







MoE-SAM: Enhancing SAM for Medical Image Segmentation with Mixture-of-Experts

Ruocheng Li^{†1}, Lei Wu^{†1,2}, Jingjun Gu¹, Qi Xu¹, Wanyi Chen¹,
Xiaoxu Cai³, and Jiajun Bu¹

¹ Zhejiang Key Laboratory of Accessible Perception and Intelligent Systems, College of Computer Science and Technology, Zhejiang University

² Hangzhou Pujian Medical Technology Co., Ltd, China

³ The Rural Development Academy, Zhejiang University, China

{lirc,shenhai1895,gjj,xuqi591,chenwanyi,xiaoxu.cai,bjj}@zju.edu.cn

Abstract. Recent adaptations of the powerful and promptable Segment Anything Model (SAM), pretrained on a large-scale dataset, have shown promising results in medical image segmentation. However, existing methods fail to fully leverage the intermediate features from SAM’s image encoder, limiting its adaptability. To address this, we introduce MoE-SAM, a novel approach that enhances SAM by incorporating a Mixture-of-Experts (MoE) during adaptation. Central to MoE-SAM is a MoE-driven feature enhancing block, which uses learnable gating functions and expert networks to select, refine, and fuse latent features from multiple layers of SAM’s image encoder. By combining these features, the model creates a more robust image embedding that captures both low-level local and high-level global information. This comprehensive embedding facilitates prompt embedding generation and mask decoding, thereby enabling more effective self-prompting segmentation. Extensive evaluations across four benchmark medical image segmentation tasks show that MoE-SAM outperforms both task-specialized models and other SAM-based approaches, achieving state-of-the-art segmentation accuracy. The code is available at: <https://github.com/Asphyxiate-Rye/E-SAM>.

Keywords: Medical Image Segmentation · SAM · Mixture-of-Experts.

1 Introduction

Medical image segmentation is crucial in modern healthcare, enabling precise diagnosis, treatment planning, and disease monitoring. With the advent of deep learning, significant progress has been made in this area, leading to ongoing improvements in accuracy and efficiency. One key recent advancement is the application of the Segment Anything Model (SAM) [14] to medical images, which has emerged as a powerful approach. SAM operates by accepting a prompt (such as a point or a bounding box) from the user and then segments the corresponding

[†] Equal contribution; ✉ corresponding author.

region in the image. Trained on a large-scale dataset of over a billion masks, SAM shows remarkable generalization to unseen data [23,22], making it a strong foundation model for tasks in fields with costly data acquisition and annotation, such as medical imaging. However, since SAM’s training set primarily consists of natural images, and given the gap between natural and medical images, applying SAM to medical images is nontrivial and requires careful adaptation [29].

Existing approaches to adapting SAM for medical image segmentation generally fall into two categories: fine-tuning the model or modifying its architecture. As a pioneering effort in fine-tuning, MedSAM [20] fine-tunes all SAM parameters on a large-scale medical image dataset. While it achieves impressive segmentation performance, this full fine-tuning approach incurs substantial computational overhead. To mitigate this, several studies have employed LoRA technology [12], which adjusts smaller, low-rank decompositions of the model’s weight matrices during fine-tuning. For instance, SAMed [29] uses LoRA to fine-tune SAM’s image encoder, while Feng et al. [8] apply LoRA to fine-tune SAM’s mask decoder. Alternatively, adapting SAM by adding extra layers or networks has also proven effective. A notable example is Medical SAM Adapter [28], which enhances SAM by inserting adapter layers into the image encoder. SAMUS [17], on the other hand, introduces a CNN-based encoder parallel to SAM’s image encoder and uses cross-encoder fusion to obtain a more robust image embedding. DeSAM [9] modifies SAM’s mask decoder by introducing a prompt-relevant IoU module and a prompt-decoupled mask module to extract multi-scale features. While these methods show efficiency, we argue that they do not fully leverage the feature embeddings from SAM’s image encoder. SAM employs a Vision Transformer (ViT) [6] as its image encoder to effectively capture semantically rich representations. As highlighted in [6], the Mean Attention Distance across different ViT layers reveals substantial variation in “receptive fields”: lower layers attend to fine-grained local details, while higher layers capture broader global context. However, most existing SAM-based approaches [27,10] rely solely on the final layer’s embedding or on four stages, thereby overlooking the fine-grained semantics embedded in earlier layers. This neglect of the hierarchical semantic representations within SAM’s encoder may limit its ability to achieve accurate segmentation.

To address the limitation outlined above, we propose MoE-SAM, a novel pipeline that integrates deep global and shallow local features from SAM’s image encoder using Mixture-of-Experts (MoE) technology to enhance SAM’s adaptation to medical image segmentation. At the core of MoE-SAM is a MoE-driven feature enhancing block, which selects, refines, and fuses features from multiple layers of SAM’s image encoder using learnable gating functions and expert networks. The resulting features are then summed with SAM’s image encoder output, generating a strengthened image embedding that facilitates mask decoding. We also propose a lightweight prompt embedding generator that creates a prompt embedding directly from the image embedding, offering a more efficient self-prompted segmentation approach. Additionally, to avoid costly full-tuning of SAM’s image encoder, we insert adapters into the encoder’s transformer blocks,

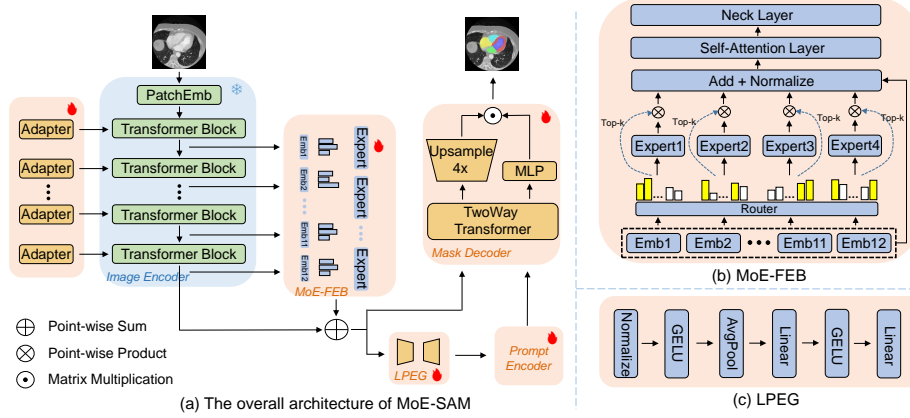


Fig. 1: The pipeline of MoE-SAM. MoE-SAM features with a set of adapters for efficient fine-tuning, (a) a MoE-driven Feature Enhancing Block (MoE-FEB) and (b) a Lightweight Prompt Embedding Generator (LPEG).

updating only adapters while freezing the pretrained encoder during training. This strategy strikes a balance between training complexity and the preservation of pre-learned knowledge. We validate our method on four public benchmark medical image datasets, demonstrating its superiority over state-of-the-art task-specific segmentation methods and SAM-based approaches. Our results also emphasize the crucial role of shallow encoder layers in segmentation performance, highlighting the importance of leveraging multi-level features.

In summary, the main contributions of this work are as follows:

- We introduce MoE-SAM, a novel adaptation of the SAM architecture for medical image segmentation. MoE-SAM leverages Mixture-of-Experts to enhance image embedding by selecting and integrating features from multiple layers of SAM’s image encoder. It also includes a lightweight prompt embedding generator that automatically learns a prompt embedding from the enhanced image embedding, enabling more efficient self-prompting segmentation. Additionally, MoE-SAM incorporates image encoder adapters for efficient fine-tuning.
- We conduct extensive experiments on four benchmark medical image datasets, demonstrating that MoE-SAM outperforms the current state-of-the-art by a large margin.

2 Method

The proposed method, MoE-SAM, adapts the pretrained vision foundation model SAM for medical image segmentation. As shown in Fig. 1, MoE-SAM consists of three main components: Image Encoder Adapters, a MoE-driven Feature Enhancing Block, and a Lightweight Prompt Embedding Generator. The adapters

are integrated into the transformer blocks of SAM’s image encoder for efficient fine-tuning. The Feature Enhancing Block, which is central to MoE-SAM, uses Mixture-of-Experts (MoE) to select, refine, and fuse features from multiple SAM encoder layers. These features are then combined with the SAM encoder output through simple addition, resulting in a more robust image embedding for prompting and mask decoding. The prompt embedding generator learns a prompt embedding from the image embedding, enabling SAM to operate in a self-prompting manner. During training, MoE-SAM freezes SAM’s pretrained image encoder and fine-tunes the adapters, feature enhancing block, prompt embedding generator, and mask decoder, achieving a balance between reducing computational overhead and preserving SAM’s pre-learned knowledge. The key components are detailed below.

Image Encoder Adapters SAM’s image encoder, ViT-B, has around 86 million parameters, making full parameter tuning highly resource-intensive. To address this, we insert an adapter layer into each Transformer block [28] of the image encoder, while keeping the encoder frozen and only tuning the adapters. Specifically, the adapter is placed in the residual path of the Transformer block, following the multi-head attention layer. Given input \mathbf{X} , the adapter layer is formulated as:

$$\text{Adapter}(\mathbf{X}) = \mathbf{X} + \sigma(\mathbf{X}W_{\text{down}})W_{\text{up}} \quad (1)$$

where $\sigma(\cdot)$ denotes the activation function, $W_{\text{down}} \in \mathbb{R}^{c \times c1}$ and $W_{\text{up}} \in \mathbb{R}^{c1 \times c}$ represent the down-projection and up-projection layers, respectively. Similar to [3], we scale the embedding output from the adapter by a factor s before it reaches the end of the residual path.

MoE-driven Feature Enhancing Block Features from different layers of SAM’s image encoder capture varying levels of semantic information: lower layers focus on local details, while higher layers capture more global context. However, existing SAM-based methods typically rely solely on the features from the final layer for mask decoding, which limits segmentation performance. To address this limitation, we propose using MoE with Expert Choice Routing Mechanism [30] to selectively combine features from multiple layers of the image encoder, generating a more robust image embedding that incorporates both local and global information. MoE was originally designed to assign different experts to distinct input samples, enabling more efficient and scalable models. In this study, we adapt MoE for feature enhancement. Specifically, we construct a set of learnable expert networks, with each expert paired with a learnable gating function. The gating function selects intermediate features from SAM’s image encoder, which are then passed to the corresponding expert for refinement by interacting with features from other encoder layers. Given a group of features $X \in \mathbb{R}^{n \times d}$ (n is the number of features, which equals the number of encoder layers, and d is the

feature dimensionality.), the gating function $H(x)$ is formulated as follows:

$$H(X) = g(X, \theta)^T + R_{\text{noise}}, \quad (2)$$

$$I = \text{TopK}(H(x), k), \quad (3)$$

$$S_I = \text{Softmax}(H(X)[I]). \quad (4)$$

Here, $g(X, \theta)$, parameterized by learnable weights θ , computes raw logits that indicate the preference of the expert for each feature. As in [26, 7], $H(x)$ includes a noise term R_{noise} to encourage exploration among experts and improve the stability of MoE training. Next, we select the k features with the largest $H(x)$ values using the TopK operation, and normalize the selected values with a softmax function. With the feature indices I , we retrieve a subset of X , denoted as X_I , which forms the input for the corresponding expert. The expert is implemented as a multilayer perceptron (MLP), with the following formulation:

$$\hat{X}_I = \text{ReLU}(X_I \cdot W_1) \cdot W_2 \quad (5)$$

Where $W_1 \in \mathbb{R}^{d \times d'}$ and $W_2 \in \mathbb{R}^{d' \times d}$ are the parameters of the Feedforward Neural Network. For features chosen by more than one expert, we obtain the final representation by summing the results of all the corresponding experts:

$$X_{\text{final}} = \sum_I S_I \cdot \hat{X}_I \quad (6)$$

Here, S_I represents the softmax value associated with the feature, weighting its contribution to the final feature representation. With the MoE operation, we obtain a set of enhanced features, each refined from an intermediate feature of the original image encoder by incorporating information from features of other layers. These enhanced features are further processed by a self-attention layer followed by a convolution-based neck layer, resulting in a strong auxiliary image embedding. The self-attention layer captures interactions across different input regions, while the neck layer refines the features for better integration of local and global information.

Lightweight Prompt Embedding Generator This module takes enhanced image embeddings as input and generates prompt embeddings for prompt decoding. Specifically, it first applies Adaptive Average Pooling to extract global contextual information, ensuring that the prompt embedding captures a comprehensive understanding of the image. Next, a two-layer bottleneck architecture (Linear-GELU-Linear) is applied, which transforms the distribution of the prompt embedding. This shift in representation moves from being image-centric to prompt-centric, achieved through compression and reconstruction. The self-prompting mechanism improves SAM’s flexibility and enables dynamic adaptation to various image inputs.

Loss Function The overall training loss function is formulated as:

$$\mathcal{L} = (1 - \lambda)\mathcal{L}_{ce} + \lambda\mathcal{L}_{dice} \quad (7)$$

where \mathcal{L}_{ce} and \mathcal{L}_{dice} denote binary cross-entropy loss and dice loss, respectively [21]. The weighting parameter λ controls the balance between these two losses and is set to 0.8 in our experiments.

3 Experiments

3.1 Experimental Setup

Datasets We conduct experiments on four public medical datasets: MMWHS (Multi-Modality Whole Heart Segmentation) [32,31], Synapse Multi-Organ CT [15], BTCV (Beyond The Cranial Vault) [15], and ACDC (Automated Cardiac Diagnosis Challenge) [1]. MMWHS consists of 20 3D cardiac CT scans, with 16 for training and 4 for testing. Synapse contains 18 training cases and 12 testing cases, covering eight abdominal organs. BTCV provides 30 labeled CT scans (24 for training and 6 for testing) spanning 13 anatomical structures. ACDC includes 150 labeled cardiac MRI cases, focusing on the right ventricle, myocardium, and left ventricle. We follow the official splits for each dataset.

Evaluation Metrics We use the widely accepted Dice Similarity Coefficient (DSC) and Hausdorff Distance (HD) as evaluation metrics.

Implementation Details To facilitate training, we apply various data augmentation techniques, including flipping, rotation, scaling, and intensity shifting. All images, except those from the Synapse CT dataset, are resized to 256×256 , while Synapse CT images are resized to 224×224 . The model is trained with a batch size of 8 using the AdamW optimizer [19], with hyperparameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, and weight decay = 0.1. The learning rate is set to 0.0005, and a warmup strategy is employed to ensure stable convergence during the early stages. In the MoE configuration, we set the number of experts to 4 and the top-k value to half of the total feature count.

3.2 Comparison with State-of-the-Art Methods

We compare the proposed MoE-SAM with state-of-the-art (SOTA) SAM adaptations for medical image segmentation across the 4 benchmark datasets. These SAM-based methods are categorized into prompt-based and prompt-free adaptations. For the prompt-based adaptations, we use a unified point sampled from the ground-truth segmentation mask as the prompt. The results are listed in Table. 1, where it is evident that MoE-SAM outperforms all other SAM adaptations by a significant margin in both DSC and HD scores. Additionally, we compare MoE-SAM with state-of-the-art task-specialized models. The results, presented in the upper half of Table. 1, show that MoE-SAM surpasses task-specialized models in 6 out of the 8 measures, highlighting its highly competitive performance.

Table 1: Comparison with state-of-the-art methods on Synapse CT, MMWHS, BTCV, and ACDC.

Type	Method	Synapse CT		MMWHS		BTCV		ACDC	
		DSC ↑	HD ↓	DSC ↑	HD ↓	DSC ↑	HD ↓	DSC ↑	HD ↓
Task-special	nnU-Net [13]	79.89	28.52	87.55	17.72	72.67	15.72	91.54	1.086
	TransUNet [2]	79.95	11.58	88.47	24.31	77.67	9.498	88.10	1.538
	Swin-UNETR [11]	80.58	15.46	88.92	14.31	78.11	7.405	89.74	1.239
	MedNeXt [24]	82.69	11.98	88.55	14.95	80.81	7.379	90.88	1.129
	Swin-UMamba [18]	83.48	8.140	88.91	15.06	80.59	5.910	90.39	1.253
prompt-based SAM	SAM (1 pt) [14]	64.94	39.83	82.11	46.94	63.84	20.36	75.15	4.311
	MedSAM (1 pt) [20]	72.45	20.43	84.53	55.94	69.14	18.49	82.11	3.720
	MSA (1 pt) [28]	77.13	25.34	85.50	35.68	72.31	17.51	83.01	2.715
	SAMUS (1 pt) [17,16]	70.55	43.65	83.98	30.74	65.12	24.58	69.66	5.559
	DeSAM (1 pt) [9]	76.77	9.704	81.54	16.34	68.08	7.263	67.26	5.391
	SAM-Med2D (1 pt) [4]	66.96	22.85	81.42	59.66	53.64	23.63	80.02	4.587
prompt-free SAM	SAMed [29]	80.42	10.77	87.05	25.32	71.23	9.010	88.83	1.429
	AutoSAM [25]	81.61	10.17	88.71	12.99	75.77	7.693	72.05	3.248
	H-SAM [5]	80.27	13.17	87.33	14.11	72.81	7.037	88.38	1.410
	MoE-SAM (Ours)	84.71	8.756	89.38	13.67	76.82	5.637	91.89	1.064

3.3 Ablation Studies

We conduct extensive ablation experiments to examine the effectiveness of the key components in MoE-SAM, including the MoE-driven Feature Enhancing Block (MoE-FEB), and the Lightweight Prompt Embedding Generator (LPEG). To validate the fine-tuning strategy based on image encoder adapters, we compare it with the widely used LoRA technique. All experiments are conducted on the MMWHS dataset, using the Dice measure for evaluation. For brevity, we refer to the components as MoE-FEB and LPEG in the following sections. The results are presented in Table 2, which clearly demonstrates that the full MoE-SAM model achieves the best Dice score. The impact of each component is discussed below.

MoE-FEB As shown in the last two rows of Table 2 (both the upper and lower halves), applying the MoE-FEB leads to a significant performance increase. Specifically, the model shows a 1.46% improvement (89.38% vs. 88.09%) when fine-tuned using adapters, and a 1.53% improvement (88.76% vs. 87.42%) when fine-tuned using LoRA. This highlights the crucial role of MoE-FEB in our method. To further explore the reason behind the performance boost, we visualize the features from different layers of SAM’s image encoder and the features enhanced by MoE-FEB in Fig. 2. As shown in Fig. 2(a), the feature maps change progressively across layers, with lower layers capturing local features (e.g., edges

Table 2: Ablation study on the impact of MoE-FEB, LPEG based on different fine-tuning strategy.

MoE-FEB	LP-EG	Fine-tuning Strategy		DSC
		Adapter	LoRA	
		✓		87.24
✓		✓		88.38
	✓	✓		88.09
✓	✓	✓		89.38
			✓	87.05
✓			✓	88.59
	✓		✓	87.42
✓	✓		✓	88.76

Method	Fusion Strategy		DSC
	Add	MoE-FEB	
SAM			82.11
SAM	✓		83.02
SAM		✓	83.60
SAM-Med2D			81.42
SAM-Med2D	✓		81.04
SAM-Med2D		✓	81.96
SAMed			87.05
SAMed	✓		87.90
SAMed		✓	88.59

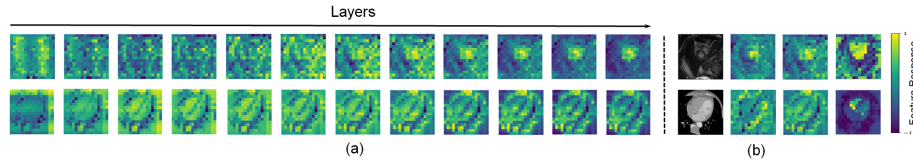


Fig. 2: Feature visualization. (a) Feature representations within the MoE-SAM Image Encoder, with the layers progressing in the direction of the arrow. (b) The first column shows the original input image. The second column presents the baseline feature map from SAM after the neck module, without any fusion. The third column presents the fused feature map, and the fourth column shows the corresponding attention map based on the fused features.

and shapes) and higher layers capturing more abstract information. After processing by MoE-FEB, the feature maps (shown in the third column of Fig. 2(b)) integrate both low-layer and high-layer information, represented by intensified color. The corresponding attention maps demonstrate the same trend. This indicates that MoE-FEB successfully integrates multi-layer features from SAM’s image encoder, resulting in a stronger image embedding for subsequent prompting and mask decoding. We further tested the MoE-FEB with different baseline models, including vanilla SAM, SAM-Med2D [20] and SAMed [29]. As shown in Table 3, our MoE-SAM achieves much better Dice score comparing with the feature fusion approach using multi-layer addition. This demonstrates the strong generalization ability of MoE-FEB.

LPEG The results also highlight the importance of the LPEG, which boosts the model’s performance from 88.38% to 89.38% with adapter-based fine-tuning, and from 88.59% to 88.76% with LoRA-based fine-tuning.

4 Conclusion

In this paper, we introduce MoE-SAM, a novel self-prompting adaptation of SAM using Mixture-of-Experts (MoE) for medical image segmentation. The method features a MoE-driven feature enhancing block, which effectively integrates features from multiple layers of SAM’s image encoder. The resulting enhanced features are combined with the encoder’s output to create a more robust image embedding, which significantly improves the subsequent prompt embedding generation and mask decoding. We extensively evaluate MoE-SAM on four public benchmark datasets, and the results demonstrate its superiority over both state-of-the-art task-specialized and SAM-based approaches.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (Grant No. 62372408) and Hangzhou Pujian Medical Technology Co., Ltd, China and ZJU-Pujian Research & Development Center of Medical Artificial Intelligence for Hepatobiliary and Pancreatic Disease.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this paper.

References

1. Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.A., Cetin, I., Lekadir, K., Camara, O., Ballester, M.A.G., et al.: Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE transactions on medical imaging* **37**(11), 2514–2525 (2018)
2. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306* (2021)
3. Chen, S., Ge, C., Tong, Z., Wang, J., Song, Y., Wang, J., Luo, P.: Adaptformer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems (NeurIPS)* **35**, 16664–16678 (2022)
4. Cheng, J., Ye, J., Deng, Z., Chen, J., Li, T., Wang, H., Su, Y., Huang, Z., Chen, J., Jiang, L., et al.: Sam-med2d. *arXiv preprint arXiv:2308.16184* (2023)
5. Cheng, Z., Wei, Q., Zhu, H., Wang, Y., Qu, L., Shao, W., Zhou, Y.: Unleashing the potential of sam for medical adaptation via hierarchical decoding. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 3511–3522 (2024)
6. Dosovitskiy, A.: An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020)
7. Fedus, W., Zoph, B., Shazeer, N.: Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research* **23**(120), 1–39 (2022)
8. Feng, W., Zhu, L., Yu, L.: Cheap lunch for medical image segmentation by fine-tuning sam on few exemplars. *arXiv preprint arXiv:2308.14133* (2023)
9. Gao, Y., Xia, W., Hu, D., Wang, W., Gao, X.: Desam: Decoupled segment anything model for generalizable medical image segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. pp. 509–519. Springer (2024)

10. Gong, S., Zhong, Y., Ma, W., Li, J., Wang, Z., Zhang, J., Heng, P.A., Dou, Q.: 3dsam-adapter: Holistic adaptation of sam from 2d to 3d for promptable medical image segmentation. arXiv e-prints pp. arXiv-2306 (2023)
11. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: International MICCAI brainlesion workshop. pp. 272–284. Springer (2021)
12. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)
13. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods* **18**(2), 203–211 (2021)
14. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4015–4026 (2023)
15. Landman, B., Xu, Z., Igelsias, J., Styner, M., Langerak, T., Klein, A.: Miccai multi-atlas labeling beyond the cranial vault—workshop and challenge. In: Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge. vol. 5, p. 12 (2015)
16. Lin, X., Xiang, Y., Yu, L., Yan, Z.: Beyond adapting sam: Towards end-to-end ultrasound image segmentation via auto prompting. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). pp. 24–34. Springer (2024)
17. Lin, X., Xiang, Y., Zhang, L., Yang, X., Yan, Z., Yu, L.: Samus: Adapting segment anything model for clinically-friendly and generalizable ultrasound image segmentation. arXiv preprint arXiv:2309.06824 (2023)
18. Liu, J., Yang, H., Zhou, H.Y., Xi, Y., Yu, L., Li, C., Liang, Y., Shi, G., Yu, Y., Zhang, S., et al.: Swin-umamba: Mamba-based unet with imagenet-based pretraining. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). pp. 615–625. Springer (2024)
19. Loshchilov, I.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
20. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. *Nature Communications* **15**(1), 654 (2024)
21. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 fourth international conference on 3D vision (3DV). pp. 565–571. Ieee (2016)
22. Osco, L.P., Wu, Q., de Lemos, E.L., Gonçalves, W.N., Ramos, A.P.M., Li, J., Junior, J.M.: The segment anything model (sam) for remote sensing applications: From zero to one shot. *International Journal of Applied Earth Observation and Geoinformation* **124**, 103540 (2023)
23. Ren, S., Luzi, F., Lahrchi, S., Kassaw, K., Collins, L.M., Bradbury, K., Malof, J.M.: Segment anything, from space? In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 8355–8365 (2024)
24. Roy, S., Koehler, G., Ulrich, C., Baumgartner, M., Petersen, J., Isensee, F., Jaeger, P.F., Maier-Hein, K.H.: Mednext: transformer-driven scaling of convnets for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). pp. 405–415. Springer (2023)

25. Shaharabany, T., Dahan, A., Giryas, R., Wolf, L.: Autosam: Adapting sam to medical images by overloading the prompt encoder. arXiv preprint arXiv:2306.06370 (2023)
26. Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., Dean, J.: Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. arXiv preprint arXiv:1701.06538 (2017)
27. Tang, Y., Yang, D., Li, W., Roth, H.R., Landman, B., Xu, D., Nath, V., Hatamizadeh, A.: Self-supervised pre-training of swin transformers for 3d medical image analysis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 20730–20740 (2022)
28. Wu, J., Ji, W., Liu, Y., Fu, H., Xu, M., Xu, Y., Jin, Y.: Medical sam adapter: Adapting segment anything model for medical image segmentation. arXiv preprint arXiv:2304.12620 (2023)
29. Zhang, K., Liu, D.: Customized segment anything model for medical image segmentation. arXiv preprint arXiv:2304.13785 (2023)
30. Zhou, Y., Lei, T., Liu, H., Du, N., Huang, Y., Zhao, V., Dai, A.M., Le, Q.V., Laudon, J., et al.: Mixture-of-experts with expert choice routing. *Advances in Neural Information Processing Systems* **35**, 7103–7114 (2022)
31. Zhuang, X., Bai, W., Song, J., Zhan, S., Qian, X., Shi, W., Lian, Y., Rueckert, D.: Multiatlas whole heart segmentation of ct data using conditional entropy for atlas ranking and selection. *Medical physics* **42**(7), 3822–3833 (2015)
32. Zhuang, X., Shen, J.: Multi-scale patch and multi-modality atlases for whole heart segmentation of mri. *Medical image analysis* **31**, 77–87 (2016)