Mask-Enhanced Segment Anything Model for Tumor Lesion Semantic Segmentation

Hairong Shi[1,2], Songhao Han[1,2], Shaofei Huang[3(✉)], Yue Liao[4], Guanbin Li[5(✉)], Xiangxing Kong[6], Hua Zhu[6], Xiaomu Wang[7], and Si Liu[1,2]

[1] Beihang University, 37 Xueyuan Road, Haidian District, Beijing, China
[2] Hangzhou Innovation Institute, Beihang University, 18 Chuanghui Street, Binjiang District, Hangzhou, China
[3] Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
nowherespyfly@gmail.com
[4] The Chinese University of Hong Kong, Hong Kong, China
[5] Sun Yat-sen University, Guangzhou, China
liguanbin@mail.sysu.edu.cn
[6] Department of Nuclear Medicine, Peking University Cancer Hospital & Institute, Peking University, Beijing, China
[7] Nanjing University, Nanjing, China

Abstract. Tumor lesion segmentation on CT or MRI images plays a critical role in cancer diagnosis and treatment planning. Considering the inherent differences in tumor lesion segmentation data across various medical imaging modalities and equipment, integrating medical knowledge into the Segment Anything Model (SAM) presents promising capability due to its versatility and generalization potential. Recent studies have attempted to enhance SAM with medical expertise by pre-training on large-scale medical segmentation datasets. However, challenges still exist in 3D tumor lesion segmentation owing to tumor complexity and the imbalance in foreground and background regions. Therefore, we introduce Mask-Enhanced SAM (M-SAM), an innovative architecture tailored for 3D tumor lesion segmentation. We propose a novel Mask-Enhanced Adapter (MEA) within M-SAM that enriches the semantic information of medical images with positional data from coarse segmentation masks, facilitating the generation of more precise segmentation masks. Furthermore, an iterative refinement scheme is implemented in M-SAM to refine the segmentation masks progressively, leading to improved performance. Extensive experiments on seven tumor lesion segmentation datasets indicate that our M-SAM not only achieves high segmentation accuracy but also exhibits robust generalization. The code is available at https://github.com/nanase1025/M-SAM.

Keywords: Tumor Lesion Segmentation · Medical Image Segmentation · Segment Anything Model

---

H. Shi and S. Han – Denote equal contribution.

1   Introduction

Tumor lesion segmentation[12] aims to identify and delineate regions of abnormal tissue in medical images, *e.g.*, computed tomography (CT) or magnetic resonance imaging (MRI). It plays a critical role in the processes of cancer diagnosis and treatment planning. Since manual segmentation is extremely labor-intensive and requires a high level of expertise, deep learning based approaches [19,16] have been introduced to improve efficiency and reduce the workload of physicians. Earlier works [26,8] mainly rely on manually designed kernels to construct the segmentation architecture. In recent years, the U-Net architecture [18] has emerged as one of the most well-known structures for medical image segmentation [25,9] given its effective extraction and utilization of multi-scale information. However, it still struggles to capture long-range spatial dependencies for data with extended sequence lengths. Additionally, applying a U-Net model trained on one specific dataset to others with distribution shifts may lead to significant performance degradation. As a result, methods based on U-Net architecture often exhibit limited performances in lesion segmentation tasks, which are characterized by imbalanced foreground and background regions and require high precision.

In natural image segmentation, the Segment Anything Model (SAM) [14] has shown exceptional versatility and impressive performance across various tasks. It is constructed on the Transformer [21] architecture, which is inherently better suited for learning long-range spatial dependencies. Due to being trained on a vast amount of data, it also shows strong generalization capabilities in segmentation. Furthermore, it contains an interactive system to prioritize regions of interest based on clinician cues, providing a more precise and flexible experience. However, due to limited knowledge in the medical imaging domain, SAM produces unsatisfactory results in medical image segmentation tasks [20].

Recently, a variety of studies [17,22,24,5,15,7,23] have sought to integrate medical expertise into SAM to enhance its capabilities for medical applications. For example, MedSAM [17] refines the decoder with a large amount of medical data and Med-SA [24] extends the SAM architecture using a lightweight and effective adaptation technique. However, these approaches involve slice-by-slice processing of volumetric images, which may lead to suboptimal performance on 3D medical images for disregarding inter-slice 3D spatial information. SAM-Med3D [22], through a comprehensive reformatting of SAM into a fully 3D architecture and extensive pre-trained on 3D medical images, has demonstrated competitive performance in general medical segmentation scenarios such as multi-organ segmentation. However, for the task of tumor lesion segmentation which features huge variability of tumor characteristics (*e.g.*, shape, size, location, and appearance) and imbalanced tissue and tumor regions, SAM-Med3D [22] still obtains less than satisfactory performances.

In this work, we propose a novel architecture named Mask-Enhanced SAM (M-SAM) to adapt SAM-Med3D to 3D tumor lesion segmentation tasks. A new Mask-Enhanced Adapter (MEA) is developed to enhance the semantic information contained in image embeddings with the positional information contained
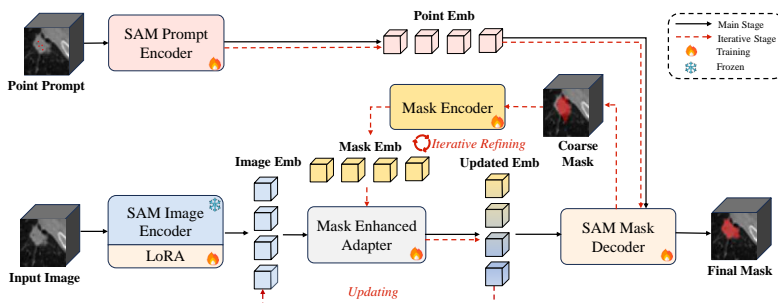
Fig. 1. Overall architecture of Mask-Enhanced SAM (M-SAM).

in coarse segmentation masks, which further helps the generation of refined segmentation masks. To reuse as many parameters of the pre-trained SAM-Med3D as possible, our MEA is designed to be plug-and-play and inserted between the image encoder and mask decoder for training. Based on the MEA module, we further design an iterative refinement scheme, which leverages segmentation results from the previous iterative stage to guide the refinement process of the next iteration. Through iterative refinement, the segmentation masks can be gradually improved, further boosting segmentation performances. We have the following three key contributions:

(1) We introduce a novel Mask-Enhanced SAM (M-SAM) architecture to explore the application of SAM in the medical domain, validating its effectiveness in tumor lesion segmentation.

(2) We propose a Mask-Enhanced Adapter (MEA) to align the positional information of the prompt with the semantic information of the input image, optimizing precise guidance for mask prediction. Based on the design of the MEA, we further implement an iterative refining scheme to refine masks, yielding improved performances.

(3) With updates to only about 20% of the parameters, our model outperforms state-of-the-art medical image segmentation methods on five tumor lesion segmentation benchmarks. Additionally, we validate the effectiveness of our method in domain transferring.

2 Method

2.1 Overall Review

The overall architecture of our M-SAM is illustrated in Fig. 1. M-SAM is built upon the architecture of SAM-Med3D [22], which has been pre-trained on large-scale medical image segmentation datasets to obtain general medical knowledge. It is composed of a ViT-based 3D image encoder to extract 3D image embeddings, a 3D prompt encoder to extract prompt embeddings, and a lightweight
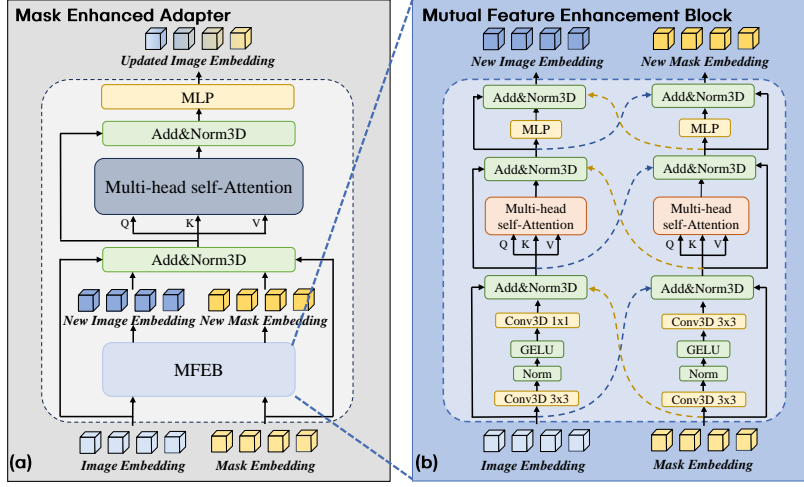
Fig. 2. Module design of Mask-Enhanced Adapter. **(a)** Detailed architecture of Mask-Enhanced Adapter. **(b)** Mutual Feature Enhancement Block.

3D mask decoder to predict segmentation results. As shown in Fig. 1, the inputs are a 3D image (CT or MRI) $I \in \mathbb{R}^{C \times H \times W \times Dp}$, where C represents the number of channels, H the height, W the width, and Dp the depth (determined by the number of image modalities, *e.g.*, the T1 and T2 modalities of MRI), and a randomly-initialized point as the initial point prompt. We first process them with SAM image encoder and SAM prompt encoder respectively, to obtain the initial image embedding $E_I^0 \in \mathbb{R}^{N_I \times D}$ and point embedding $E_P^0 \in \mathbb{R}^{N_P \times D}$. Here, N_I and N_P denote the dimensions of the image and point embeddings, respectively, and D represents the feature dimensionality. We also initialize a null mask $M \in \mathbb{R}^{C \times H \times W \times Dp}$ consisting of all zeros and feed it into a mask encoder to obtain the initial mask embedding $E_M^0 \in \mathbb{R}^{N_I \times D}$. Afterward, the image embedding E_I^0 and mask embedding E_M^0 are fed into our Mask-Enhanced Adapter to obtain the updated image embedding $\hat{E}_I^0$. By feeding $\hat{E}_I^0$ and E_P^0 into the SAM mask decoder, coarse segmentation mask M^1 can be produced, which can also be used to guide the update of image embeddings in MEA for the next stage of segmentation refinement. The image embeddings are also updated in the iterative process accordingly. Through multiple stages of iterative refinement, it is possible to improve the segmentation masks in a coarse-to-fine manner continuously, thus boosting segmentation performance.

2.2 Mask-Enhanced Adapter

Our MEA is proposed to aggregate the image embedding with corresponding mask, so that the updated image embedding can perceive position priors of the lesion regions. The details of the MEA are presented in Fig. 2(a). For simplicity,

we omit superscripts and used an iteration stage for explanation in this section. Given the image embedding E_I and the mask embedding E_M as input, we first feed them into a Mutural Feature Enhancement Block (MFEB) to aggregate their features with each other. Concretely, as shown in Fig. 2(b), the MFEB consists of two parallel 3D Transformer blocks, each with E_I and E_M as inputs. Then, we modify the residual connections in the original Transformer block to mutual residual connections, which facilitates their fusion. New image embedding E_I' and E_M' with the same shape as the original ones are produced as the output of MFEB. Afterwards, the original image and mask embeddings are added to the new ones as residual connections, and the sum is normalized using layer normalization. To incorporate the mask information into the image embedding and make it aware of the foreground regions, the normalized embeddings E_I'' and E_M'' are then transmitted into a multi-head attention layer for fusion as follows:

$$[Q, K, V] = (E_I'' + E_M'')W_{QKV} \tag{1}$$
$$\bar{E}_I = MHSA(Q, K, V), \tag{2}$$

where $W_{QKV} \in \mathbb{R}^{D \times 3D}$ are learnable parameters of linear projections, D represents the feature dimensionality and $MHSA(\cdot)$ represents Multi-Head Self-Attention module. After adding E_I'' to $\bar{E}_I$ and processing them with layer normalization, we apply an MLP layer to the normalized embedding to obtain the updated image embedding $\hat{E}_I$ as the output of MEA.

2.3   Iterative Refinement

Based on the design of our MEA, it is possible to refine the predicted segmentation masks iteratively, thus obtaining more accurate segmentation boundaries progressively. As shown in Fig. 1, the M-SAM makes predictions in a coarse-to-fine manner, beginning with the initial image embedding E_I^0 extracted by SAM image encoder, point embedding E_P^0 extracted by the SAM prompt encoder, and mask embedding E_M^0 derived from an all-zero mask. We use the superscript to denote the stage number of iterative refinement in this section. The updated image embedding $\hat{E}_I^0$ is obtained by feeding E_I^0 and E_M^0 into MEA, and then fed into SAM mask decoder with E_P^0 to produce the first coarse mask M^0, which is then used to generate the mask embedding E_M^1 for the next stage. Concurrently, the new image embedding E_I^1 is updated with $\hat{E}_I^0$ and the new prompt embedding E_P^1 is also generated based on the last segmentation mask, yielding the refined segmentation results for the next stage. In this way, iterative refinement is carried out continuously, resulting in increasingly accurate segmentation results and the segmentation result of the last stage is taken as the final output.

Following SAM-Med3D [22], to simulate the clinical scenario of interactive segmentation, one point per iteration is randomly sampled: from the foreground in the 1st iteration, and from the error region between the coarse mask and ground truth in the subsequent 9 iteration, totaling 10 points in 10 iterations. The updated prompt embedding E_P is thus generated by feeding the newly sampled point into the prompt encoder at each iteration. This point sampling

strategy is used for both training and inference in our experiments, while in real clinical use, it operates interactively with physicians.

2.4   Loss Functions

The overall loss function is the combination of Dice Loss [1] and Cross-Entropy Loss with coefficients w_0 and w_1, and it is formulated as follows:

$$\mathcal{L}_{DiceCE} = w_0 \mathcal{L}_{CE} + w_1 \mathcal{L}_{Dice}, \tag{3}$$

where

$$\mathcal{L}_{Dice} = 1 - \frac{2}{M} \sum_{j=1}^{M} \frac{\sum_{i=1}^{N} p_{ij} g_{ij}}{\sum_{i=1}^{N} p_{ij}^2 + \sum_{i=1}^{N} g_{ij}^2}, \tag{4}$$

$$\mathcal{L}_{CE} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{M} g_{ij} \log(p_{ij}), \tag{5}$$

where N represents the number of pixels, M is the number of classes, g_{ij} is a binary indicator which is set to 1 if class j is the correct classification for pixel i, and p_{ij} is the predicted probability that pixel i belongs to class j. w_0 and w_1 are both set as 0.5 in our experiments.

3   Experiments and Results

3.1   Experimental Setting

Implementation Details. We use the AdamW optimizer with an initial learning rate of $8e-4$ and train for 200 epochs and the batch size was set to 4, with a weight decay of 0.1. In our dataset transform process, we employ a crop-or-pad strategy to standardize all images to a resolution of $128 \times 128 \times 128$. This involves zero-padding for images with any dimension falling short of 128 and applying cropping for dimensions exceeding 128. As for the data augmentation, we use RandomFlip along all three spatial axes and perform z-score normalization on each medical image data. All the experiments are implemented in PyTorch and trained on one NVIDIA Tesla V100 GPU.

Datasets and Evaluation Metrics. We employ seven distinct segmentation tasks to comprehensively demonstrate the advantages of M-SAM. 1) Brain Tumor Segmentation Challenge 2021 (BraTS21) [3], 2) Kidney Tumor Segmentation Challenge 2019 Dataset (KiTS19) [12], 3) Medical Segmentation Decathlon Lung (MSD Lung) [2], 4) Medical Segmentation Decathlon Pancreas (MSD Pancreas) [2], 5) Liver Tumor Segmentation (LiTS) [4], 6) Medical Segmentation Decathlon Hepatic (MSD Hepatic) [2], 7) Lung100*. We adopt the Dice Similarity Coefficient (DSC) and IoU as the evaluation metrics to compare the performance of our method with other methods. More details of the datasets and evaluation metrics can be found in our supplementary materials.

Table 1. Comparison with U-Net-based methods and other SAM-based methods. All methods underwent parameter updates on five datasets respectively. **Bold data** indicate the highest DSC metric among all methods compared. <u>Underline data</u> indicate the second highest.

Method	Cat.	Param(M)	Tunable Param(M)	BraTS21	KiTS19	Lung	Pancreas	LiTS
TransUNet [6]	*U-Net-based*	96	96	89.62	80.75	75.21	76.30	86.50
UNETR [11]		104	104	89.65	84.10	73.29	73.65	81.48
SwinUNETR [11]		138	138	90.48	87.36	75.55	70.71	84.00
nnFormer [27]		151	151	90.42	89.09	77.95	<u>78.65</u>	<u>89.83</u>
nnU-Net [13]		16	16	<u>91.23</u>	<u>89.88</u>	74.31	76.52	87.97
Med-SA [24]	*SAM-based*	636	13	90.50	87.19	73.26	76.47	83.67
SAM3D [5]		91.88	1.88	72.90	80.36	71.42	71.26	82.27
SAM-Med3D [22]		101	101	86.45	86.65	<u>78.32</u>	75.76	88.71
Ours		118	25	**92.08**	**93.50**	**81.62**	**80.49**	**89.95**

3.2 Main Result

Comparison with State-of-the-Art methods. As shown in the Table 1, our method outperforms all other methods on the aforementioned datasets. Notably, compared to the existing methods, our method achieves an average improvement of approximately 2%. This demonstrates the effectiveness of our method across multiple types of tumor lesion datasets. Additionally, it can be observed from the table that U-Net-based methods tend to perform better on individual datasets. However, the performance of such methods varies significantly across different datasets. For example, SwinUNETR [10] achieves a Dice score of over 90% on BraTS21, but only achieves a score of 70.7% on the Pancreas dataset, which is nearly 10% lower than our method. For other SAM-based methods, the Dice scores are generally lower than those of UNet-based methods. In terms of the number of parameters, our model has fewer parameters than the average of the methods shown in the table, and the tunable parameter count also ranks fourth among these methods. These experimental results indicate that our method can surpass these state-of-the-art methods with lower computational costs, effectively leveraging the advantages of SAM in the field of tumor lesion segmentation.

Transfer Results between Different Datasets. To further validate our method's generalizability, we performed transfer experiments from source to target datasets without training on the latter. We compared our method against two SAM-based and two UNet-based methods, as shown in Table 2. The comparison involved transfers between both public datasets and a mix of public and private datasets. Our method consistently surpassed the others, demonstrating minimal performance degradation post-transfer—3.23% for public dataset transfers and 4.9% for transfers involving private datasets. In contrast, UNet-based methods experienced over 10% degradation, underscoring SAM-based methods' superior domain adaptability. Furthermore, our method outperformed SAM-Med3D [22] in both pre- and post-transfer accuracy, with a marginal reduction in performance degradation—0.75% and 0.72%, respectively. This indicates our mod-

Table 2. Transfer Results between datasets. Δ represents the Dice scores difference before and after the transfer. **Rate** denotes Δ as a proportion of the pre-transfer scores. ***** represents private dataset.

Source	Target	Method	Before	After	Δ	Rate(%)
		U-Net [18]	61.54	50.77	-10.77	-17.50
LiTS Tumor	MSD Hepatic Tumor	UNETR [11]	62.19	51.82	-10.37	-16.67
		SAM-Med3D [22]	77.64	75.29	**-2.35**	-3.03
		Ours	**81.97**	**79.58**	-2.39	**-2.92**
		U-Net [18]	61.08	52.23	-8.85	-14.49
MSD Hepatic Tumor	LiTS Tumor	UNETR [11]	64.94	57.15	-7.79	-12.00
		SAM-Med3D [22]	85.01	80.82	-4.19	-4.93
		Ours	**86.89**	**83.81**	**-3.08**	**-3.54**
		UNETR [11]	65.16	52.99	-12.17	-18.68
MSD Lung	Lung100*	U-Net [18]	73.22	62.32	-10.9	-14.89
		SAM-Med3D [22]	78.53	74.35	**-4.18**	-5.32
		Ours	**82.17**	**77.90**	-4.27	**-5.20**
		UNETR [11]	71.99	60.16	-11.83	-16.43
Lung100*	MSD Lung	U-Net [18]	74.77	65.10	-9.67	-12.93
		SAM-Med3D [22]	78.32	73.70	-4.62	-5.90
		Ours	**81.62**	**77.87**	**-3.75**	**-4.59**

ule's ability to enhance accuracy and maintain SAM's generalization capabilities across diverse datasets.

3.3   Ablation Study

Table 3. Ablation on training setting. ***** represents private dataset.

Setting Variants	Tunable Param(%)	KiTS19		BraTS2021		Lung100*	
		DSC	IoU	DSC	IoU	DSC	IoU
baseline	0	73.28	57.82	83.30	71.38	10.07	5.30
w/o MEA	8.09	86.75	76.59	84.35	72.94	77.96	63.88
backbone full-ft	100	**92.67**	**86.34**	**92.15**	**85.44**	82.09	69.62
M-SAM(Ours)	21.41	92.58	86.19	92.08	85.32	**82.17**	**69.74**

We conduct ablation study on KiTS19, BraTS2021 and Lung100* datasets to validate the effectiveness of our method. As illustrated in Table 3, after fine-tuning with LoRA, our method requires updating only 21% of the parameters to achieve results nearly identical or even better than full-parameter fine-tuning. This efficient fine-tuning approach significantly reduces the requirements for computational resources. With the incorporation of our proposed Mask-Enhanced Adapter module, Dice and IoU metrics on public datasets further increase by over 3.5% and 6%, respectively. On our private Lung100* dataset, the performance improvement reaches approximately 2.8% for Dice Score and 4% for IoU respectively.

4   Conclusion

In this work, we introduce Mask-Enhanced SAM (M-SAM), a novel architecture designed specifically for 3D tumor lesion segmentation. In M-SAM, the Mask-Enhanced Adapter (MEA) enhances the semantic information of medical images by incorporating positional data from coarse segmentation masks, which assists in generating more precise segmentation masks. Additionally, we implement an iterative refinement scheme in M-SAM to progressively improve segmentation mask quality, resulting in enhanced performance. Extensive experiments on seven tumor lesion segmentation datasets demonstrate that M-SAM outperforms existing state-of-the-art methods and robust generalization capabilities of the proposed architecture.

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

References

1. Abdollahi, A., Pradhan, B., Alamri, A.: Vnet: An end-to-end fully convolutional neural network for road extraction from high-resolution remote sensing data. IEEE Access **8**, 179424–179436 (2020)
2. Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., Ronneberger, O., Summers, R.M., et al.: The medical segmentation decathlon. Nature communications **13**(1), 4128 (2022)
3. Baid, U., Ghodasara, S., Mohan, S., Bilello, M., Calabrese, E., Colak, E., Farahani, K., Kalpathy-Cramer, J., Kitamura, F.C., Pati, S., et al.: The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. arXiv preprint arXiv:2107.02314 (2021)
4. Bilic, P., Christ, P., Li, H.B., Vorontsov, E., Ben-Cohen, A., Kaissis, G., Szeskin, A., Jacobs, C., Mamani, G.E.H., Chartrand, G., et al.: The liver tumor segmentation benchmark (lits). Medical Image Analysis **84**, 102680 (2023)
5. Bui, N.T., Hoang, D.H., Tran, M.T., Le, N.: Sam3d: Segment anything model in volumetric medical images. arXiv preprint arXiv:2309.03493 (2023)
6. 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)
7. Chen, T., Zhu, L., Ding, C., Cao, R., Zhang, S., Wang, Y., Li, Z., Sun, L., Mao, P., Zang, Y.: Sam fails to segment anything?–sam-adapter: Adapting sam in underperformed scenes: Camouflage, shadow, and more. arXiv preprint arXiv:2304.09148 (2023)

8. Dou, Q., Yu, L., Chen, H., Jin, Y., Yang, X., Qin, J., Heng, P.A.: 3d deeply supervised network for automated segmentation of volumetric medical images. Medical image analysis **41**, 40–54 (2017)

9. Gordienko, Y., Gang, P., Hui, J., Zeng, W., Kochura, Y., Alienin, O., Rokovyi, O., Stirenko, S.: Deep Learning with Lung Segmentation and Bone Shadow Exclusion Techniques for Chest X-Ray Analysis of Lung Cancer, p. 638–647 (Jan 2019). https://doi.org/10.1007/978-3-319-91008-6_63, http://dx.doi.org/10.1007/978-3-319-91008-6_63

10. 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)

11. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: Unetr: Transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 574–584 (2022)

12. Heller, N., Isensee, F., Maier-Hein, K.H., Hou, X., Xie, C., Li, F., Nan, Y., Mu, G., Lin, Z., Han, M., et al.: The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge. Medical image analysis **67**, 101821 (2021)

13. Isensee, F., Petersen, J., Klein, A., Zimmerer, D., Jaeger, P.F., Kohl, S., Wasserthal, J., Koehler, G., Norajitra, T., Wirkert, S., et al.: nnu-net: Self-adapting framework for u-net-based medical image segmentation. arXiv preprint arXiv:1809.10486 (2018)

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. arXiv preprint arXiv:2304.02643 (2023)

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. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. Medical image analysis **42**, 60–88 (2017)

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. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015)

19. Shen, D., Wu, G., Suk, H.I.: Deep learning in medical image analysis. Annual review of biomedical engineering **19**, 221–248 (2017)

20. Shi, P., Qiu, J., Abaxi, S.M.D., Wei, H., Lo, F.P.W., Yuan, W.: Generalist vision foundation models for medical imaging: A case study of segment anything model on zero-shot medical segmentation. Diagnostics **13**(11), 1947 (2023)

21. 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)

22. Wang, H., Guo, S., Ye, J., Deng, Z., Cheng, J., Li, T., Chen, J., Su, Y., Huang, Z., Shen, Y., et al.: Sam-med3d. arXiv preprint arXiv:2310.15161 (2023)

23. Wei, X., Cao, J., Jin, Y., Lu, M., Wang, G., Zhang, S.: I-medsam: Implicit medical image segmentation with segment anything. arXiv preprint arXiv:2311.17081 (2023)

24. 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)
25. Zeng, G., Yang, X., Li, J., Yu, L., Heng, P.A., Zheng, G.: 3D U-net with Multi-level Deep Supervision: Fully Automatic Segmentation of Proximal Femur in 3D MR Images, p. 274–282 (Jan 2017). https://doi.org/10.1007/978-3-319-67389-9_32, http://dx.doi.org/10.1007/978-3-319-67389-9_32
26. Zhang, W., Li, R., Deng, H., Wang, L., Lin, W., Ji, S., Shen, D.: Deep convolutional neural networks for multi-modality isointense infant brain image segmentation. NeuroImage **108**, 214–224 (2015)
27. Zhou, H.Y., Guo, J., Zhang, Y., Yu, L., Wang, L., Yu, Y.: nnformer: Interleaved transformer for volumetric segmentation. arXiv preprint arXiv:2109.03201 (2021)