Auto-Prompting SAM for Mobile Friendly 3D Medical Image Segmentation

Chengyin Li [1], **Prashant Khanduri**[1], **Yao Qiang**[1], **Rafi Ibn Sultan**[1], **Indrin Chetty**[2], **Dongxiao Zhu**[1]

[1] Department of Computer Science, Wayne State University, Detroit, MI 48202 USA, dzhu@wayne.edu
[2] Henry Ford Health System, Detroit, MI 48202 USA

Abstract

The Segment Anything Model (SAM) has rapidly been adopted for segmenting a wide range of natural images. However, recent studies have indicated that SAM exhibits subpar performance on 3D medical image segmentation tasks. In addition to the domain gaps between natural and medical images, disparities in the spatial arrangement between 2D and 3D images, the substantial computational burden imposed by powerful GPU servers, and the time-consuming manual prompt generation impede the extension of SAM to a broader spectrum of medical image segmentation applications. To address these challenges, in this work, we introduce a novel method, AutoSAM Adapter, designed specifically for 3D multi-organ CT-based segmentation. We employ parameter-efficient adaptation techniques in developing an automatic prompt learning paradigm to facilitate the transformation of the SAM model's capabilities to 3D medical image segmentation, eliminating the need for manually generated prompts. Furthermore, we effectively transfer the acquired knowledge of the AutoSAM Adapter to other lightweight models specifically tailored for 3D medical image analysis, achieving state-of-the-art (SOTA) performance on medical image segmentation tasks. Through extensive experimental evaluation, we demonstrate the AutoSAM Adapter as a critical foundation for effectively leveraging the emerging ability of foundation models in 2D natural image segmentation for 3D medical image segmentation.

Introduction

Recently, computer vision foundation models like the Segment Anything Model (SAM) have further pushed the frontiers of image segmentation (Kirillov et al. 2023). SAM has demonstrated impressive performance and generalizability on a variety of semantic segmentation tasks (Zhang et al. 2023b), bringing new promise to medical image segmentation where the current approaches are limited by the quantity and quality of the segmentation masks. In contrast to existing custom-designed transformer models, such as UNETR (Hatamizadeh et al. 2022), SwinUNETR (Tang et al. 2022), and FocalUNETR (Li et al. 2023), that are trained only with a few patient samples and masks, foundation models including SAM are trained with billions of images and millions of masks. Such large-sized foundation models have also been observed to generalize to medical image segmentation tasks but with relatively worse performance compared to the state-of-the-art (SOTA) models in medical image segmentation (Gong et al. 2023).

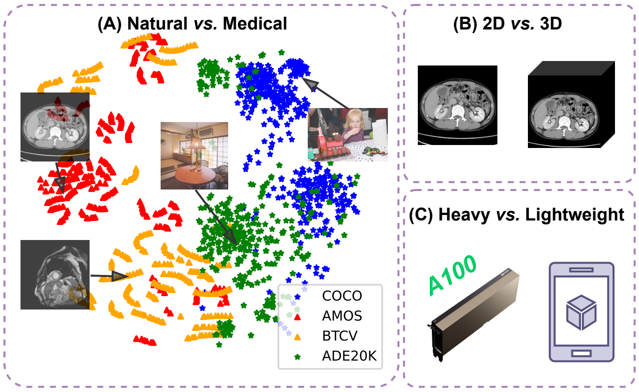


Figure 1: Challenges of using SAM for medical image segmentation, (A) T-SNE plot of embeddings encoded by SAM's image encoder for medical image datasets AMOS (Ji et al. 2022), and BTCV (Landman et al. 2015), and for natural image datasets ADE20K (Zhou et al. 2017) and COCO (Lin et al. 2014), (B) 2D image vs 3D volumetric input, and (C) heavy vs lightweight computing requirements.

Recent efforts have attempted to extend the success of SAM to medical image segmentation tasks including (Zhang and Jiao 2023; Wu et al. 2023; Ma and Wang 2023; Gong et al. 2023; Shaharabany et al. 2023). However, the demonstrated performance has exhibited reduced precision and stability, particularly in more intricate segmentation tasks characterized by smaller sizes, irregular shapes, and lower contrast properties of medical images in comparison to natural images (Gong et al. 2023). Adapting the original SAM architecture, which is rooted in 2D natural images, to effectively harness the 3D spatial information inherent in volumetric medical data poses a significant challenge. Novel approaches must be devised to bridge the gap between natural and medical image segmentation tasks, opening doors for the development of cutting-edge segmentation techniques. A few substantial issues (please see Fig. 1) that need to be addressed in developing a SAM-based medical image segmentation framework are, (A) encompassing the oversight of substantial domain disparities between natural and medical images (Fig. 1(A)), (B) extracting 3D spatial information from volumetric medical images effectively (Fig. 1(B)), and (C) the high computational demands even during inference (Fig. 1(C)). Furthermore, SAM's reliance on labor-intensive manually generated prompts (Gao et al. 2023; Shaharabany et al. 2023) hampers its successful application, particularly in multi-organ medical image segmentation tasks.

As healthcare becomes increasingly patient-centered and portable imaging devices like Computed Tomography (CT) and Magnetic Resonance Imaging (MRI) become more accessible, point-of-care tests (POCT) hold significant potential to enhance treatment effectiveness and efficiency by providing diagnoses at the patient's location. Especially in time-sensitive scenarios, POCT can substantially improve diagnosis and treatment processes, resulting in smoother and more efficient experiences for both patients and caregivers. Notably, portable 3D medical image segmentation techniques drive the functionality of POCTs, demanding the development of highly compressed models without compromising the segmentation performance.

To address the above issues, we introduce a novel AutoSAM Adapter method for a transition of SAM from 2D to 3D for medical image segmentation. Initially, we design intricate modifications for the image encoder at the input level, enabling the original 2D transformer to adeptly accommodate volumetric inputs while optimizing the reusability of pre-trained weights with a parameter-efficient adaptation method. Subsequently, at the prompt encoder level, we design an automatic prompt encoder module that takes the extracted feature maps from the previous image encoder as input and automatically learns the required prompts for the following mask encoder. This design effectively removes the time-consuming manual prompt generation process, especially, for multi-organ medical image segmentation tasks. Additionally, we prioritize a lightweight design for the mask decoder at the output level, emphasizing multi-layer aggregation. Through extensive experimentation on CT-based multi-organ segmentation datasets, inclusive of comprehensive comparisons with state-of-the-art approaches including nn-UNet (Isensee et al. 2021), as well as recent adapters in the field, our results exhibit a significant performance improvement over existing techniques. Finally, we apply knowledge distillation (KD) to transfer the learned knowledge from AutoSAM Adapter to other lightweight models like SwinUNETR (Tang et al. 2022) for the efficiency-aware POCT use scenario. The main contributions of this work are summarized below.

- To tackle the domain gap between 2D and 3D inputs, we introduce a 3D adaptor to extract spatial information for volumetric segmentation in medical images.
- To the best of our knowledge, this is the first work to adapt the SAM model for 3D-based multi-organ segmentation that can automatically learn prompts without the laborious manual prompting process.
- To facilitate mobile-friendly use scenarios for POCT, we employ KD to train lightweight models tailored for point-of-care medical image segmentation applications.
- Extensive experiments and analysis on the AMOS and BTCV CT-based multi-organ segmentation datasets demonstrate that the proposed AutoSAM Adapter and its lightweight version achieve superior performance on medical image segmentation tasks compared to SOTA.

Related Work

Foundation Computer Vision Models. With advancements in deep learning models, most contemporary vision frameworks adhere to the pre-training and fine-tuning paradigm (Min et al. 2021). Recently, computer vision researchers have shown substantial interest in large and adaptable foundational models, capitalizing on pre-training techniques such as self-supervised learning (Jing and Tian 2020), contrastive learning (Wang and Qi 2022), and language-vision pre-training (Radford et al. 2021), among others. Notably, the SAM model (Kirillov et al. 2023), recently pre-trained on a dataset of over 11 million images, has emerged as a versatile foundational model for natural image segmentation. SAM demonstrates impressive zero-shot capabilities in segmenting diverse subjects in real-world environments, using an interactive and prompt-driven approach. Additionally, SEEM (Zou et al. 2023), another contemporaneous effort to SAM, introduces a more comprehensive prompting scheme to facilitate semantic-aware open-set segmentation. Furthermore, DINOv2 (Oquab et al. 2023) focuses on scaling up the pre-training of a ViT model in terms of data and model size. This approach aims to generate versatile visual features that simplify the fine-tuning of the downstream tasks.

Parameter-efficient Model Fine-tuning. Given the extensive utilization of foundational models, the concept of parameter-efficient fine-tuning has garnered significant attention. Existing methods for efficient fine-tuning can be categorized into three groups (Ding et al. 2023). Addition-based methods that involve incorporating lightweight adapters (Pan et al. 2022; Wang et al. 2023) or prompts (Liu et al. 2023; Jia et al. 2022) into the original model, with the sole focus on adjusting these parameters; Specification-based methods (Zaken, Ravfogel, and Goldberg 2021; Guo, Rush, and Kim 2020) that concentrate on selecting a small subset of the original parameters for tuning; and reparameterization-based methods (Hu et al. 2021) that leverage low-rank matrices to approximate parameter updates. In recent times, a few researchers have extended pre-trained image models to encompass video comprehension (Pan et al. 2022) or volumetric segmentation (Wang et al. 2023). Nevertheless, these methods treat the additional dimension as a “word group” and employ specialized modules to aggregate information along the word dimension. In contrast, in our work, we consider all three dimensions as isotropic and directly adapt the trained transformer block to capture 3D patterns.

SAM-based Medical Image Segmentation. This line of work primarily focuses on enhancing SAM through fine-tuning for specific segmentation datasets, aiming to mitigate the noticeable performance drop of SAM on medical images. MedSAM (Ma and Wang 2023) specifically concentrates on refining the SAM decoder by employing prompts generated from label masks across more than 30 medical image datasets. The outcome demonstrates improved performance compared to zero-shot predictions using prompts. Zhang et al. (Zhang and Liu 2023) opt for a low-rank fine-tuning approach, focusing on the SAM encoder. By combining this strategy with SAM decoder training, they tailor SAM for abdominal segmentation tasks. Junde Wu et al.(Wu et al. 2023) follow a distinct path, wherein they freeze SAM's weights and incorporate a trainable adaptation module within SAM to mitigate the need for complete re-training. Despite some progress, these methods either ignore the 3D pattern of medical images or require a laborious manual prompt generation process, which restricts the full potential of SAM from being realized in the medical image segmentation domain.

Lightweight Models for POCT. Due to the limited computation power in POCT usage scenarios, directly using the SAM model for medical segmentation is not feasible. Instead of using the tailored network architectures like Mobile ViT (Mehta and Rastegari 2021), MobileNet (Howard et al. 2017), and EfficientNet (Tan and Le 2019), compressing large models during training stages into lightweight ones is a promising strategy. One popular approach for in-training model compression is knowledge distillation (KD) (Hinton, Vinyals, and Dean 2015) where a full teacher model is trained on the cloud or an on-premise GPU cluster, and a student model is trained at the mobile device with the "knowledge" distilled via the soft labels from the teacher model. Thus the student model is trained to mimic the outputs of the teacher model as well as to minimize the cross-entropy loss between the true labels and predictive probabilities (soft labels). KD yields compact student models that demonstrate outstanding performance in a wide range of real-world applications, e.g., COVID-MobileXpert (Li, Li, and Zhu 2020), on-device text classification (Qiang et al. 2022), etc. Recently, the technique has also been applied to develop mobile SAM (Zhang et al. 2023a).

Methodology

In this section, we explain how to modify the original SAM architecture developed for 2D natural images to work with 3D volumetric medical images for segmentation tasks. We first provide a brief overview of the SAM framework (as shown in Fig. 2), followed by a detailed explanation of the adjustments made to the image encoder, auto prompt generator, and mask decoder.

The SAM Architecture

SAM (Kirillov et al. 2023) is a prompt-driven image segmentation framework, known for its impressive performance and generalization ability in segmenting natural images. The SAM architecture comprises an image encoder, prompt encoder, and mask decoder. The image encoder utilizes the Vision Transformer (ViT) to transform original images into discrete embeddings. The prompt encoder converts diverse prompts into compact embeddings, achieved by combining fixed positional encoding and adaptable prompt-specific embedding. The mask decoder integrates a prompt self-attention module and bidirectional cross-attention modules for prompt-to-image and image-to-prompt attention. After attention processes, the feature map undergoes upsampling and passes through a multi-layer perceptron (MLP) to generate segmentation masks. However, the model's design suits 2D natural image segmentation, struggling with 3D volumetric medical imagery due to slice-wise predictions that disregard inter-slice spatial context. Performance on medical images also falters due to domain disparities between medical and natural images. Therefore, achieving an effective performance with SAM on medical-imaging tasks requires tailored adaptation and fine-tuning.

Handling 3D Volumetric Inputs

The original SAM model is built upon the 2D Vision Transformer (ViT), excelling in capturing 2D image patterns but facing limitations with 3D medical imaging like CT and MRI. These modalities produce volumetric 3D data, challenging the 2D ViT's processing ability. Common medical imaging workflows analyze images slice by slice, integrating information using spatial adaptors or temporal modules, yet the core architecture remains 2D-centric.

However, for medical image analysis, the 2D-centric approach falls short due to the inherent 3D nature of volumetric medical images, which have uniform spatial resolutions across dimensions. To address this, we propose an adaptation strategy (Fig. 2A and Fig. 2B) with two aims: enabling the model to directly learn 3D spatial patterns and maintaining continuity by inheriting most parameters from the pre-trained model, introducing easily adjustable incremental parameters:

- **Positional Encoding:** The pre-trained ViT model has a lookup table of size $C \times H \times W$ with positional encoding. Furthermore, we initialize a tunable lookup table of size $C \times D$ with zeros. To obtain the positional encoding of a 3D point (d, h, w), we add the embedding from the frozen lookup table with (h, w) to the embedding from the tunable lookup table with (d).
- **Patch Embedding:** We utilize a combination of $1 \times k \times k$ and $k \times 1 \times 1$ 3D convolutions to approximate the effect of a $k \times k \times k$ convolution (e.g., $k = 14$). The $1 \times k \times k$ convolution is initialized with the weights from a pre-trained 2D convolution and remains unaltered during the fine-tuning phase. As for the newly introduced $k \times 1 \times 1$ 3D convolution, we apply depth-wise convolution to decrease the number of parameters that need adjustment. This approach helps in managing the complexity of the model.
- **Attention Block:** The attention blocks can be directly adjusted to accommodate 3D features. In the case of 2D inputs, the size of the queries is $[B, HW, C]$, which can be effortlessly modified to $[B, DHW, C]$ for 3D inputs, while retaining all the pre-trained weights. We implement sliding-window mechanisms akin to those in the SwinUNETR (Tang et al. 2022) to mitigate the memory impact resulting from the increase in dimensions. This

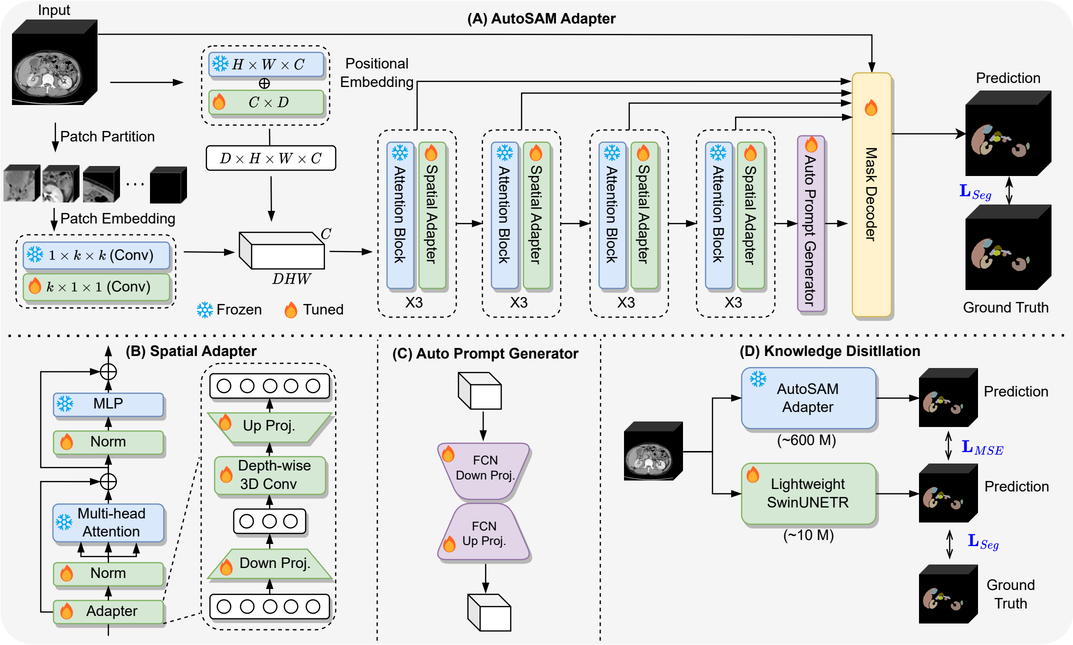


Figure 2: (A) The overall AutoSAM Adapter design, (B) the architecture of spatial adapter module, (C) the architecture of Auto Prompt Generator, and (D) the pipeline of deriving lightweight SwinUNETR from AutoSAM Adapter through the knowledge distillation process.

approach aids in optimizing the model's performance while managing memory requirements.

- **Bottleneck:** Given that convolution layers are generally easier to optimize than transformers, we replace 2D convolutions in the bottleneck with 3D counterparts and train them from scratch to improve performance.

By making the above adjustments, we can smoothly transition the 2D ViT into a 3D ViT, reusing most parameters. However, fully fine-tuning the 3D ViT can be resource-intensive. To address this, we propose using a lightweight adapter approach for efficient fine-tuning. The adapter comprises a down-projection linear layer and an up-projection linear layer, represented as $Adapter(\mathbf{X}) = \mathbf{X} + Act(\mathbf{X}W_{down})W_{up}$. Here, $\mathbf{X} \in \mathbb{R}^{N \times C}$ is the original feature representation, $W_{down} \in \mathbb{R}^{C \times N'}$ and $W_{up} \in \mathbb{R}^{N' \times C}$ are down-projection and up-projection layers, and $Act(\cdot)$ is the activation function (e.g., $ReLu$). To enhance 3D spatial awareness, we include a depth-wise 3D convolution after the down-projection layer, as shown in Fig. 2B. This enhancement improves the adapter's utilization of 3D spatial cues.

Throughout the training phase, we exclusively adjust the parameters of convolutions, spatial adapters, and normalization layers, while maintaining all other parameters in a frozen state. This frozen approach enhances memory efficiency during training. Fine-tuning the adapter and normalization layers aids in bridging the gap between natural images and medical images, enabling the model to adapt more effectively to the medical image domain.

Auto Prompt Generator

The original SAM model utilizes positional embedding to represent the prompt, applying it to both the prompt and the image. This guarantees that prompt and image embeddings for the same position share identical positional encoding. Subsequently, the prompt embedding engages in cross-attention with the image embedding, evolving from positional to semantic attributes. However, this cross-attention, though effective in 2D settings, can trigger over-smoothing issues when extended to 3D feature maps. Adapting to 3D can significantly inflate token numbers, leading to a uniform probability distribution.

Prompt-based segmentation might not be suitable for real-world applications due to two main reasons. Firstly, it becomes time-consuming for multi-class prompts. In situations involving multiple classes, generating prompts becomes a time-intensive task. Many public medical image segmentation challenges necessitate simultaneous segmentation of multiple classes. Precisely specifying prompts for each class can be challenging, especially for small or closely located organs or tissues. Additionally, note that segmentation performance heavily relies on the quality of provided prompts, however, prompt quality is difficult to control since crafting accurate prompts demands domain-specific expertise, which might not be universally available. This limitation hampers the effectiveness of prompt-based approaches, particularly for non-expert users.

In pursuit of these objectives, we propose to use an Auto Prompt Generator instead of positional encoding to represent the prompt. The whole process is illustrated in Fig. 2C. Instead of using manually generated points or bounding boxes, we directly take the output feature map after the last block of attention and spatial adapter operation. This Auto Prompt Generator follows a fully convolutional neural (FCN) based encoder-decoder design that resembles 3D UNet (Ronneberger, Fischer, and Brox 2015). This generator boasts a lightweight structure, leveraging 3D-based convolution operations, and can be effortlessly learned from scratch. This enables precise prompt generation tailored to different medical segmentation tasks. Notably, it eliminates the need for additional manually generated prompts, simplifying and expediting the multi-class medical image segmentation tasks.

Lightweight Mask Decoder

The mask decoder in SAM is intentionally lightweight, employing stacks of convolution layers. We update this design by replacing all 2D convolutions with 3D convolutions, enabling direct 3D mask generation. The initial decoder, devoid of progressive upsampling or skip connections, is effective for natural images where object sizes are generally substantial, and boundaries are distinct. Nonetheless, in the context of volumetric medical image segmentation, it's widely recognized that U-shaped networks featuring skip connections at multiple levels are crucial (Isensee et al. 2021). This is due to the fact that medical image objects are often diminutive, and their boundaries are frequently indistinct. Consequently, such images demand networks capable of preserving higher-resolution details for improved discrimination, making the adoption of U-shape architectures with skip connections imperative.

To tackle this challenge while maintaining a lightweight design, we utilize a multi-layer aggregation mechanism (Zheng et al. 2021) in our decoder. Here, the encoder's intermediate outputs are concatenated, enriching the mask feature map without compromising model efficiency. For enhanced resolution information, we upsample the mask feature map to match the original resolution. This upsampled map is concatenated with the original image and fused using another 3D convolution, generating the final mask. This strategy seamlessly integrates high-resolution details and original image data into the mask-generation process. We simplify the original SAM by removing multi-masks generation and ambiguity awareness, aiming to fine-tune it for a specific downstream task. The mask decoder's backbone predominantly consists of lightweight 3D convolutional layers, known for their optimization-friendliness. Hence, we train all parameters from scratch.

Knowledge Distillation

Despite our efforts to design significantly lightweight modules compared to the original SAM, there remains a challenge in reducing the initial weight complexity within SAM's ViT encoder segment. This encoder component holds a significant portion of the model's parameters, making it difficult to seamlessly integrate the AutoSAM Adapter into POCT.

Inspired by the simplicity of KD techniques and the availability of medical segmentation-specific model designs, we take an additional step forward. Our objective is to transfer the accumulated knowledge from the larger AutoSAM Adapter (around 600 million parameters) to a much smaller SwinUNETR model (around 10 million parameters). This approach aims to bridge the gap between complex models and resource-efficient solutions, fostering advancements in practical medical image segmentation within the academic realm.

Loss Function

For training the AutoSAM Adapter (as shown in Fig. 2A), a combination of Dice loss and Cross-Entropy loss is used to assess the alignment between the predicted mask and the ground truth on a pixel-wise basis. The objective function for the segmentation head is defined as follows:

$$\mathcal{L}_{seg} = \mathcal{L}_{dice}(\hat{p}_i, g_i) + \mathcal{L}_{ce}(\hat{p}_i, g_i), \tag{1}$$

where $\hat{p}_i$ represents the predicted probabilities from the main task, and g_i represents the ground truth mask for an input volume i. The predicted probabilities, $\hat{p}_i$, result from applying the AutoSAM Adapter to the input 3D volume for the main task.

Regarding the KD process (as illustrated in Fig. 2D), we adopt the following formulation:

$$\mathcal{L}_{tol} = \lambda\mathcal{L}_{seg} + (1-\lambda)\mathcal{L}_{mse}, \tag{2}$$

where λ serves as a hyperparameter regulating how much the lightweight SwinUNETR model should learn from both the prediction mask generated by the AutoSAM Adapter and the ground truth and $\mathcal{L}_{mse} = \frac{1}{N}\sum_{i=1}^{N}(\hat{p}_i, g_i)^2$. This approach enables the transfer of knowledge from the AutoSAM Adapter to the SwinUNETR model while striking a balance between the two information sources.

Experiments

Datasets and Evaluation Metrics

BTCV Dataset. Beyond the Cranial Vault (BTCV) abdomen challenge dataset (Landman et al. 2015) includes 30 subjects with abdominal CT scans. In this dataset, 13 organs are annotated by interpreters under the supervision of radiologists at Vanderbilt University Medical Center. Each CT scan is acquired during the portal venous contrast enhancement phase and consists of 80 to 225 slices. These slices have dimensions of 512x512 pixels, and the slice thickness ranges from 1 to 6 mm. The multi-organ segmentation task is framed as a 13-class segmentation challenge with 24 scans for training and 6 scans for testing.

AMOS Dataset. We also employ the publicly accessible AMOS2022 dataset (Ji et al. 2022). It consists of 200 multi-contrast abdominal CT scans for training and 100 scans for testing, sourced from AMOS 2022. These scans are annotated for sixteen anatomies, enabling assessment of abdominal multi-organ segmentation.

Evaluation Metrics. We utilize the Dice coefficient and the Normalized Surface Distance (NSD) (Nikolov et al. 2018) as metrics to evaluate the segmentation performance. The NSD metric quantifies the agreement between ground

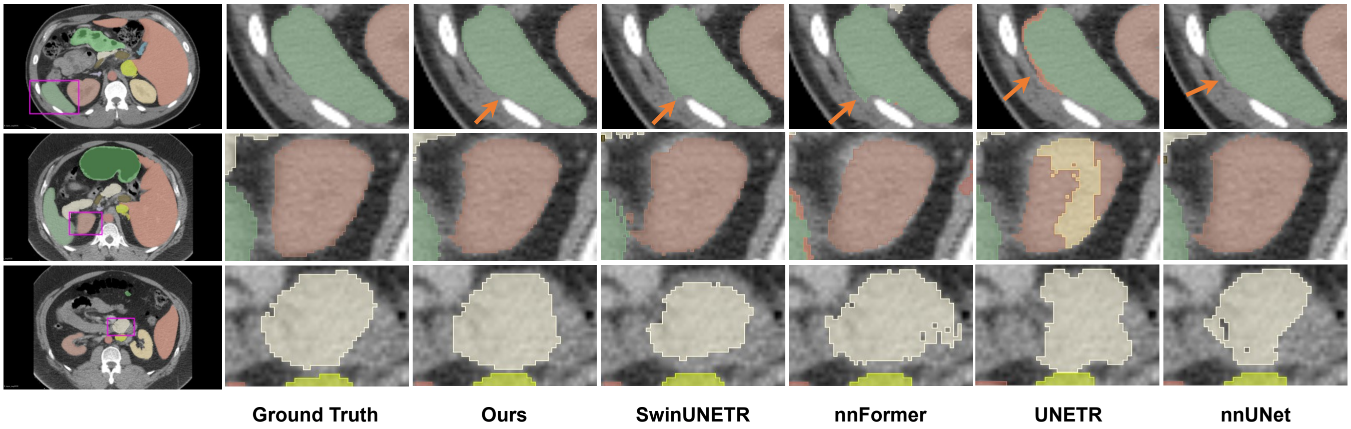


Figure 3: Qualitative visualizations of the proposed AutoSAM Adaptor (ours) and baseline methods. Three representative subjects are demonstrated. Regions of evident improvements are enlarged to show better details of spleen (light green), left kidney (light red), and pancreas (beige).

Model	Tuned Params.	BTCV		AMOS	
		mDice(%)	mNSD(%)	mDice(%)	mNSD(%)
nnUNet	31.18M	84.34	73.21	87.43	77.12
nnFormer	150.14M	83.51	71.65	84.52	70.06
UNETR	93.02M	85.47	74.35	77.24	60.58
SwinUNETR	62.83M	86.58	75.26	86.19	74.83
Ours	26.53M	**87.15**	**78.83***	**88.65***	**79.41***

Table 1: Comparison of the overall performance of four SOTA approaches to AutoSAM ADapter (ours) on BTCV and AMOS datasets, respectively. The best results are presented in bold font. (*: $p < 0.01$, with Wilcoxon signed-rank test to all SOTA approaches)

truth and predicted surfaces, considering a fixed tolerance. Unlike comparing two volumes, this metric assesses the overlap between surface structures.

Implementation Details

We implement our approach and establish baseline comparisons using both PyTorch and MONAI frameworks. All experiments and comparisons employ SAM-B, utilizing ViT-B as the backbone for the image encoder. Model training is conducted with a batch size of 1 on NVIDIA A100 GPUs, utilizing the AdamW optimizer (Loshchilov and Hutter 2017). A learning rate scheduler with exponential decay, incorporating 5 epochs of warmup and a maximum of 200 epochs, is employed. The initial learning rate is set at $5e^{-4}$, with a momentum of 0.9 and weight decay of $1e^{-5}$. A Houndsfield unit (HU) range of $[-125, 275]$ is normalized to the interval $[0, 1]$ for the BTCV dataset (Tang et al. 2022). Following the procedure outlined in (Ji et al. 2022), HU values for each scan in the AMOS dataset are clipped to the range $[-991, 362]$. Subsequently, truncated voxel values are normalized by subtracting 50 and dividing by 141. For both datasets, all CT scans are interpolated into an isotropic voxel spacing of $[1.0 \times 1.0 \times 1.5]$ mm, and each CT scan is then cropped to a $128 \times 128 \times 128$ input patch for 3D models. Data augmentation includes random flip, rotation, and intensity scaling with probabilities of 0.1, 0.1, and 0.2, respectively. During training, foreground and background patches are randomly sampled at a $1 : 1$ ratio. Our method's performance is evaluated by comparing it against SOTA volumetric segmentation approaches.

Comparison with SOTA

We extensively compare our model with the SOTA 3D medical image segmentation approaches, including the most recent Transformer-based methods including UNETR (Hatamizadeh et al. 2022), SwinUNETR (Tang et al. 2022), and nnFormer (Zhou et al. 2023), as well as CNN-based methods such as nnUNet (Isensee et al. 2021). As reported in Table 1, we observe that the proposed AutoSAM Adapter outperforms all other SOTA methods in both Dice and NSD metrics for both BTCV and AMOS datasets. Distinct improvements can be specifically observed for the BTCV dataset, e.g., $1\% \sim 3\%$ improvement for the average Dice score and $3\% \sim 7\%$ improvement for the average NSD metric. With the increase of training samples for the AMOS dataset compared with the BTCV dataset, the proposed AutoSAM Adapter can even achieve better overall performance, i.e., $1\% \sim 14\%$ improvement for Dice and $2\% \sim 19\%$ improvement for NSD. This is also evident from Fig. 3, which shows qualitative comparisons over the predicted masks for different segmentation models. The AutoSAM Adapter demonstrates visually better mask prediction results with more accurate boundaries over the competing SOTA approaches. In contrast to the original SAM, which demonstrates subpar performance on medical image segmentation tasks compared to the SOTA (Mazurowski et al. 2023), our design shows significant improvements.

Comparison with Existing Adapters

We further compare our adaptation strategy with existing adaptation methods for multi-class medical image segmentation, which include 2D adaptations such as MedSAM (Ma

Model	Tuned Option	BTCV		AMOS	
		mDice(%)	mNSD(%)	mDice(%)	mNSD(%)
SAM point	None	54.86	-	49.31	-
MedSAM point	P&M	80.32	66.31	70.0	58.45
MedSAM point	Full	84.32	73.63	82.65	70.13
Ours	P&M	**87.15**	**78.83**	**88.65**	**79.41**

Table 2: Comparison with existing existing adaptation methods. The best results are bolded. P&M: only fine-tuning the prompt encoder and mask decoder, Full: fully fine-tuning, and None: no fine-tuning.

and Wang 2023). Other methods are not considered since they either do not have publicly available code or are not designed for multi-organ segmentation. MedSAM can be implemented with full fine-tuning or can do partial fine-tuning by only updating the prompt encoder and mask decoder. We also compare our AutoSAM Adapter with the original SAM. All the pre-trained weights are from SAM-B.

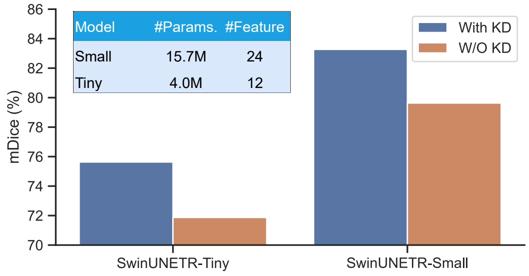
The outcomes are meticulously outlined in Table 2, underscoring how our adaptation strategy surpasses all existing methods. Notably, our approach outshines the second-best technique by a margin of 3% in terms of BTCV segmentation Dice, 5% for NSD, and 6% concerning AMOS segmentation Dice, accompanied by a substantial 9% for NSD. Impressively, it even outperforms the complete fine-tuning variant of MedSAM, even when considering parameter-efficient fine-tuning. These results effectively validate our hypothesis that parameters pre-trained on 2D images can be effectively harnessed to grasp 3D spatial features with only minor adjustments. Moreover, our approach of treating all dimensions equivalently emerges as a superior strategy compared to interpreting the depth dimension as a distinct group in the context of medical image segmentation.

Lightweight Models via Knowledge Distillation

To address the requirement for lightweight models in the context of POCT, we have taken a further step towards compressing the AutoSAM Adapter into lightweight SwinUNETRs (specifically, the tiny or small versions) using a straightforward KD process, as illustrated in Fig. 2D. For the experiments, the tuning parameter λ has been set to 0.5 for the KD learning process. Other training strategies remain consistent with those employed in optimizing the AutoSAM Adapter. The outcomes pertaining to BTCV's average Dice scores, both with and without the utilization of KD, are shown in Fig. 4. When compared to their KD-absent counterparts, models incorporating KD demonstrate a marked improvement in the average Dice scores. Specifically, SwinUNETR-Tiny (with a feature size of 12 and 4.0M parameters) displays an approximate 4% enhancement, while SwinUNETR-Small (with a feature size of 24 and 15.7M parameters) exhibits a comparable advancement.

Ablation Study

Effects of Auto Prompt Generator. Given that the Auto Prompt Generator leverages feature maps from the final stage of the attention block in the image encoder, these feature maps can be employed as direct inputs for the mask decoder. We conducted a comparative analysis of performance with and without the utilization of the Auto Prompt Generator on the BTCV dataset. The outcomes are presented in Table 3, revealing noteworthy enhancements in both Dice (approximately 2%) and NSD (around 3%) metrics when the Auto Prompt Generator is employed.



Model	#Params.	#Feature
Small	15.7M	24
Tiny	4.0M	12

Figure 4: The comparison of Dice metric on BTCV dataset with/without using the KD for lightweight models.

Setting	mDice(%)	mNSD(%)
With Auto Prompt Generator	**87.15**	**78.83**
Without Auto Prompt Generator	85.23	74.12

Table 3: Comparison with/without Auto Prompt Encoder.

Effects of λ for Knowledge Distillation. By tuning λ in Equation (2), we can change the weight for the SwinUNETR-small learn from the ground truth of BTCV dataset or from the AutoSAM Adapter teacher. With the increase of λ, the performance increases first and decreases after $\lambda = 0.5$ (Table 4).

λ	0	0.2	0.5	0.8	1.0
mDice (%)	78.54	82.45	**83.27**	81.14	79.63

Table 4: The impact of λ for the KD process.

Conclusion

In this study, we proposed AutoSAM Adapter architecture for 3D-based multi-organ medical image segmentation. Adapting the Segment Anything Model (SAM) from 2D to 3D medical images poses challenges in domain differences, spatial disparities, computation demands, and manual prompt generation complexities. Our approach overcomes these hurdles through parameter-efficient adaptation techniques and an automatic prompt generation framework to simplify SAM's application for 3D-based multi-organ segmentation tasks. The knowledge gained through a KD process also enhances the performance of other lightweight 3D medical image segmentation models. With comprehensive experiments, we validated the effectiveness of the proposed AutoSAM Adapter, thereby establishing a sturdy foundation for the advancement of image segmentation within the intricate landscape of medical imaging.

References

Ding, N.; Qin, Y.; Yang, G.; Wei, F.; Yang, Z.; Su, Y.; Hu, S.; Chen, Y.; Chan, C.-M.; Chen, W.; et al. 2023. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5(3): 220–235.

Gao, Y.; Xia, W.; Hu, D.; and Gao, X. 2023. DeSAM: Decoupling Segment Anything Model for Generalizable Medical Image Segmentation. *arXiv preprint arXiv:2306.00499*.

Gong, S.; Zhong, Y.; Ma, W.; Li, J.; Wang, Z.; Zhang, J.; Heng, P.-A.; and Dou, Q. 2023. 3DSAM-adapter: Holistic Adaptation of SAM from 2D to 3D for Promptable Medical Image Segmentation. *arXiv preprint arXiv:2306.13465*.

Guo, D.; Rush, A. M.; and Kim, Y. 2020. Parameter-efficient transfer learning with diff pruning. *arXiv preprint arXiv:2012.07463*.

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

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Howard, A. G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; and Adam, H. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

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

Isensee, F.; Jaeger, P. F.; Kohl, S. A.; Petersen, J.; and Maier-Hein, K. H. 2021. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2): 203–211.

Ji, Y.; Bai, H.; Ge, C.; Yang, J.; Zhu, Y.; Zhang, R.; Li, Z.; Zhanng, L.; Ma, W.; Wan, X.; et al. 2022. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *Advances in Neural Information Processing Systems*, 35: 36722–36732.

Jia, M.; Tang, L.; Chen, B.-C.; Cardie, C.; Belongie, S.; Hariharan, B.; and Lim, S.-N. 2022. Visual prompt tuning. In *European Conference on Computer Vision*, 709–727. Springer.

Jing, L.; and Tian, Y. 2020. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(11): 4037–4058.

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.; and 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*, volume 5, 12.

Li, C.; Qiang, Y.; Sultan, R. I.; Bagher-Ebadian, H.; Khanduri, P.; Chetty, I. J.; and Zhu, D. 2023. FocalUNETR: A Focal Transformer for Boundary-aware Segmentation of CT Images. arXiv:2210.03189.

Li, X.; Li, C.; and Zhu, D. 2020. COVID-MobileXpert: On-device COVID-19 patient triage and follow-up using chest X-rays. In *2020 IEEE international conference on bioinformatics and biomedicine (BIBM)*, 1063–1067. IEEE.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Liu, P.; Yuan, W.; Fu, J.; Jiang, Z.; Hayashi, H.; and Neubig, G. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9): 1–35.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Ma, J.; and Wang, B. 2023. Segment anything in medical images. *arXiv preprint arXiv:2304.12306*.

Mazurowski, M. A.; Dong, H.; Gu, H.; Yang, J.; Konz, N.; and Zhang, Y. 2023. Segment anything model for medical image analysis: an experimental study. *Medical Image Analysis*, 102918.

Mehta, S.; and Rastegari, M. 2021. Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer. *arXiv preprint arXiv:2110.02178*.

Min, B.; Ross, H.; Sulem, E.; Veyseh, A. P. B.; Nguyen, T. H.; Sainz, O.; Agirre, E.; Heintz, I.; and Roth, D. 2021. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*.

Nikolov, S.; Blackwell, S.; Zverovitch, A.; Mendes, R.; Livne, M.; De Fauw, J.; Patel, Y.; Meyer, C.; Askham, H.; Romera-Paredes, B.; et al. 2018. Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy. *arXiv preprint arXiv:1809.04430*.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

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

Qiang, Y.; Kumar, S. T. S.; Brocanelli, M.; and Zhu, D. 2022. Tiny rnn model with certified robustness for text classification. In *2022 International Joint Conference on Neural Networks (IJCNN)*, 1–8. IEEE.

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*, 8748–8763. PMLR.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, 234–241. Springer.

Shaharabany, T.; Dahan, A.; Giryes, R.; and Wolf, L. 2023. AutoSAM: Adapting SAM to Medical Images by Overloading the Prompt Encoder. *arXiv preprint arXiv:2306.06370*.

Tan, M.; and Le, Q. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, 6105–6114. PMLR.

Tang, Y.; Yang, D.; Li, W.; Roth, H. R.; Landman, B.; Xu, D.; Nath, V.; and Hatamizadeh, A. 2022. 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*, 20730–20740.

Wang, W.; Shen, J.; Chen, C.; Jiao, J.; Zhang, Y.; Song, S.; and Li, J. 2023. Med-Tuning: Exploring Parameter-Efficient Transfer Learning for Medical Volumetric Segmentation. *arXiv preprint arXiv:2304.10880*.

Wang, X.; and Qi, G.-J. 2022. Contrastive learning with stronger augmentations. *IEEE transactions on pattern analysis and machine intelligence*, 45(5): 5549–5560.

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

Zaken, E. B.; Ravfogel, S.; and Goldberg, Y. 2021. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. *arXiv preprint arXiv:2106.10199*.

Zhang, C.; Han, D.; Qiao, Y.; Kim, J. U.; Bae, S.-H.; Lee, S.; and Hong, C. S. 2023a. Faster Segment Anything: Towards Lightweight SAM for Mobile Applications. *arXiv preprint arXiv:2306.14289*.

Zhang, C.; Liu, L.; Cui, Y.; Huang, G.; Lin, W.; Yang, Y.; and Hu, Y. 2023b. A Comprehensive Survey on Segment Anything Model for Vision and Beyond. *arXiv preprint arXiv:2305.08196*.

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

Zhang, Y.; and Jiao, R. 2023. How Segment Anything Model (SAM) Boost Medical Image Segmentation? *arXiv preprint arXiv:2305.03678*.

Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P. H.; et al. 2021. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6881–6890.

Zhou, B.; Zhao, H.; Puig, X.; Fidler, S.; Barriuso, A.; and Torralba, A. 2017. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 633–641.

Zhou, H.-Y.; Guo, J.; Zhang, Y.; Han, X.; Yu, L.; Wang, L.; and Yu, Y. 2023. nnFormer: Volumetric Medical Image Segmentation via a 3D Transformer. *IEEE Transactions on Image Processing*.

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