

# SAMora: Enhancing SAM through Hierarchical Self-Supervised Pre-Training for Medical Images

Shuhang Chen<sup>1†</sup> Hangjie Yuan<sup>1†</sup> Pengwei Liu<sup>1</sup> Hanxue Gu<sup>2</sup>  
Tao Feng<sup>3</sup> Dong Ni<sup>1\*</sup>

<sup>1</sup>Zhejiang University <sup>2</sup>Duke University <sup>3</sup>Tsinghua University

## Abstract

The Segment Anything Model (SAM) has demonstrated significant potential in medical image segmentation. Yet, its performance is limited when only a small amount of labeled data is available, while there is abundant valuable yet often overlooked hierarchical information in medical data. To address this limitation, we draw inspiration from self-supervised learning and propose SAMora, an innovative framework that captures hierarchical medical knowledge by applying complementary self-supervised learning objectives at the image, patch, and pixel levels. To fully exploit the complementarity of hierarchical knowledge within LoRAs, we introduce HL-Attn, a hierarchical fusion module that integrates multi-scale features while maintaining their distinct characteristics. SAMora is compatible with various SAM variants, including SAM2, SAMed, and H-SAM. Experimental results on the Synapse, LA, and PROMISE12 datasets demonstrate that SAMora outperforms existing SAM variants. It achieves state-of-the-art performance in both few-shot and fully supervised settings while reducing fine-tuning epochs by 90%. The code is available at <https://github.com/ShChen233/SAMora>.

## 1. Introduction

The Segmentation Anything Model (SAM) [32] stands as one of the most versatile and comprehensive foundational models in the field of image segmentation, gaining widespread recognition for its adaptability and high performance across a broad range of applications. Its effectiveness has been demonstrated in diverse domains, including satellite imagery analysis [49] and autonomous driving [10], where robust and accurate segmentation is critical.

However, SAM’s performance is notably less impressive when directly applied to medical images [14, 62]. The inherent complexity of medical imaging, coupled with

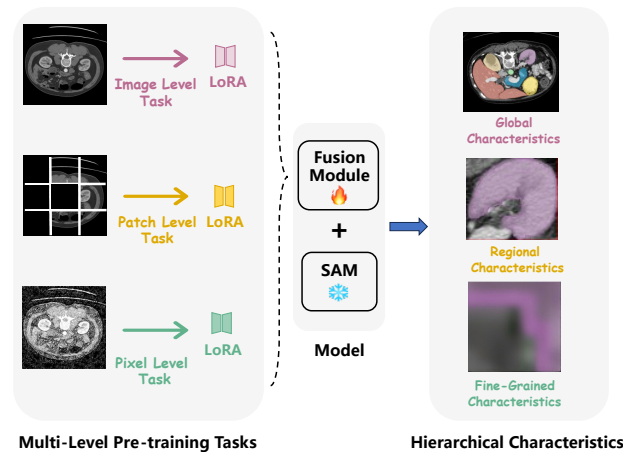


Figure 1. **The Hierarchical Characteristics of Multi-Level Pre-training Tasks on Medical Images.** The abundant hierarchical characteristics inherent in vast amounts of unlabeled data, when effectively fused, can significantly enhance the segmentation performance of SAM.

the model’s reliance on prompts, poses significant challenges for its direct application in this domain. Additionally, the scarcity of training medical images—due to legal restrictions and annotation difficulty further exacerbates these challenges, leading to suboptimal segmentation results in medical imaging tasks. To overcome this limitation, researchers have turned to prompt-free fine-tuning techniques, leveraging the available labeled data to adapt SAM more effectively to the medical domain. Efforts such as SAMed [66] and H-SAM [15] have demonstrated that fine-tuning SAM with domain-specific data significantly enhances its ability to capture the unique patterns inherent in medical images. These advancements have paved the way for more accurate and reliable medical image segmentation; however, they often *overlook the vast amounts of unlabeled data available*, leaving a wealth of potentially valuable information untapped.

Recent advancements in self-supervised learning (SSL), such as masked autoencoder (MAE) [23] and contrastive

<sup>†</sup> These authors contributed equally to this work.

\* Corresponding author: dni@zju.edu.cn.

learning [48], have gained substantial attention for their ability to leverage unlabeled data without the need for manual annotation. Studies have demonstrated that models trained using self-supervised learning techniques can even surpass the performance of those trained exclusively on labeled data [19]. Given the abundance of unlabeled data in the medical imaging domain, this approach presents a significant opportunity to enhance SAM’s adaptability to the complex and diverse characteristics of medical images. These promising developments have inspired us to explore the integration of self-supervised learning into fine-tuning the SAM model.

Furthermore, we observed that medical images exhibit hierarchical and multi-scale structures [30, 35], with each scale providing unique features that are critical for accurate diagnosis. As depicted in Fig. 1, The hierarchical characteristics of the multi-level pre-training tasks on the medical image show that the image-level patterns capture global characteristics, which is a broad overview of the anatomical region and essential for identifying general structures and context. The patch-level patterns allow for a more detailed examination of specific regions called regional characteristics, highlighting finer anatomical features, while the pixel-level patterns offer fine-grained characteristics with the highest resolution, enabling the detection of subtle tissue variations. We hypothesize that *these hierarchical patterns are complementary to each other and that combining them with self-supervised learning on unlabeled data can significantly improve segmentation performance.*

In response to these insights, we propose **SAMora**, a prompt-free fine-tuned SAM model that incorporates multiple LoRA experts, which leverage large amounts of unlabeled medical image data via two stages. The first stage focuses on enhancing SAM model with hierarchical medical knowledge, with the use of hierarchical self-supervised pre-training. Specifically, we pre-train LoRA experts for the SAM model in the image, patch, and pixel levels, which are used for follow-up fine-tuning. Notably, the LoRA experts for image and patch levels are learned by distilling from continually pre-trained teacher models (*i.e.*, SimCLRv2 [12] and MAE [23]) that enables more effective medical knowledge awareness.

The second stage involves fine-tuning with labeled data, during which we introduce a hierarchical attention mechanism, HL-Attn (Hierarchical LoRA Attention). HL-Attn leverages the hierarchical properties at each level, **first** progressively integrating features from lower to higher levels. By adaptively integrating and refining knowledge from multiple levels of medical representations, it effectively captures the hierarchical characteristics of each level. **Fig. 6 in Appendix B.4 shows that hierarchical fusion is key to the model’s ability to handle complex medical imaging tasks.**

We perform comprehensive experiments on multi-organ segmentation datasets (*i.e.*, Synapse, left atrial (LA), and PROMISE2012 datasets) in both fully supervised and few-shot settings, which demonstrates SAMora’s consistent superiority over existing prompt-free SAM counterparts. Moreover, SAMora is compatible with different prompt-free SAM variants, such as SAMed [66] (by default), H-SAM [15] where the decoder has been modified from SAMed. Notably, we also apply the SAMora on SAM2 [44], which is a novel segment anything model proposed recently, called **SAMora-2**. Specifically, on the Synapse dataset, SAMora and its variant achieve a Mean Dice boost of **4.09%** using only 10% of the training data and a Mean Dice boost of 2.10% when utilizing the whole dataset while consuming only **10%** of the fine-tuning epochs compared with other SAM counterparts.

The core contributions can be summarized as follows: **1)** We propose to integrate three hierarchical levels of self-supervised knowledge from unlabeled medical images to existing SAM variants by pre-training LoRA experts. **2)** We propose the HL-Attn module to adaptively fuse hierarchical medical knowledge, ensuring that the model fully exploits the information available across different scales. **3)** SAMora is comparable with different SAM variants and achieves state-of-the-art (SOTA) performance on the LA, PROMISE12, and Synapse datasets in fully supervised and few-shot settings with only 10% of the fine-tuning epochs, highlighting its efficiency and effectiveness in medical image segmentation.

## 2. Related Works

### 2.1. SAM and Related Fine-tuning Approaches

Recently, the Segment Anything Model (SAM) has gained significant attention as a robust foundation model for image segmentation [32, 59]. However, the adaptation of SAM from general-purpose image segmentation to more specialized domains, such as medical imaging, presents considerable challenges [29, 40, 52, 61].

Prompt-based SAM variants, which leverage user-defined prompts to guide the segmentation process, have shown promise in improving the model’s performance on specific tasks. For instance, SAM-Path [65] and MedSAM [39], have demonstrated improved accuracy when providing manual prompts. However, the complexity of images and the need for clinical expertise make manual annotation impractical, while prompts introduce ambiguity due to varying interpretations and inconsistent capture of object structures.

Hence, prompt-free fine-tuning is proposed without the need for user-defined prompts. For instance, SAMed [66] employs Low-Rank Adaptation (LoRA) [26] to fine-tune the SAM model with labeled image data. Similarly, the

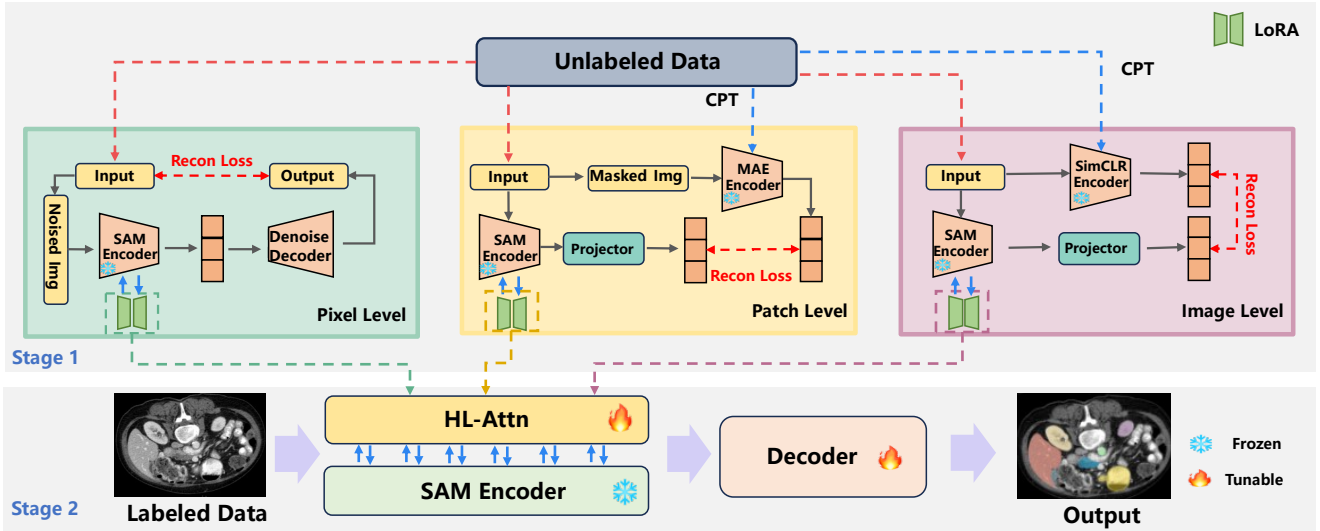


Figure 2. **The Overview of SAMora.** The training process of SAMora is divided into two stages. Stage 1 involves self-supervised pre-training using different LoRA experts across hierarchical levels. Each level employs a distinct self-supervised learning method: SimCLRv2 for the image level, MAE for the patch level, and denoising autoencoder for the pixel level. Continual Pre-Training (CPT) is applied to adapt the teacher models (SimCLRv2 and MAE) to the medical imaging domain. Stage 2 focuses on fine-tuning with labeled data, where the SAM encoder and LoRA experts remain frozen, and only the HL-Attn and Decoder components are tuned. The projector is a trainable dimension-alignment module.

Medical-SAM-Adapter [13] leverages model adapters [17] to fine-tune SAM. However, given the limited amount of labeled data, which constrains the model’s ability to learn all domain-specific information, we introduced unlabeled data by pre-training before fine-tuning.

## 2.2. Self-Supervised Learning (SSL)

Self-Supervised Learning (SSL) has emerged as a paradigm for utilizing unlabeled data in pre-training models due to its ability to learn robust feature representations without the need for manual annotations. Among the various SSL approaches, contrastive learning methods such as SimCLR [11], MoCo [22], and InfoNCE [55] have been widely adopted for their effectiveness in distinguishing between similar and dissimilar samples by contrasting positive pairs with negative pairs, which is suited for capturing global relationships and broader patterns.

Additionally, reconstruction-based approaches have gained considerable attention for their ability to capture features at finer scales. masked autoencoder (MAE) [23], denoising autoencoders [18, 54], and other related techniques [2, 5] such as I-JEPA focuses on reconstructing corrupted or missing parts of the data. MAE, for example, emphasizes the reconstruction of missing patches in an image, which is ideal for learning intermediate features at the patch level, where understanding localized structures is key. Unfortunately, few researchers have explored the combination of different SSL methods to fully leverage their unique characteristics for improving downstream task performance.

## 2.3. Multi-LoRA Fusion

The primary goal of multi-LoRA fusion techniques is to enhance model performance by effectively combining the outputs of different LoRA experts [42, 67], each of which may be specialized for different tasks or datasets.

Linear Arithmetic Composition (LAC) [27, 51, 64] involves a simple linear combination of the outputs from each LoRA, such as LoraHub [28]. However, this approach often fails to preserve the unique characteristics of each LoRA block. On the other hand, tuning-based composition [20] has been developed specifically for the vision-and-language domain. However, this method notably limited the flexibility of LoRA fusion, primarily due to its reliance on manually-designed masks.

To address these limitations, the Mixture of LoRA Experts (MOLE) [56] introduces a gating mechanism that dynamically adjusts the contribution of each LoRA block based on the input data. However, MOLE feeds different experts with the same or similar sampled subsets. *In contrast, our method introduces different levels of image features—ranging from image-level to patch-level—into the model, preventing capturing duplicated information across various model components.*

## 3. Method

### 3.1. Model Overview

We propose SAMora, which leverages self-supervised learning and explores hierarchical feature fusion as a potential enhancement, as shown in Fig. 2. SAMora follows a

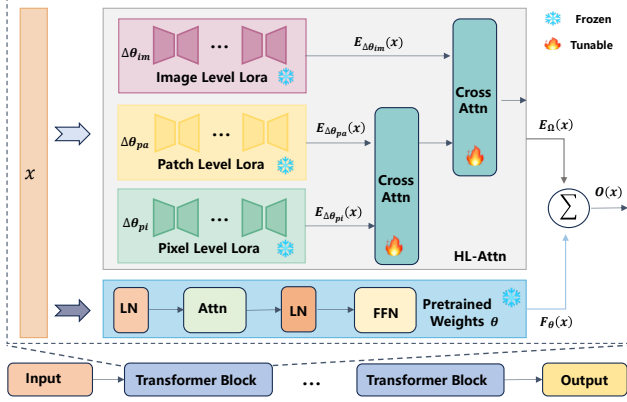


Figure 3. **The Structure of HL-Attn.** Note that self-attention is not visualized in this figure.

two-stage process:

In stage 1, we pre-train the SAM model using LoRA, capitalizing on unlabeled medical image data abundance. Specifically, we employ multiple self-supervised learning methods: contrastive learning for image-level features, MAE for patch-level features, and denoising for pixel-level features. The LoRA experts are then passed into stage 2, where we fine-tune the SAM model using a small amount of labeled data. To effectively integrate the three LoRA experts, we propose HL-Attn (Hierarchical LoRA Attention), a hierarchical cross-attention mechanism that fuses features across different levels. This process involves sequentially merging features from the pixel level, patch level, and image level through multiple layers of cross-attention. During training, we freeze the SAM encoder and the pre-trained LoRA weights, allowing only the HL-Attn and Decoder weights to be updated. This approach ensures that the model benefits from the robust feature representations learned during pre-training while optimizing the integration of multi-level features for improved performance.

### 3.2. SAMora: Self-Supervised Pre-Training Stage

To capture the unique characteristics at three levels effectively, we employ different self-supervised learning methods tailored to each scale, as illustrated in Fig. 3.

**Image-Level Pre-Training: Contrastive Learning.** At the image level, the focus is on capturing global structures that are crucial for identifying broad anatomical features. Hence, we utilize SimCLRv2 [12].

As shown in Fig. 2, the pretext task at the image level employs a teacher-student framework [25] to distill knowledge from the teacher model into the student model. In this setup, the SimCLRv2 encoder functions as the teacher network, transferring its learned representations, while the SAM encoder, augmented with LoRA, serves as the student network that receives and integrates this distilled knowledge. Specifically, the teacher network is initialized using ResNet50 (2X+SK) [21] weights.

Since these weights were originally trained on the ImageNet dataset, their performance in the medical domain is limited. To address this, we further **continual pre-train** [16, 46] the SimCLRv2 weights on a dataset of 100,000 unlabeled medical images to better align the model’s representations.

Given a mini-batch of augmented examples, the contrastive loss between a pair of positive examples  $i, j$  (which are augmentations of the same image) is followed by SimCLRv2.

After fine-tuning, we freeze the weights of the SimCLRv2 model. During training, the SAM encoder remains frozen, with only the corresponding LoRA weights being updated. Additionally, following the approach used in EfficientSAM [59], we employ a reconstruction loss to optimize.

**Patch-Level Pre-Training: MAE.** At the patch level, we identify smaller anatomical regions or organs. To achieve this, we utilize a masked autoencoder (MAE) that helps the model learn the relationships between patches by reconstructing randomly masked sections of the input images. This process enhances the model’s ability to capture intermediate features critical for detailed analysis. Specifically, we initialize the MAE encoder with ViT-Large weights, originally trained on the ImageNet dataset. However, similar to the approach at the image level, we recognize that these weights may lack domain-specific information relevant to medical images. Therefore, before the distillation process, we perform continual pre-training of the MAE encoder on a dataset of 100,000 unlabeled medical images.

Next, we also use a teacher-student framework with the MAE encoder guiding the SAM encoder with LoRA by minimizing the reconstruction loss.

**Pixel-Level Pre-Training: Denoising.** At the pixel level, the focus is on capturing fine-grained details, such as subtle tissue variations, critical for downstream tasks like segmentation. To achieve this, we utilize a denoising autoencoder, training the model to remove noise from input images. Given the relatively straightforward nature of the denoising task and the lack of large-scale pre-trained weights, we combine the SAM encoder with the U-Net decoder as our denoising model, which is optimized by the reconstruction loss.

**Loss Functions of Self-Supervised Pre-Training Stage.** Despite the differences in detailed implementation across three levels, the underlying principle across all these models remains consistent: the core of each loss function is fundamentally based on reconstruction loss.

$$\mathcal{L}_{\text{recon}} = \frac{1}{n} \sum_{i=1}^n \|F(x_i) - G(x_i)\|^2 \quad (1)$$

where  $n$  is the number of data iteration,  $F(*)$  and  $G(*)$  represent functions specific to the task. Specifically, at the image and patch levels,  $F(*)$  represents the teacher network for input  $x_i$ , and  $G(*)$  is the student network. At pixel level,

$F(*) = \mathbb{1}(*)$  indicates that no processing is applied to the input image,  $G(*)$  represents the denoising autoencoder.

### 3.3. SAMora: Fine-Tuning Stage

We fine-tune the overview SAM model during this stage, focusing on the decoder, using a smaller labeled dataset. Furthermore, we design an effective fusion strategy to combine the features from multiple LoRA experts and ensure that each block’s strengths are utilized to their fullest potential.

**Fusing Multi-LoRA.** Previous fusion approaches, such as LAC, tend to diminish the unique characteristics of each individual LoRA block, while MoLE does not fully exploit the distinct features present at different hierarchical levels, limiting the model’s ability to capture multi-level representations effectively. Thus, we propose a hierarchical fusion module (HL-Attn) based on a cross-attention mechanism to fuse the features sequentially.

Referring to Fig. 3, consider a transformer block of SAM encoder, parameterized by  $\theta$  (encompassing both the multi-head attention layer and the feed-forward neural network), and multiple LoRA experts  $\Omega = \{\Delta\theta_{im}, \Delta\theta_{pa}, \Delta\theta_{pi}\}$ , which indicates image-level LoRA, patch-level LoRA, pixel-level LoRA respectively. When given an input  $\mathbf{x} \in \mathbb{R}^{L \times d}$ , the output of the pre-trained model block  $\theta$  is presented as  $\mathbf{F}_\theta \in \mathbb{R}^{L \times d}$ :

$$\mathbf{x}'_\theta = \mathbf{x} + f_{\text{Attn}}(\text{LN}(\mathbf{x}) \mid \theta) \quad (2)$$

$$\mathbf{F}_\theta(\mathbf{x}) = \mathbf{x}'_\theta + f_{\text{FFN}}(\text{LN}(\mathbf{x}'_\theta) \mid \theta) \quad (3)$$

where  $L$  and  $d$  indicate the sequence length and the dimension of  $\mathbf{x}$ , respectively.  $f_{\text{Attn}}(\cdot)$  and  $f_{\text{FFN}}(\cdot)$  denotes the multi-head attention layer and feed-forward neural network, respectively. LN refers to layer normalization. The output of each LoRA is presented as  $\mathbf{E}_{\Delta\theta_i}(\mathbf{x}) \in \mathbb{R}^{L \times d}$ .

$$\mathbf{x}'_{\Delta\theta_i} = \mathbf{x} + f_{\text{Attn}}(\text{LN}(\mathbf{x}) \mid \Delta\theta_i) \quad (4)$$

$$\mathbf{E}_{\Delta\theta_i}(\mathbf{x}) = \mathbf{x}'_{\Delta\theta_i} + f_{\text{FFN}}(\text{LN}(\mathbf{x}'_{\Delta\theta_i}) \mid \Delta\theta_i) \quad (5)$$

After that, we apply HL-Attn to fuse the outputs of multiple LoRA.

$$\mathbf{E}_\Omega(\mathbf{x}) = f_{\text{HL-Attn}}(\mathbf{x}_{\Delta\theta_{im}}, \mathbf{x}_{\Delta\theta_{pa}}, \mathbf{x}_{\Delta\theta_{pi}}) \quad (6)$$

Finally, the final output of this block is computed by adding the output of the HL-Attn to the output of the pre-trained block:

$$\mathbf{O}(\mathbf{x}) = \mathbf{F}_\theta(\mathbf{x}) + \mathbf{E}_\Omega(\mathbf{x}). \quad (7)$$

**Sequential Fusion.** Given that our approach involves hierarchically fusing LoRA experts from three different levels, it is essential to determine the optimal fusion order. Drawing on insights from DINOv2 [41], we prioritize the fusion of patch-level and pixel-level features, which capture

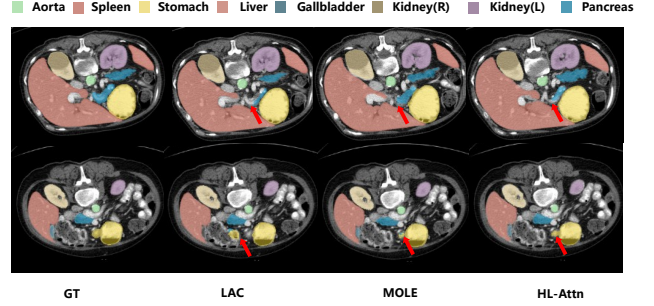


Figure 4. The Performance of SAMora on Synapse Dataset.

more fine-grained image information, before incorporating the broader, image-level features.

**Cross-Attention.** It is particularly well-suited by using cross-attention for this hierarchical fusion strategy because it selectively fuses information across different levels [36, 50]. Cross-attention mechanisms can facilitate a more effective fusion of diverse representations by dynamically focusing on the most relevant features at each hierarchical level. Specifically, for each fusion step, the feature from the higher-level LoRA block is used as the query (Q), while the features from the lower-level LoRA experts are utilized as the key (K) and value (V). The cross-attention mechanism can then be computed as follows:

$$f_{\text{Cr-Attn}}(Q_H, K_L, V_L) = \text{softmax}\left(\frac{Q_H K_L^T}{\sqrt{d_k}}\right) V_L \quad (8)$$

where the  $Q_H = W_q \cdot \mathbf{E}_{\Delta\theta_H}(\mathbf{x})$ ,  $K_L = W_k \cdot \mathbf{E}_{\Delta\theta_L}(\mathbf{x})$ ,  $V_L = W_v \cdot \mathbf{E}_{\Delta\theta_L}(\mathbf{x})$ ,  $\mathbf{E}_{\Delta\theta_H}(\mathbf{x})$  is the feature from higher level LoRA, while the  $\mathbf{E}_{\Delta\theta_L}(\mathbf{x})$  is the feature from lower level LoRA,  $d_k$  is the dimension of the key vectors.

**Fine-tuning with Labeled Data.** Unlike other prompt-free SAM variants, our approach results in distinct LoRA experts after self-supervised stage training, requiring only fine-tuning of the Decoder. Specifically, we freeze the SAM encoder and the pre-trained LoRA weights, allowing the HL-Attn module and the subsequent Decoder to be fine-tuned. By simplifying the training process in this way, we can focus on optimizing the later stages of the model, reducing overall complexity.

**Flexibility.** Although SAMora is primarily built upon the SAMed architecture, it maintains flexibility in adapting to other prompt-free SAM variants. For instance, we combined SAMora with H-SAM and SAMed-2, resulting in two new models, H-SAMora and SAMora-2. Notably, SAMed-2 follows a similar approach to SAMed by applying LoRA to fine-tune the SAM2 model. This adaptability allows SAMora to inherit and integrate improvements from various SAM-based models while ensuring efficient fine-tuning in different medical imaging tasks.

**Loss Functions of Fine-Tuning Stage.** We trained SAMora, H-SAMora and SAMora-2 using the respective loss functions from SAMed, H-SAM and SAMora-

Table 1. **Performance Comparison of SAM and SAM2 Variants on Synapse Dataset.** Bold numbers indicate the best performance. By default, we utilize SAM as our base model. † indicates H-SAM based model; \* indicates SAM2 based model. *The full table is provided in the Appendix B.7.*

Training Set	Method	Spleen	Right Kidney	Left Kidney	Gallbladder	Liver	Stomach	Aorta	Pancreas	Mean Dice ↑	HD ↓
10%	AutoSAM [34]	68.80	77.44	76.53	24.87	88.06	52.70	75.19	34.58	55.69	31.67
	SAMed [66]	85.82	82.25	82.62	63.15	92.72	67.20	78.72	52.12	75.57	23.02
	SAMora (Ours)	<b>88.04</b>	<b>83.41</b>	<b>86.07</b>	<b>67.33</b>	<b>94.27</b>	<b>69.20</b>	<b>82.85</b>	<b>64.13</b>	<b>79.41</b>	<b>15.68</b>
	SAMed-2*	86.61	83.01	84.56	61.51	91.07	69.02	77.99	52.09	76.68	18.93
	SAMora-2* (Ours)	<b>87.81</b>	<b>85.73</b>	<b>86.35</b>	<b>68.30</b>	<b>93.78</b>	<b>75.24</b>	<b>81.12</b>	<b>63.62</b>	<b>80.24</b>	<b>16.27</b>
	H-SAM [15]	90.21	84.16	85.65	70.70	94.29	76.10	85.54	56.17	80.35	15.54
H-SAMora† (Ours)	<b>92.46</b>	<b>85.13</b>	<b>86.71</b>	<b>73.15</b>	<b>95.82</b>	<b>81.85</b>	<b>88.56</b>	<b>72.72</b>	<b>84.34</b>	<b>11.63</b>	
Fully Supervised	MERIT [43]	92.01	84.85	87.79	74.40	95.26	85.38	87.71	71.81	84.90	13.22
	SAMed [66]	87.77	69.11	80.45	79.95	94.80	72.17	88.72	82.06	81.88	20.64
	SAMora (Ours)	<b>89.27</b>	<b>74.05</b>	<b>81.04</b>	<b>81.51</b>	<b>94.97</b>	<b>74.53</b>	<b>88.87</b>	<b>82.42</b>	<b>83.33</b>	<b>14.57</b>
	SAMed-2*	88.63	68.63	81.22	80.33	95.18	71.00	87.63	81.93	82.12	12.76
	SAMora-2* (Ours)	<b>91.78</b>	<b>75.85</b>	<b>82.02</b>	<b>83.52</b>	<b>95.49</b>	<b>75.11</b>	<b>87.11</b>	<b>82.26</b>	<b>84.14</b>	<b>10.28</b>
	H-SAM [15]	93.34	89.93	91.88	73.49	95.72	87.10	89.38	71.11	86.49	8.18
H-SAMora† (Ours)	<b>94.62</b>	<b>91.45</b>	<b>93.00</b>	<b>76.55</b>	<b>96.51</b>	<b>89.95</b>	<b>89.55</b>	<b>77.09</b>	<b>88.59</b>	<b>7.09</b>	

2, specifically leveraging dice loss and cross-entropy loss ( $L_{dice}$  and  $L_{ce}$ ) as follows,

$$\mathcal{L} = \lambda_{ce} \mathcal{L}_{ce} + \lambda_{dice} \mathcal{L}_{dice} \quad (9)$$

where the  $L_{dice}$  and  $L_{ce}$  denote dice loss and cross-entropy loss, respectively. And the  $\lambda_{ce}$  and  $\lambda_{dice}$  are set to 0.2 and 0.8, respectively.

## 4. Experiment

### 4.1. Experimental Setup and Evaluation Metrics

**Datasets.** For pre-training with unlabeled data, we sampled 100,000 CT images from the Amos22 [31], LiTS [6], KiTS [24], and Decathlon Challenge [1] datasets. The pre-processing steps followed the Fed-MENU [60], and finally, the input images were resized to 224×224 before pre-training. Furthermore, we conduct fine-tuned experiments on Synapse [33], LA [8], and PROMISE12 [37] datasets.

**Baselines.** We benchmarked our approach against several prompt-free SAM variants, including MA-SAM [9], I-MedSAM [53], AutoSAM [34], SAM Adapter [14], SAMed [66], and H-SAM [15]. Additionally, we also compared our model with several SOTA methods that are not based on SAM. These include SwinUnet [7], TransDeepLab [3], DAE-Former [4], and MERIT [43]. Furthermore, we also compare SAMora and its variants against semi-supervised methods across various datasets, including UA-MT [63], SS-Net [58], MC-Net [57] and DTC [38]. We followed the data split and preprocessing protocols from H-SAM [15]. The corresponding part of the dataset was designated as labeled data, while the remaining portion was

treated as unlabeled data to facilitate semi-supervised training.

**Evaluation Metrics.** We utilize the Dice coefficient [45] and the average Hausdorff distance (HD) [47] as evaluation metrics.

$$\text{Dice} = \frac{2|A \cap B|}{|A| + |B|}, \quad \text{HD} = \frac{1}{N} \sum_{i=1}^N d_H(A_i, B_i) \quad (10)$$

**Evaluation Protocol.** To ensure a fair comparison, the segmentation result is evaluated on the complete test volumes, following the protocol established by H-SAM [15].

- The Synapse dataset consists of 3,779 contrast-enhanced axial abdominal CT images, with 2,212 slices used in the training set. To effectively demonstrate the efficiency of SAMora, we also fine-tuned the model using only 10% of the training data. Following the H-SAM protocol, we evaluated the segmentation of eight abdominal organs: aorta, gallbladder, spleen, left kidney, right kidney, liver, pancreas, and stomach.
- The left atrial (LA) dataset is derived from the 2018 Atrial Segmentation Challenge [8]. We strictly follow H-SAM for data split and data pre-processing. Specifically, we only keep (4/100)(5%) scans as labeled data to fine-tune, followed by H-SAM.
- PROMISE2012 dataset is derived from the Prostate MR Image Segmentation 2012 [37]. We strictly follow the data split and pre-processing methods of H-SAM. Specifically, we only keep 7.5% (3/40) scans as labeled data to fine-tune followed by H-SAM.

**Implementation Details.** All implementations use PyTorch, with all models trained on eight NVIDIA RTX A100 GPUs. The SAM (ViT-B) and SAM2 (hiera-base-plus)

Table 2. Comparison of SAM Variants against Semi-Supervised Methods across Various Datasets. *The full table is provided in the Appendix B.7.*

Method	10% Synapse	5% LA	7.5% PROMISE12
SS-Net [58]	56.74	<b>86.33</b>	<b>73.19</b>
MC-Net [57]	<b>61.20</b>	83.59	72.66
SAMed [66]	75.57	87.72	86.00
SAMora (Ours)	<b>79.41</b>	<b>90.13</b>	<b>88.44</b>
SAMed-2	76.68	87.91	86.50
SAMora-2 (Ours)	<b>80.24</b>	<b>91.04</b>	<b>89.27</b>
H-SAM [15]	80.35	89.22	87.27
H-SAMora (Ours)	<b>84.34</b>	<b>92.46</b>	<b>90.14</b>

backbone are utilized throughout the entire training process. We combine data augmentation techniques, including elastic deformation, rotation, and scaling. The training loss is a combination of Cross-Entropy loss and Dice loss. For all LoRA experts used in this paper, we adopt the same settings as in SAMed and H-SAM in which the rank of LoRA is set to 4. For fairness in comparison, we use the same image resolution of 224×224 on the Synapse dataset, aligning with other SAM variants and SOTA methods.

## 4.2. Results

Tab. 1 and Tab. 2 compare different SAM variants across various datasets, evaluating their performance using Mean Dice(%) and HD as the primary metrics.

**Firstly**, SAMora demonstrates exceptional few-shot transferability, achieving impressive results even when fine-tuned on only a fraction of the available data. Compared to other prompt-free SAM variants and other SOTA models, SAMora, SAMora-2, and H-SAMora achieved remarkable Mean Dice scores of 79.41%, 80.24%, and 84.34%, respectively. Furthermore, Tab. 2 shows that SAMora and its variants also achieve the SOTA performance against other SOTA semi-supervised methods across the 10% Synapse, the 5% LA and 7.5% PROMISE12 dataset.

**Secondly**, when evaluated on the full Synapse dataset (100% Synapse), SAMora and its variants continued to show improvements, We also conducted *statistical validation* to confirm the significance of our performance improvements in Appendix B.6.

**Thirdly**, Tab. 3 shows the number of fine-tuning epochs and the total parameter among SAMora and its variants. The results indicate that compared to other prompt-free SAM variants, SAMora and H-SAMora demonstrate significantly achieving high performance with considerably 10% training epochs.

**Visualization.** Fig. 4 shows the performance of our proposed model, SAMora, on the Synapse dataset. These results highlight the effectiveness of SAMora, further emphasizing its potential as a robust solution for medical image

Table 3. Comparison of Fine-tuning Efficiency and Performance of SAMora, SAMora-2, H-SAMora.

Method	Fine-tuning Epochs	Total Parameter (M)	Mean Dice (%)
SAMed	200	108.8	75.57
SAMora (Ours)	20	118.5	<b>79.41</b>
SAMed-2	200	99.1	76.68
SAMora-2 (Ours)	20	109.7	<b>80.24</b>
H-SAM	300	112.3	80.35
H-SAMora (Ours)	30	122.0	<b>84.34</b>

Table 4. Effectiveness of Different Multiple LoRA experts Fusion Strategies. *The full table is provided in the Appendix B.3.*

Image-level LoRA	Patch-level LoRA	Pixel-level LoRA	Fusion Module	Mean Dice (%)
✓	✓	✓	LAC [64]	82.41
✓	✓	✓	MOLE [56]	83.91
✓	✓	✓	LoRAHub [28]	81.07
✓	✓	✓	HL-Attn (ours)	<b>84.34</b>
1	1	2	HL-Attn (ours)	84.21
1	2	1	HL-Attn (ours)	83.86
2	1	1	HL-Attn (ours)	<b>84.34</b>

segmentation tasks. *Fig. 6 illustrates the complementary nature of multiple LoRA modules.*

## 4.3. Model Analysis

**Different LoRA Fusion.** As shown in Tab. 4 (top) and Fig. 4, we compared against two alternative LoRA fusion modules: Linear Arithmetic Composition (LAC) [27, 64] and Mixture of LoRA Experts (MOLE) [56]. LAC exhibits the lowest performance, confirming that simple linear weighting is prone to diminishing each trained LoRA’s unique characteristics. In contrast, HL-Attn not only better preserves the individual characteristics of each LoRA but also fuses them hierarchically, leading to superior task performance. *The details are further elaborated in Tab. 10 in Appendix B.2.*

**Sequences of Fusion.** To illustrate the importance of the fusion sequence in HL-Attn, as shown in Tab. 4 (bottom), we conducted experiments evaluating the impact of different LoRA block fusion orders on H-SAMora’s performance on the Synapse dataset. Here, we compare various fusion sequences, where ”1” represents earlier fusion and ”2” indicates later fusion of LoRA experts. The results show that the sequence where the LoRA experts obey the fusion strategy of 2-1-1 yields the highest Dice of 84.34%, while the 1-2-1 sequence yields the lowest Dice of 83.86%. The result demonstrates that the hierarchical fusion strategy indeed influences the performance of models, suggesting that the high-level spatial structures are relatively complex. This underscores the need for an effective approach to fusing various hierarchical features to fully capture the intricacies of these structures.

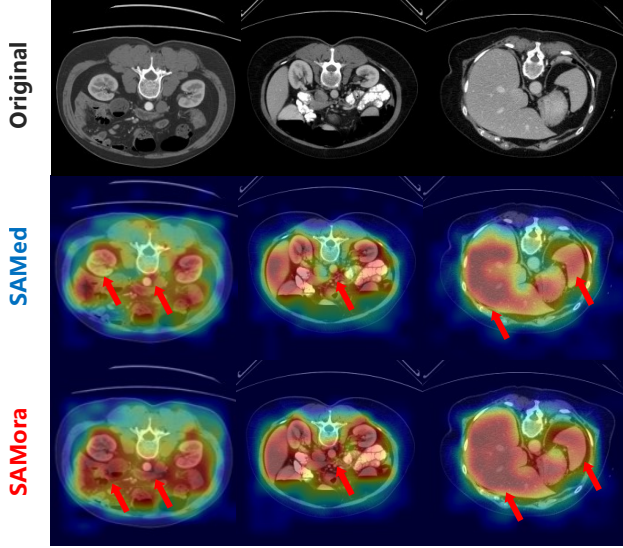


Figure 5. **The Visual Heatmaps between SAMed and SAMora.** The heatmaps display regions of interest with varying levels of relevance, where red denotes areas of high attention, yellow indicates moderate attention, and blue represents low or no attention Table 5. **Ablation Analysis of Multiple LoRA experts on 10% Synapse.** “Scratch” means the model is trained from scratch, while “T-S” indicates the model is trained by the Teacher-Student framework. *The full table is provided in the Appendix B.7.*

Image-level LoRA	Patch-level LoRA	Pixel-level LoRA	Mean Dice (%)
Scratch	✗	✗	77.20
T-S (w/o CPT)	✗	✗	77.31
T-S (w/ CPT)	✗	✗	<b>78.03</b>
✗	Scratch	✗	76.54
✗	T-S (w/o CPT)	✗	77.19
✗	T-S (w/ CPT)	✗	<b>78.81</b>
✗	✗	Scratch	76.97

**Visual Interpretation.** To enhance the interpretability of the model and improve transparency in segmentation, Fig. 5 visualizes the attention maps between SAMed and SAMora.

On the one hand, the focus regions of SAMed are somewhat scattered, failing not only to fully capture all relevant areas but also covering some irrelevant regions. On the other hand, SAMora exhibits a more focused attention on critical anatomical structures, accurately covering all target organ regions.

#### 4.4. Ablation Studies

**Effectiveness of each LoRA.** In Tab. 5, we performed ablation studies on 10% Synapse dataset to evaluate the contribution of each LoRA block at different levels—image, patch, and pixel—to the overall performance of models. By selectively retaining only one LoRA block during fine-tuning, we observed the distinct impact of each level. The results indicate that the patch-level LoRA achieved the highest Mean Dice score of 83.02%, demonstrating its sig-

Table 6. **Performance Comparison of Different Models.**

Model	MA-SAM	I-MedSAM	SAMora	SAMora-2	H-SAMora
AMOS22	82.70	86.26	90.51	90.77	<b>91.84</b>
BTCV	83.12	85.67	89.87	91.06	<b>92.51</b>
Synapse	72.69	75.11	79.41	80.24	<b>84.34</b>
Inference Time	4.3	3.4	<b>3.1</b>	4.2	5.7

nificant contribution to the model’s effectiveness in capturing intermediate-level features.

**Effectiveness of Teacher-Student Framework.** We employed distillation techniques based on the Teacher-Student framework to pre-train the SAM encoder at the image and patch level in stage 1. Tab. 5 presents a comparison between the distillation-based training approach and direct training of the SAM encoder.

In particular, the results indicate that when using the teacher-student framework (T-S) with CPT, the model achieves higher performance (78.81% and 83.02%) compared to the models trained directly without distillation. This suggests that distilling knowledge based on Teacher-Student framework can effectively capture more nuanced and hierarchical features within medical images.

**Effectiveness of Continual Pre-Training.** Furthermore, Tab. 5 also highlights the significant impact of CPT during the distillation process. The results clearly demonstrate that models utilizing CPT exhibit substantial improvements compared to those without it, underscoring the importance of this step in enhancing the model’s performance. *Tab. 12 in Appendix B.5 also shows the effectiveness of the CPT.*

**Comparison of computational efficiency.** Tab.6 presents a performance comparison of different models. Inference time is measured in seconds (s). The results demonstrate that SAMora and its variants achieve an optimal balance between segmentation accuracy and efficiency, consistently outperforming I-MedSAM [53] and MA-SAM [9] across all datasets. Notably, SAMora achieves superior segmentation accuracy while maintaining a lower inference time, reinforcing its practical applicability in real-world scenarios.

## 5. Conclusion

In this paper, we propose to integrate three hierarchical levels of self-supervised knowledge. Additionally, we designed an HL-Attn fusion module to effectively fuse hierarchical medical knowledge. Our experiments on the LA, PROMISE12, and Synapse datasets in fully supervised and few-shot settings demonstrate that SAMora and its variants consistently outperform other strategies, achieving SOTA performance with a mean Dice score of 84.34%.

## Acknowledgement

The authors would like to acknowledge the financial support from National Science Foundation China grant No. 62173298.

## References

- [1] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature communications*, 13(1):4128, 2022. 6
- [2] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15619–15629, 2023. 3
- [3] Reza Azad, Moein Heidari, Moein Shariatnia, Ehsan Khodapanah Aghdam, Sanaz Karimijafarbigloo, Ehsan Adeli, and Dorit Merhof. Transdeeplab: Convolution-free transformer-based deeplab v3+ for medical image segmentation. In *International Workshop on PRedictive Intelligence In MEDicine*, pages 91–102. Springer, 2022. 6
- [4] Reza Azad, René Arimond, Ehsan Khodapanah Aghdam, Amirhossein Kazerouni, and Dorit Merhof. Dae-former: Dual attention-guided efficient transformer for medical image segmentation. In *International Workshop on PRedictive Intelligence In MEDicine*, pages 83–95. Springer, 2023. 6
- [5] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 3
- [6] Patrick Bilic, Patrick Christ, Hongwei Bran Li, Eugene Vorontsov, Avi Ben-Cohen, Georgios Kaissis, Adi Szeskin, Colin Jacobs, Gabriel Efrain Humpire Mamani, Gabriel Chartrand, et al. The liver tumor segmentation benchmark (lits). *Medical Image Analysis*, 84:102680, 2023. 6
- [7] 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. In *European conference on computer vision*, pages 205–218. Springer, 2022. 6
- [8] Chen Chen, Wenjia Bai, and Daniel Rueckert. Multi-task learning for left atrial segmentation on ge-mri. In *Statistical Atlases and Computational Models of the Heart. Atrial Segmentation and LV Quantification Challenges: 9th International Workshop, STACOM 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers 9*, pages 292–301. Springer, 2019. 6
- [9] Cheng Chen, Juzheng Miao, Dufan Wu, Aoxiao Zhong, Zhiling Yan, Sekeun Kim, Jiang Hu, Zhengliang Liu, Lichao Sun, Xiang Li, et al. Ma-sam: Modality-agnostic sam adaptation for 3d medical image segmentation. *Medical Image Analysis*, 98:103310, 2024. 6, 8
- [10] Li Chen, Penghao Wu, Kashyap Chitta, Bernhard Jaeger, Andreas Geiger, and Hongyang Li. End-to-end autonomous driving: Challenges and frontiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1
- [11] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 3
- [12] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255, 2020. 2, 4
- [13] Tianrun Chen, Lanyun Zhu, Chaotao Deng, Runlong Cao, Yan Wang, Shangzhan Zhang, Zejian Li, Lingyun Sun, Ying Zang, and Papa Mao. Sam-adapter: Adapting segment anything in underperformed scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3367–3375, 2023. 3
- [14] Tianrun Chen, Lanyun Zhu, Chaotao Ding, Runlong Cao, Yan Wang, Zejian Li, Lingyun Sun, Papa Mao, and Ying Zang. Sam fails to segment anything?—sam-adapter: Adapting sam in underperformed scenes: Camouflage, shadow, medical image segmentation, and more. *arXiv preprint arXiv:2304.09148*, 2023. 1, 6
- [15] Zhiheng Cheng, Qingyue Wei, Hongru Zhu, Yan Wang, Liangqiong Qu, Wei Shao, and Yuyin Zhou. Unleashing the potential of sam for medical adaptation via hierarchical decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3511–3522, 2024. 1, 2, 6, 7
- [16] Andrea Cossu, Antonio Carta, Lucia Passaro, Vincenzo Lomonaco, Tinne Tuytelaars, and Davide Bacciu. Continual pre-training mitigates forgetting in language and vision. *Neural Networks*, 179:106492, 2024. 4
- [17] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2): 581–595, 2024. 3
- [18] Lovedeep Gondara. Medical image denoising using convolutional denoising autoencoders. In *2016 IEEE 16th international conference on data mining workshops (ICDMW)*, pages 241–246. IEEE, 2016. 3
- [19] Hanxue Gu, Haoyu Dong, Jichen Yang, and Maciej A. Mazurowski. How to build the best medical image segmentation algorithm using foundation models: a comprehensive empirical study with segment anything model, 2024. 2
- [20] Yuchao Gu, Xintao Wang, Jay Zhangjie Wu, Yujun Shi, Yunpeng Chen, Zihan Fan, Wuyou Xiao, Rui Zhao, Shuning Chang, Weijia Wu, et al. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024. 3
- [21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4
- [22] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 3
- [23] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable

- vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 1, 2, 3
- [24] Nicholas Heller, Niranjana Sathianathan, Arveen Kalapara, Edward Walczak, Keenan Moore, Heather Kaluzniak, Joel Rosenberg, Paul Blake, Zachary Rengel, Makinna Oestreich, et al. The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes. *arXiv preprint arXiv:1904.00445*, 2019. 6
- [25] Chengming Hu, Xuan Li, Dan Liu, Xi Chen, Ju Wang, and Xue Liu. Teacher-student architecture for knowledge learning: A survey. *arXiv preprint arXiv:2210.17332*, 2022. 4
- [26] 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. *arXiv preprint arXiv:2106.09685*, 2021. 2
- [27] Chengsong Huang, Qian Liu, Bill Yuchen Lin, Tianyu Pang, Chao Du, and Min Lin. Lorahub: Efficient cross-task generalization via dynamic lora composition. *arXiv preprint arXiv:2307.13269*, 2023. 3, 7
- [28] Chengsong Huang, Qian Liu, Bill Yuchen Lin, Tianyu Pang, Chao Du, and Min Lin. Lorahub: Efficient cross-task generalization via dynamic lora composition. *arXiv preprint arXiv:2307.13269*, 2023. 3, 7
- [29] Yuhao Huang, Xin Yang, Lian Liu, Han Zhou, Ao Chang, Xinrui Zhou, Rusi Chen, Junxuan Yu, Jiongqun Chen, Chaoyu Chen, et al. Segment anything model for medical images? *Medical Image Analysis*, 92:103061, 2024. 2
- [30] Xiangzuo Huo, Gang Sun, Shengwei Tian, Yan Wang, Long Yu, Jun Long, Wendong Zhang, and Aolun Li. Hifuse: Hierarchical multi-scale feature fusion network for medical image classification. *Biomedical Signal Processing and Control*, 87:105534, 2024. 2
- [31] Yuanfeng Ji, Haotian Bai, Chongjian Ge, Jie Yang, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhanng, Wanling Ma, Xiang Wan, et al. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *Advances in neural information processing systems*, 35:36722–36732, 2022. 6
- [32] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 1, 2
- [33] Bennett Landman, Zhoubing Xu, J Igelsias, Martin Styner, Thomas Langerak, and Arno Klein. Miccai multi-atlas labeling beyond the cranial vault—workshop and challenge. In *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, page 12, 2015. 6
- [34] Chengyin Li, Prashant Khanduri, Yao Qiang, Rafi Ibn Sultan, Indrin Chetty, and Dongxiao Zhu. Auto-prompting sam for mobile friendly 3d medical image segmentation. *arXiv e-prints*, pages arXiv–2308, 2023. 6
- [35] Cong Lin, Yinjie Chen, Siling Feng, and Mengxing Huang. A multibranch and multiscale neural network based on semantic perception for multimodal medical image fusion. *Scientific Reports*, 14(1):17609, 2024. 2
- [36] Hezheng Lin, Xing Cheng, Xiangyu Wu, and Dong Shen. Cat: Cross attention in vision transformer. In *2022 IEEE international conference on multimedia and expo (ICME)*, pages 1–6. IEEE, 2022. 5
- [37] Geert Litjens, Robert Toth, Wendy Van De Ven, Caroline Hoeks, Sjoerd Kerkstra, Bram Van Ginneken, Graham Vincent, Gwenael Guillard, Neil Birbeck, Jindang Zhang, et al. Evaluation of prostate segmentation algorithms for mri: the promise12 challenge. *Medical image analysis*, 18(2):359–373, 2014. 6
- [38] Xiangde Luo, Jieneng Chen, Tao Song, and Guotai Wang. Semi-supervised medical image segmentation through dual-task consistency. In *Proceedings of the AAAI conference on artificial intelligence*, pages 8801–8809, 2021. 6
- [39] Jun Ma, Yao Zhang, Song Gu, Cheng Zhu, Cheng Ge, Yichi Zhang, Xingle An, Congcong Wang, Qiyan Wang, Xin Liu, et al. Abdomenct-1k: Is abdominal organ segmentation a solved problem? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6695–6714, 2021. 2
- [40] 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, 2023. 2
- [41] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 5
- [42] Oleksiy Ostapenko, Zhan Su, Edoardo Maria Ponti, Laurent Charlin, Nicolas Le Roux, Matheus Pereira, Lucas Caccia, and Alessandro Sordani. Towards modular llms by building and reusing a library of loras. *arXiv preprint arXiv:2405.11157*, 2024. 3
- [43] Md Mostafijur Rahman and Radu Marculescu. Multi-scale hierarchical vision transformer with cascaded attention decoding for medical image segmentation. In *Medical Imaging with Deep Learning*, pages 1526–1544. PMLR, 2024. 6
- [44] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 2
- [45] Reuben R Shamir, Yuval Duchin, Jinyoung Kim, Guillermo Sapiro, and Noam Harel. Continuous dice coefficient: a method for evaluating probabilistic segmentations. *arXiv preprint arXiv:1906.11031*, 2019. 6
- [46] Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. Ernie 2.0: A continual pre-training framework for language understanding. In *Proceedings of the AAAI conference on artificial intelligence*, pages 8968–8975, 2020. 4
- [47] Abdel Aziz Taha and Allan Hanbury. An efficient algorithm for calculating the exact hausdorff distance. *IEEE transactions on pattern analysis and machine intelligence*, 37(11):2153–2163, 2015. 6
- [48] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good

- views for contrastive learning? *Advances in neural information processing systems*, 33:6827–6839, 2020. 2
- [49] Fedra Trujillano, Gabriel Jimenez, Edgar Manrique, Najat F Kahamba, Fredros Okumu, Nombre Apollinaire, Gabriel Carrasco-Escobar, Brian Barrett, and Kimberly Fornace. Using image segmentation models to analyse high-resolution earth observation data: new tools to monitor disease risks in changing environments. *International Journal of Health Geographics*, 23(1):13, 2024. 1
- [50] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 5
- [51] Hanqing Wang, Bowen Ping, Shuo Wang, Xu Han, Yun Chen, Zhiyuan Liu, and Maosong Sun. Lora-flow: Dynamic lora fusion for large language models in generative tasks. *arXiv preprint arXiv:2402.11455*, 2024. 3
- [52] Xiaobao Wei, Jiajun Cao, Yizhu Jin, Ming Lu, Guangyu Wang, and Shanghang Zhang. I-medsam: Implicit medical image segmentation with segment anything. *arXiv preprint arXiv:2311.17081*, 2023. 2
- [53] Xiaobao Wei, Jiajun Cao, Yizhu Jin, Ming Lu, Guangyu Wang, and Shanghang Zhang. I-medsam: Implicit medical image segmentation with segment anything. In *European Conference on Computer Vision*, pages 90–107. Springer, 2024. 6, 8
- [54] Ruxue Wen, Hangjie Yuan, Dong Ni, Wenbo Xiao, and Yaoyao Wu. From denoising training to test-time adaptation: Enhancing domain generalization for medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 464–474, 2024. 3
- [55] Chuhan Wu, Fangzhao Wu, and Yongfeng Huang. Rethinking infonce: How many negative samples do you need? *arXiv preprint arXiv:2105.13003*, 2021. 3
- [56] Xun Wu, Shaohan Huang, and Furu Wei. Mixture of lora experts. *arXiv preprint arXiv:2404.13628*, 2024. 3, 7
- [57] Yicheng Wu, Minfeng Xu, Zongyuan Ge, Jianfei Cai, and Lei Zhang. Semi-supervised left atrium segmentation with mutual consistency training. In *Medical image computing and computer assisted intervention—MICCAI 2021: 24th international conference, Strasbourg, France, September 27–October 1, 2021, proceedings, part II 24*, pages 297–306. Springer, 2021. 6, 7
- [58] Yicheng Wu, Zhonghua Wu, Qianyi Wu, Zongyuan Ge, and Jianfei Cai. Exploring smoothness and class-separation for semi-supervised medical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 34–43. Springer, 2022. 6, 7
- [59] Yunyang Xiong, Bala Varadarajan, Lemeng Wu, Xiaoyu Xiang, Fanyi Xiao, Chenchen Zhu, Xiaoliang Dai, Dilin Wang, Fei Sun, Forrest Iandola, et al. Efficientsam: Leveraged masked image pretraining for efficient segment anything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16111–16121, 2024. 2, 4
- [60] Xuanang Xu, Hannah H Deng, Jamie Gateno, and Pingkun Yan. Federated multi-organ segmentation with inconsistent labels. *IEEE transactions on medical imaging*, 42(10):2948–2960, 2023. 6
- [61] Yunqiu Xu, Linchao Zhu, and Yi Yang. Mc-bench: A benchmark for multi-context visual grounding in the era of mllms. *arXiv preprint arXiv:2410.12332*, 2024. 2
- [62] Sihan Yang, Haixia Bi, Hai Zhang, and Jian Sun. Sam-unet: Enhancing zero-shot segmentation of sam for universal medical images. *arXiv preprint arXiv:2408.09886*, 2024. 1
- [63] Lequan Yu, Shujun Wang, Xiaomeng Li, Chi-Wing Fu, and Pheng-Ann Heng. Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. In *Medical image computing and computer assisted intervention—MICCAI 2019: 22nd international conference, Shenzhen, China, October 13–17, 2019, proceedings, part II 22*, pages 605–613. Springer, 2019. 6
- [64] Jinghan Zhang, Junteng Liu, Junxian He, et al. Composing parameter-efficient modules with arithmetic operation. *Advances in Neural Information Processing Systems*, 36: 12589–12610, 2023. 3, 7
- [65] Jingwei Zhang, Ke Ma, Saarthak Kapse, Joel Saltz, Maria Vakalopoulou, Prateek Prasanna, and Dimitris Samaras. Sam-path: A segment anything model for semantic segmentation in digital pathology. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 161–170. Springer, 2023. 2
- [66] Kaidong Zhang and Dong Liu. Customized segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.13785*, 2023. 1, 2, 6, 7
- [67] Ziyu Zhao, Leilei Gan, Guoyin Wang, Wangchunshu Zhou, Hongxia Yang, Kun Kuang, and Fei Wu. Loraretriever: Input-aware lora retrieval and composition for mixed tasks in the wild. *arXiv preprint arXiv:2402.09997*, 2024. 3