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

Supplementary Material

Table 7. Stage1 Setting.

Config	image-level	patch-level	pixel-level
Optimizer	LARS	AdamW	AdamW
Base learning rate	0.075	1.5e-4	1e-4
Batch size	512	512	512
Weight decay	1e-4	0.05	0.05
Warmup period	30	30	10
epoch nums	80	60	30

Table 8. Stage2 Setting.

Config	SAMora	SAMora-2	H-SAMora
Optimizer	AdamW	AdamW	AdamW
Base learning rate	5e-3	5e-3	2.5e-3
Batch size	32	32	32
Weight decay	0.1	0.1	0.1
Warmup period	25	25	25
epoch nums	20	25	30

In this supplementary material, we first provide more implementation details for training strategies and datasets (Sec. A). Then, we conduct more additional ablation studies (Sec. B) to validate the effectiveness of each component in our proposed method. Finally, we discuss SAMora's limitations and potential directions for future work (Sec. C).

A. Implementation Details

A.1. Training strategy

We provide the training strategy and hyperparameter settings as supplementary material.

In Stage 1, we perform pretraining for image-level, patch-level, and pixel-level tasks using different models: SimCLRv2 (ResNet50 (2X+SK)) for the image-level task, MAE (ViT-Large) for the patch-level task, and a denoising model (U-net model) for the pixel-level task. As shown in Table 6, for image-level task, We adopt warmup during training, The learning rate is linearly increased for the first 5% of epochs, and then decayed with a cosine decay schedule where the weight decay is $1e^{-4}$, followed by SimCLRv2 [7]. For the patch-level and pixel-level tasks, we use the AdamW optimizer. The optimizer momentum is set to 0.9 and 0.95 for the patch-level task, and 0.9 and 0.99 for the pixel-level task, respectively.

The training loss is a combination of Dice loss and Mean Squared Error (MSE) loss. As indicated in Tab. 7, the weights for these losses are set to 0.9 for Dice loss and 0.1 for MSE loss. In our two-stage hierarchical structure, each stage applies a weighted loss, controlled by a parameter that gradually decreases through exponential decay, starting from 0.4 and reaching 0 over 300 epochs.

In Tab. 8, we present the settings for Stage 2 across various backbones. For SAMora, SAMed (ViT-B) serves as the backbone. The loss weights are assigned as 0.2 for cross-entropy and 0.8 for Dice loss. For the warmup configuration, the initial learning rate is set to 0.005, with a warmup period of 250 steps, and the total number of iterations is 18,600. Notably, the learning rate adjustment strategy is described as follows:

$$lr = \begin{cases} T\frac{I_{lr}}{WP}, & T <= WP, \\ I_{lr}\left(1 - \frac{T-WP}{MI}\right), & T > WP. \end{cases} \quad (1)$$

Where I_{lr} represents the initial learning rate, while T, WP, and MI denote the training iterations, warmup period, and maximum iterations, respectively.

SAMora-2 uses SAM2 (hiera-base-plus) as the backbone, with the only difference being that the number of epochs is set to 25. All other training parameters remain the same as those of SAMora. The configuration of H-SAMora follows the guidelines of H-SAM [9].

A.2. Additional datasets information

We detail the dataset settings. Firstly, the unlabeled data that we use to pre-train is sampled from the Amos22 [14], LiTS [4], KiTS [11], and Decathlon Challenge [1] datasets.

- AMOS22 [14] is a large-scale dataset that provides 500 CT and 100 MRI scans with voxel-level annotations for 15 abdominal organs, supporting both CT-only and cross-modality segmentation tasks across diverse clinical scenarios.
- The LiTS dataset [4] focuses on liver and liver tumor segmentation. It comprises 201 abdominal CT volumes, helping to tackle challenges such as lesion variability and segmentation complexity, making it a widely used benchmark for medical imaging algorithms.
- The KiTS dataset [11] emphasizes kidney and kidney tumor segmentation. Its 2019 release, KiTS19, includes 300 CT cases, collected from patients who underwent nephrectomy, and is designed to support automated kidney and tumor segmentation research through comprehensive annotations.
- The Decathlon Challenge dataset [1] offers a broad range of segmentation tasks across multiple organs, aiming to advance generalization in medical image analysis. It provides an opportunity to test algorithms on various

Table 9. Ablation study on rank size of LoRA layers.

Method	Rank = 1	Rank = 4	Rank = 16
SAMed	69.12	75.57	69.03
SAMora (Ours)	75.26	79.41	76.88
SAMed-2	69.89	76.68	73.54
SAMora-2 (Ours)	75.53	80.24	76.12
H-SAM	72.14	80.35	77.14
H-SAMora (Ours)	**78.91**	**84.34**	**80.57**

anatomical regions and imaging scenarios, making it ideal for benchmarking segmentation models across different tasks.

In Stage 2, we utilize the Synapse dataset from the MICCAI 2015 Multi-Atlas Abdomen Labeling Challenge. For the fully supervised training setup, we adhere to the H-SAM framework to evaluate the segmentation performance across eight abdominal organs: the aorta, gallbladder, spleen, left kidney, right kidney, liver, pancreas, and stomach.

In addition to the fully supervised setup, we also implement a few-shot learning scenario. For this, we adopt a slice-based data selection strategy, randomly sampling 10% of the training data (221 slices) from different subjects within the complete training set, which consists of 2,212 axial slices.

A.3. Preprocessing and augmentation strategies for training datasets

To improve the generalization ability of the model and enhance the robustness of training, we follow the preprocessing and data augmentation strategies adopted in TransUNet [6], SAMed [27], H-SAM [9].

The original medical images are first resampled to a uniform spatial resolution to mitigate variations caused by different imaging protocols. Following TransUNet, for 3D volumetric data, each volume is processed in a slice-by-slice manner, where the slices are extracted along the axial plane. The extracted 2D slices are then normalized to zero mean and unit variance to ensure consistent intensity distributions across different datasets.

To prevent overfitting and improve the diversity of training samples, we employ several data augmentation techniques:

- **Random rotation:** Each image slice is randomly rotated by an angle within $[-15°, 15°]$ to simulate different orientations.
- **Random flipping:** Horizontal and vertical flipping are applied with a probability of 0.5 to introduce spatial variability.
- **Scaling:** The images are randomly scaled within the range $[0.9, 1.1]$ to enhance robustness to size variations.
- **Elastic deformation:** Spatially elastic transformations are applied to simulate realistic deformations in medical images.
- **Contrast and brightness adjustment:** To account for variations in image acquisition settings, we randomly adjust the contrast and brightness of images.

These augmentation strategies ensure that the model learns from diverse image distributions while preserving anatomical structures.

All preprocessing and augmentation operations are implemented using standard deep learning libraries, and applied online during training to maximize variability in training samples.

Table 10. Effectiveness of HL-Attn compared to HCAT.

Model	Mean Dice(%)	Inference Time(s)
HCAT	**84.40**	4.2
HL-Attn (ours)	84.34	**3.1**

B. Additional analysis

B.1. Ablation study on the LoRA component

We also conduct our additional ablation studies on 10% Synapse dataset. In the Tab. 9, we compare the effectiveness of the layers of LoRA component among these models and their variants. From the result, we found that all models and their variants, the best performance is achieved when the rank increases to 4, while the performance drops when the rank increases to 16.

Furthermore, the model incorporating multiple LoRA experts exhibits a smaller performance gap compared to the original model at different rank values, suggesting that the proposed mechanisms enhance the model's robustness to variations in the rank parameter.

B.2. Additional study on the HL-Attn

Our work focuses on proposing an innovative multi-level framework that integrates existing methods in a novel way to address specific challenges in medical image analysis. While we build upon widely recognized techniques like MAE and SimCLRv2, leveraging strong foundations is common and necessary in advancing research. The novelty of HL-Attn lies in the hierarchical design and effective combination of these methods, with a focus on simplicity and adaptability. Even with a straightforward fusion strategy, our approach demonstrates significant gains. To further validate the effectiveness of our method, we conducted experiments on the hierarchical cross-attention transformer (HCAT). The results (Tab. 10), demonstrate that HL-Attn achieves comparable mean Dice scores with a reduction in inference time, highlighting the efficiency of our framework.

Table 11. **Full Effectiveness of Different Multiple LoRA experts Fusion Strategies**

Image-level LoRA	Patch-level LoRA	Pixel-level LoRA	Fusion Module	10% Synapse	5% LA	7.5% PROMISE12
✓	✓	✓	LAC [26]	82.41	90.01	88.97
✓	✓	✓	MOLE [21]	83.91	91.59	89.44
✓	✓	✓	LoRAHub [12]	81.07	88.31	87.43
✓	✓	✓	HL-Attn (ours)	**84.34**	**92.46**	**90.14**
1	1	2	HL-Attn (ours)	84.21	92.10	89.95
1	2	1	HL-Attn (ours)	83.86	91.80	89.72
2	1	1	HL-Attn (ours)	**84.34**	**92.46**	**90.14**

B.3. Effectiveness of Different Multiple LoRA experts Fusion Strategies

We have further supplement our experiments on the LA and PROMISE12 datasets to provide a more comprehensive assessment of SAMora's segmentation performance. The results in Tab. 11 show that HL-Attn outperforms other fusion strategies across both datasets, achieving the highest mean Dice scores. This demonstrates the effectiveness of our proposed method in enhancing segmentation performance across different medical imaging tasks.

B.4. Complementarity of multiple LoRAs

Furthermore, the Fig. 6 illustrates the complementarity across the three LoRA levels. It shows that individual levels fail to capture certain structural details, while the fusion image effectively integrates these features, resulting in improved overall accuracy. This highlights how the hierarchical fusion leverages distinct strengths from each level. These visual results demonstrate that the modifications to the model architecture have successfully guided the network to concentrate on the most relevant features.

B.5. Clarification of Training Time

The CPT process can be seen as an equivalent fine-tuning phase for SimCLRv2 and MAE. For CPT, we sampled 100,000 images from datasets like AMOS and employed a comprehensive pre-training process integrating SimCLRv2 and MAE to effectively learn hierarchical features. As shown in the table, models with shorter CPT durations demonstrate that SAMora can balance efficiency and performance. H-SAMora-T1, which excludes CPT and performs minimal pre-training, achieves a Mean Dice of 80.72, slightly outperforming H-SAM. H-SAMora-T2, with a reduced CPT duration of 0.8 hours, improves further to 80.97. The full CPT version, H-SAMora, achieves the highest Mean Dice of 84.34, highlighting the benefits of a complete pre-training process. These results confirm SAMora's adaptability to different resource constraints, as even shorter CPT durations deliver significant improvements, while the default CPT duration maximizes performance and demonstrates the framework's full potential. The detailed training configurations and results will be presented in the revised manuscript.

Table 12. Results of different training time of CPT

Model	CPT	Pre-Training	Fine-Tuning	Mean Dice
H-SAM	-	-	2	80.35
H-SAMora-T1	-	1.8	0.1	80.72
H-SAMora-T2	0.8	1.3	0.1	80.97
H-SAMora	12.7	13.4	0.1	**84.34**

B.6. Statistical validation

To address this concern, we conducted statistical validation to confirm the significance of our performance improvements on the Synapse dataset. We performed a paired t-test on the mean Dice scores of SAMora, SAMed, SAMora-2, SAMed-2, H-SAMora, and H-SAM. The results show highly significant differences, such as H-SAMora versus H-SAM with a p-value of 1.7×10^{-8} and a 95% confidence interval of [0.0329, 0.0407]. Similarly, SAMora outperforms SAMed with a p-value of 0.0167 and SAMora-2 outperforms SAMed-2 with a p-value of 0.0281. These statistical tests validate the robustness and significance of the reported improvements, and the detailed analysis will be included in the revised manuscript.

Although SAMora performs well in most medical image segmentation tasks, its performance may degrade when handling noisy or low-quality images. Future research could focus on improving the model's robustness to such challenging image quality issues.

B.7. Complete Experimental Results

This section presents the full experimental results only partially included in the main text, providing a more comprehensive evaluation of the proposed method. Table 13 offers a detailed performance comparison of SAM and SAM2 variants on the Synapse dataset, where bold numbers indicate the best performance. Table 14 extends the comparison by benchmarking various SAM variants against multiple semi-supervised methods across different datasets. Additionally, Table 15 provides a complete ablation analysis of multiple LoRA experts on the 10% Synapse dataset, where "Scratch" refers to models trained from scratch. At the same time, "T-S" denotes training using a teacher-student

Figure 6. Complementarity of multiple LoRAs.

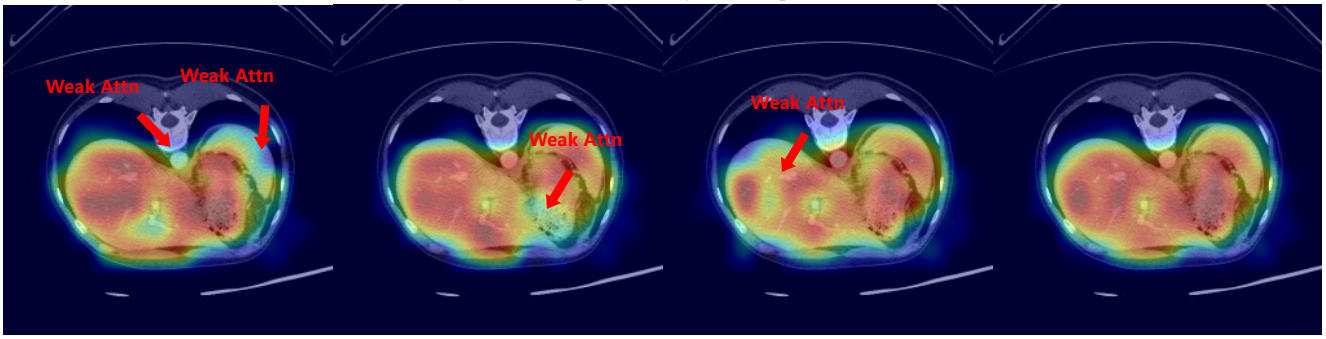

Table 13. **Full Performance Comparison of SAM and SAM2 Variants on Synapse Dataset.** **Bold** numbers indicate the best performance. By default, we utilize SAM as our base model. † indicates H-SAM based model; * indicates SAM2 based model.

Training Set	Method	Spleen	Right Kidney	Left Kidney	Gallbladder	Liver	Stomach	Aorta	Pancreas	Mean Dice ↑	HD ↓
10%	AutoSAM [15]	68.80	77.44	76.53	24.87	88.06	52.70	75.19	34.58	55.69	31.67
	SAM Adapter [8]	72.42	68.38	66.77	22.38	89.69	53.15	66.74	26.76	58.28	54.22
	SAMed [27]	85.82	82.25	82.62	63.15	92.72	67.20	78.72	52.12	75.57	23.02
	SAMora (Ours)	**88.04**	**83.41**	**86.07**	**67.33**	**94.27**	**69.20**	**82.85**	**64.13**	**79.41**	**15.68**
	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)	**87.81**	**85.73**	**86.35**	**68.30**	**93.78**	**75.24**	**81.12**	**63.62**	**80.24**	**16.27**
	H-SAM [9]	90.21	84.16	85.65	70.70	94.29	76.10	85.54	56.17	80.35	15.54
	H-SAMora† (Ours)	**92.46**	**85.13**	**86.71**	**73.15**	**95.82**	**81.85**	**88.56**	**72.72**	**84.34**	**11.63**
Fully Supervised	TransUNet [6]	81.87	85.08	77.02	63.16	94.08	75.62	87.23	55.86	77.49	31.69
	UNETR [10]	85.60	85.00	84.52	56.30	94.57	70.46	89.80	60.47	78.35	18.59
	SwinUnet [5]	85.47	66.53	83.28	79.61	94.29	56.58	90.66	76.60	79.13	21.55
	TransDeepLab [2]	86.04	69.16	84.08	79.88	93.53	61.19	89.00	78.40	80.16	21.25
	DAE-Former [3]	88.96	72.30	86.08	**80.88**	94.98	65.12	91.94	79.19	82.43	17.46
	MERIT [19]	**92.01**	84.85	**87.79**	74.40	95.26	85.38	87.71	71.81	84.90	13.22
	nnFormer [28]	86.57	90.51	86.25	70.17	**96.84**	**86.83**	92.04	**83.35**	86.57	10.63
	UNETR++ [20]	87.54	**95.77**	87.18	71.25	96.42	86.01	**92.52**	81.10	**87.22**	**7.53**
	SAM Adapter [8]	83.68	**79.00**	79.02	57.49	92.67	69.48	77.93	43.07	72.80	33.08
	SAM3D [24]	84.29	85.64	**86.31**	49.81	95.42	**76.11**	**89.57**	69.32	79.56	17.87
	SAMed [27]	87.77	69.11	80.45	79.95	94.80	72.17	88.72	82.06	81.88	20.64
	SAMora (Ours)	**89.27**	74.05	**81.04**	**81.51**	**94.97**	74.53	88.87	**82.42**	**83.33**	**14.57**
	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)	**91.78**	**75.85**	**82.02**	**83.52**	**95.49**	**75.11**	**87.11**	**82.26**	**84.14**	**10.28**
	H-SAM [9]	93.34	89.93	91.88	73.49	95.72	87.10	89.38	71.11	86.49	8.18
	H-SAMora† (Ours)	**94.62**	**91.45**	**93.00**	**76.55**	**96.51**	**89.95**	**89.55**	**77.09**	**88.59**	**7.09**

framework. These tables collectively reinforce the findings and conclusions drawn in the main text, offering more profound insights into the effectiveness of the proposed approach.

C. Limitation and Future Work

Despite the promising results of SAMora, several limitations need to be addressed in future research.

While SAMora reduces the reliance on labeled data through self-supervised learning, it still requires some labeled data for fine-tuning. Therefore, further exploration of fully unsupervised data is needed. On the other hand, we observe that weakly labeled data, compared to fully labeled data, has been widely applied in research due to its lower cost and reduced need for manual annotation, which makes it more scalable and practical in real-world applications [16, 18]. Consequently, future work will explore integrating weakly labeled data to enhance SAMora's per-

Table 14. **Full Comparison of SAM Variants against Semi-Supervised Methods across Various Datasets**

Method	10% Synapse	5% LA	7.5% PROMISE12
nnUnet [13]	-	64.02	84.22
UA-MT [25]	-	82.26	65.05
SS-Net [23]	56.74	86.33	73.19
MC-Net [22]	61.20	83.59	72.66
DTC [17]	-	81.25	63.44
AutoSAM [15]	55.69	74.73	68.40
SAM Adapter [8]	58.28	82.79	75.45
SAMed [27]	75.57	87.72	86.00
SAMora (Ours)	**79.41**	**90.13**	**88.44**
SAMed-2	76.68	87.91	86.50
SAMora-2 (Ours)	**80.24**	**91.04**	**89.27**
H-SAM [9]	80.35	89.22	87.27
H-SAMora (Ours)	**84.34**	**92.46**	**90.14**

Table 15. **Full 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

Image-level LoRA	Patch-level LoRA	Pixel-level LoRA	Model	Mean Dice (%)
Scratch	✗	✗	SAMora	77.20
T-S (w/o CPT)	✗	✗	SAMora	77.31
T-S ($w/$ CPT)	✗	✗	SAMora	**78.03**
Scratch	✗	✗	H-SAMora	82.09
T-S (w/o CPT)	✗	✗	H-SAMora	82.17
T-S ($w/$ CPT)	✗	✗	H-SAMora	**82.65**
✗	**Scratch**	✗	SAMora	76.54
✗	**T-S** (w/o CPT)	✗	SAMora	77.19
✗	**T-S** ($w/$ CPT)	✗	SAMora	**78.81**
✗	**Scratch**	✗	H-SAMora	81.67
✗	**T-S** (w/o CPT)	✗	H-SAMora	82.04
✗	**T-S** ($w/$ CPT)	✗	H-SAMora	**83.02**
✗	✗	**Scratch**	SAMora	76.97
✗	✗	**Scratch**	H-SAMora	81.58

formance, allowing it to better generalize across a broader range of medical image segmentation tasks.

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. 1

[2] 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. 4

[3] 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. 4

[4] 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. 1

[5] 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. 4

[6] 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, 4

[7] 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. 1

[8] 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. 4, 5

[9] 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, 4, 5

[10] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 574–584, 2022. 4

[11] Nicholas Heller, Niranjan Sathianathen, 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. 1

[12] 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

[13] Fabian Isensee, Paul F Jaeger, Simon AA 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, 2021. 5

[14] 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. 1

[15] 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. 4, 5

[16] Han Liu, Zhoubing Xu, Riqiang Gao, Hao Li, Jianing Wang, Guillaume Chabin, Ipek Oguz, and Sasa Grbic. Cosst: Multi-organ segmentation with partially labeled datasets using comprehensive supervisions and self-training. *IEEE Transactions on Medical Imaging*, 2024. 4

[17] 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. 5

[18] Olivier Petit, Nicolas Thome, and Luc Soler. Iterative confidence relabeling with deep convnets for organ segmentation with partial labels. *Computerized Medical Imaging and Graphics*, 91:101938, 2021. 4

[19] 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. 4

[20] Abdelrahman M Shaker, Muhammad Maaz, Hanoona Rasheed, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Unetr++: delving into efficient and accurate 3d medical image segmentation. *IEEE Transactions on Medical Imaging*, 2024. 4

[21] Xun Wu, Shaohan Huang, and Furu Wei. Mixture of lora experts. *arXiv preprint arXiv:2404.13628*, 2024. 3

[22] 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. 5

[23] 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. 5

[24] Yunhan Yang, Xiaoyang Wu, Tong He, Hengshuang Zhao, and Xihui Liu. Sam3d: Segment anything in 3d scenes. *arXiv preprint arXiv:2306.03908*, 2023. 4

[25] 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. 5

[26] 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

[27] Kaidong Zhang and Dong Liu. Customized segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.13785*, 2023. 2, 4, 5

[28] Hong-Yu Zhou, Jiansen Guo, Yinghao Zhang, Lequan Yu, Liansheng Wang, and Yizhou Yu. nnformer: Interleaved transformer for volumetric segmentation. *arXiv preprint arXiv:2109.03201*, 2021. 4