AFTer-SAM: Adapting SAM with Axial Fusion Transformer for Medical Imaging Segmentation

Xiangyi Yan Shanlin Sun Kun Han Thanh-Tung Le Haoyu Ma Chenyu You Xiaohui Xie
University of California, Irvine Yale University
{xiangyy4, shanlins, khan7, thanhtul, haoyum3, xhx}@uci.edu
chenyu.you@yale.edu

Abstract

The Segmentation Anything Model (SAM) has demonstrated effectiveness in various segmentation tasks. However, its application to 3D medical data has posed challenges due to its inherent design for both 2D and natural images. While there have been attempts to apply SAM to medical images on a slice-by-slice basis, the outcomes have been less than optimal. In this study, we introduce AFTer-SAM, an adaptation of SAM designed for volumetric medical image segmentation. By incorporating an Axial Fusion Transformer, AFTer-SAM is capable of capturing both intra-slice details and inter-slice contextual information, essential for accurate medical image segmentation. Given the potential computational challenges of training this enhanced model, we utilize Low-Rank Adaptation (LoRA) to efficiently fine-tune the weights of the Axial Fusion Transformer. This ensures a streamlined training process without compromising on performance. Our results indicate that AFTer-SAM offers significant improvements in volumetric medical image segmentation, suggesting a promising direction for the application of large pre-trained models in medical imaging.

1. Introduction

Medical image segmentation serves as a critical nexus between cutting-edge clinical procedures and technological evolution. Its role is indispensable in numerous clinical applications, from diagnosis to treatment planning [10]. However, the inherent variability in medical images, characterized by diverse organ shapes and intricate tissue structures, poses significant challenges.

In the past, atlas-based methods were the mainstay for medical image segmentation. While groundbreaking for their era, these methods struggled to accommodate the vast anatomical differences arising from factors such as tumor growths, surgical alterations, and imaging inconsistencies. The advent of deep learning brought forth the prowess of Convolutional Neural Networks (CNNs), with models like U-Net [26] leading the charge. While these models excelled in feature extraction, they often fell short in capturing essential long-range interactions, especially as medical images became more complex.

The rise of transformer architectures [31], inspired by their success in the Natural Language Processing (NLP) domain, presented a potential solution to this limitation. Transformers, with their innate ability to model long-range dependencies, appeared tailor-made for the challenges of medical image segmentation.

Enter the "Segment Anything" (SAM) model [19], a paradigm shift aiming to equip models with the versatility to segment a wide variety of images. SAM's adaptability, showcased by its impressive zero-shot capabilities, makes it a potent tool for diverse and novel image distributions.

However, as models like SAM grow in complexity, the challenge of adaptation, especially in large-scale pre-training scenarios, becomes pronounced. Comprehensive fine-tuning of such models can be resource-intensive. This is where techniques like Low-Rank Adaptation (LoRA) [14] come into play. By judiciously freezing certain pre-trained weights and introducing trainable rank decomposition matrices, LoRA significantly reduces the number of trainable parameters, optimizing memory usage and computational efficiency. This approach ensures that the model remains agile and precise across different tasks.

In this context, we present AFTer-SAM, a novel fusion of the SAM model with axial fusion transformer [37] techniques. This model is designed to harness both the adaptability of SAM and the depth-aware insights of transformers, making it a formidable tool for medical image segmentation. Our key contributions are:

- We introduced the Axial Fusion Transformer to SAM, enhancing its capability from 2D to 3D segmentation. This allows for the adept fusion of both intra-slice and inter-slice information.

- We successfully adapted the heavily trained SAM parameters from natural images to medical imaging. This approach eliminates the need for costly fine-tuning while preserving the original SAM's segmentation prowess.

- We conducted comprehensive evaluations on extensive medical image datasets, demonstrating that AFTer-SAM surpasses the performance of prior state of the art methods.

2. Related Work

2.1. Transformers for medical image segmentation

Medical image segmentation has undergone significant transformation with the rise of deep learning [35,36]. While conventional techniques have their value, the versatility and adaptability of deep learning models, particularly when paired with transformers, have established new standards in the field [40,42].

TransUNet [4] was a trailblazing attempt to merge the capabilities of Transformers with the U-Net [26] structure for medical image segmentation. Its proficiency in handling long-range dependencies is notable. However, its direct use for volumetric medical data poses computational challenges. Furthermore, its primary focus on 2D slices might overlook the comprehensive context offered by 3D volumes. CoTr [34] brought forth the deformable self-attention mechanism, targeting a reduction in the computational demands of 3D self-attention. Yet, its dependence on 3D patches for inputs might result in potential data loss. Moreover, the restricted locations over which CoTr processes self-attention could miss capturing the nuanced details present in medical images. Expanding on the Swin Transformer's concepts, [30] introduced a self-supervised learning strategy using Swin UNETR for 3D medical image analysis. By devising specialized proxy tasks to understand human anatomical patterns, the team effectively pre-trained their model on an extensive CT image dataset. The model's prowess is further confirmed by its top-tier performance in medical image segmentation tasks. This method highlights the promise of self-supervised learning in medical imaging. However, from a computational perspective, SwinUnetR demands significant resources for training from scratch and poses challenges when adapting to new body regions. Lastly, [37] made a notable contribution by flawlessly fusing convolutional layers with transformer architectures. The axial fusion technique it employs excels in capturing both intra-slice and inter-slice relationships. In the study by [37], a CNN encoder is utilized to process the input, driven by two primary considerations. Firstly, the ViT [8] encoder, when trained on limited medical image datasets, tends to overfit. However, this challenge can be substantially mitigated by leveraging the dense visual representation capabilities of SAM. Secondly, GPU memory constraints previously hindered the end-to-end training of expansive ViT encoders. While memory remains a significant concern for volumetric medical imaging, the introduction of adapters has facilitated improvements in this area, enabling the effective use of a ViT-base encoder.

2.2. Parameter-efficient fine-tuning

Compared to general fine-tuning [38,39,41], Parameter-Efficient Fine-Tuning (PEFT) [12] has emerged as a cornerstone in the machine learning domain, acclaimed for its capability to hasten training while judiciously utilizing computational resources. By selectively updating parameters, PEFT accelerates convergence and adeptly curtails overfitting, especially in data-scarce environments. This paradigm enables the nimble adaptation of expansive pre-trained models to novel tasks, circumventing burdensome computational demands, and rendering it indispensable across diverse applications.

Adapter tuning seamlessly integrates diminutive, task-specific modules into pre-trained models, leaving the original weights untouched. This methodology strikes a harmonious balance between adaptability and efficiency, facilitating task-specific nuances while safeguarding the model's foundational knowledge. Notably, its parameter footprint is minimal, introducing only a fraction of the model's total parameters, ensuring that multiple tasks can be catered to without mutual interference. Low-Rank Adaptation (LoRA) [14] introduces trainable rank decomposition matrices into each layer of a pre-trained model, leading to a significant reduction in trainable parameters for downstream tasks. In stark contrast to traditional full fine-tuning, LoRA can curtail the number of trainable parameters by a factor of 10,000 and trim the GPU memory requirement by threefold. Conversely, traditional fine-tuning, which modifies extensive portions of the model, often grapples with overfitting and the potential erosion of pre-trained insights.

BitFit [43], with its focus on bias adjustments, might occasionally lack the requisite flexibility for intricate tasks, teetering on the brink of underfitting. The linear probe [1] strategy, commonly employed in self-supervised learning, appends a singular linear layer atop a static backbone to evaluate the quality of learned representations. Its efficacy is inherently tethered to the quality of the backbone's representations, potentially faltering in deciphering intricate task-specific intricacies. Prompting [17], which steers models using meticulously crafted input prompts, can oscillate in its efficacy across diverse models and demands nuanced expertise in prompt design.

In our endeavor, we gravitate towards the Adapter methodology for its equilibrium of flexibility and efficiency in task-specific modifications without compromising pre-trained insights. Specifically, we harness LoRA for its unparalleled parameter efficiency coupled with its competitive performance.

2.3. SAM in Medical Imaging

The introduction of the Segment Anything Model (SAM) has catalyzed a series of explorations in the medical imaging domain. A predominant theme in these endeavors

has been the evaluation of SAM's zero-shot capabilities, often through the lens of diverse prompts tailored for medical imaging.

Hu et al. [13] delved into SAM's prowess in multi-phase liver tumor segmentation. Their findings underscored SAM's potential as a robust annotation tool, albeit with performance nuances specific to the task at hand. Mazurowski et al. [23] embarked on an extensive evaluation spanning 19 medical imaging datasets. Their results illuminated the variability in SAM's performance, contingent on the dataset and task, with certain prompts amplifying its accuracy. Liu et al. [21] unveiled the Segment Any Medical Model (SAMM), a novel extension that synergizes SAM with 3D Slicer (a medical image processing software). This fusion facilitated near real-time segmentation across a plethora of medical imaging modalities. Deng et al. [7] spotlighted SAM's zero-shot segmentation acumen in digital pathology, emphasizing its adeptness in segmenting expansive connected objects and pinpointing challenges in dense instance object segmentation. Mattjie et al. [22] further corroborated SAM's zero-shot capabilities, showcasing its competitive edge against contemporary state-of-the-art models in select scenarios. Cheng et al.

While many have leveraged SAM's inherent capabilities, there have also been efforts to enhance performance through training, including fine-tuning and adaptation, on medical images rather than solely relying on inference.

[6] undertook a comprehensive evaluation, and accentuated the pivotal role of apt prompts in optimizing SAM's performance. Their exploration also ventured into adapter-based fine-tuning of SAM, albeit in a 2D context. Several studies, including those by Hu et al. [15] and Shi et al. [28], have elucidated the intricacies of tailoring SAM for medical images. Their consensus advocates for fine-tuning SAM, underscoring significant performance enhancements. Paranjape et al. [25] introduced AdaptiveSAM, a nimble tuning methodology for SAM tailored for surgical scene segmentation, showcasing its adaptability to novel datasets. However, this approach remains anchored in 2D.

Medical SAM Adapter (MSA) [33] was unveiled as a means to tailor the Segment Anything Model (SAM) specifically for medical image segmentation. Rather than extensively fine-tuning the entirety of SAM, the authors incorporated an Adapter module, which resulted in enhanced performance across a spectrum of medical imaging techniques. While MSA retains the need for a prompt during inference, akin to SAM, it's noteworthy that our proposed AFTer-SAM has been optimized to function efficiently without any additional input. In the study by Gong et al. [11], the 3DSAM-adapter is presented as a technique to transition SAM from 2D to 3D, with a particular emphasis on its proficiency in tumor segmentation tasks. While both our approach and the 3DSAM-adapter share the overarching objective of adapting SAM for 3D medical image segmentation, there are notable methodological distinctions. The 3DSAM-adapter begins with a 2D feature map, which is then subjected to temporal convolution for the fusion of temporal information. This is subsequently reshaped back to a 2D format and processed through attention blocks. In contrast, our methodology initiates with a 2D feature map that is directly channeled into attention blocks, with the fusion of temporal information being integrated at the axial adaptation stage. Furthermore, while the 3DSAM-adapter's design restricts the use of adapters to within the 2D slice, focusing predominantly on intra-slice information, our approach embraces a comprehensive strategy, incorporating adapters both within the 2D slice (intra-slice) and across slices (inter-slice or axial). This distinction underscores our innovation in the spatial-temporal transformer adapter design, emphasizing a nuanced approach to fusing spatial and temporal information.

Other explorations have spanned diverse data modalities, encompassing surgical scenes [27, 32], nuclei [27], and Optical Coherence Tomography (OCT) [9], further testifying to SAM's versatility in the medical domain.

3. Method

Figure 1 provides an in-depth view of AFTer-SAM. We have chosen not to alter the foundational design of SAM, which comprises a ViT encoder for detailed image feature extraction and a lightweight mask decoder for precise pixel-level segmentation. To enhance the encoding of high-level semantic information—both within individual slices and across neighboring ones—we've integrated the axial fusion transformer. The intricacies of each module will be further expounded upon in the subsequent subsections.

3.1. 3D CT Scan Representation

Given a 3D CT scan, denoted as $\mathbf{s} \in \mathbb{R}^{C\times H\times W\times D}$, it consists of a series of 2D slices along the axial axis. Each slice has a height H, width W, and a single channel, $C = 1$. This scan can be expressed as $\mathbf{s} = \mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_D$, with each $\mathbf{s}_d \in \mathbb{R}^{C\times H\times W}$. For every 2D slice, represented as $\mathbf{s}_i$, we select its N_A neighboring slices along the axial axis at an interval of N_f. This results in a set $\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_D\}$, where each $\mathbf{x}_d \in \mathbb{R}^{C\times H\times W\times N_A}$. For each group of sampled neighboring slices, $\mathbf{x}_d$, it can be defined as $\mathbf{x}_d = \mathbf{s}_{a_0}, \mathbf{s}_{a_1}, ..., \mathbf{s}_{aN_A}$. Here, $a_n = d - N_f \times (\frac{N_A}{2} - n)$, and n ranges from 0 to N_A.

3.2. Visual Representation with SAM

We mainly follow the SAM original model design to maintain its best ability to provide visual representation. Given an input neighboring slice group $\mathbf{x}_d \in \mathbb{R}^{C\times H\times W\times N_A}$, the ViT encoder $\mathcal{E}^{\text{ViT}}$ provides a corresponding feature map group $\mathbf{g}_d = \{\mathbf{g}_d^0, \mathbf{g}_d^1, ..., \mathbf{g}_d^{N_A}\}$, where $\mathbf{g}_d^n$ denotes the feature map for slice n and $\mathbf{g}_d^n \in \mathbb{R}^{C\times H\times W}$.

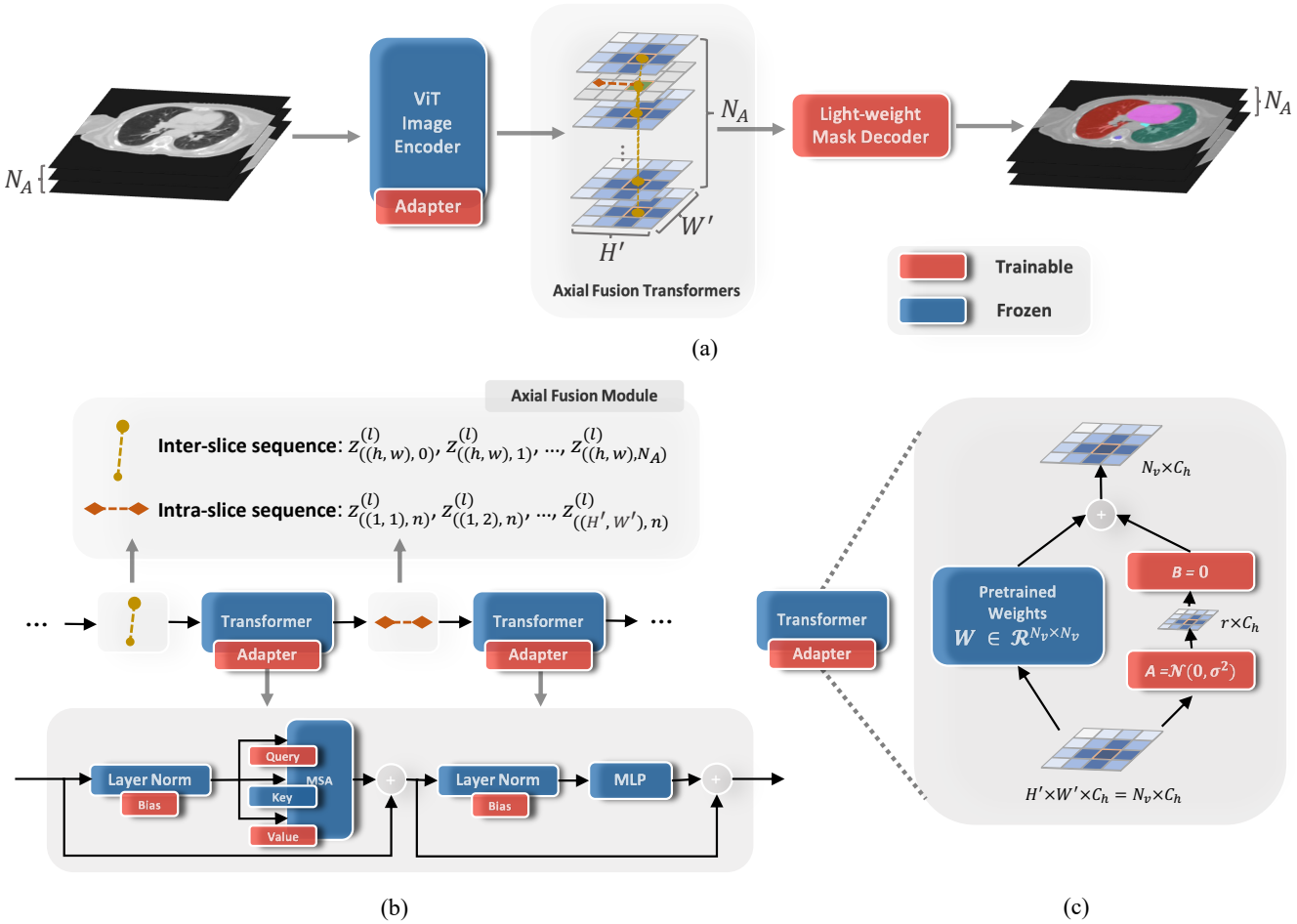


Figure 1. Overview of AFTer-SAM. (a) AFTer-SAM's architecture begins by encoding the neighboring slice group, denoted as $\mathbf{x}_d$, using the ViT encoder, resulting in a feature map group, $\mathbf{g}$. This is followed by the application of the axial fusion transformer to $\mathbf{g}$. The final step involves feeding the feature group, which has been enriched with both intra-slice and inter-slice cues, into the Light-weight mask decoder to produce the segmentation. (b) This panel delves into the axial fusion mechanism, illustrating the specifics of how the transformers are adapted. AFTer-SAM employs distinct processes to fuse inter-slice and intra-slice information. (c) A detailed breakdown of the fine-tuning process for transformers using low-rank adaptation is provided.

Here we have H, W and C denoting the height, width and number of channels. We take the final feature map group $\mathbf{g}_d = \mathcal{E}^{\text{ViT}}(\mathbf{x}_d)$ as input to the axial fusion transformer. For simplicity, we denote it as $\mathbf{g}$, where $\mathbf{g} \in \mathbb{R}^{C' \times N_v \times N_A}$. C' is the embeddings' length and $N_v = H' \times W'$ is the number of embeddings provided by $\mathcal{E}^{\text{ViT}}$, which is flattened from embedding's height H' by its width W'.

In our approach, we predominantly adhere to the original design of SAM to leverage its optimal visual representation capabilities. Given an input group of neighboring slices, represented as $\mathbf{x}_d \in \mathbb{R}^{C \times H \times W \times N_A}$, the ViT encoder, denoted as $\mathcal{E}^{\text{ViT}}$, produces a corresponding group of feature maps, $\mathbf{g}_d = \mathbf{g}_d^0, \mathbf{g}_d^1, ..., \mathbf{g}_d^{N_A}$. Here, each $\mathbf{g}_d^n$ signifies the feature map for the nth slice, with dimensions $\mathbf{g}_d^n \in \mathbb{R}^{C \times H \times W}$. The terms H, W, and C represent the height, width, and number of channels, respectively.

Subsequently, the final group of feature maps, $\mathbf{g}_d = \mathcal{E}^{\text{ViT}}(\mathbf{x}_d)$, is fed into the axial fusion transformer. For the sake of clarity, we refer to this as $\mathbf{g}$, with dimensions $\mathbf{g} \in \mathbb{R}^{C' \times N_v \times N_A}$. Here, C' denotes the length of the embeddings, while $N_v = H' \times W'$ represents the total number of embeddings offered by $\mathcal{E}^{\text{ViT}}$. This is derived by flattening the embeddings from a height of H' and a width of W'.

3.3. Axial Fusion Transformer

Upon obtaining fine-grained features through the ViT encoder, we further employ the Axial Fusion Transformer encoder. This step is crucial for capturing high-level semantic information, both within individual slices and across neighboring slices along the axial axis. In contrast, [37] opted for a CNN encoder for initial input handling, driven by two primary considerations. The first is the tendency of the ViT encoder to overfit, especially when exposed to limited medical image datasets. However, this challenge is significantly

alleviated by leveraging the robust visual representation capabilities of SAM. The second concern revolves around the GPU memory constraints that previously impeded the end-to-end training of expansive ViT encoders. While these memory challenges remain for volumetric medical imaging, the introduction of adapters has brought about notable improvements, rendering the use of a ViT-base encoder increasingly viable. The specifics of our approach are detailed in the subsequent subsections:

In line with the approach of [37], we define the initial representation $\mathbf{z}^{(0)}_{((h,w),n)} \in \mathbb{R}^{C'}$, bypassing the linear projection step found in the original ViT [8]. Specifically, $\mathbf{z}^{(0)}_{((h,w),n)} = \mathbf{g}_{((h,w),n)} + \mathbf{e}^{pos}_{((h,w),n)}$, where $\mathbf{e}^{pos}_{((h,w),n)} \in \mathbb{R}^{C'}$ is a learnable positional embedding. This embedding encodes two facets of the vector's location: its position (h, w) within an individual feature map $\mathbf{g}_n$, and its relative position n among the feature maps in the group $\mathbf{g}$.

The sequence $\mathbf{z}^{(0)}_{((h,w),n)}$, spanning $(h, w) = (1, 1), \ldots, (H', W')$ and $n = 0, 1, \ldots, N_A$, serves as the input to the Transformer. Analogous to sequences of embedded words in natural language processing (NLP) transformers, this sequence plays a pivotal role in our model. It's worth noting that, in our actual code implementation, the height and width dimensions are flattened. This means the vector can alternatively be denoted as $\mathbf{z}^{(0)}_{(p,n)}$, where $p = W' \cdot (h-1) + w$. However, for clarity in this exposition, we retain the (h, w) notation.

3.4. Attention Adaptation

The axial fusion transformer is structured into L blocks. Within each block l, the query, key, and value vectors for every location are derived from the representation $\mathbf{z}^{(l-1)}_{((h,w),n)}$, which is the output of the preceding block.

To finetune the transformers efficiently, we employ the low rank adaptation (LoRA) technique [14]. In this adaptation, only the queries and values undergo modification:

$$\mathbf{q}^{(l,a)}_{((h,w),n)} = (W^{(l,a)}_Q + B^{(l,a)}_Q A^{(l,a)}_Q) \cdot \mathrm{LN}\left(\mathbf{z}^{(l-1)}_{((h,w),n)}\right),$$
$$\mathbf{v}^{(l,a)}_{((h,w),n)} = (W^{(l,a)}_V + B^{(l,a)}_V A^{(l,a)}_V) \cdot \mathrm{LN}\left(\mathbf{z}^{(l-1)}_{((h,w),n)}\right),$$
$$\mathbf{k}^{(l,a)}_{((h,w),n)} = W^{(l,a)}_K \cdot \mathrm{LN}\left(\mathbf{z}^{(l-1)}_{((h,w),n)}\right).$$

Here, $\mathrm{LN}()$ stands for LayerNorm [2]. The symbol a ranges over the set $\{1, 2, \ldots, \mathcal{M}\}$, indicating the index for multiple attention heads. The total count of these attention heads is represented by $\mathcal{M}$. Consequently, the dimensionality for each attention head is given by $C_h = C'/\mathcal{M}$. The matrices $A \in \mathbb{R}^{r \times N_v}$ and $B \in \mathbb{R}^{N_v \times r}$ encompass trainable parameters.

Self-attention weights are derived using the dot-product mechanism. For a given query at location $((h, w), n)$, the self-attention weights, denoted as $\boldsymbol{\alpha}^{(l,a)}_{((h,w),n)}$, and falling within the dimensionality $\mathbb{R}^{(H' \cdot W') \cdot N_A}$, are computed as:

$$\boldsymbol{\alpha}^{(l,a)}_{((h,w),n)} = \mathrm{SoftMax}\left(\frac{\mathbf{q}^{(l,a)}_{((h,w),n)}}{\sqrt{C_h}}^{\top} \cdot \mathbf{k}^{(l,a)}_{((h,w)',n')}\right),$$

where $(h, w)'$ spans the set $\{(1, 1), ..., (H', W')\}$ and n' ranges from 0 to N_A. It's noteworthy that the computational demand diminishes significantly when attention is restricted to a singular feature map or solely along the axial axis. Specifically, when attention is confined to a single feature map, a mere $H' \cdot W'$ query-key comparisons are executed, utilizing only the keys from the identical feature map as the query:

$$\boldsymbol{\alpha}^{(l,a)\textbf{intra}}_{((h,w),n)} = \mathrm{SoftMax}\left(\frac{\mathbf{q}^{(l,a)}_{((h,w),n)}}{\sqrt{C_h}}^{\top} \cdot \mathbf{k}^{(l,a)}_{((h,w)',n)}\right),$$

with $(h, w)'$ again spanning the set $\{(1, 1), ..., (H', W')\}$.

To obtain the encoding $\mathbf{z}^{(l)}_{((h,w),n)}$ for block l, an initial step involves computing the weighted sum of value vectors, leveraging the self-attention coefficients from each attention head. Subsequent to this, vectors from all attention heads are concatenated. This aggregated data is then subjected to a linear projection via a fully connected layer (FC) and is processed through a multi-layer perceptron (MLP) equipped with layer normalization (LN). Post each operation, residual connections are incorporated.

Due to the limit of memory, computing self-attention over a 3D space is not feasible. Replacing it with 2D attention applied only on one single slice can certainly reduce the computational cost. However, such a model ignores to capture information among neighboring slices, which is naturally provided by a 3D volume. We apply axial fusion mechanism for computing attention along the axial axis, where the attention along the axial axis and the attention within a single slice are separately applied one after the other. By fusing the axial information this way, we firstly compute attention along the axial with all the channels at the same position at (h, w):

$$\boldsymbol{\alpha}^{(l,a)\textbf{inter}}_{((h,w),n)} = \mathrm{SoftMax}(\frac{\mathbf{q}^{(l,a)}_{((h,w),n)}}{\sqrt{C_h}}^{\top} \cdot \mathbf{k}^{(l,a)}_{((h,w),n')}),$$

where $n' \in \{1, ..., N_A\}$. The encoding $\mathbf{z}'^{(l)\textbf{inter}}_{((h,w),n)}$ using axial attention is then fed back for single slice attention computation instead of directly being passed to the MLP. In other words, new key/query/value vectors are obtained from $\mathbf{z}'^{(l)\textbf{inter}}_{((h,w),n)}$ and the single slice attention is then computed. Finally, the resulting vector $\mathbf{z}'^{(l)\textbf{intra}}_{((h,w),n)}$ is passed to the MLP to compute the final encoding $\mathbf{z}^{(l)}_{((h,w),n)}$ at position $((H, W), n)$ by block l. The final fused encoding for the feature map group $\mathbf{g}$ is $\mathbf{z}^{(L)} \in \mathbb{R}^{C \times H' \times W' \times N_A}$.

We learn distinct query/key/value matrices $\left\{W^{(l,a)}_{Q^{\text{slice}}}, W^{(l,a)}_{K^{\text{slice}}}, W^{(l,a)}_{V^{\text{slice}}}\right\}$ and $\left\{W^{(l,a)}_{Q^{\text{axial}}}, W^{(l,a)}_{K^{\text{axial}}}, W^{(l,a)}_{V^{\text{axial}}}\right\}$

Methods	Average	Eso	Trachea	Spinal Cord	Lung(L)	Lung(R)	Heart
U-Net [26]	91.18	78.85	90.72	89.37	97.31	96.37	94.46
nnUNet-2D [16]	89.74	78.82	88.32	86.61	96.03	96.65	92.01
nnUNet-3D [16]	91.63	81.18	89.32	91.21	97.68	97.74	92.66
Attention U-Net [24]	90.19	76.35	88.14	89.43	97.65	97.87	91.68
TransUNet [4]	91.38	78.27	91.45	88.36	97.63	97.84	94.74
Swin-Unet [3]	91.26	78.98	91.20	88.64	97.64	97.79	93.30
CoTr [34]	91.39	79.06	91.55	88.67	97.47	97.65	93.92
AFTer-UNet [37]	92.32	81.47	91.76	90.12	97.80	97.90	94.86
Swin UNETR [30]	91.63	79.64	91.28	88.89	97.64	97.69	94.66
MSA [33]	91.50	77.91	91.19	90.12	97.70	97.85	94.23
AFTer-SAM	**92.75**	**82.34**	**91.89**	**91.24**	**97.85**	**97.98**	**95.17**

Table 1. Dice scores of different methods on in-house Thorax-85 dataset.

over dimensions within one single slice and among slices along the axial axis. Note that compared to the $(H' \cdot W') \cdot N_A$ comparisons each vector needed by the self-attention model, our approach performs only $(H' \cdot W') + N_A$ comparisons per vector.

3.5. Representation Decoding

The decoder in AFTer-SAM, denoted as $\mathcal{D}^{\text{ViT}}$, closely mirrors the design principles of SAM, emphasizing a lightweight structure. We have opted to exclude the prompt encoder, enabling end-to-end trainability for our model. The rationale behind this decision, substantiated by experimental outcomes, is elaborated upon in our ablation study section. With the omission of the prompt encoder, the mask decoder retains only the image-to-token attention module. Given the lightweight nature of the mask decoder and its minimal parameter count, there's a consideration about the optimal strategy: either adapting or fully fine-tuning it. Our empirical findings lean towards full fine-tuning of the streamlined mask decoder. We delve deeper into the reasoning behind this decision in our ablation study discussion.

Given a series of sampled neighboring slice groups $\mathbf{x}$ as inputs, the model produces D segmentation map groups $\mathbf{y} = \{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_D\}$ as outputs, where $\mathbf{y}_d \in \mathbb{R}^{C_{\text{cls}} \times H \times W \times N_A}$ and C_{cls} denotes the number of organ classes. From each segmentation map group, only the middle segmentation map $\mathbf{y}_d^{\frac{N_A}{2}}$ is retained, discarding its neighbors. This process is repeated for all D segmentation map groups, which are then concatenated to produce the final 3D prediction corresponding to the 3D scan.

3.6. Loss Function

The model's loss function is a combination of dice loss and cross-entropy loss.

Methods	Avg	Eso	Trachea	Aorta	Heart
U-Net	89.97	80.07	91.23	94.73	93.83
Att U-Net	90.47	81.25	90.82	94.74	95.07
TransUNet	91.50	81.41	94.05	94.48	96.07
Swin-Unet	91.29	81.06	93.27	94.82	96.02
CoTr	91.41	81.53	94.03	94.06	96.01
AFTer-UNet	92.10	82.98	94.20	94.92	96.31
Swin UNETR	92.06	82.82	93.84	95.15	96.42
MSA	92.16	82.98	94.11	94.94	96.62
AFTer-SAM	**92.42**	**83.21**	**94.27**	**95.41**	**96.78**

Table 2. Dice scores of different methods on SegTHOR dataset.

4. Experiments

4.1. Dataset

In alignment with the experimental setup of [37], we evaluate our approach on two datasets:

Thorax-85 An in-house dataset introduced in [5], Thorax-85 contains 85 3D thorax CT images. We present the average DSC for 6 thorax organs (eso, trachea, spinal cord, left lung, right lung, and heart), based on a split of 60 training cases and 25 test cases.

SegTHOR Originating from the 2019 Challenge on Segmentation of THoracic Organs at Risk in CT Images [20], SegTHOR includes 40 3D thorax CT scans. We determine the average DSC for 4 thorax organs (eso, trachea, aorta, and heart) with a dataset division of 30 training cases and 10 validation cases.

Evaluation Metric: Consistent with prior studies [33, 37], we employ the Sørensen–Dice coefficient (DSC) as our evaluation metric. The DSC quantifies the overlap between the predicted mask $\mathbf{m}_p$ and the ground truth mask $\mathbf{m}_g$. It is formulated as: $\text{DSC}(\mathbf{m}_p, \mathbf{m}_g) = \frac{2|\mathbf{m}_p \cap \mathbf{m}_g|}{|\mathbf{m}_p| + |\mathbf{m}_g|}$.

4.2. Implementation Details

All images undergo resampling to achieve a spacing of 2.5mm × 1.0mm × 1.0mm, corresponding to the depth, height, and width dimensions of the 3D volume. During

training, we incorporate an elastic transform to counteract overfitting. The model, AFTer-SAM, is trained using the Adam optimizer [18] with a momentum set at 0.9 and a weight decay of 10^{-4}. The training spans 550 epochs, with a learning rate of 10^{-4} for the initial 500 epochs and 10^{-5} for the concluding 50 epochs. Within each epoch, from every 3D CT scan s, only a single random slice group $\mathbf{x}_d$ is selected, as opposed to selecting all available slice groups. The configuration parameters are set as follows: number of axial fusion transformers $L = 6$, attention heads $\mathcal{M} = 8$, neighboring slices $N_A = 8$, and sampling frequency $N_f = 1$.

	Box	1 Point	3 Points	Text	w.o.
Eso	81.20	81.72	81.67	81.64	**82.34**
Trachea	**91.92**	91.46	91.50	91.49	91.89
Spinal	90.38	90.05	90.40	90.35	**91.24**
Lung(L)	97.53	**97.95**	97.77	97.58	97.85
Lung(R)	98.18	97.60	97.65	97.63	**97.98**
Heart	94.56	95.14	94.90	94.59	**95.17**
Average	92.30	92.32	92.32	92.21	**92.75**

Table 3. Ablation study on prompt encoder. Retaining it was unnecessary, as it only offered marginal and inconsistent improvements to the results.

	LoRA	Finetune
Eso	82.03	**82.34**
Trachea	91.68	**91.89**
Spinal Cord	90.99	**91.24**
Lung(L)	97.41	**97.85**
Lung(R)	97.82	**97.98**
Heart	94.70	**95.17**
Average	92.44	**92.75**

Table 4. Ablation study on mask decoder.

4.3. Quantitative Results

Table 1 shows the performance comparison of AFTer-UNet with previous work on Thorax-85. We ran the following representative algorithms: U-Net [26], Attention U-Net [24], nnU-Net [16], TransUNet [4], Swin-Unet [3], CoTr [34], AFTer-UNet [37], Swin UNETR [30] and MSA [33]. U-Net is a well-established medical image segmentation baseline algorithm. Attention U-Net [24] is a multi-organ segmentation framework that uses gated attention to filter out irrelevant responses in the feature maps. nnU-Net [16] is a self-adaptive medical image semantic segmentation framework that wins the first in the Medical Segmentation Decathlon(MSD) challenge [29]. TransUNet [4] presents the first study which explores the potential of transformers in the context of 2D medical image segmentation. Swin-Unet [3] explores using pure transformer modules on 2D medical image segmentation tasks, without any convolutional layers. CoTr [34] firstly explores transformer modules for 3D medical image segmentation. Swin UNETR [30] is an extension of the UNETR model that incorporates the Swin Transformer architecture, leveraging the hierarchical nature of Swin Transformers. AFTer-UNet [37] combines the U-Net architecture with axial attention, focusing on relevant features in the axial plane of medical images. MSA [33] introduces the Medical SAM Adapter, designed to be integrated into existing segmentation networks. The above-mentioned works cover a wide range of algorithms for multi-organ segmentation and should provide a comprehensive and fair comparison to our method on the in-house Thorax-85 dataset.

From the in-house Thorax-85 dataset (Table 1), for organs with regular shapes like the left lung, AFTer-UNet, Swin UNETR, and MSA achieve scores of 97.80%, 97.64%, and 97.70% respectively. AFTer-SAM, however, slightly edges out with a score of 97.85%. Similarly, for the right lung, while AFTer-UNet scores 97.90%, Swin UNETR achieves 97.69%, and MSA gets 97.85%, AFTer-SAM tops them with a score of 97.98%.

The distinction becomes more pronounced for organs with more anatomical variance. For the esophagus, AFTer-UNet achieves 81.47%, Swin UNETR scores 79.64%, and MSA gets 77.91%. In contrast, AFTer-SAM outperforms all three with a score of 82.34%. Similarly, for the trachea, AFTer-SAM's score of 91.89% is superior to AFTer-UNet's 91.76%, Swin UNETR's 91.28%, and MSA's 91.19%.

Transitioning to the SegTHOR dataset, the trend continues. For the heart, AFTer-UNet, Swin UNETR, and MSA achieve scores of 96.31%, 96.42%, and 96.62% respectively. AFTer-SAM, however, leads with a score of 96.78%. For the esophagus, AFTer-SAM's score of 83.21% is a significant improvement over AFTer-UNet's 82.98%, Swin UNETR's 82.82%, and even surpasses MSA's 82.98%.

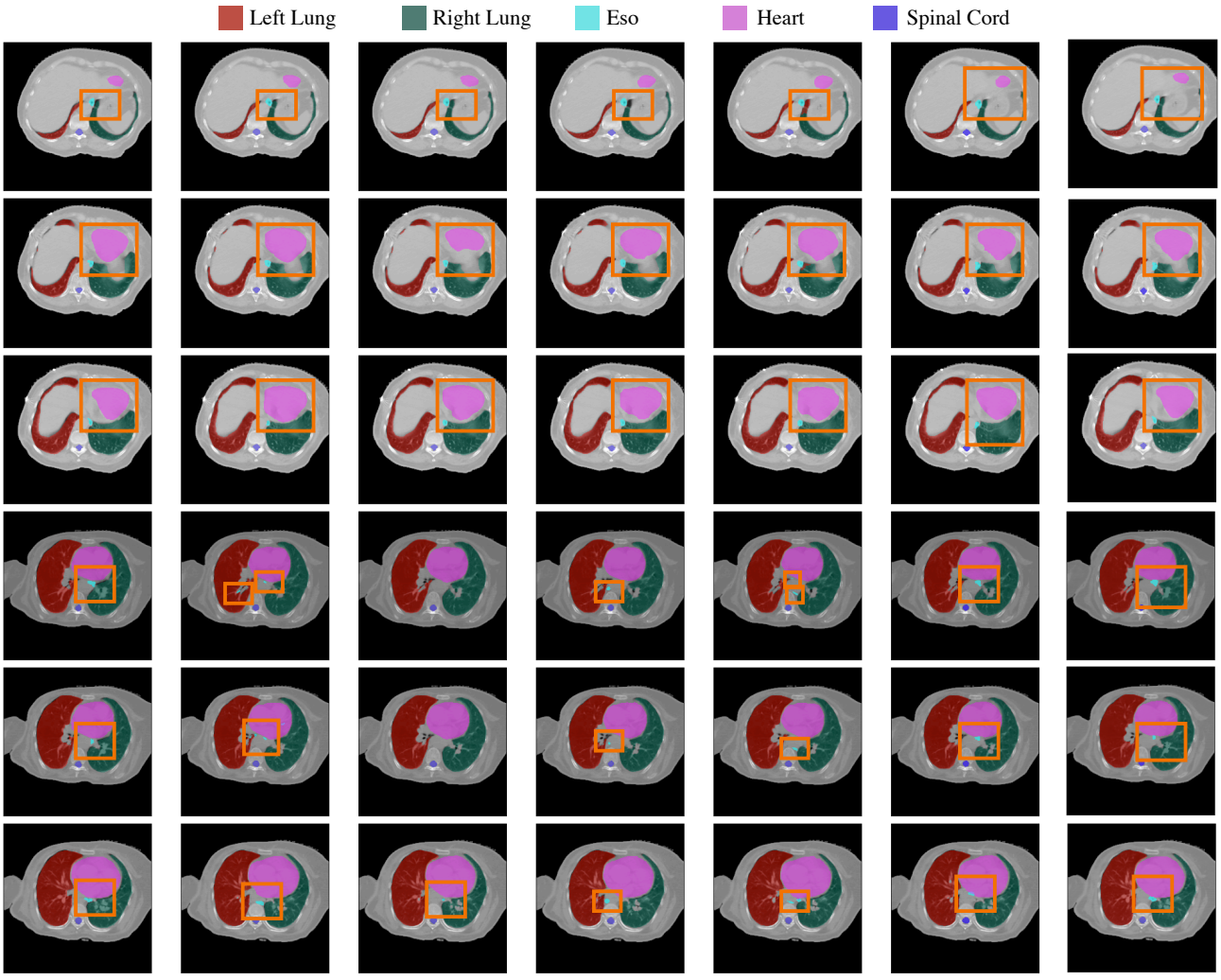
For the aorta, AFTer-SAM's performance of 95.41% is notably higher than AFTer-UNet's 94.92%, Swin UNETR's 95.15%, and MSA's 94.94%. This consistent outperformance across multiple organs in both datasets underscores AFTer-SAM's robustness and capability.

In conclusion, while AFTer-UNet, Swin UNETR, and MSA are formidable models in their own right, AFTer-SAM consistently achieves top-tier results across both datasets. Its ability to handle both regular-shaped organs and those with intricate structures positions it as a leading solution for multi-organ segmentation tasks.

4.4. Ablation study on Thorax-85

We further conduct extensive ablation studies on Thorax-85 to explore the impact of excluding the prompt encoder on end-to-end trainability and the optimal strategy for adapting or fully fine-tuning the lightweight mask decoder.

Prompt encoder In the initial SAM paper, the prompt encoder (PE) was designed to accept various inputs such as



(a) Ground Truth (b) U-Net (c) nnUNet-3D (d) TransUNet (e) CoTr (f) AFTer-UNet (g) AFTer-SAM

Figure 2. Qualitative results of different approaches on Thorax-85 dataset. (a) shows the ground truth of the CT slice. (b)-(f) show the results of previous methods. (g) shows the results of AFTer-SAM. The regions in orange rectangles indicate the effectiveness to our model.

points, boxes, or text to enhance segmentation outcomes. However, during our adaptation phase, we show that retaining it was unnecessary, as it only offered marginal and inconsistent improvements to the results in 3.

Mask decoder In table 4, we evaluate the decision of whether to fully finetune the mask decoder or simply adapt it. Based on the results, finetuning the entire mask decoder proves to be the more effective approach. This preference can be attributed to the mask decoder's lightweight design and its limited number of parameters.

4.5. Qualitative Results

Fig.2 displays the qualitative outcomes of various methods on the Thorax-85 dataset. AFTer-SAM demonstrates superior effectiveness in comparison to other methods. A notable area of focus is the esophagus, which is particularly challenging to segment in the Thorax-85 dataset due to its significant anatomical variance and elongated form. Earlier methods often produced segmentation maps that varied significantly between slices, an outcome that is not logical. In contrast, AFTer-SAM (as seen in the last column) offers consistent and precise predictions.

5. Conclusion

In this study, we presented AFTer-SAM, a comprehensive end-to-end solution for medical image segmentation. Our proposed framework employs Low Rank Adaptation to refine SAM and leverages Axial Fusion Transformers to seamlessly integrate both intra-slice and inter-slice contextual data, thereby guiding the segmentation outcome. Experimental results underscore the superior performance of our model in comparison to prior methodologies.

References

[1] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes, 2018. 2

[2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, 2016. 5

[3] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation, 2021. 6, 7

[4] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L. Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021. 2, 6, 7

[5] Xuming Chen, Shanlin Sun, Narisu Bai, Kun Han, Qianqian Liu, Shengyu Yao, Hao Tang, Chupeng Zhang, Zhipeng Lu, Qian Huang, Guoqi Zhao, Yi Xu, Tingfeng Chen, Xiaohui Xie, and Yong Liu. A deep learning-based auto-segmentation system for organs-at-risk on whole-body computed tomography images for radiation therapy. *Radiotherapy and Oncology*, 160:175–184, July 2021. 6

[6] Dongjie Cheng, Ziyuan Qin, Zekun Jiang, Shaoting Zhang, Qicheng Lao, and Kang Li. Sam on medical images: A comprehensive study on three prompt modes, 2023. 3

[7] Ruining Deng, Can Cui, Quan Liu, Tianyuan Yao, Lucas W. Remedios, Shunxing Bao, Bennett A. Landman, Lee E. Wheless, Lori A. Coburn, Keith T. Wilson, Yaohong Wang, Shilin Zhao, Agnes B. Fogo, Haichun Yang, Yucheng Tang, and Yuankai Huo. Segment anything model (sam) for digital pathology: Assess zero-shot segmentation on whole slide imaging, 2023. 3

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 2, 5

[9] Botond Fazekas, José Morano, Dmitrii Lachinov, Guilherme Aresta, and Hrvoje Bogunović. Samedoct: Adapting segment anything model (sam) for retinal oct, 2023. 3

[10] Eli Gibson, Francesco Giganti, Yipeng Hu, Ester Bonmati, Steve Bandula, Kurinchi Gurusamy, Brian Davidson, Stephen P Pereira, Matthew J Clarkson, and Dean C Barratt. Automatic multi-organ segmentation on abdominal ct with dense v-networks. *IEEE transactions on medical imaging*, 37(8):1822–1834, 2018. 1

[11] Shizhan Gong, Yuan Zhong, Wenao Ma, Jinpeng Li, Zhao Wang, Jingyang Zhang, Pheng-Ann Heng, and Qi Dou. 3dsam-adapter: Holistic adaptation of sam from 2d to 3d for promptable medical image segmentation, 2023. 3

[12] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp, 2019. 2

[13] Chuanfei Hu, Tianyi Xia, Shenghong Ju, and Xinde Li. When sam meets medical images: An investigation of segment anything model (sam) on multi-phase liver tumor segmentation, 2023. 3

[14] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. 1, 2, 5

[15] Xinrong Hu, Xiaowei Xu, and Yiyu Shi. How to efficiently adapt large segmentation model(sam) to medical images, 2023. 3

[16] Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnU-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, Dec. 2020. 6, 7

[17] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning, 2022. 2

[18] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, Conference Track Proceedings*, 2015. 7

[19] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023. 1

[20] Z. Lambert, C. Petitjean, B. Dubray, and S. Ruan. Segthor: Segmentation of thoracic organs at risk in ct images, 2019. 6

[21] Yihao Liu, Jiaming Zhang, Zhangcong She, Amir Kheradmand, and Mehran Armand. Samm (segment any medical model): A 3d slicer integration to sam, 2023. 3

[22] Christian Mattjie, Luis Vinicius de Moura, Rafaela Cappelari Ravazio, Lucas Silveira Kupssinskü, Otávio Parraga, Marcelo Mussi Delucis, and Rodrigo Coelho Barros. Zero-shot performance of the segment anything model (sam) in 2d medical imaging: A comprehensive evaluation and practical guidelines, 2023. 3

[23] Maciej A. Mazurowski, Haoyu Dong, Hanxue Gu, Jichen Yang, Nicholas Konz, and Yixin Zhang. Segment anything model for medical image analysis: An experimental study. *Medical Image Analysis*, 89:102918, oct 2023. 3

[24] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018. 6, 7

[25] Jay N. Paranjape, Nithin Gopalakrishnan Nair, Shameema Sikder, S. Swaroop Vedula, and Vishal M. Patel. Adaptivesam: Towards efficient tuning of sam for surgical scene segmentation, 2023. 3

[26] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 1, 2, 6, 7

[27] Tal Shaharabany, Aviad Dahan, Raja Giryes, and Lior Wolf. Autosam: Adapting sam to medical images by overloading the prompt encoder, 2023. 3

[28] Xiaoyu Shi, Shurong Chai, Yinhao Li, Jingliang Cheng, Jie Bai, Guohua Zhao, and Yen-Wei Chen. Cross-modality attention adapter: A glioma segmentation fine-tuning method for sam using multimodal brain mr images, 2023. 3

[29] Amber L Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram Van Ginneken, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063*, 2019. 7

[30] Yucheng Tang, Dong Yang, Wenqi Li, Holger R. Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20698–20708, 2022. 2, 6, 7

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 1

[32] An Wang, Mobarakol Islam, Mengya Xu, Yang Zhang, and Hongliang Ren. Sam meets robotic surgery: An empirical study on generalization, robustness and adaptation, 2023. 3

[33] Junde Wu, Yu Zhang, Rao Fu, Huihui Fang, Yuanpei Liu, Zhaowei Wang, Yanwu Xu, and Yueming Jin. Medical sam adapter: Adapting segment anything model for medical image segmentation, 2023. 3, 6, 7

[34] Yutong Xie, Jianpeng Zhang, Chunhua Shen, and Yong Xia. Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation, 2021. 2, 6, 7

[35] Xiangyi Yan, Junayed Naushad, Shanlin Sun, Kun Han, Hao Tang, Deying Kong, Haoyu Ma, Chenyu You, and Xiaohui Xie. Representation recovering for self-supervised pre-training on medical images. In *WACV*, pages 2685–2695, 2023. 2

[36] Xiangyi Yan, Junayed Naushad, Chenyu You, Hao Tang, Shanlin Sun, Kun Han, Haoyu Ma, James Duncan, and Xiaohui Xie. Localized region contrast for enhancing self-supervised learning in medical image segmentation, 2023. 2

[37] Xiangyi Yan, Hao Tang, Shanlin Sun, Haoyu Ma, Deying Kong, and Xiaohui Xie. After-unet: Axial fusion transformer unet for medical image segmentation, 2021. 1, 2, 4, 5, 6, 7

[38] Chenyu You, Weicheng Dai, Fenglin Liu, Yifei Min, Haoran Su, Xiaoran Zhang, Xiaoxiao Li, David A Clifton, Lawrence Staib, and James S Duncan. Mine your own anatomy: Revisiting medical image segmentation with extremely limited labels. *arXiv preprint arXiv:2209.13476*, 2022. 2

[39] Chenyu You, Weicheng Dai, Yifei Min, Fenglin Liu, Xiaoran Zhang, Chen Feng, David A Clifton, S Kevin Zhou, Lawrence Hamilton Staib, and James S Duncan. Rethinking semi-supervised medical image segmentation: A variance-reduction perspective. *arXiv preprint arXiv:2302.01735*, 2023. 2

[40] Chenyu You, Weicheng Dai, Yifei Min, Lawrence Staib, and James S Duncan. Implicit anatomical rendering for medical image segmentation with stochastic experts. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2023. 2

[41] Chenyu You, Jinlin Xiang, Kun Su, Xiaoran Zhang, Siyuan Dong, John Onofrey, Lawrence Staib, and James S Duncan. Incremental learning meets transfer learning: Application to multi-site prostate mri segmentation. In *International Workshop on Distributed, Collaborative, and Federated Learning*. Springer, 2022. 2

[42] Chenyu You, Ruihan Zhao, Fenglin Liu, Siyuan Dong, Sandeep Chinchali, Ufuk Topcu, Lawrence Staib, and James Duncan. Class-aware adversarial transformers for medical image segmentation. *Advances in Neural Information Processing Systems*, 2022. 2

[43] Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models, 2022. 2