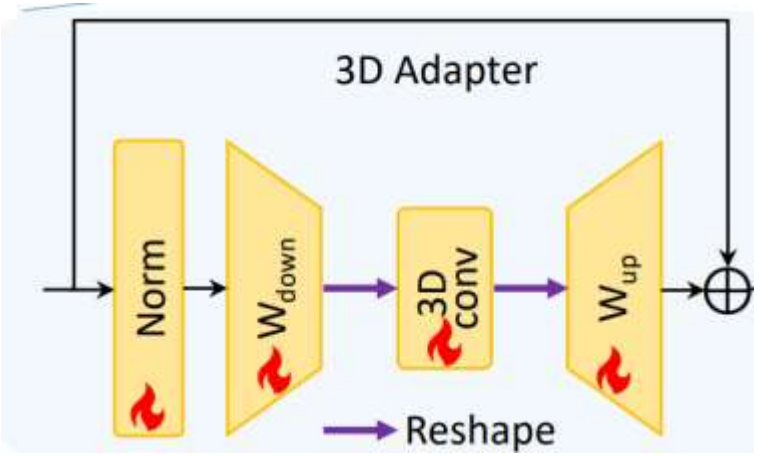
2.3 Incorporating volumetric or temporal information



Each 3D adapter consists of a normalization layer, a linear down-projection layer, a 3D convolutional layer followed by an activation layer, and a linear up-projection layer.

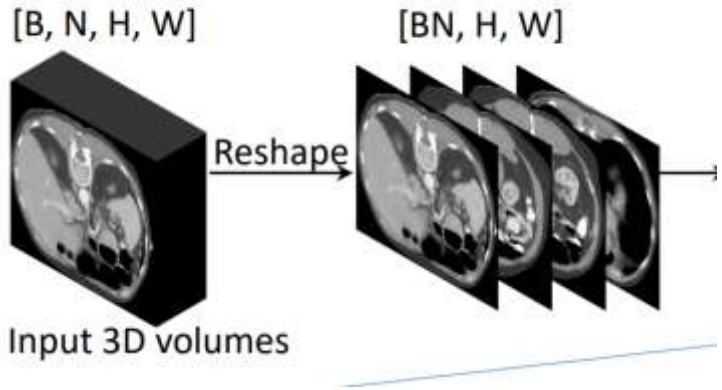
3D adapter can be expressed as:

$$\mathrm{3DAdapter}(\boldsymbol{M}) = \boldsymbol{M} + \sigma(\mathrm{Conv3D}(\mathrm{Norm}(\boldsymbol{M}) \cdot \boldsymbol{W}_{\mathrm{down}}))\boldsymbol{W}_{\mathrm{up}},$$



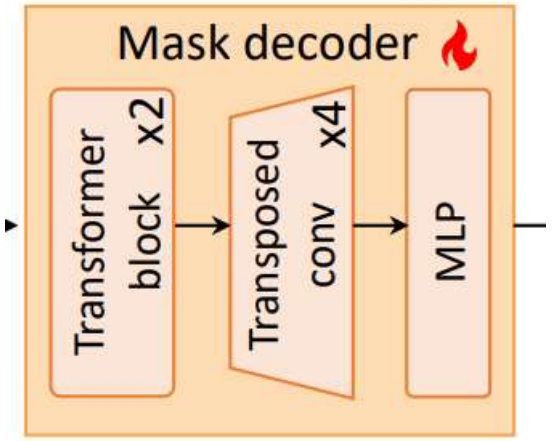
To make the 3D adapters compatible with the 2D SAM backbone, for the network inputs, we extract a set of adjacent slices.

$$\boldsymbol{x} = \{x_{i-\frac{N-1}{2}}, ..., x_i, ..., x_{i+\frac{N-1}{2}}\}_{i=1}^{B}, \boldsymbol{x} \in \mathbb{R}^{B \times N \times H \times W}.$$



2.4 Adapting mask decoder

- Explore two approaches include “progressive upsampling” and “multi-scale fusion ”.
- Progressive upsampling: With each layer up-samples the feature maps by a factor of 2, the four transposed convolutional layers progressively restore feature maps to their original input resolution.
- Multi-scale fusion: creating a design resembling a “U-shaped” network, a concept akin to that of U-Net.



3.Experiments

3.1Datasets

- Task1: The Beyond the Cranial Vault (BTCV) challenge dataset contains **30 CT volumes** with manual annotations for **13 abdominal organs.**
- Task2: We perform **prostate segmentation** on 6 **MRI** data sources, Site A to F, that were collected from **NIC-ISBI13** , **I2CVB**, and **PROMISE12** datasets.
- Task3: The 2018 MICCAI Robotic Scene Segmentation Challenge(EndoVis18) dataset comprises 19 sequences. The dataset encompasses the surgical scene,with 12 classes annotated for various anatomical structures and robotic instruments.

- Task4: The Pancreas Tumor Segmentation task within 2018 MICCAI Medical Segmentation Decathlon Challenge (MSDPancreas) dataset contains **281 CT scan** with annotations for **pancreas and tumor**. Each scan comprises 37 to 751 slices with an axial size of 512×512.
- Use the Multi-Modality Abdominal MultiOrgan Segmentation Challenge **(AMOS 22)** dataset for the evaluation of model generalization. This dataset contains **abdominal CT and MRI** data that were acquired from different patients. **300 CT scans** and **60 MRI scans** in the training and validation sets of AMOS 22 are used for our **generalization evaluation**.

3.2 Implementation details

- Every five consecutive slices were taken as the network inputs.
- For **data augmentation**, we applied a range of transformations including random rotation, flip, erasing, shearing, scaling, translation, posterization, contrast adjustment, brightness modification, and sharpness enhancement.
- Employed ViT H as the backbone of the image encoder and conducted a total of 400 epochs of training.
- Framework was implemented in PyTorch 2.0 using 8 NVIDIA A100 GPUs.

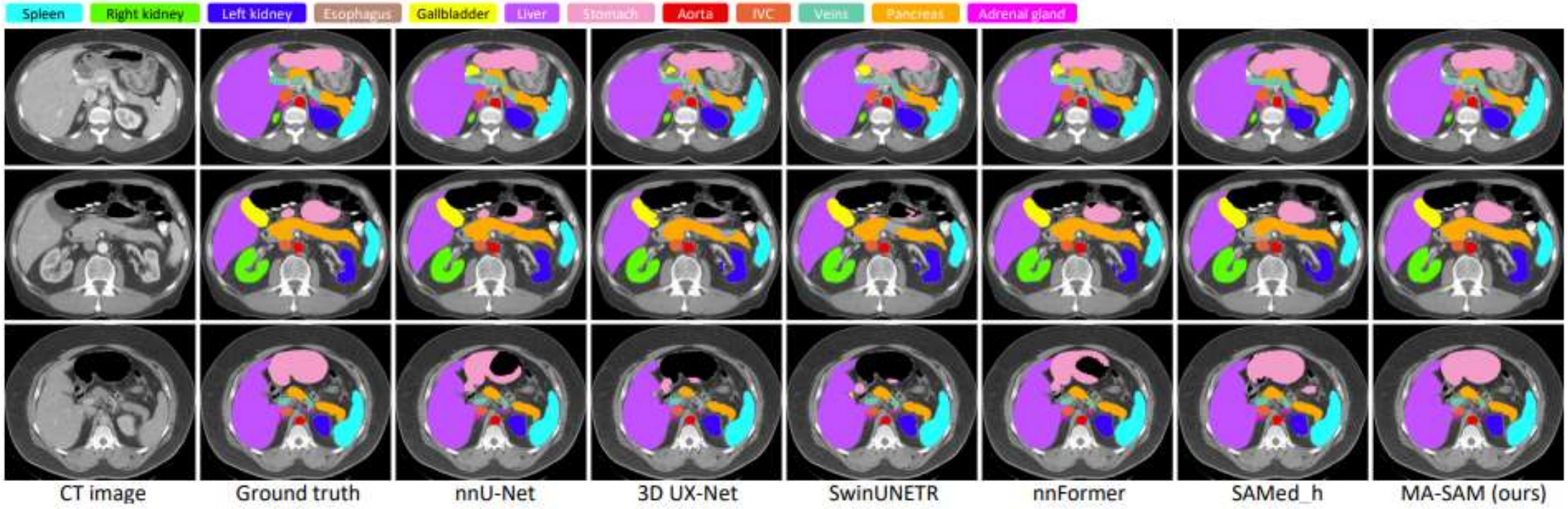
3.3 Comparison with SOTA methods

BTCV datasets:

Methods	Spleen	R.Kd	L.Kd	GB	Eso.	Liver	Stomach	Aorta	IVC	Veins	Pancreas	AG	Average
Dice [%] ↑													
nnU-Net (Isensee et al., 2021)	**97.0**	**95.3**	95.3	63.5	77.5	**97.4**	89.1	90.1	**88.5**	79.0	**87.1**	**75.2**	86.3
3D UX-Net (Lee et al., 2023)	94.6	94.2	94.3	59.3	72.2	96.4	73.4	87.2	84.9	72.2	80.9	67.1	81.4
SwinUNETR (Tang et al., 2022b)	95.6	94.2	94.3	63.6	75.5	96.6	79.2	89.9	83.7	75.0	82.2	67.3	83.1
nnFormer (Zhou et al., 2023a)	93.5	94.9	95.0	64.1	79.5	96.8	90.1	89.7	85.9	77.8	85.6	73.9	85.6
SAMed_h (Zhang and Liu, 2023)	95.3	92.1	92.9	62.1	75.3	96.4	90.2	87.6	79.8	74.2	77.9	61.0	82.1
MA-SAM (Ours)	96.7	95.1	**95.4**	**68.2**	**82.1**	96.9	**92.8**	**91.1**	87.5	**79.8**	86.6	73.9	**87.2**
HD [%] ↓													
nnU-Net (Isensee et al., 2021)	1.07	**1.19**	1.19	7.49	8.56	**1.14**	4.84	14.11	**2.87**	5.67	**2.31**	**2.23**	4.39
3D UX-Net (Lee et al., 2023)	3.17	1.59	1.26	4.53	13.92	1.75	19.72	12.53	3.47	9.99	3.70	4.11	6.68
SwinUNETR (Tang et al., 2022b)	1.21	1.41	1.37	2.25	5.82	1.70	13.75	5.92	4.46	7.58	3.53	3.40	4.37
nnFormer (Zhou et al., 2023a)	78.03	1.41	1.43	3.00	4.92	1.38	4.24	7.53	4.02	6.53	2.96	2.76	9.95
SAMed_h (Zhang and Liu, 2023)	1.37	33.53	1.84	6.27	4.84	1.77	7.49	**4.97**	7.28	6.87	10.00	6.49	7.73
MA-SAM (Ours)	**1.00**	**1.19**	**1.07**	**1.59**	**3.77**	1.36	**3.87**	5.29	3.12	**3.25**	3.93	2.57	**2.67**

SAMed_h: ViT_H version of SAMed, R.Kd: Right kidney, L.Kd: Left kidney, GB: Gall ladder, Eso.: Esophagus, IVC: Inferior vena cava, AG: Adrenal gland

- Comparison of **abdominal multi-organ segmentation results** generated from **MA-SAM** method and other **SOTA** methods on BTCV dataset.



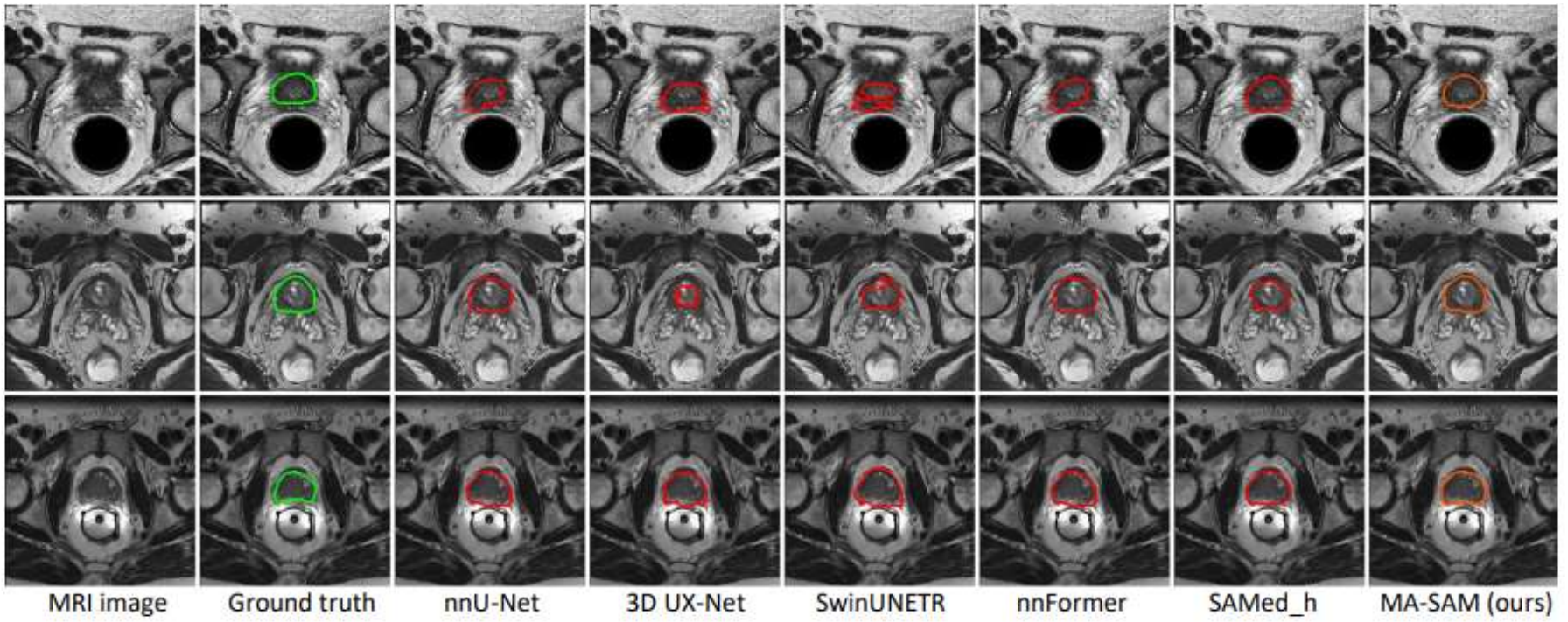
- Qualitative visualization of segmentation results generated from our **MA-SAM** method and other SOTA methods on BTCV dataset. Abdominal organs are denoted in different colors as shown in the corresponding color bar.

Six prostate MRI datasets:

Methods	Site A	Site B	Site C	Site D	Site E	Site F	Average	Site A	Site B	Site C	Site D	Site E	Site F	Average
	Dice [%] ↑							HD [%] ↓						
nnU-Net (Isensee et al., 2021)	93.3	89.2	89.5	86.5	91.0	90.2	90.0	1.74	2.34	3.61	2.98	2.74	1.80	2.54
3D UX-Net (Lee et al., 2023)	91.8	86.0	88.3	70.4	85.9	88.4	85.1	1.95	3.20	4.37	9.61	5.07	2.67	4.48
SwinUNETR (Tang et al., 2022b)	88.7	88.0	88.4	71.5	84.7	84.6	84.3	3.27	3.02	4.37	8.59	5.24	2.82	4.55
nnFormer (Zhou et al., 2023a)	93.6	90.1	89.5	86.8	91.9	90.6	90.4	1.73	2.11	3.54	2.93	2.75	2.08	2.52
SAMed_h (Zhang and Liu, 2023)	94.6	89.5	88.6	87.9	**92.7**	91.3	90.8	1.14	3.90	**3.10**	3.00	2.61	1.67	2.57
MA-SAM (Ours)	**95.3**	**92.7**	**90.4**	**91.3**	**92.7**	**93.1**	**92.6**	**1.00**	**1.54**	3.29	**1.80**	**2.56**	**1.47**	**1.94**

SAMed_h: ViT_H version of SAMed

- Comparison of prostate(前列腺) segmentation results generated from **MA-SAM** method and other **SOTA** methods on six prostate MRI datasets.



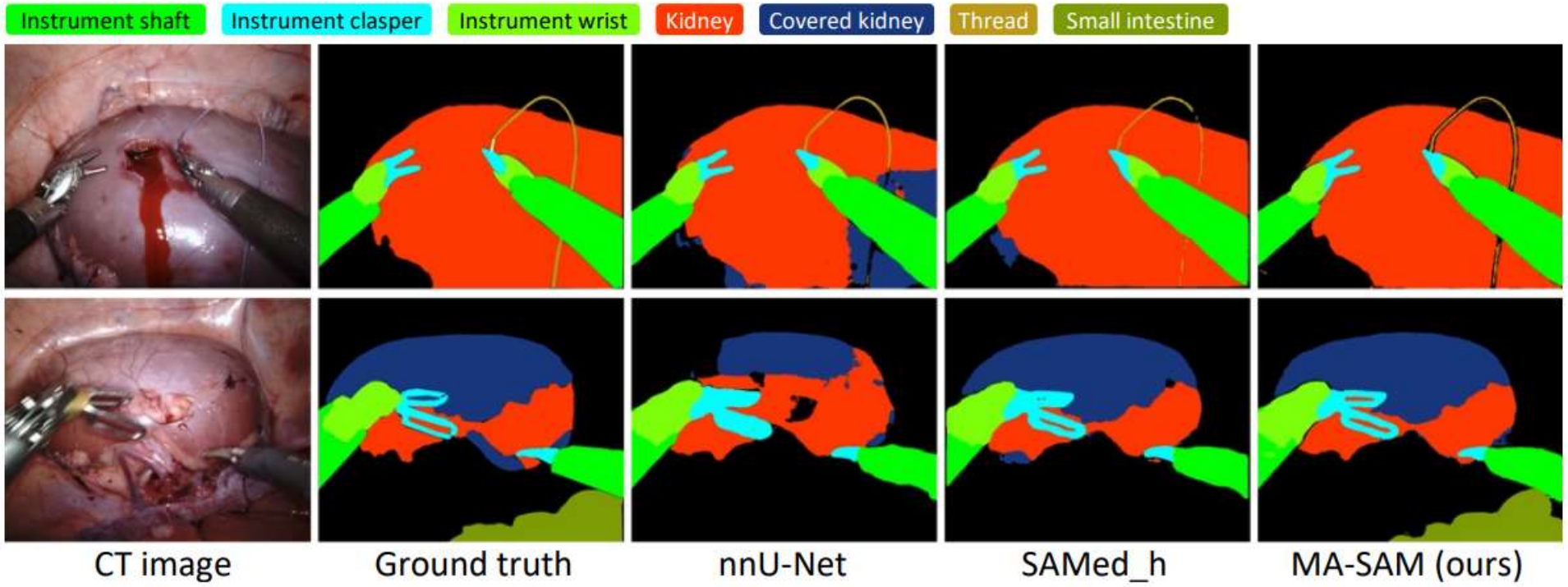
Qualitative visualization of segmentation results generated from our MA-SAM method and other state-of-the-art methods on prostate MRI datasets.

Endovis18 dataset

Comparison of segmentation results from different methods for surgical scene segmentation on Endovis18 dataset

Methods	mIoU	Sequence (mIoU)				Dice
		Seq 1	Seq 2	Seq 3	Seq 4	
NCT (Shvets et al., 2018)	58.5	65.8	55.5	76.5	36.2	-
UNC (Ren et al., 2020)	60.7	63.3	57.8	81.4	37.3	-
OTH (Chen et al., 2018)	62.1	69.1	57.5	82.9	39.0	-
Noisy-LSTM (Wang et al., 2021)	60.4	67.0	56.3	81.8	36.4	69.1
STswinCL (Jin et al., 2022)	63.6	67.0	63.4	83.7	40.3	72.0
nnU-Net (Isensee et al., 2021)	58.7	65.7	57.5	81.3	30.4	67.1
SAMed_h (Zhang and Liu, 2023)	66.5	68.7	60.7	84.3	52.3	74.7
MA-SAM (Ours)	**69.2**	**73.4**	**64.5**	**85.4**	**53.4**	**77.0**

Qualitative visualization of segmentation results generated from different methods for surgical video data. Classes are denoted in different colors.

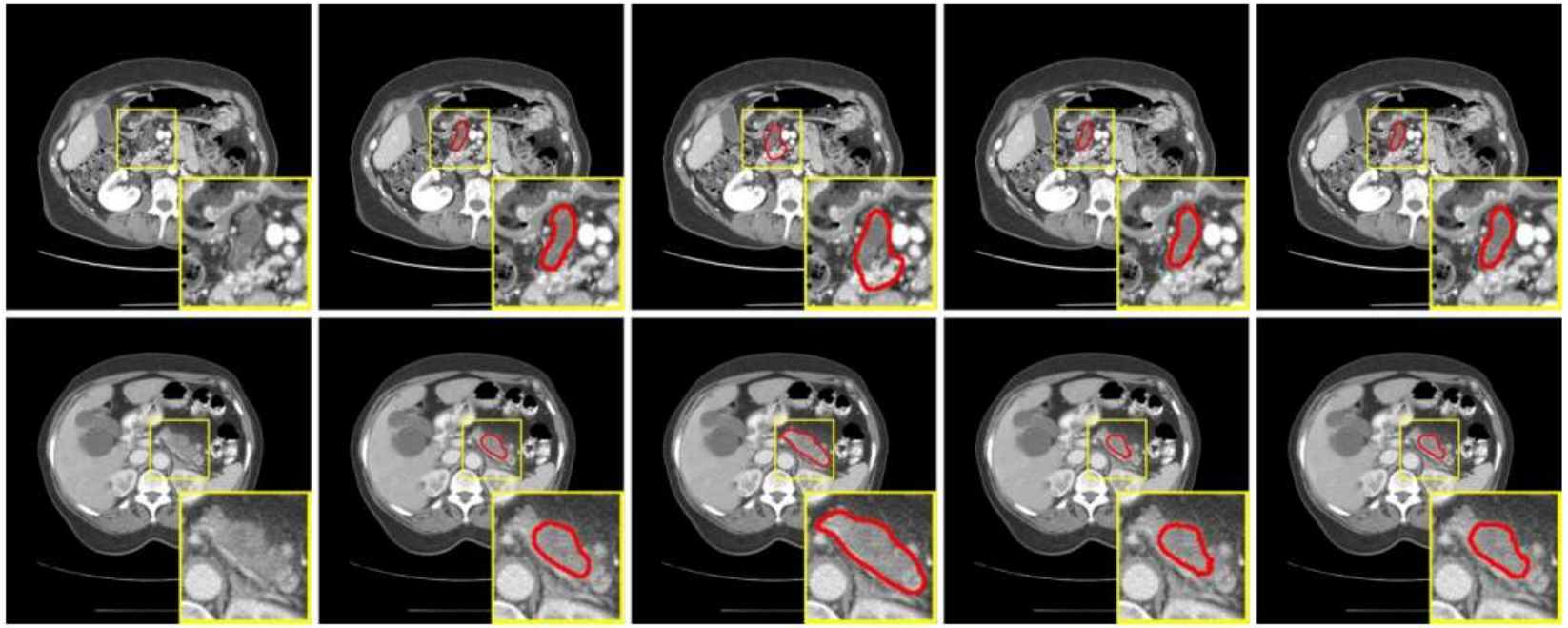


Pancreas tumor(胰腺肿瘤)

Methods	Dice ↑	NSD ↑
nnU-Net (Isensee et al., 2021)	41.6	62.5
3D UX-Net (Lee et al., 2023)	34.8	52.6
SwinUNETR (Tang et al., 2022b)	40.6	60.0
nnFormer (Zhou et al., 2023a)	36.5	54.0
3DSAM-adapter (automatic) (Gong et al., 2023)	30.2	45.4
3DSAM-adapter (10 pts/scan) (Gong et al., 2023)	57.5	79.6
MA-SAM (automatic)	40.2	59.1
MA-SAM (1 tight 3D bbx/scan)	**80.3**	**97.9**
MA-SAM (1 relaxed 3D bbx/scan)	74.7	97.1

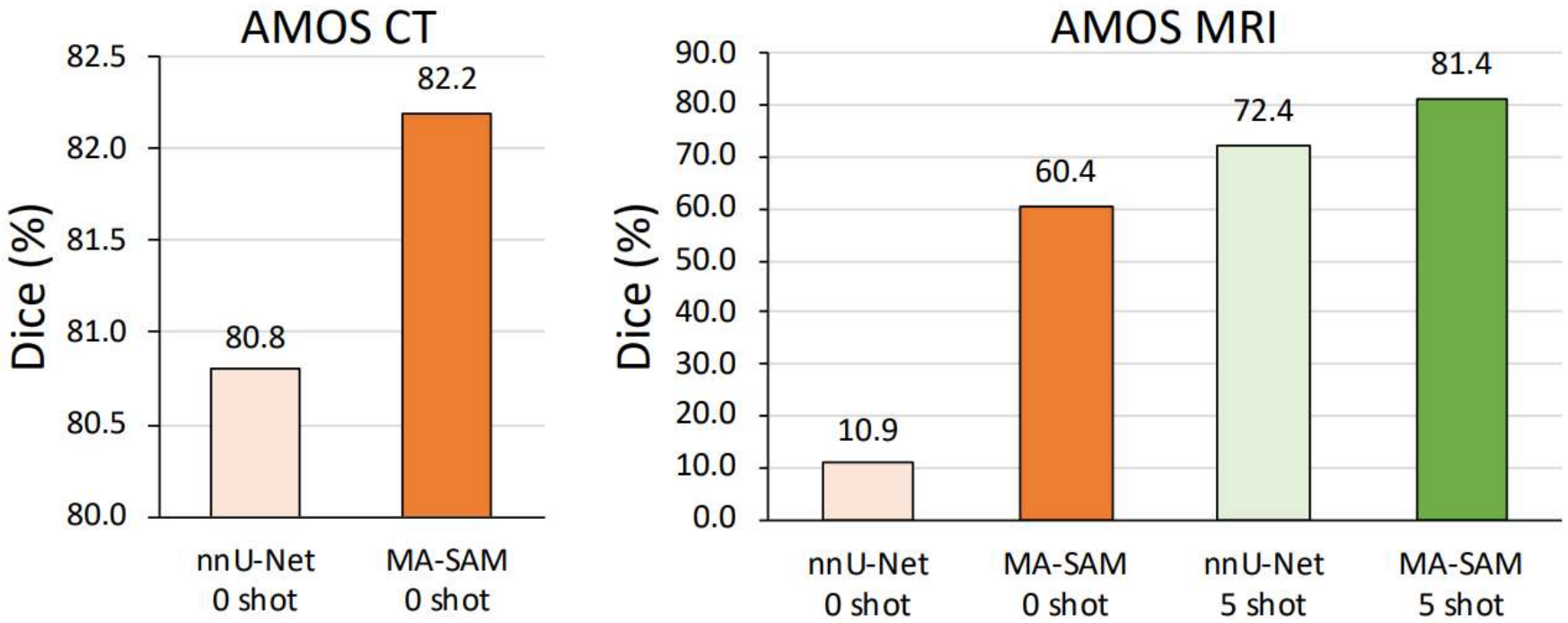
Comparison of segmentation results from different methods for pancreas tumor segmentation in CT images

Qualitative visualization of segmentation results generated from different methods for pancreas tumor segmentation.



All automated segmentation models struggle to correctly depict the pancreatic tumour region, by adding prompts in the form of a tight 3D bounding box in the model, our approach significantly improves the Dice score from 41.6% to 80.35%

3.4 Generalization performance



Comparison of zero-shot and five-shot generalization performance of nnU-Net and our MA-SAM model on AMOS CT and MRI data.

Comparison of generalization performance of nnU-Net and our MASAM model with SOTA domain generalization methods on prostate datasets.

Methods	Site B	Site C	Site D	Site E	Site F	Average
nnU-Net (Isensee et al., 2021)	72.0	69.6	84.7	42.5	82.9	70.3
TTST* (Karani et al., 2021)	86.0	74.8	81.0	74.0	80.9	79.3
TASD* (Liu et al., 2022)	**87.1**	**76.4**	82.5	76.0	83.2	81.1
MA-SAM (Ours)	86.7	66.6	**88.6**	**79.1**	**89.5**	**82.1**

Note: $*$ means the method uses domain generalization techniques.

3.4 Ablation analysis

Table 9: Ablation on each key component in our method. The markers • and ○ denote whether a specific component is used or not.

SAM weights	Full FT	FacT	3D Adapters	Dice [%] ↑
○	●	○	○	72.2
●	○	○	○	70.4
●	●	○	○	85.3
●	○	●	○	85.1
●	○	○	●	86.4
●	○	●	●	87.2

- FacT is capable of delivering performance on par with full fine-tuning, by adjusting a small portion of weight increments.

Table 6: Comparison of model performance with different mask decoder designs.

Decoder design	Dice [%]
SAM mask decoder	84.4
Progressive up-sampling	85.1
Multi-scale fusion	84.5

- Progressive up-sampling strategy yields superior results, no significant improvements were observed with the multi-scale fusion strategy

Table 7: Comparison of model performance with different network backbones.

Backbone	Dice [%]
ViT_B	82.5
ViT_L	84.1
ViT_H	85.1

- There is a noticeable improvement in Dice performance as the model size increases from ViT B to ViT H.

Model	Patch Size	Layers	Hidden Size D	MLP size	Heads	Params
ViT-Base	16x16	12	768	3072	12	86M
ViT-Large	16x16	24	1024	4096	16	307M
ViT-Huge	14x14	32	1280	5120	16	632M

Table 8: Comparison of model performance with different position of 3D adapters.

Position	Dice [%]
Before MHSA	86.7
After MHSA	86.8
Before & after MHSA	87.2

- The configuration with two 3D adapters positioned both before and after MHSA yields superior performance for final model.

Table 10: The change of Dice score for our method with different ranks.

	$r = 4$	$r = 8$	$r = 16$	$r = 32$	$r = 64$
MA-SAM	81.4	82.7	84.6	85.1	85.3

- With an increase in rank, there is a corresponding improvement in average Dice performance, but the performance tends to saturate when r ≥ 32.

4. Conclusion

- MA-SAM exploits FacT to efficiently update a small set of weight increments and inject a set of 3D adapters designed to extract critical volumetric or temporal information in medical video data of medical images during fine-tuning.
- The general applicability and effectiveness of MA-SAM has been validated for **four medical image segmentation tasks** across **three imaging modalities.**
- MA-SAM also demonstrates excellent **generalization capabilities,** with significant advantages in particularly challenging tumor segmentation when using prompts.

Thank You!