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

Cheng Chen[a], Juzheng Miao[b], Dufan Wu[a], Zhiling Yan[c], Sekeun Kim[a], Jiang Hu[a], Aoxiao Zhong[d,a], Zhengliang Liu[e,a], Lichao Sun[c], Xiang Li[a], Tianming Liu[e], Pheng-Ann Heng[b], Quanzheng Li[a,*]

[a] *Center of Advanced Medical Computing and Analysis, Massachusetts General Hospital and Harvard Medical School, Boston, MA 02114, USA*
[b] *Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong, China*
[c] *Department of Computer Science and Engineering, Lehigh University, Bethlehem, PA 18015, USA*
[d] *Harvard John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, USA*
[e] *School of Computing, The University of Georgia, Athens, GA 30602, USA*

ABSTRACT

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: https://github.com/cchen-cc/MA-SAM.

1. Introduction

The rise of foundation models (Bommasani et al., 2021) that are trained on vast and diverse datasets has catalyzed a paradigm shift in intelligent model development. Driven by their remarkable generalization and few-shot learning capability, it has become increasingly appealing to adapt a pre-trained large model to a diversity of downstream tasks, as opposed to the traditional approach of crafting and training distinct task-specific models from scratch. The Segment Anything Model (SAM) (Kirillov et al., 2023) is a recently developed visual foundation model for promptable image segmentation, pre-trained over 1 billion masks on 11 million natural images. Thanks to its large-scale training data and general model architecture, SAM has demonstrated impressive zero-shot performance on various tasks in the context of natural images. Given these merits, a natural question arises: can SAM be directly extended to address the critical medical image segmentation tasks, a domain that has been struggling with limited availability of high-quality images and labels essential for training deep models? However, due to the significant domain gap between natural images and medical images, the latest works on evaluating SAM on medical images have shown that SAM's zero-shot capability, regardless of whether prompts are employed, falls short for direct deployment on medical images (Huang et al., 2023; He et al., 2023; Wald et al., 2023). In these assessments, SAM obtains inferior performance when compared to state-of-the-art (SOTA) medical image segmentation models, and even encounters complete failure in some challenging tasks.

Based on these evaluations, it becomes evident that fine-tuning is an essential step for applying SAM to medical images. But why are we inclined to adapt SAM for medical image tasks? This can be attributed to three potential advantages associated with SAM. Firstly, SAM's training dataset consists of an extensive collection of images. Acquiring a similarly large-scale training dataset in the context of medical applications is extremely challenging. Although SAM's training data only comprises natural images, it is not restricted to any specific medical imaging modality. If SAM fine-tuning proves effective for one type of medical imaging, there is a good chance that the same approach could be applicable to other modalities as well. Secondly, after fine-tuning, SAM as a pre-trained large models may possess potential for robust generalization, which is of great importance for effectively deploying intelligent models in critical medical applications. Thirdly, SAM's prompt design provides a convenient solution for semi-automatic segmentation in tackling difficult tasks, such as tumor segmentation. In these aspects, SAM provides a general-purpose foundation model with the potential to be adapted across diverse medical imaging modalities, offering good generalization capability for both fully-automatic and semi-automatic segmentation.

Efforts to adapt SAM for medical applications are rapidly

[*] Corresponding author: li.quanzheng@mgh.harvard.edu

growing, with the majority of these approaches relying on SAM's prompt design (Cheng et al., 2023; Wu et al., 2023a; Deng et al., 2023a; Dai et al., 2023). However, providing suitable prompts for segmenting each object within medical data is non-trivial. For example, consider an abdominal CT volume containing multiple organs, even providing a basic point prompt for each organ in every slice demands substantial efforts. Moreover, in cases where segmentation objects present relatively regular shapes and locations, automatic segmentation methods already obtain encouraging results, obviating the need for prompts in semi-automatic segmentation. In the context of SAM adaptation for automatic medical image segmentation, some recent studies employ parameter-efficient transfer learning (PETL) techniques, such as LoRA (Hu et al., 2021) or Adapters (Houlsby et al., 2019), showing promising performance in automatic segmentation (Zhang and Liu, 2023; Wang et al., 2023a). However, these methods focus on pure 2D adaptation, overlooking the valuable third-dimensional information inherently present in medical images. This includes 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, which efficiently and effectively captures the volumetric or temporal information in medical data. For the fine-tuning of image encoder, we leverage the PETL technique called FacT (Jie and Deng, 2023), which is based on tensorization-decomposition to enhance the tuning efficiency. Such fine-tuning approach retains the pre-trained weights to a large extent and only updates lightweight weight increments, ensuring the preservation of general knowledge necessary for object segmentation and reducing the number of parameters that need to be adjusted. To bridge the gap between 2D natural images and volumetric or video medical data, we further incorporate a set of 3D adapters into each transformer block of the image encoder to extract the valuable third-dimensional information. For the adaptation of the lightweight mask decoder, we employ full fine-tuning and modify its original architecture with a simple yet effective progressive up-sampling mechanism to recover the prediction resolution. We demonstrate the efficacy of our SAM adaptation framework on multiple medical imaging modalities in tackling various segmentation tasks. By comparing with multiple SOTA methods, our automatic segmentation demonstrates superior performance and remarkable generalization capability. Our main contributions are highlighted as follows:

- We propose a parameter-efficient fine-tuning method to adapt SAM to volumetric and video medical data. Our method effectively incorporates the essential third-dimensional information from medical images into the 2D network backbone via lightweight 3D adapters.

- We demonstrate that our SAM adaptation can be applied to various medical imaging modalities, including CT, MRI, and surgical video data, for anatomy, surgical scene, and tumor segmentation. Without using any prompt, our automatic segmentation consistently outperforms competitive SOTA methods by a large margin.

- We validate that after fine-tuning on medical images, the obtained models present outstanding generalization capability, showing even superior performance than SOTA domain generalization approaches.

- We show that by further leveraging prompts, our method achieves impressive results in challenging tumor segmentation task, surpassing nnU-Net by 38.7% in Dice score.

2. Related work

2.1. Vision foundation models

Foundation models has recently been actively developed in computer vision, although to a lesser extent compared to their prevalence in natural language processing. Pioneering vision foundation models learn directly from vast image-text pairs sourced from the web in a self-supervised manner. Representative works CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) leverage contrastive learning techniques to train both text and image encoders. However, these models primarily excel in tasks that involve mapping images to text, such as classification. Later on, Florence (Yuan et al., 2021) incorporates universal visual-language representations, showing adaptability to more diverse computer vision tasks. One of the latest developments is SAM (Kirillov et al., 2023), a vision foundation model for general-purpose image segmentation. By pre-training on 1 billion masks, SAM demonstrates impressive zero-shot capability across numerous image segmentation tasks. Concurrently, SegGPT (Wang et al., 2023c) and SEEM (Zou et al., 2023) have also emerged for general image segmentation, but are pre-trained on relatively smaller datasets compared to SAM.

2.2. Parameter-efficient transfer learning

With the remarkable performance exhibited by large models, the paradigm of pre-training large foundation models and subsequently fine-tuning for specific downstream tasks has gained increasing popularity. As the pre-trained large models continue to grow in scale, the research on PETL has emerged to achieve effective and efficient adaptation by optimizing only a small subset of model parameters while keeping substantial amount of parameters fixed. PETL techniques has been originally proposed in natural language processing, and can be categorized into three main groups (Lialin et al., 2023), including additive methods, selective methods, and reparameterization-based methods. Additive methods, such as Adapters (Houlsby et al., 2019), aim to augment the existing pre-trained model by introducing additional parameters or layers, and then fine-tuning only these newly introduced components (He et al., 2022; Liu et al., 2023). Selective methods focus on updating a few selected influential layers or internal structure within the model (Gheini et al., 2021; Zaken et al., 2022). Reparameterization-based methods, such as LoRA (Hu et al., 2021) and FacT (Jie and Deng, 2023), leverage low-rank representations to minimize the number of trainable parameters, demonstrating robust and SOTA performance across various PETL tasks. Recently, PETL has also been actively studied in computer vision, enabling the effective adaptation of vision foundation models to a wide range of downstream tasks (Zhou et al., 2022; Jia et al., 2022; Pan et al., 2022; Wang et al., 2023b).

2.3. Adapting SAM in medical imaging

Attracted by SAM's outstanding zero-shot performance in natural images, a plethora of evaluation studies quickly emerged in various medical image segmentation tasks (Huang et al., 2023; He et al., 2023; Wald et al., 2023; Zhou et al., 2023b; Deng et al., 2023b; Hu and Li, 2023; Cheng et al., 2023; Zhang et al., 2023). However, due to the large domain gap between natural and medical images, directly applying SAM to medical applications typically resulted in unsatisfactory performance. For example, He et al. (He et al., 2023) assessed SAM's segmentation accuracy on 12 medical imaging datasets and observed that SAM's zero-shot performance lagged significantly behind models trained on domain-specific medical images, with performance gap as large as 70% in Dice in some tasks. Similar observations were reported in (Huang et al., 2023), even when using different types of prompts. These findings suggest the necessity of task-specific fine-tuning to adapt SAM for medical images for a better segmentation performance.

Subsequently, attention has shifted from evaluation to adaptation of SAM to medical images (Zhang and Liu, 2023; Biswas, 2023; Wu et al., 2023b; Li et al., 2023; Feng et al., 2023). Driven by the improvements observed with the use of prompts, a majority of works leverage SAM's prompt design during fine-tuning (Cheng et al., 2023; Deng et al., 2023a; Dai et al., 2023; Yue et al., 2023). For instance, SAM-Med2D (Cheng et al., 2023) adopted more comprehensive prompts involving points, bounding boxes, and masks to tailor SAM for 2D medical images, and conducted comprehensive evaluations. MSA (Wu et al., 2023a) employed point prompts and the Adapter technique to integrate medical domain knowledge into the SAM model. However, creating prompts for each 2D slice of 3D medical data is labor-intensive. In the case of SAM adaptation for fully automatic medical image segmentation (Hu et al., 2023; Paranjape et al., 2023), SAMed (Zhang and Liu, 2023) and Wang et al. (Wang et al., 2023a) adopted LoRA for fine-tuning, showing superior performance than multiple 2D medical image segmentation methods. However, these methods do not take into account the critical 3D volumetric or temporal information, which is well-known to be valuable for enhancing medical image segmentation performance.

3. Methodology

In this section, we first briefly introduce the overview of SAM architecture, then introduce our method for the parameter-efficient fine-tuning of image encoder, the incorporation of volumetric or temporal information, and the adaptation of mask decoder, respectively. An overview of our framework for effective SAM adaptation is illustrated in Fig. 1.

3.1. Overview of SAM

SAM is a promptable segmentation architecture consisting of three main components, i.e., the image encoder, the prompt encoder, and the mask decoder. The image encoder employs the Vision Transformer (ViT) (Dosovitskiy et al., 2020) as the backbone, extracting essential features of the images with a set of transformer blocks. The prompt encoder takes in various types of prompts including points, boxes, or texts, and encodes these inputs into prompts embeddings to facilitate the segmentation task. The mask decoder is designed to be lightweight, which computes the cross-attention between embeddings of image and prompts, and utilizes transposed convolutional layers and multi-layer perception to generate segmentation masks. When applying to medical images, the model's performance largely degrades since medical images present distinct texture and objects from natural images. This highlights the necessity for task-specific fine-tuning of SAM to address such challenges.

3.2. Parameter-efficient fine-tuning of image encoder

In order to effectively extract image features, SAM's image encoder comprises a substantial portion of network parameters. Fine-tuning all these weights is computationally intensive. Previous research has shown that PETL techniques can achieve adaptation performance similar to full fine-tuning but with significantly fewer network parameters updated (Hu et al., 2021; Pan et al., 2022). In this regard, we adopt FacT (Jie and Deng, 2023), a SOTA PETL technique, that can obtain comparable or superior performance compared to other PETL methods while introducing a smaller number of trainable parameter.

Based on the common observations that the transformer-based models tend to be redundant in rank, FacT assumes the dense weight increment matrices $\Delta\boldsymbol{W}$ used for fine-tuning can be approximated by a set of low-rank factors with cross-layer weight sharing. Following the tensor decomposition in FacT, we decompose the weight increments $\Delta\boldsymbol{W}$ for each layer into three factors $\boldsymbol{U} \in \mathbb{R}^{d\times r}$, $\boldsymbol{V} \in \mathbb{R}^{d\times r}$, and $\boldsymbol{\Sigma} \in \mathbb{R}^{r\times r}$, where d denotes the feature dimensions in ViT, r stands for the rank of these factors with $r << d$. It is worth noting that the two factors, $\boldsymbol{U}$ and $\boldsymbol{V}$, are shared across all layers, while the factor $\boldsymbol{\Sigma}$ is unique for each layer. The weight increments can then be calculated using the following equation:

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

where s denotes a hyper-parameter for adjusting the learning rate of factors. We fix s as 1 in our experiments and tune the overall learning rate with the optimizer to achieve a similar scaling effect. The FacT weight increments are applied to the query and value transformations within each transformer block, while all the other weights initialized from SAM remain frozen, as empirically there were no obvious improvements observed when applying FacT to other layers. With the FacT weight increments, the query and value transformations become:

$$\boldsymbol{W}_{q/v} = \boldsymbol{W}_0 + s \cdot \boldsymbol{U}\boldsymbol{\Sigma}_{q/v}\boldsymbol{V}^T, \tag{2}$$

where $\boldsymbol{W}_{q/v}$ denotes the query or value transformation after fine-tuning, $\boldsymbol{W}_0$ represents the SAM pre-trained weights.

3.3. Incorporating volumetric or temporal information

SAM is initially pre-trained on 2D images, yet medical imaging typically involves more than two dimensions. For example,

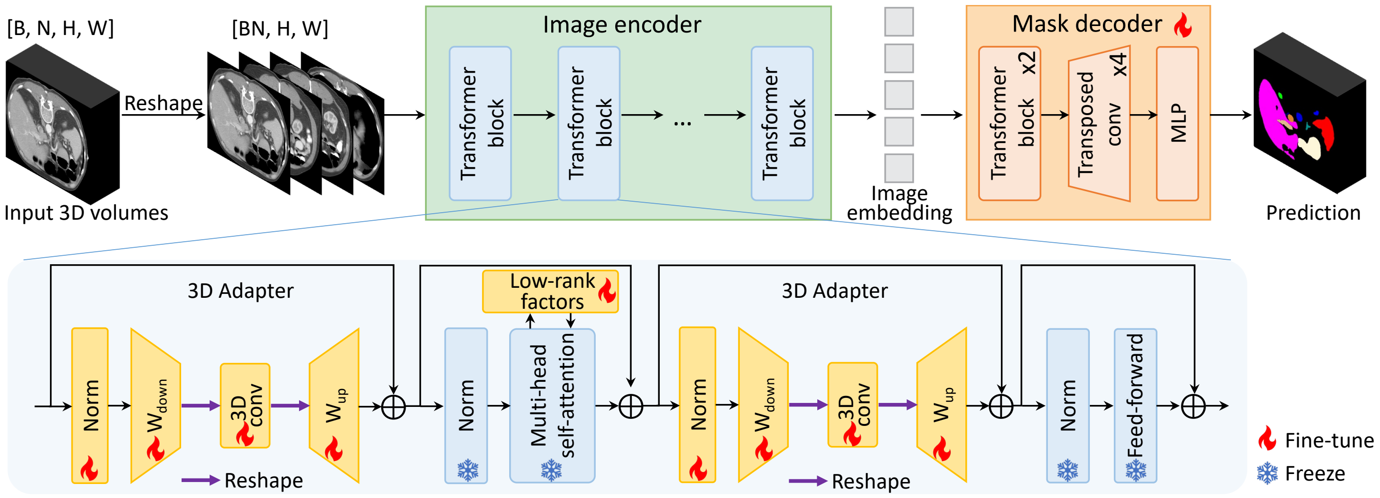

Fig. 1: The overview of our proposed modality-agnostic SAM adaptation framework (MA-SAM) for medical image segmentation. The image encoder is updated through a parameter-efficient fine-tuning strategy with FacT. The volumetric or temporal information is effectively incorporated via a set of 3D adapters. The mask decoder is fully fine-tuned and modified to recover the prediction resolution. Reshape operations are used to make 3D operations compatible with the 2D backbone.

volumetric CT and MRI data contain crucial 3D spatial information for depicting anatomical structures or lesions, and surgical video data possesses valuable temporal relations between frames. Incorporating this volumetric or temporal knowledge inherent in medical imaging data, is pivotal for the successful transfer learning of SAM in medical applications. To address this key challenge, we propose to integrate a series of 3D adapters into the 2D transformer blocks within the SAM architecture. These adapters serve the purpose of extracting the essential volumetric or temporal insights needed for medical image analysis. By incorporating these adapters, we bridge the gap between the inherent complexities of medical imaging data and SAM's pre-trained 2D backbone, enabling it to effectively handle multidimensional medical data.

Specifically, as shown in Fig. 1, 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. The core extraction of volumetric or temporal information primarily resides within the 3D convolutional layer. The purpose of the down-projection layer is to reduce the dimensionality of the original d-dimensional features into a more compact c-dimensional representation, so as to control the number of newly introduced parameters. Conversely, the up-projection layer restores the feature dimensions. With $\boldsymbol{M}$ denoting feature maps, the 3D adapter can be expressed as:

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

where Norm denotes the layer normalization, σ denotes the activation function, $\boldsymbol{W}_{\text{down}} \in \mathbb{R}^{d\times c}$ and $\boldsymbol{W}_{\text{up}} \in \mathbb{R}^{c\times d}$ denote the linear down- and up-projection layer respectively, and Conv3D denotes the 3D convolutional layer with a kernel size of $3\times1\times1$ to specifically extract the third dimensional information.

To make the 3D adapters compatible with the 2D SAM backbone, for the network inputs, we extract a set of adjacent slices $\boldsymbol{x} = \{x_{i-\frac{N-1}{2}}, ..., x_i, ..., x_{i+\frac{N-1}{2}}\}_{i=1}^{B}, \boldsymbol{x} \in \mathbb{R}^{B\times N\times H\times W}$. Here, B denotes the batch size, N denotes the number of adjacent slices, and $H \times W$ denotes the slice dimensions. Before the inputs are passed into the SAM backbone, a reshape operation is applied to transform $\boldsymbol{x} \in \mathbb{R}^{B\times N\times H\times W}$ into $\boldsymbol{x} \in \mathbb{R}^{BN\times H\times W}$ by merging the adjacent slices into the batch dimension. Then for the feature maps, prior to feeding into the 3D convolutional layer of a 3D adapter, they are reshaped from $[BN, H/16, W/16, c]$ to $[B, c, N, H/16, W/16]$. Here $H/16$ and $W/16$ denote the spatial dimensions of feature maps, which are down-sampled by 16 times because of the patch embedding process in transformer. After the 3D convolutional operation, the shape of feature maps are changed back again. In this way, the volumetric or temporal information can be effectively extracted within a 2D network backbone. For each transformer block, we incorporate two 3D adapters before and after the attention layers, as empirically superior performance can be obtained with such a design.

3.4. Adapting mask decoder

The mask decoder within original SAM comprises only two transformer layers, two transposed convolutional layers, and a single multilayer perception layer. Considering its lightweight architecture, it is feasible to apply full fine-tuning on the complete mask decoder for effective adaptation on medical images. During the patch embedding process of the transformer backbone within SAM's image encoder, each 16×16 patch is embedded as a feature vector, leading to 16×16 times down-sampling of the inputs. The SAM mask decoder utilizes two consecutive transposed convolutional layers to up-sample the feature maps by 4 times, yet the final predictions generated by SAM remain 4 times lower in resolution than the original shapes. Nevertheless, since many anatomical structures or lesions in medical images are quite small, achieving a higher resolution is often necessary to ensure improved discrimination in the context of medical imaging (Ronneberger et al., 2015).

To address this issue, we explore two approaches to tailor the mask decoder for enhanced suitability in medical image segmentation. For the first approach, termed as "progressive upsampling", we introduce modest adjustments to the SAM decoder by integrating two additional transposed convolutional operations. 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. The second approach, termed as “multi-scale fusion”, entails creating a design resembling a “U-shaped” network (Ronneberger et al., 2015). This involves connecting the multi-scale feature maps of the image encoder with corresponding stages of the mask decoder using skip connections, a concept akin to that of U-Net. To achieve this, we uniformly divide the image encoder into four stages, establishing connections between the feature maps of each stage and those of the decoder through a series of up-sampling and convolutional operations. In our experiments, we have observed that the gradual up-sampling mechanism yields superior outcomes compared to multi-layer feature aggregation, showing the efficacy and simplicity of the progressive up-sampling approach.

4. Experiments

We extensively evaluate our method on four medical image segmentation tasks, covering three types of medical imaging modalities from 10 datasets, i.e., abdominal multi-organ segmentation in CT, prostate segmentation in MRI, and surgical scene segmentation in surgical video. We first conduct comparison with SOTA medical image segmentation methods and SAM fine-tuning methods, and then provide generalization evaluation and in-depth ablation studies to analyze our method.

4.1. Datasets and evaluation metrics

Task1: The Beyond the Cranial Vault (BTCV) challenge dataset (Landman et al., 2015) contains 30 CT volumes with manual annotations for 13 abdominal organs. Each CT scan contains 85 to 198 slices with the slice thickness varying from 2.5 *mm* to 5.0 *mm*. The axial size is 512 × 512 for all scans, but with in-plane resolution ranging from 0.54 × 0.54 mm^2 to 0.98 × 0.98 mm^2. We use the same data split as (Tang et al., 2022a), which contains 24 cases for training and 6 cases for testing.

Task2: We perform prostate segmentation on 6 MRI data sources (Liu et al., 2020), i.e., Site A to F, that were collected from NIC-ISBI13 (Bloch et al., 2015), I2CVB (Lemaître et al., 2015), and PROMISE12 (Litjens et al., 2014) datasets. The case number for each site is 30, 30, 19, 13, 12, 12 respectively, which were randomly divided into 80% and 20% for training and testing. These MRI scans from different sites were acquired with varying imaging protocols and present heterogeneous data distributions, thus were commonly used in previous domain generalization studies (Liu et al., 2022).

Task3: The 2018 MICCAI Robotic Scene Segmentation Challenge (EndoVis18) dataset (Allan et al., 2020) comprises 19 sequences, captured using the da Vinci X or Xi system. Each sequence contains either 149, 249, or 250 frames at a resolution of 1280 × 1024. The dataset encompasses the surgical scene, with 12 classes annotated for various anatomical structures and robotic instruments. The dataset is officially split into 15 sequences for training and 4 sequences for testing.

Task4: The Pancreas Tumor Segmentation task within 2018 MICCAI Medical Segmentation Decathlon Challenge (MSD-Pancreas) dataset (Antonelli et al., 2022) contains 281 CT scans with annotations for pancreas and tumor. Each scan comprises 37 to 751 slices with an axial size of 512×512. We follow (Gong et al., 2023) to utilize only tumor labels in our experiments and employ the same data split as in their work.

In addition, we use the Multi-Modality Abdominal Multi-Organ Segmentation Challenge (AMOS 22) dataset (Ji et al., 2022) for the evaluation of model generalization. This dataset contains abdominal CT and MRI data that were acquired from different patients. Each scan was annotated with 15 organs, but we focus on the 12 organs that overlap with the BTCV dataset. 300 CT scans and 60 MRI scans in the training and validation sets of AMOS 22 are used for our generalization evaluation.

For data pre-processing, the intensity values of each CT scan in BTCV and MSD-Pancreas datasets were truncated within the range of [-200, 250] Hounsfield Units (HU) and [-50, 200] HU respectively. The intensity of each MRI scan was truncated at the 99th percentile. Each CT or MRI scan was normalized to zero mean and unit variance. For surgical video data, each frame was normalized to [0, 1] range. We resized all images to 512 × 512 for the axial plane of CT and MRI data, as well as for each frame of surgical video sequences. For model evaluation, we employ the common metrics, i.e., the Dice score and Hausdorff Distance (HD) to assess pixel-wise segmentation accuracy and the segmentation boundary quality respectively. We also report the mean intersection-over-union (mIoU) for the EndoVis18 dataset and the normalized surface distance (NSD) for the MSD-Pancreas dataset to align with previous studies.

4.2. Implementation details

The fine-tuning process was supervised using a hybrid segmentation loss, which combines the cross-entropy loss and Dice loss as: $\mathcal{L}_{\text{seg}} = \alpha\mathcal{L}_{\text{ce}} + \beta\mathcal{L}_{\text{Dice}}$. The weighting factors α and β were set as 0.2 and 0.8 following (Zhang and Liu, 2023), except for the surgical video data for which only the Dice loss was utilized. 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. Our model was trained using Adam optimizer with a batch size of 24. As in (Zhang and Liu, 2023), we adopted a warmup training strategy to increase the learning rate linearly to the specific value and then exponentially decrease it towards the end of training to stabilize the training. We employed ViT_H as the backbone of the image encoder and conducted a total of 400 epochs of training, ensuring that the model converged effectively. Our framework was implemented in PyTorch 2.0 using 8 NVIDIA A100 GPUs.

4.3. Comparison with SOTA methods

For CT and MRI datasets, we extensively compare our method with various SOTA 3D medical image segmentation methods including CNN-based approaches **nnU-Net** (Isensee et al., 2021), which is a U-Net (Ronneberger et al., 2015) based self-configuring framework, showing robust performance on various medical image segmentation competitions, and **3D UX-Net** (Lee et al., 2023), which is a very recent large kernel volumetric ConvNet for 3D medical image segmentation, as well as

Table 1: Comparison of abdominal multi-organ segmentation results generated from our MA-SAM method and other state-of-the-art methods on BTCV dataset.

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

Table 2: Comparison of prostate segmentation results generated from our MA-SAM method and other state-of-the-art methods on 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

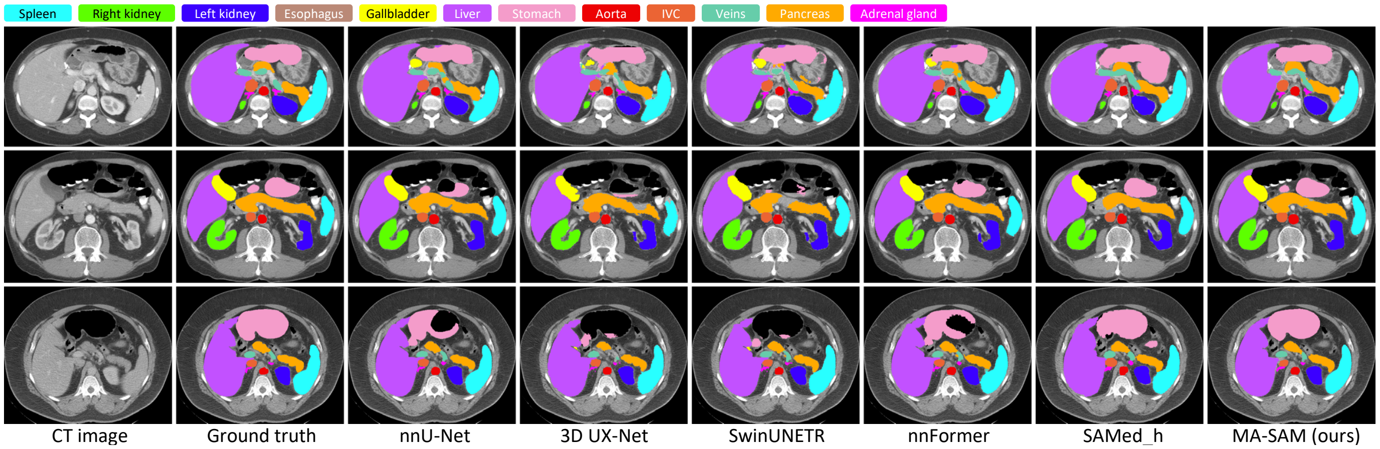

Fig. 2: Qualitative visualization of segmentation results generated from our MA-SAM method and other state-of-the-art methods on BTCV dataset. Abdominal organs are denoted in different colors as shown in the corresponding color bar.

transformer-based methods **SwinUNETR** (Tang et al., 2022b), which is a 3D transformer-based model with a hierarchical encoder, and **nnFormer** (Zhou et al., 2023a), which is a model combining local and global volume-based self-attention mechanism. We also compare our method with the most recent SAM adaptation methods **SAMed_h** (Zhang and Liu, 2023), which is an automatic 2D medical image segmentation model for organ segmentation, and **3DSAM-adapter** (Gong et al., 2023), which is a promptable 3D medical image segmentation model for tumor segmentation. For surgical video data, we compare our method with SOTA surgical scene segmentation methods, **NCT** (Shvets et al., 2018), **UNC** (Ren et al., 2020), and **OTH** (Chen et al., 2018), which are the top-three approaches reported in the challenge, **Noisy-LSTM** (Wang et al., 2021) which uses ConvLSTM to learn temporal cues, **STswinCL** (Jin et al., 2022) which is a transformer-based model capturing intra- and inter-video relations, and **nnU-Net**. For all comparison experiments, the dataset splits remain consistent across all the methods.

Table 3: 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**

Table 4: Comparison of segmentation results from different methods for pancreas tumor segmentation in CT images.

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

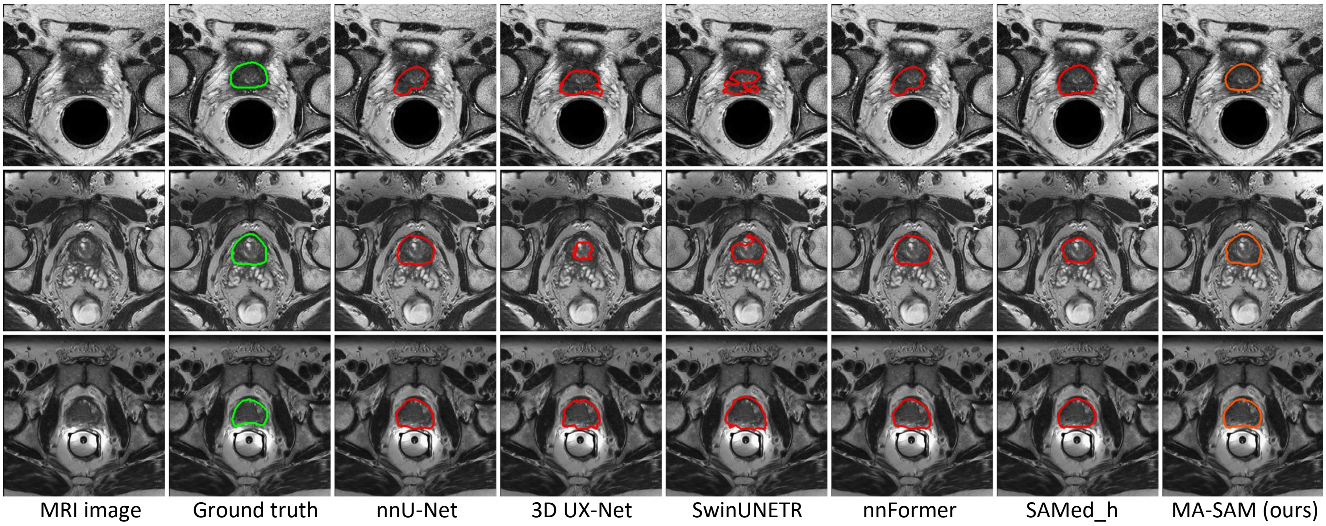

Fig. 3: Qualitative visualization of segmentation results generated from our MA-SAM method and other state-of-the-art methods on prostate MRI datasets. The prostate boundary is delineated in green for ground truth, in orange for our method, and in red for other methods, respectively.

Table 1 to Table 4 present comparative results for the four different tasks: abdominal multi-organ segmentation in CT data, prostate MRI segmentation across 6 sites, scene segmentation in surgical video, and tumor segmentation in CT data, respectively. When prompts are not specified, all methods generate results automatically without using any prompt. With our dedicatedly designed fine-tuning strategy for SAM, our method consistently and significantly outperforms other comparison approaches across all the four tasks. In terms of fully automatic segmentation for the first three tasks, our method improves the Dice score by 0.9%, 2.6%, 5% compared to the second-best performing approach, respectively. Notably, nnU-Net proves to be a strong competitor, showing robust segmentation performance across CT and MRI datasets. However, in surgical scene segmentation, nnU-Net obtains lower results compared to methods specifically tailored for processing surgical videos. Our method demonstrates strong performance across both volumetric and video medical data, indicating the potential of unifying the network architecture in these two domains of medical imaging, where previous methods were developed separately. When comparing with the pure 2D SAM fine-tuning method SAMed_h, which employs the same network backbone as ours, our method also achieves significantly better results, demonstrating the benefits of incorporating volumetric or temporal information for 3D medical image segmentation. The visual comparison results are presented in Fig. 2 to Fig. 4.

Pancreas tumor segmentation presents a substantial challenge due to the irregular contours and unclear margins of pancreas tumors in CT scans. As can be seen in Table 4 and Fig. 5, all automatic segmentation models struggle to correctly delineate pancreas tumor regions, obtaining merely a 41.6% Dice score for the best-performing model. We consider in such a demanding segmentation task, the use of prompts become valuable. By adding prompts in the form of one tight 3D bounding box per volume into the model, our method remarkably boosts the Dice score from 41.6% to 80.35%, demonstrating the effectiveness of leveraging prompts for tumor segmentation. However, if allowing 5% relaxation on the tightness of provided bounding box, the performance drops to 74.7%, showing the effect of prompts quality on segmentation performance. Our method also significantly outperforms the recent holistic 3D SAM adaptation method 3DSAM-adapter, with 10% Dice improvement when using automatic segmentation. This can be attributed to our method's effective incorporation of third-dimensional information into model fine-tuning, as well as its substantial utilization of pre-trained weights from SAM to retain general discriminative knowledge. We notice that our method's automatic segmentation performance falls slightly behind nnU-Net on tumor

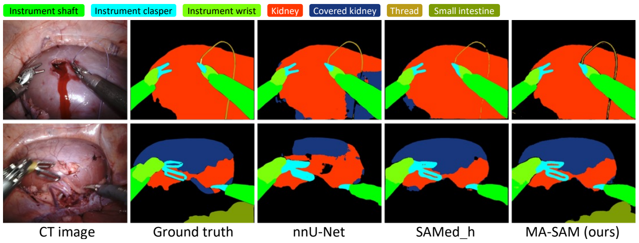

Fig. 4: Qualitative visualization of segmentation results generated from different methods for surgical video data. Classes are denoted in different colors.

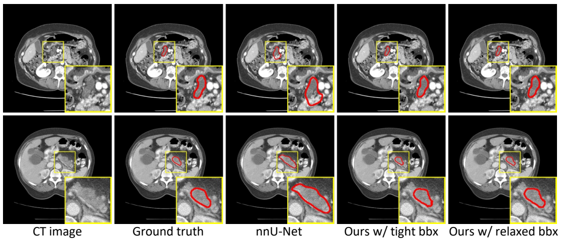

Fig. 5: Qualitative visualization of segmentation results generated from different methods for pancreas tumor segmentation.

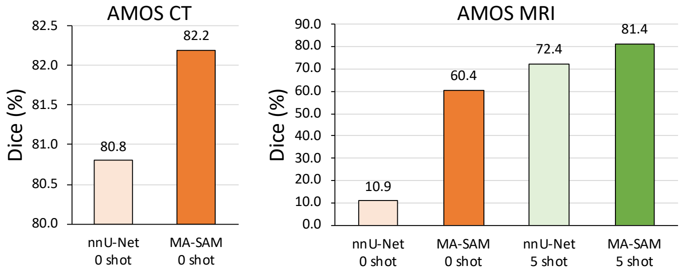

Fig. 6: Comparison of zero-shot and five-shot generalization performance of nnU-Net and our MA-SAM model on AMOS CT and MRI data.

Table 5: Comparison of generalization performance of nnU-Net and our MA-SAM 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.

segmentation. This observation might indicate that SAM fine-tuning might be less effective for objects with ill-defined margins and small sizes, as these characteristics differ from the natural images on which SAM was originally trained.

4.4. Generalization evaluation

One of the most appealing advantage of foundation models lies in their impressive generalization capability. To investigate the generalization of our models adapted from SAM, we first compare the zero-shot and few-shot capability of nnU-Net and our method by applying models trained on the BTCV CT scans to the AMOS22 CT and MRI scans. In Fig. 6, "nnU-Net 0 shot" and MA-SAM 0 shot" denote that the models trained on BTCV data are directly used to perform inference on AMOS22 images, and "nnU-Net 5 shot" and MA-SAM 5 shot" denote the models are further fine-tuned with 5 additional training cases from the AMOS22 dataset. From the results, we can see that our method exhibits better zero-shot and few-shot segmentation performance on AMOS22 CT and MRI images, demonstrating higher generalization capability. Especially for MRI images, nnU-Net encounters complete failure in the zero-shot context, obtaining only 10.9% Dice score, while our model still retains the performance of 60.4% Dice score. In the five-shot context, our method also shows 9% improvements than nnU-Net, further underscoring the advantages of generalization.

We also compare the model generalization on prostate MRI segmentation. In Table 5, the results of nnU-Net and our models are obtained by directly applying the models fine-tuned on Site A to make predictions for Site B to F. We include two recent SOTA test-time domain generalization methods into comparison, i.e., TTST (Karani et al., 2021) and TASD (Liu et al., 2022), which employ additional domain generalization techniques when performing predictions on each specific site. The results demonstrate that our method not only outperforms nnU-Net by a large margin when generalizing to different sites, but also achieves superior performance than SOTA domain generalization approaches. All these results on AMOS22 dataset and different prostate MRI datasets underscore the impressive generalization capability of our method, which is an importance characteristic for critical medical applications.

4.5. Ablation analysis of our method

We conduct extensive ablation experiments on the BTCV dataset to investigate several key aspects regarding our SAM fine-tuning strategy: 1) effectiveness of each important component in our method, 2) effect of mask decoder design, 3) influence of network backbone, 4) choice of location for 3D adapters, 5) choice of rank for parameter-efficient fine-tuning.

1) Effectiveness of each component: We first validate the contribution of key components within our method, i.e., SAM's pre-trained weights, the parameter-efficient fine-tuning strategy with FacT, and the incorporation of 3D information with 3D adapters. In Table 9, the "Full FT" model denotes whether full fine-tuning is used for the image encoder, since the mask decoder is fully fine-tuned for all the models. By comparing the results between the first and third rows in Table 9, we observe a substantial 13.1% improvement in Dice when the model is initialized with SAM's pre-trained weights. This underscores the benefits of utilizing SAM's original weights that were pre-trained on a large-scale and diverse dataset. The second row shows that if the entire image encoder remains frozen without being fine-tuned, and only the mask decoder is updated, the performance is unsatisfactory. This suggests that due to the significant difference in image texture between natural and medical images, the image encoder trained solely on natural images struggles to extract essential features from medical images. Moreover, Table 9 demonstrates that FacT is capable of delivering performance on par with full fine-tuning, by adjusting a small portion of weight increments. The models equipped with 3D adapters achieve superior performance, validating the

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

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

Backbone	Dice [%]
ViT_B	82.5
ViT_L	84.1
ViT_H	85.1

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

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

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

importance of incorporating the third-dimensional information for medical image segmentation.

2) Effect of mask decoder design: We compare the performance of different mask decoder designs, including the original SAM mask decoder, the progressive up-sampling strategy, and the multi-scale fusion strategy. Table 6 shows that the straightforward progressive up-sampling strategy yields superior results, validating its simplicity and effectiveness. These results demonstrate the importance of recovering prediction resolution for medical images, which often contain small objects. However, no significant improvements were observed with the multi-scale fusion strategy. This might because of the extensive modifications it introduces to the original SAM decoder, resulting in less effective utilization of the pre-trained SAM weights.

3) Influence of network backbone: We conduct experiments with different network backbones, i.e., ViT_B, ViT_L, and ViT_H, to assess their impact on the performance of our method. As can be observed in Table 7, there is a noticeable improvement in Dice performance as the model size increases from ViT_B to ViT_H, signifying the advantages of using a larger model size to enhance overall model performance.

4) Choice of location for 3D adapters: We perform ablation experiments to investigate the placement of 3D adapters within our model. Specifically, we compare the performance when incorporating a 3D adapter in one of three locations: before the multi-head self-attention block (MHSA), after MHSA, or in both of these positions. As demonstrated in Table 8, the configuration with two 3D adapters positioned both before and after MHSA yields superior performance for our final model.

5) Choice of rank: We investigate how the model's performance changes with varying decomposition rank r, by considering the rank value from the set {4, 8, 16, 32, 64}. As expected, Table 10 shows that with an increase in rank, there is a corresponding improvement in average Dice performance, but the performance tends to saturate when $r \geq 32$. We thus set $r = 32$ in our experiments to seek a balance between performance gains and the number of parameters introduced.

5. Discussion

Foundation models, like the Segment Anything Model (SAM), have revolutionized intelligent model development by offering robust generalization and few-shot learning capabilities. SAM has demonstrated impressive zero-shot performance for natural image tasks. However, applying SAM directly to medical image segmentation has proven ineffective due to the substantial domain differences. To address this problem, in this work, we propose a parameter-efficient fine-tuning method to adapt SAM to various medical imaging modalities. Our method leverages FacT to efficiently update a small portion of weight increments and injects a set of designed 3D adapters to extract crucial volumetric or temporal knowledge of medical images during fine-tuning. The general applicability and effectiveness of our method has been validated on four medical image segmentation tasks across three imaging modalities. Our model also demonstrates outstanding generalization capability, as well as significant advantage in particularly challenging tumor segmentation when prompts are used.

One significant motivation for adapting SAM to medical images is its pre-training on a vast and diverse dataset, which is difficult to achieve in the filed of medical imaging. This makes SAM's adaptation generally applicable to various medical imaging modalities. In medical applications, there are recent efforts trying to pre-train modality-specific foundation models. However, these models are often constrained to a specific medical imaging modality and challenging to extend to others. For example, models pre-trained with chest x-ray data may face difficulties when applied to MRI data. By leveraging SAM's pre-trained weights, we are able to train a large-scale segmentation network, such as ViT_H, for medical image segmentation, even when limited data, such as just 5 imaging scans are used. Our experiments have demonstrated the benefits of increasing the model size, raising the intriguing question of how performance evolves with further increases in model size. Can we achieve improved accuracy or enhanced generalization with larger models for medical images? Exploring these possibilities holds great interest.

Using promptable segmentation is less meaningful for tasks that can already achieve satisfactory results with SOTA medical image segmentation methods. Prompts prove particularly beneficial and valuable when dealing with challenging tumor segmentation tasks, as demonstrated in our experiments as well as other SAM fine-tuning works. However, crafting effective prompts demands a substantial amount of effort. As shown in Table 4, the performance of promptable segmentation drops as the quality of prompts declines. Given the challenges associated with manual prompt creation, there is considerate room for future exploration in automating this process. It would be interesting and valuable to investigate methods for generating suitable prompts automatically or study how to train an accurate segmentation model with noisy or imperfect prompts. This would enhance the practicality of promptable segmentation in scenarios where manual prompt creation is challenging.

6. Conclusion

We present an effective SAM adaptation framework that is general and can be applied to diverse medical image segmentation tasks across different modalities. Our method roots in the parameter-efficient fine-tuning strategy and successfully incorporates the volumetric or temporal information of medical images during fine-tuning. Without using any prompt, our method with automatic segmentation outperforms various SOTA 3D medical image segmentation methods by a large margin. Our model has demonstrated outstanding generalization capability, which is crucial for successful deployment of intelligent model across medical datasets. We have also shown the substantial advantage of the prompt mode, which is particularly valuable in tackling challenging tumor segmentation task. Our method holds significant promise as a general segmentation framework that can be applied to various medical imaging modalities for both fully automatic and promptable segmentation.

References

Allan, M., Kondo, S., Bodenstedt, S., Leger, S., Kadkhodamohammadi, R., Luengo, I., Fuentes, F., Flouty, E., Mohammed, A., Pedersen, M., et al., 2020. 2018 robotic scene segmentation challenge. arXiv preprint arXiv:2001.11190 .

Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., Ronneberger, O., et al., 2022. The medical segmentation decathlon. Nature communications 13, 4128.

Biswas, R., 2023. Polyp-sam++: Can a text guided sam perform better for polyp segmentation? arXiv preprint arXiv:2308.06623 .

Bloch, N., Madabhushi, A., Huisman, H., Freymann, J., Kirby, J., Grauer, M., Enquobahrie, A., Jaffe, C., Clarke, L., Farahani, K., 2015. Nci-isbi 2013 challenge: automated segmentation of prostate structures. The Cancer Imaging Archive 370.

Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., et al., 2021. On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258 .

Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation, in: Proceedings of the European conference on computer vision (ECCV), pp. 801–818.

Cheng, D., Qin, Z., Jiang, Z., Zhang, S., Lao, Q., Li, K., 2023. Sam on medical images: A comprehensive study on three prompt modes. arXiv preprint arXiv:2305.00035 .

Dai, H., Ma, C., Liu, Z., Li, Y., Shu, P., Wei, X., Zhao, L., Wu, Z., Zhu, D., Liu, W., et al., 2023. Samaug: Point prompt augmentation for segment anything model. arXiv preprint arXiv:2307.01187 .

Deng, G., Zou, K., Ren, K., Wang, M., Yuan, X., Ying, S., Fu, H., 2023a. Sam-u: Multi-box prompts triggered uncertainty estimation for reliable sam in medical image. arXiv preprint arXiv:2307.04973 .

Deng, R., Cui, C., Liu, Q., Yao, T., Remedios, L.W., Bao, S., Landman, B.A., Wheless, L.E., Coburn, L.A., Wilson, K.T., et al., 2023b. Segment anything model (sam) for digital pathology: Assess zero-shot segmentation on whole slide imaging. arXiv preprint arXiv:2304.04155 .

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 .

Feng, W., Zhu, L., Yu, L., 2023. Cheap lunch for medical image segmentation by fine-tuning sam on few exemplars. arXiv preprint arXiv:2308.14133 .

Gheini, M., Ren, X., May, J., 2021. Cross-attention is all you need: Adapting pretrained transformers for machine translation, in: Moens, M., Huang, X., Specia, L., Yih, S.W. (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021, Association for Computational Linguistics. pp. 1754–1765.

Gong, S., Zhong, Y., Ma, W., Li, J., Wang, Z., Zhang, J., Heng, P.A., Dou, Q., 2023. 3dsam-adapter: Holistic adaptation of sam from 2d to 3d for promptable medical image segmentation. arXiv preprint arXiv:2306.13465 .

He, J., Zhou, C., Ma, X., Berg-Kirkpatrick, T., Neubig, G., 2022. Towards a unified view of parameter-efficient transfer learning, in: The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022, OpenReview.net.

He, S., Bao, R., Li, J., Grant, P.E., Ou, Y., 2023. Accuracy of segment-anything model (sam) in medical image segmentation tasks. arXiv preprint arXiv:2304.09324 .

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S., 2019. Parameter-efficient transfer learning for nlp, in: International Conference on Machine Learning, PMLR. pp. 2790–2799.

Hu, C., Li, X., 2023. When sam meets medical images: An investigation of segment anything model (sam) on multi-phase liver tumor segmentation. arXiv preprint arXiv:2304.08506 .

Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., 2021. Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 .

Hu, X., Xu, X., Shi, Y., 2023. How to efficiently adapt large segmentation model (sam) to medical images. arXiv preprint arXiv:2306.13731 .

Huang, Y., Yang, X., Liu, L., Zhou, H., Chang, A., Zhou, X., Chen, R., Yu, J., Chen, J., Chen, C., et al., 2023. Segment anything model for medical images? arXiv preprint arXiv:2304.14660 .

Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H., 2021. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods 18, 203–211.

Ji, Y., Bai, H., Yang, J., Ge, C., Zhu, Y., Zhang, R., Li, Z., Zhang, L., Ma, W., et al., 2022. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. arXiv preprint arXiv:2206.08023 .

Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T., 2021. Scaling up visual and vision-language representation learning with noisy text supervision, in: International conference on machine learning, PMLR. pp. 4904–4916.

Jia, M., Tang, L., Chen, B.C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.N., 2022. Visual prompt tuning, in: European Conference on Computer Vision, Springer. pp. 709–727.

Jie, S., Deng, Z.H., 2023. Fact: Factor-tuning for lightweight adaptation on vision transformer, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 1060–1068.

Jin, Y., Yu, Y., Chen, C., Zhao, Z., Heng, P.A., Stoyanov, D., 2022. Exploring intra-and inter-video relation for surgical semantic scene segmentation. IEEE Transactions on Medical Imaging 41, 2991–3002.

Karani, N., Erdil, E., Chaitanya, K., Konukoglu, E., 2021. Test-time adaptable neural networks for robust medical image segmentation. Medical Image Analysis 68, 101907.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al., 2023. Segment anything. arXiv preprint arXiv:2304.02643 .

Landman, B., Xu, Z., Igelsias, J., Styner, M., Langerak, T., Klein, A., 2015. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge, in: Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge, p. 12.

Lee, H.H., Bao, S., Huo, Y., Landman, B.A., 2023. 3d ux-net: A large kernel volumetric convnet modernizing hierarchical transformer for medical image segmentation, in: The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023, OpenReview.net.

Lemaître, G., Martí, R., Freixenet, J., Vilanova, J.C., Walker, P.M., Meriaudeau, F., 2015. Computer-aided detection and diagnosis for prostate cancer based on mono and multi-parametric mri: a review. Computers in biology and medicine 60, 8–31.

Li, C., Khanduri, P., Qiang, Y., Sultan, R.I., Chetty, I., Zhu, D., 2023. Autoprompting sam for mobile friendly 3d medical image segmentation. arXiv preprint arXiv:2308.14936 .

Lialin, V., Deshpande, V., Rumshisky, A., 2023. Scaling down to scale up: A guide to parameter-efficient fine-tuning. arXiv preprint arXiv:2303.15647 .

Litjens, G., Toth, R., van de Ven, W., Hoeks, C., Kerkstra, S., van Ginneken, B., Vincent, G., Guillard, G., Birbeck, N., Zhang, J., et al., 2014. Evaluation of prostate segmentation algorithms for mri: the promise12 challenge. Medical image analysis 18, 359–373.

Liu, Q., Chen, C., Dou, Q., Heng, P.A., 2022. Single-domain generalization in medical image segmentation via test-time adaptation from shape dictionary, in: AAAI, pp. 1756–1764.

Liu, Q., Dou, Q., Yu, L., Heng, P.A., 2020. Ms-net: Multi-site network for improving prostate segmentation with heterogeneous mri data. IEEE Transactions on Medical Imaging .

Liu, X., Zheng, Y., Du, Z., Ding, M., Qian, Y., Yang, Z., Tang, J., 2023. Gpt understands, too. AI Open .

Pan, J., Lin, Z., Zhu, X., Shao, J., Li, H., 2022. St-adapter: Parameter-efficient image-to-video transfer learning. Advances in Neural Information Processing Systems 35, 26462–26477.

Paranjape, J.N., Nair, N.G., Sikder, S., Vedula, S.S., Patel, V.M., 2023. Adaptivesam: Towards efficient tuning of sam for surgical scene segmentation. arXiv preprint arXiv:2308.03726 .

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al., 2021. Learning transferable visual models from natural language supervision, in: International conference on machine learning, PMLR. pp. 8748–8763.

Ren, X., Ahmad, S., Zhang, L., Xiang, L., Nie, D., Yang, F., Wang, Q., Shen, D., 2020. Task decomposition and synchronization for semantic biomedical image segmentation. IEEE Transactions on Image Processing 29, 7497–7510.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Springer. pp. 234–241.

Shvets, A.A., Rakhlin, A., Kalinin, A.A., Iglovikov, V.I., 2018. Automatic instrument segmentation in robot-assisted surgery using deep learning, in: 2018 17th IEEE international conference on machine learning and applications (ICMLA), IEEE. pp. 624–628.

Tang, Y., Yang, D., Li, W., Roth, H.R., Landman, B., Xu, D., Nath, V., Hatamizadeh, A., 2022a. Self-supervised pre-training of swin transformers for 3d medical image analysis, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 20730–20740.

Tang, Y., Yang, D., Li, W., Roth, H.R., Landman, B., Xu, D., Nath, V., Hatamizadeh, A., 2022b. Self-supervised pre-training of swin transformers for 3d medical image analysis, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 20730–20740.

Wald, T., Roy, S., Koehler, G., Disch, N., Rokuss, M.R., Holzschuh, J., Zimmerer, D., Maier-Hein, K., 2023. Sam. md: Zero-shot medical image segmentation capabilities of the segment anything model, in: Medical Imaging with Deep Learning, short paper track.

Wang, A., Islam, M., Xu, M., Zhang, Y., Ren, H., 2023a. Sam meets robotic surgery: An empirical study on generalization, robustness and adaptation. arXiv preprint arXiv:2308.07156 .

Wang, B., Li, L., Nakashima, Y., Kawasaki, R., Nagahara, H., Yagi, Y., 2021. Noisy-lstm: Improving temporal awareness for video semantic segmentation. IEEE Access 9, 46810–46820.

Wang, W., Shen, J., Chen, C., Jiao, J., Zhang, Y., Song, S., Li, J., 2023b. Medtuning: Exploring parameter-efficient transfer learning for medical volumetric segmentation. arXiv preprint arXiv:2304.10880 .

Wang, X., Zhang, X., Cao, Y., Wang, W., Shen, C., Huang, T., 2023c. Seggpt: Segmenting everything in context. arXiv preprint arXiv:2304.03284 .

Wu, J., Fu, R., Fang, H., Liu, Y., Wang, Z., Xu, Y., Jin, Y., Arbel, T., 2023a. Medical sam adapter: Adapting segment anything model for medical image segmentation. arXiv preprint arXiv:2304.12620 .

Wu, Q., Zhang, Y., Elbatel, M., 2023b. Self-prompting large vision models for few-shot medical image segmentation. arXiv preprint arXiv:2308.07624 .

Yuan, L., Chen, D., Chen, Y.L., Codella, N., Dai, X., Gao, J., Hu, H., Huang, X., Li, B., Li, C., et al., 2021. Florence: A new foundation model for computer vision. arXiv preprint arXiv:2111.11432 .

Yue, W., Zhang, J., Hu, K., Xia, Y., Luo, J., Wang, Z., 2023. Surgicalsam: Efficient class promptable surgical instrument segmentation. arXiv preprint arXiv:2308.08746 .

Zaken, E.B., Goldberg, Y., Ravfogel, S., 2022. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models, in: Muresan, S., Nakov, P., Villavicencio, A. (Eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022, Association for Computational Linguistics. pp. 1–9.

Zhang, K., Liu, D., 2023. Customized segment anything model for medical image segmentation. arXiv preprint arXiv:2304.13785 .

Zhang, L., Liu, Z., Zhang, L., Wu, Z., Yu, X., Holmes, J., Feng, H., Dai, H., Li, X., Li, Q., et al., 2023. Segment anything model (sam) for radiation oncology. arXiv preprint arXiv:2306.11730 .

Zhou, H.Y., Guo, J., Zhang, Y., Han, X., Yu, L., Wang, L., Yu, Y., 2023a. nnformer: Volumetric medical image segmentation via a 3d transformer. IEEE Transactions on Image Processing .

Zhou, K., Yang, J., Loy, C.C., Liu, Z., 2022. Conditional prompt learning for vision-language models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16816–16825.

Zhou, T., Zhang, Y., Zhou, Y., Wu, Y., Gong, C., 2023b. Can sam segment polyps? arXiv preprint arXiv:2304.07583 .

Zou, X., Yang, J., Zhang, H., Li, F., Li, L., Gao, J., Lee, Y.J., 2023. Segment everything everywhere all at once. arXiv preprint arXiv:2304.06718 .