3D-SAutoMed: Automatic Segment Anything Model for 3D Medical Image Segmentation from Local-Global Perspective

Junjie Liang[1,2], Peng Cao[1,2,3(✉)], Wenju Yang[1,2], Jinzhu Yang[1,2,3], and Osmar R. Zaiane[4]

[1] Computer Science and Engineering, Northeastern University, Shenyang, China
[2] Key Laboratory of Intelligent Computing in Medical Image of Ministry of Education, Northeastern University, Shenyang, China
[3] National Frontiers Science Center for Industrial Intelligence and Systems Optimization, Shenyang, China
caopeng@cse.neu.edu.cn
[4] Amii, University of Alberta, Edmonton, Alberta, Canada

Abstract. 3D medical image segmentation is critical for clinical diagnosis and treatment planning. Recently, with the powerful generalization, the foundational segmentation model SAM is widely used in medical images. However, the existing SAM variants still have many limitations including lack of 3D-aware ability and automatic prompts. To address these limitations, we present a novel SAM-based segmentation framework in 3D medical images, namely 3D-SAutoMed. We respectively propose the Inter- and Intra-slice Attention and Historical slice Information Sharing strategy to share local and global information, so as to enable SAM to be 3D-aware. Meanwhile, we propose a Box Prompt Generator to automatically generate prompt embedding, leading full automation in SAM. Our results demonstrate that 3D-SAutoMed outperforms advanced universal methods and SAM variants on both metrics and across BTCV, CHAOS and SegTHOR datasets. Particularly, a large improvement of HD score is achieved, e.g. 44% and 20.7% improvement compared with the best result in the other SAM variants on the BTCV and SegTHOR dataset, respectively.

Keywords: Automatic SAM · Local-global 3D-aware · Medical image segmentation

1 Introduction

The Segment Anything Model (SAM) [12] is a foundational segmentation model trained on over 11 million images and 1 billion masks, aiming to segment any object in any image. By employing prompt-driven strategy, SAM demonstrates outstanding zero-shot segmentation performance on unseen datasets and tasks. In medical image segmentation, a limitation is the lack of sufficient data for training models to learn robust representations. Recent studies [10, 8] have shown that

directly applying SAM to medical image segmentation tasks does not yield good performance. This is attributed to the fact that the majority of SAM's training data consists of RGB-based natural images, which exhibit a significant discrepancy compared to medical images (typically MRI, CT scans, etc.). To solve it, some studies [21, 22, 17] attempted to finetune SAM on medical images, achieving remarkable results. Despite these advancements, there are still two significant challenges making the extension of SAM to 3D medical image segmentation tasks potentially problematic.

SAM lacks 3D spatial information perception. SAM is a segmentation model designed for 2D images. However, most medical image data are typically 3D. Most existing improvements based on SAM [21, 22, 17] do not consider adaptation for 3D medical image segmentation, thereby overlooking the 3D spatial information inherent in volume. These methods inevitably limit the accuracy of segmentation and affect the continuity of 3D segmentation results.

SAM requires prompts in each slice when segmenting 3D images. SAM is a prompt-based interactive segmentation model. However, providing prompts for each slice in the entire volume is labor-intensive. In order to reduce manual participation, [15, 18] generate the current slice's box prompt by using an extra detection model or directly using previous slice prediction. However, the quality of the prompts generated by the above methods is coarse and depends on the accuracy of the segmentation model.

To address the aforementioned challenges, we propose a novel segmentation framework based on SAM for 3D medical image segmentation, namely 3D-SAutoMed. 3D-SAutoMed aims to leverage the pre-trained weights of the foundational model SAM and adapt it to automatic 3D medical image segmentation scenarios. Firstly, we apply the LoRA [9] technique in the encoder to learn medical image-specific feature representations during finetuning. Simultaneously, to leverage 3D spatial information, we introduce inter- and intra-slice attention after the encoder to facilitate feature interaction between adjacent slices, enabling local 3D-aware. Meanwhile, to better leverage the global information inherent in the entire volume, we propose a Historical slice Information Sharing (HIS) strategy. In this strategy, we store global information and continuously reuse and update it during the iterative segmentation process to achieve global 3D-aware. Second, we propose a box prompt generator to automatically generate box prompts, motivated by DETR-based methods [2, 20, 16]. We use the bounding box obtained from the previous slice segmentation result to initialize the anchor query of the generator, and predict the target location to obtain the box prompt embedding for the current slice. At the same time, we introduce semantic embedding into the box prompt generator to enhance the role of prompts in semantic segmentation tasks. Our main contributions can be summarized as follows:

- **Enable SAM to be 3D-aware.** We introduce Inter- and Intra-slice Attention and Historical slice Information Sharing strategy to share information from the local and global perspectives, respectively.

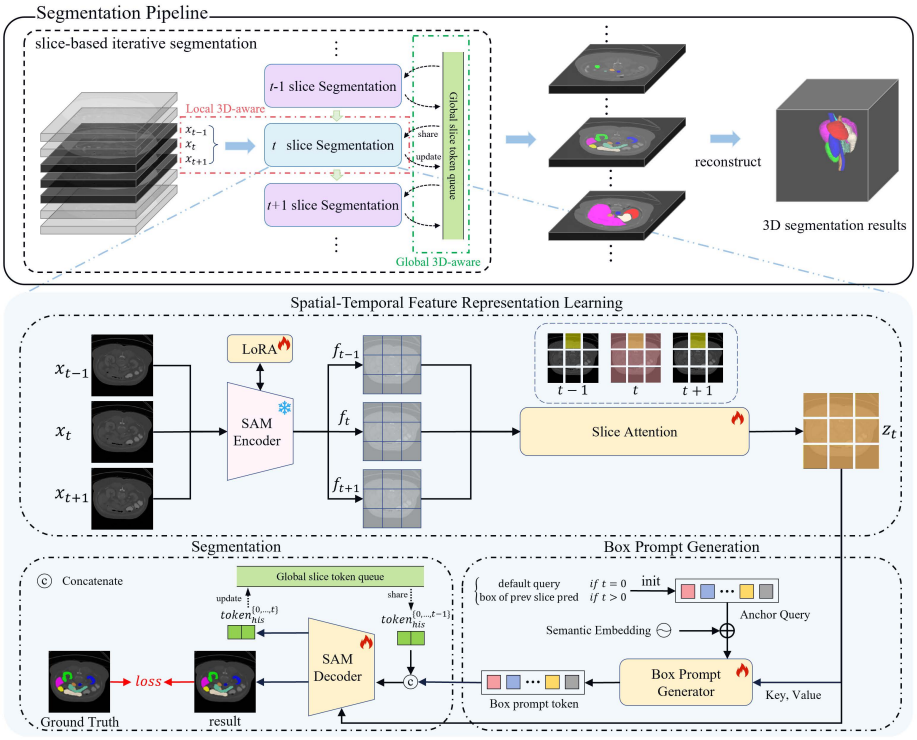

Fig. 1. The overall segmentation pipeline of 3D-SAutoMed for 3D medical image.

– **Enable SAM to generate prompt automatically.** We propose a Box Prompt Generator to automatically generate prompt embedding, leading to fully automatic segmentation. In addition, by replacing the segmentation head and introducing semantic embedding in Box Prompt Generator, we enable SAM to automatically perform semantic segmentation with adaptive box prompts, eliminating the requirement for manual prompts.
– We conducted extensive experiments on multiple datasets, demonstrating that our 3D-SAutoMed can achieve remarkable performance in 3D medical image segmentation. Particularly, we found that box prompts effectively enhance the model in segmenting the boundary of the target, leading to a 44% and 20.7% HD improvement compared to the previous best results on BTCV and SegTHOR dataset.

2 Method

2.1 overview of 3D-SAutoMed

The overall segmentation pipeline of 3D-SAutoMed for 3D medical images is shown in Fig. 1, a slice-based iterative segmentation framework. In the iterative

segmentation, we start with the middle slice of the volume and perform iterative segmentation to both ends in parallel. Compared with other segmentation methods treating each slice independently, our highlight is that we continuously transmit the information of historical slices in the iterative process, leading to global 3D-aware. The overview of 3D-SAutoMed in t slice segmentation includes three stages. **In the spatial-temporal representation learning stage**, we first utilize the encoder of SAM to extract feature representation of the current t and adjacent slices, and we apply LoRA technique in encoder to reduce the discrepancy between medical and neural image during finetuning. To capture the 3D spatial information, we introduce Inter- and Intra-slice Attention and HIS strategy. The former is responsible for sharing local information of adjacent slices, and the latter focuses on sharing global information in the entire volume. **In the box prompt generation stage**, we propose a Box Prompt Generator to automatically generate prompt embedding for guiding semantic segmentation, motivated by DAB-DETR [16]. With the anchor query initialized by the previous slice segmentation prediction, the generator generates prompt embedding of current slice via predicting the target location. **In the segmentation stage**, we extend the segmentation head of SAM's decoder for adapting semantic segmentation. Moreover, we introduce semantic embedding in Box Prompt Generator to enhance the role of the prompt in the semantic segmentation scene. We now delve into the details of each part.

2.2 3D Slices Information Sharing

Some previous studies [5, 7] have demonstrated that utilizing 3D spatial information can effectively enhance the segmentation performance of medical images. In order to transform the coarse-grained output of the 2D SAM's encoder into a richer and fine-grained spatial-temporal feature space, we incorporate the Inter- and Intra-slice Attention and HIS strategy.

Local information sharing via Inter- and Intra-slice Attention. Directly Considering all slices to capture the spatial information of the entire volume is prohibitively expensive. Therefore, we choose three adjacent slices $\{x_{t-1}, x_t, x_{t+1}\}$ to model the local 3D spatial information. Inter- and Intra-slice Attention modules individually facilitate spatial information interaction between and within slices through Inter-slice Attention and Intra-slice Attention.

$$z_2^t = \text{Intra-SA}(z_1^t) + z_1^t, \quad z_1^t = \text{Inter-SA}([f^{t-1}, f^t, f^{t+1}]) + f^t \tag{1}$$

$$z^t = \text{MLP}(z_2^t) + z_2^t \tag{2}$$

where f^t denotes the image encoding feature of the t slice obtained by SAM encoder. Specifically, both Inter-SA and Intra-SA are performed by multi-head self-attention [19]. As shown in Fig. 1, Intra-SA focuses on all features in the current slice (colored in red), while Inter-SA on matching features with the same location in adjacent slices (colored in yellow).

Global information sharing via HIS. Local 3D spatial features can be captured by combining adjacent slices [3, 6], leading to a better segmentation

performance. However, these methods overlook the beneficial information from other slices in volume, as they treat segmentation of each slice independently. The main intuition of using all slices in volume is to capture the inter-slice continuity of the organ as a solid object. As shown in Fig. 1, we continuously and recursively transfer the historical slice information to promote global information sharing in the process of iterative segmentation, which is realized by the Historical slice Information Sharing (HIS) strategy. We first define a historical information token $tok_{his} \in \mathrm{R}^{K \times C}$ for storing the information of historical slices generated by previous segmentation. tok_{his} is concatenated with the output token and then fed into the SAM decoder for interaction. After two-way cross-attention calculations in the SAM decoder, the current t slice image feature is updated with the historical slice information, and the historical information token is updated with the current slice image feature. During the training phase, we conducted iterative training. Specifically, when training t slice, we first select the previous n (random from 1-3) slices for iterative inference to obtain the tok_{his} of t slice. Furthermore, considering that historical information features might contain redundancy and noise, we add an information filter at the end to filter out irrelevant information. This process can be represented as follows:

$$mask^t, tok_{his}^{\{0,\ldots,t\}} = \text{SAM-Decoder}\left(z^t,\ \text{cat}\left(tok_{his}^{\{0,\ldots,t-1\}}, tok_{out}\right)\right) \quad (3)$$

where z^t and $mask^t$ denote the encoding feature and the segmentation result of t slice, respectively. $tok_{his}^{\{0,\ldots,t\}}$ and tok_{out} represent the historical information token of $\{0, ..., t\}$ slice and the output token, respectively. The filter consists of MLP and Normalization.

2.3 Box Prompt Generator

The role of prompt in enhancing the segmentation performance of SAM is undeniable. To achieve full automation, we propose a Box Prompt Generator to automatically generate the prompt embedding for each segmentation step, as shown in Fig. 1. We primarily utilize box prompts. Differing from the original SAM, we allocate a box prompt token for each category, to make our SAM adapt to multi-class semantic segmentation. Following the principle of continuity in adjacent slices in 3D images, we utilize the bounding box obtained from the previous segmentation result to initialize the generator's object query, i.e., anchor query. Since there is no previous slicing result to initialize the query at t=0, we additionally define a learnable default query to make the initial prediction. Subsequently, the prompt embedding for the current slice is generated through a cross-attention based Transformer. In addition, we introduce semantic embedding into the box prompt generator to enhance the role of prompts in multi-class semantic segmentation. The semantic embedding is intended to allow each query to perceive the category for which it is responsible. Specifically, we add a corresponding learnable semantic token for each query separately. These

semantic tokens are randomly initialized and optimized during training.

$$SeQ_{anc} = \text{SeEmbed}\left(Q_{anc}\right), \qquad Q_{anc} = \begin{cases} \text{init}\left(bbox^{pre}\right) & t > 0 \\ \text{default query} & t = 0 \end{cases} \tag{4}$$

$$E_{pmt} = \text{Transformer}\left(SeQ_{anc},\ K, V\right) \tag{5}$$

where $bbox^{pre}$ and Q_{anc} denote the bounding box of previous slice result and anchor query, respectively. SeQ_{anc} represents the semantic anchor query. K, V denote the Key and Value obtained by image encoding feature. E_{pmt} represents the box prompt embedding. To our knowledge, this is the first SAM-based multi-class semantic segmentation work that does not discard prompting.

3 Experiments and Results

3.1 Datasets and Evaluation Metric

Datasets. In this study, we utilized the BTCV [14], CHAOS [11], and SegTHOR [13] datasets to validate the performance of our method. For the BTCV dataset, it includes 30 cases in CT modality, with each volume containing 85-198 slices. The CHAOS and SegTHOR dataset contain 20 MR volumes and 40 CT scans, respectively. All datasets were evaluated using 5-fold cross-validation.

Evaluation Metric. We employed the widely used Dice Similarity Coefficient (DSC) and Hausdorff Distances (HD) as evaluation metrics. The DSC directly reflects the quality of semantic segmentation, and HD focuses on the boundary details of the predicted masks.

3.2 Implementation Details

We implement our method in PyTorch framework with an NVIDIA RTX A6000 GPU. We selected the SAM-b as our baseline model for development in this study. In training phase, we use the hybrid segmentation loss consists of cross-entropy loss and Dice loss. We train our model using AdamW optimizer with learning rate=0.0001, weight decay=0.0001, and batch size=8. Code is available at https://github.com/BossZard/AutoSAM25D.

3.3 Results

Quantitative and qualitative results. Our method was compared with advanced universal medical image segmentation methods (TransUNet [4] and SwinUNet [1]) and other SAM-based methods (SAM [12], SAMed [22], MA-SAM [3], and Slide-SAM [18]). Table 1 and 2 show the results of comparative segmentation methods on the BTCV and CHAOS datasets, respectively. Our method achieves the best average DSC and HD score on both datasets. These results quantitatively demonstrate the superiority of our method. Notably, due to the guidance of the adaptive prompts, our method achieves significant improvements in some

Table 1. Performance comparison of existing methods on the BTCV dataset. N_{int} represents the number of interactions.

SAM variants			Methods	DSC								Avg.	
Prompt	N_{int}	3D-aware		Spleen	R. Kid	L. Kid	Gall.	Liver	Stom.	Aorta	Panc.	DSC↑	HD↓
Point	N	✗	SAM	69.73	67.33	71.13	33.37	73.54	76.43	51.49	41.13	60.51	14.60
✗	0	✗	SAMed	93.54	92.76	93.58	53.01	94.88	84.32	85.10	74.36	83.94	9.46
✗	0	Local	MA-SAM	96.13	**94.09**	**94.96**	78.07	96.46	85.14	88.27	77.30	88.80	5.78
BBox	1	Local	Slide-SAM	**96.20**	93.96	94.79	78.44	**96.46**	86.87	87.31	78.78	89.10	5.40
BBox	0	Local&Global	**Ours**	95.16	93.97	94.08	**80.78**	95.93	**90.54**	**91.23**	79.05	**90.09**	**3.00**
Universal methods			TransUNet	95.21	93.58	94.12	77.30	96.23	71.91	84.83	75.36	86.07	5.36
			SwinUNet	96.08	92.37	93.2	76.65	96.18	84.36	87.35	**79.37**	88.19	7.84

Table 2. Performance comparison of existing methods on the CHAOS dataset.

Methods	DSC				Avg.	
	R. Kid	L. Kid	Liver	Spleen	DSC↑	HD↓
TransUNet	90.72	89.84	83.25	85.31	87.28	7.89
SwinUNet	**92.48**	92.15	87.14	86.82	89.64	4.72
SAMed	88.27	88.64	86.38	86.73	87.50	5.23
MA-SAM	90.52	**91.21**	88.20	89.31	89.81	4.47
Slide-SAM	91.89	90.98	88.42	90.76	90.51	5.16
Ours	91.27	91.19	**90.12**	**91.82**	**91.10**	**4.26**

organs that are relatively difficult to segment on the BTCV dataset, such as the Gallbladder and Stomach. The zero-shot segmentation results of the original SAM still have a significant gap compared to other sophisticated models. Moreover, compared to other SAM variants, our method achieves remarkable performance under the advantages of fully automatic and semantic-aware. From case 1 in Fig. 2, our method and Slide-SAM show the advantage of the prior prompts. However, the hard prompt of Slide-SAM relies heavily on the accuracy of the previous slice's segmentation. Through the box prompt generator, our method can implicitly correct the box prompt for the current slice, effectively alleviating this dependence. Furthermore, the 3D visualization results of case 2 and case 3 in Fig. 2 indicate that the segmentation result of our method has better continuity compared with other SAM variants, demonstrating the effectiveness of local and global 3D-aware mechanisms.

Table 3 reports the results on the SegTHOR dataset. From the table, our method achieves the best average DSC and HD scores, outperforming the suboptimal method with 1.82% and 0.76mm, respectively. We find that our method has obvious advantages in segmenting organs that span multiple slices, such as esophagus, trachea and aorta. This can be attributed to the effectiveness of our proposed historical slice information sharing strategy, as well as the box prompts generation. Moreover, from the results on these datasets, our method achieves remarkable HD scores due to the effective guidance on the location and boundaries of the target provided by the box prompt.

Ablation Study. In Table 4, we conducted an ablation study for different component in our method on BTCV dataset. We observed that due to the stronger spatial correlation between adjacent slices, the improvement brought by the local

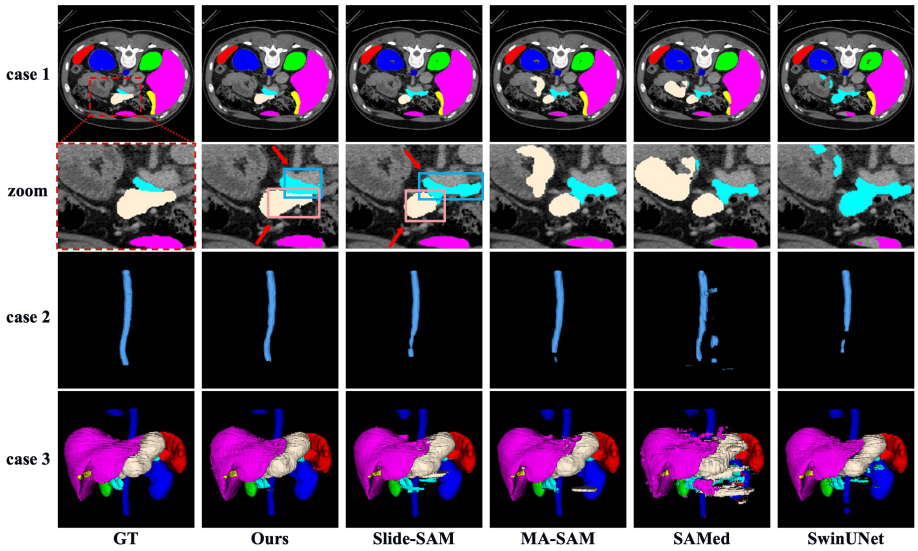

Fig. 2. Qualitative results of different methods. Case 1: 2D view of a slice. The blue and orange boxes indicate the box prompt (generated by the prediction result of the previous slice). Case 2: 3D view of aorta. Case 3: 3D view of all categories.

Table 3. Performance comparison of existing methods on the SegTHOR dataset.

Methods	DSC				Avg.	
	Eso.	Heart	Trach.	Aorta	DSC↑	HD↓
TransUNet	75.71	92.04	84.29	88.54	85.14	5.30
SwinUNet	75.62	91.97	85.46	89.36	85.60	4.95
SAMed	70.54	89.36	82.61	86.90	82.35	5.98
MA-SAM	78.02	**94.65**	88.95	**92.98**	88.65	3.67
Slide-SAM	77.15	93.42	87.93	91.26	87.44	4.52
Ours	**86.83**	92.75	**89.90**	92.42	**90.47**	**2.91**

Table 4. Ablation study for different components in our method on the BTCV dataset.

Direction of iterative segmentation	Local 3D-aware	Global 3D-aware	Prompt Generator	DSC↑	HD↓
Mid-to-Ends (from the middle slice to the ends)			✓	86.27	8.52
	✓		✓	89.39	4.51
	✓	✓		89.16	5.29
	✓	✓	✓	**90.09**	**3.00**
End-to-End (from the top slice to the bottom slice)	✓	✓	✓	89.35	3.81

3D-aware is higher than global 3D-aware. Furthermore, we analyzed the impact of different directions of information transfer during the iterative segmentation process on the segmentation results. We compared the segmentation results between iterating from the middle slice in parallel towards both ends and from one end to the other. The results demonstrate that the former achieves better segmentation results, as the middle slice typically contains richer organ feature information, effectively providing richer guidance information.

4 Conclusion

In this work, we propose 3D-SAutoMed, a novel SAM-based method to solve two limitations of existing SAM variants for 3D medical image segmentation. To enable SAM to be 3D-aware, we propose the Inter- and Intra-slice Attention and HIS strategy to share local information in adjacent slices and global information in the entire volume, respectively. To enable SAM to generate prompts automatically, we designed a Box Prompt Generator to generate prompt embedding automatically. Experimental results on three datasets demonstrate the superiority of our method compared with advanced methods and other SAM variants.

Acknowledgments. This research was supported by the National Natural Science Foundation of China (No.62076059) and the Science and Technology Joint Project of Liaoning province (2023JH2/101700367, ZX20240193)

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

References

1. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation. In: European conference on computer vision. pp. 205–218. Springer (2022)
2. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020)
3. Chen, C., Miao, J., Wu, D., Yan, Z., Kim, S., Hu, J., Zhong, A., Liu, Z., Sun, L., Li, X., et al.: Ma-sam: Modality-agnostic sam adaptation for 3d medical image segmentation. arXiv preprint arXiv:2309.08842 (2023)
4. 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)
5. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: learning dense volumetric segmentation from sparse annotation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19. pp. 424–432. Springer (2016)
6. 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 preprint arXiv:2306.13465 (2023)
7. 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)
8. He, S., Bao, R., Li, J., Grant, P.E., Ou, Y.: Accuracy of segment-anything model (sam) in medical image segmentation tasks. arXiv preprint arXiv:2304.09324 (2023)

9. 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)
10. Huang, Y., Yang, X., Liu, L., Zhou, H., Chang, A., Zhou, X., Chen, R., Yu, J., Chen, J., Chen, C., et al.: Segment anything model for medical images? Medical Image Analysis **92**, 103061 (2024)
11. Kavur, A.E., Gezer, N.S., Barış, M., Aslan, S., Conze, P.H., Groza, V., Pham, D.D., Chatterjee, S., Ernst, P., Özkan, S., et al.: Chaos challenge-combined (ct-mr) healthy abdominal organ segmentation. Medical Image Analysis **69**, 101950 (2021)
12. 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. pp. 4015–4026 (2023)
13. Lambert, Z., Petitjean, C., Dubray, B., Kuan, S.: Segthor: Segmentation of thoracic organs at risk in ct images. In: 2020 Tenth International Conference on Image Processing Theory, Tools and Applications (IPTA). pp. 1–6. IEEE (2020)
14. 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)
15. Lei, W., Wei, X., Zhang, X., Li, K., Zhang, S.: Medlsam: Localize and segment anything model for 3d medical images. arXiv preprint arXiv:2306.14752 (2023)
16. Liu, S., Li, F., Zhang, H., Yang, X., Qi, X., Su, H., Zhu, J., Zhang, L.: Dab-detr: Dynamic anchor boxes are better queries for detr. arXiv preprint arXiv:2201.12329 (2022)
17. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024)
18. Quan, Q., Tang, F., Xu, Z., Zhu, H., Zhou, S.K.: Slide-sam: Medical sam meets sliding window. In: Medical Imaging with Deep Learning (2024)
19. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
20. Wang, Y., Zhang, X., Yang, T., Sun, J.: Anchor detr: Query design for transformer-based detector. In: Proceedings of the AAAI conference on artificial intelligence. vol. 36, pp. 2567–2575 (2022)
21. Wu, J., Fu, R., Fang, H., Liu, Y., Wang, Z., Xu, Y., Jin, Y., Arbel, T.: Medical sam adapter: Adapting segment anything model for medical image segmentation. arXiv preprint arXiv:2304.12620 (2023)
22. Zhang, K., Liu, D.: Customized segment anything model for medical image segmentation. arXiv preprint arXiv:2304.13785 (2023)