Curriculum Prompting Foundation Models for Medical Image Segmentation

Xiuqi Zheng, Yuhang Zhang, Haoran Zhang, Hongrui Liang, Xueqi Bao, Zhuqing Jiang, and Qicheng Lao

Beijing University of Posts and Telecommunications, Beijing, China
qicheng.lao@bupt.edu.cn

Abstract. Adapting large pre-trained foundation models, e.g., SAM, for medical image segmentation remains a significant challenge. A crucial step involves the formulation of a series of specialized prompts that incorporate specific clinical instructions. Past works have been heavily reliant on a singular type of prompt for each instance, necessitating manual input of an ideally correct prompt, which is less efficient. To tackle this issue, we propose to utilize prompts of different granularity, which are sourced from original images to provide a broader scope of clinical insights. However, combining prompts of varying types can pose a challenge due to potential conflicts. In response, we have designed a coarse-to-fine mechanism, referred to as curriculum prompting, that progressively integrates prompts of different types. Through extensive experiments on three public medical datasets across various modalities, we demonstrate the effectiveness of our proposed approach, which not only automates the prompt generation process but also yields superior performance compared to other SAM-based medical image segmentation methods. Code will be available at: https://github.com/AnnaZzz-zxq/Curriculum-Prompting.

Keywords: Medical image segmentation · SAM · Prompt engineering · Curriculum learning.

1 Introduction

Medical image segmentation is a critical area of research within medical image analysis. It plays a vital role in identifying and delineating various tissues or lesions, thereby significantly enhancing the efficiency and accuracy of medical diagnosis [4]. Recently, with the advent of large-scale foundation models for segmentation such as SAM [16], the field of medical image segmentation has seen rapid development. SAM enables the generation of masks for regions of interest through interactive prompting, making it well-suited for universal medical image segmentation tasks. Several studies [11,24,12] have already explored the application of SAM in medical image segmentation. However, due to the substantial differences between natural and medical images, SAM struggles to achieve optimal segmentation performance across medical image datasets. One strategy to

enhance SAM's performance in medical image segmentation involves integrating medical knowledge through specialized prompts. However, the manual generation of such prompts incurs high labor costs and yields diverse prompt quality.

To address the aforementioned challenges, this paper introduces an automated approach for identifying an optimal prompt for SAM-based medical image segmentation. Unlike conventional methods that rely on a single prompt and necessitate manual intervention, our proposed methodology leverages multiple prompt types to integrate a diverse range of image-specific details and clinical knowledge into the network. However, combining diverse knowledge domains presents a non-trivial challenge. Inspired by curriculum learning [1], which is motivated by the cognitive learning strategies of humans gradually acquiring knowledge from simple to complex tasks, we propose *curriculum prompting*, which employs prompts that have progressively increasing granularity to systematically address segmentation challenges of varying difficulty levels, starting from coarse to fine-grained levels, to mitigate conflicts across different prompt domains. Specifically, we use mask prompts as an intermediary to gradually combine box and point prompts, refining the initial coarse mask prompt into a fine-tuned version. Unlike conventional SAM-based medical image segmentation methods that depend solely on a single prompt and necessitate the manual provision of an absolutely correct prompt, our approach significantly reduces the need for manual intervention, enabling the automatic generation of optimal prompts for SAM-based medical image segmentation based only on input medical images. In summary, our paper makes three significant contributions:

- Automated Prompt Generation: We propose a novel approach to automatically generate optimal prompts for SAM-based medical image segmentation, eliminating the need for manual intervention and providing more image-specific details and clinically specific knowledge to the network.
- Curriculum Prompting Method: Our method integrates prompts of varying domains in a progressive manner, starting from coarse to fine-grained levels, which helps mitigate conflicts when simultaneously using multiple prompts from different domains.
- Improved Segmentation Results: The combined effect of automated prompt generation and curriculum prompting leads to significantly improved segmentation results on three public medical datasets across various modalities, outperforming existing SAM-based methods qualitatively and quantitatively.

2 Methodology

2.1 Overview

Given an image $I \in R^{H \times W \times 3}$ with spatial resolution $H \times W$, large foundation models for segmentation, e.g., SAM, typically adopt an image encoder for extracting the image embedding $\mathbf{e}$ from the image I, transform the prompt input P through a prompt encoder Enc, and finally generate a segmentation mask S
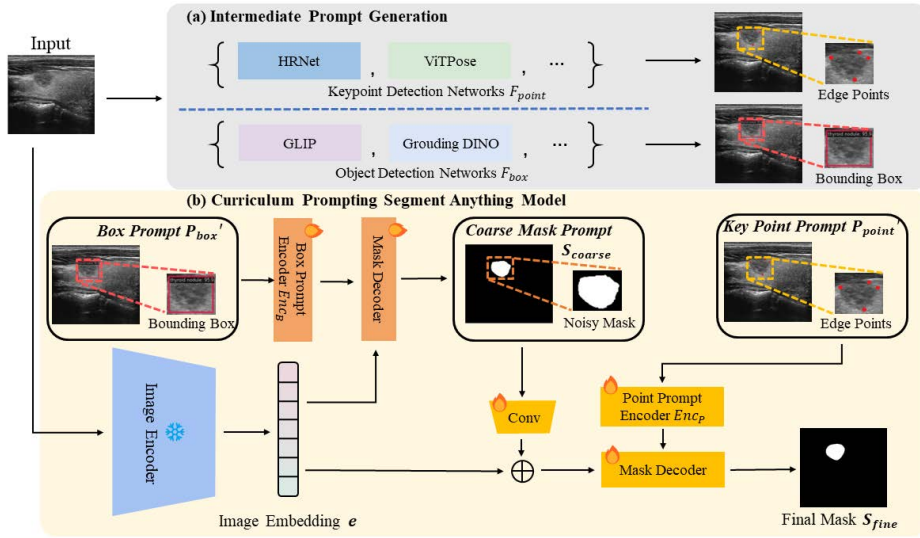
Fig. 1. Overview of Curriculum Prompting: (a) Intermediate Prompt Generation, which prepares prompts for SAM; (b) Curriculum Prompting SAM, first utilizing self-generated box prompts to obtain coarse masks, and then acquire refined masks with self-generated point prompts and coarse masks (as mask prompts).

through a mask decoder Dec, formulated as:

$$S = Dec(\mathbf{e}, Enc(P)), \tag{1}$$

where P can be in the form of various types, such as point prompt $P_{point} = [x, y]$, where x, y denotes the coordinates of the point, box prompt $P_{box} = [x_1, y_1, x_2, y_2]$, composed of coordinates of the top-left and bottom-right corners of the bounding box, and mask prompt $P_{mask} \in R^{H \times W}$.

Prompts play a crucial role during the segmenting process, where high-quality prompts enable SAM to produce accurate segmentation masks [13,3,5]. However, existing methods only utilize a single type of prompt, which contains limited information and often requires manual interventions.

Our proposed curriculum prompting adheres to a straightforward idea, which aims to progressively combine different types of prompts in a coarse-to-fine way. We begin with the initial prompt P_1 to assist SAM in segmentation tasks. Subsequently, the intermediate prediction generated by P_1 is fed back together with an auxiliary prompt as supplementary into SAM, initiating a recursive process. This cycle continues n steps until a satisfactory segmentation result is achieved, and our empirical observations indicate that a notably improved result can be

obtained when $n = 2$. This process can be described as:

$$
\begin{aligned}
P_2 &= Dec(\mathbf{e}, Enc(P_1)), \\
P_3 &= Dec(\mathbf{e}, Enc(P_2, P_2')), \\
&\quad ..., \\
S &= Dec(\mathbf{e}, Enc(P_n, P_n')),
\end{aligned}
\tag{2}
$$

where P_n' denotes an auxiliary prompt apart from P_n as a supplementary.

In summary, we design a curriculum prompting mechanism to first address intermediate easy segmentation tasks and acquire initial coarse masks with self-generated prompts, and then add more refined prompts to tackle harder segmentation tasks and obtain the ultimate mask, to improve the overall performance.

2.2   Coarse Prompting

During the coarse prompting phase, we aim to segment most of the foreground pixels which is an easier task compared to the fine-grained segmentation with a single step. We utilize prompts that are relatively coarse but contain sufficient information to obtain an initial coarse mask. Since empirical observations suggest that two different types of prompts, e.g., box prompt and point prompt, may conflict with each other [13,3], in this work, we choose to employ a single type of prompt as our coarse prompt. Compared to point prompts, box prompts encompass more significant information, indicating the precise location of the object and the potential intensity features within a specified limited area. Thus, we consider self-generated box prompts as coarse prompts for initial masks.

To break through the limitation of SAM requiring manual prompts, we intend to directly and automatically derive prompts from the original image. We generate box prompts with large pre-trained object detection models, e.g. Grounding DINO [20] or GLIP [17]. We fine-tune the pre-trained model with the given medical data and obtain the self-generated box prompts P_{box}' as follows,

$$
P_{box}' = F_{box}(I, T),
\tag{3}
$$

where F_{box} denotes the chosen object detection model, I denotes the input image and T denotes the text prompt if required for the model.

Following the acquisition of box prompts, a series of post-processing steps (e.g. NMS) are undertaken. We fine-tune SAM's prompt encoder with ground-truth bounding boxes, employing a combination of Dice Loss and BCE Loss as our loss function. Then, we acquire coarse masks S_{coarse} utilizing these self-generated box prompts and the input image embedding $\mathbf{e}$,

$$
S_{coarse} = Dec(\mathbf{e}, Enc_B(P_{box}')),
\tag{4}
$$

where Enc_B denotes the prompt encoder fine-tuned with bounding boxes.

2.3 Fine-grained Prompting

Having acquired the coarse masks, we further aim to employ more refined prompts to tackle a harder fine-grained segmentation task and guide SAM in generating the final mask. As indicated in [15], SAM struggles with precise edge segmentation, making the enhancement of edge delineation a more complex task compared to segmenting most of the foreground pixels.

Thus, we adopt edge points as additional prompts to unleash SAM's full ability for segmentation. Similar to the process of box prompt generation, we employ a keypoint detection network (e.g. HRNet [26] or ViTPose [27]) to generate point prompts. We obtain the self-generated point prompts P'_{point} as follows:

$$P'_{point} = F_{point}(I), \tag{5}$$

where F_{point} denotes the keypoint detection network.

However, utilizing multiple types of prompts synergistically requires careful design. As numerous studies have indicated [13,23,28], the simultaneous use of point and box prompts can paradoxically lead to a decrease in performance. One speculation about the cause of this contradiction is due to the structure of SAM's prompt encoder. In SAM's prompt encoder Enc, point prompts P_{point} and box prompts P_{box} are processed through a series of steps and then concatenated into a sparse embedding, which is fed into the mask decoder Dec. During this process, different types of prompts may influence each other.

The question then arises: how can we effectively incorporate the guidance of point prompts while leveraging the information from box prompts? The answer lies in employing an additional type of prompt - the mask prompt, as a bridge to combine both box prompts and point prompts. This is where we take advantage of the coarse masks S_{coarse} obtained in Section 2.2.

While point embeddings and box embeddings influence each other, the mask prompts P_{mask} will only be transformed into a dense embedding through convolutions and summed with the image embedding $\mathbf{e}$ without interacting with the sparse embedding. Thus, we employ self-generated point prompts P'_{point} on the basis of coarse masks S_{coarse} as mask prompts to achieve refined segmentation.

Similar to the process described in Section 2.2, the SAM model we use has undergone fine-tuning with medical images, and edge points and coarse masks served as prompts. Then final masks S_{fine} are acquired as follows:

$$S_{fine} = Dec(\mathbf{e}, Enc_P(S_{coarse}, P'_{point})), \tag{6}$$

where Enc_P denotes the prompt encoder that is fine-tuned with edge point prompts and mask prompts, and P'_{point} is obtained by Eq.(5).

3 Experiments and Results

3.1 Dataset

We evaluate our proposed method on three public medical image datasets across various modalities, including thyroid nodule segmentation dataset TN3K [10],

Table 1. Comparisons with traditional task-specific and SAM-based medical image segmentation methods. "*" denotes results reported by the referenced paper. "-" means results are unavailable caused by dataset being used during training.

Method	Kvasir (Endoscopy)		TN3K (Ultrasound)		QaTa-COV19 (X-ray)	
	mDice(%)	mIoU(%)	mDice(%)	mIoU(%)	mDice(%)	mIoU(%)
CaraNet [21]	92.050	86.890	72.647	62.746	73.887	63.517
TRFE+ [10]	42.819	29.517	83.300*	71.380*	45.719	32.835
LViT-T[19]	77.899	67.519	76.871	66.573	77.207	67.178
fine-tuned SAM [16]	81.848	74.191	50.791	39.771	48.794	61.504
nnSAM [18]	91.176	85.946	82.797	74.027	78.943	69.452
SAM-Med2D (9 Points) [5]	-	-	64.740	55.760	76.431	66.083
MedSAM (Box) [22]	86.473	78.046	81.126	69.464	-	-
Grounded SAM [25]	93.340	89.029	81.600	73.986	78.625	68.616
Ours	**93.670**	**89.442**	**84.430**	**76.367**	**79.826**	**70.265**

polyp segmentation dataset Kvasir [14], and pulmonary lesion segmentation dataset QaTa-COV19 [7]. The TN3K dataset includes 3493 ultrasound images with pixel-wise thyroid nodule annotations; The Kvasir dataset contains 1000 endoscopic images and their corresponding polyp ground-truth masks; The QaTa-COV19 dataset consists of 9258 chest X-ray radiographs with pneumonia segmentation masks. We follow the same dataset split as [10,9,19], respectively.

3.2   Experiment Settings and Metrics

Our method finetunes four distinct models, ensuring each model builds upon previous outputs. We fine-tune the object detection and keypoint detection network through the MMDetection [2] and MMPose [6] framework. Specifically, we select Grounding DINO [20] and HRNet [26] for box and point prompt generation, respectively. Specifically, we use 8 edge points as point prompts. In terms of fine-tuning SAM, we initialize the model with the pre-trained weight of SAM's ViT-H version [8]. We employ an AdamW optimizer with a learning rate of 0.0001 and a batch size of 4. Our model is implemented using PyTorch and trained and evaluated on an Nvidia RTX4090 24GB GPU. We adopt two commonly used metrics to quantitatively evaluate our proposed method, Dice (dice coefficient) and IoU (Intersection over Union).

3.3   Results

Our Proposed Approach Outperforms the Baselines on All Three Datasets. We compare our method with SOTA task-specific methods and SAM-based foundation models. CaraNet [21], TRFE+ [10] and LViT-T [19] are three SOTA methods on the Kvasir, TN3K and QaTa-COV19 datasets, respectively.
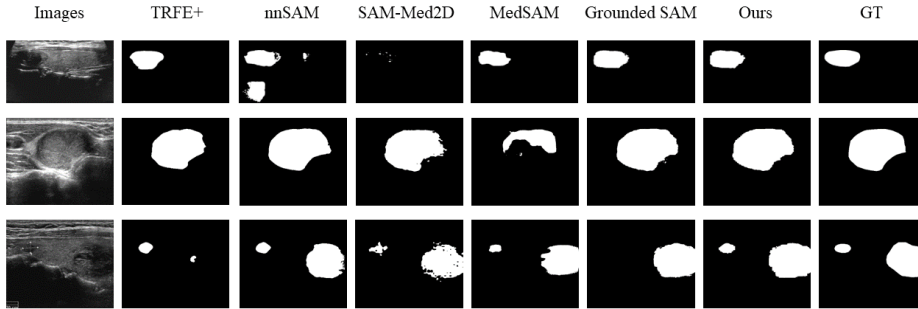
Fig. 2. Qualitative comparisons between our curriculum prompting SAM and other segmentation methods on the TN3K dataset, including SOTA task-specific method TRFE+, and other SAM-based segmentation models.
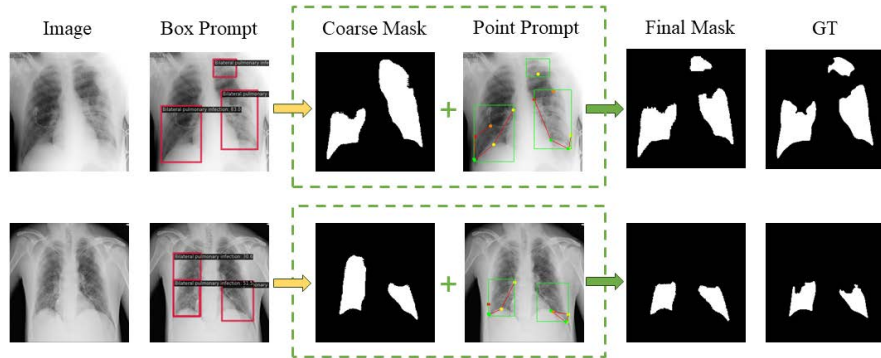

Fig. 3. The process of mask generation through our proposed curriculum prompting.

Additionally, five SOTA foundation models are chosen for comparison, including the vanilla SAM [16], nnSAM [18], SAM-Med2D [5], MedSAM [22] and Grounded SAM [25]. Note that we standardize the text prompt to the name or a simple description of the target lesion, such as polyp, thyroid nodule, or bilateral pulmonary infection, for fine-tuning LViT-T [19] and Grounded SAM [25].

Table. 1 summarizes the quantitative results. Notably, our method consistently achieves the best performance on all three tasks with average IoU scores of 89.442%, 76.367%, and 70.265%. Compared to SAM-Med2D and MedSAM which require extra point prompts or box prompts derived from labels, our method outperforms them by a large margin (e.g., mean IoU $> 6.9\%$) without human intervention. This validates the effectiveness of our proposed method by integrating multiple prompts in a coarse-to-fine manner.

We present qualitative results in Fig. 2, where the segmentation masks of the thyroid nodules from different methods are shown. As seen in the figure, our method can precisely locate the target lesion and yields more accurate and smooth edge delineation, compared to other baselines.

Visualization of Curriculum Prompting Process. As shown in Fig. 3, during the first coarse phase when only the box prompt is used, SAM is capable of segmenting the majority of the foreground pixels. Through curriculum prompting, with the addition of edge points guidance on this basis, SAM can discern where the edges of the target are, as well as accurately distinguish between two target areas when they are nearby, instead of merging the masks into one large area. Moreover, it can be observed that the edges of the final mask have become smoother, with fewer isolated dots that are not connected to the larger area, which is very common in masks generated by SAM.

Table 2. Negative or positive prompts.

Label	Metric	Result
negative(0)	mDice (%)	84.430
	mIoU (%)	76.367
positive(1)	mDice (%)	84.259
	mIoU (%)	76.192

Table 3. Ablation studies on TN3K.

Point	BBox	Mask	mDice (%)	mIoU (%)
✓			70.300	61.127
	✓		81.600	73.986
	✓	✓	81.660	74.099
✓	✓		79.466	71.454
✓	✓	✓	84.430	76.367

Edge Points Served as Negative Prompts Can Better Improve SAM's Performance. As SAM struggles with precise edge segmentation, we introduce point prompts to provide extra details, especially focusing on the lesion edges. These points can act as either positive or negative prompts. Table 2 demonstrates that labeling edge points as negative (label = 0) can better enhance the segmentation result. We theorize that negative prompts give more detailed guidance, clearly marking non-foreground areas. In contrast, positive prompts may not add valuable information, as the model might already identify these areas as foreground, diminishing their impact on edge definition. Thus, we label the point prompts as negative in all our experiments.

Ablation Study. There are three different types of prompts used in our study yielding seven unique combinations. We perform ablation studies on five scenarios on the TN3K dataset, detailed in Table 3. Given that mask prompts result from SAM's inference using box prompts, we exclude unavailable scenarios including solely utilizing mask prompts and utilizing both point and mask prompts due to their dependency on box prompts for mask generation. When segmenting solely with 8 edge points, SAM fails to achieve a satisfactory result, whereas, when using self-generated boxes, SAM is already capable of achieving relatively good segmentation. We can observe a decline when simultaneously using point and box prompts, compared to using box prompts alone. The results show that when utilizing three prompt types in the proposed curriculum manner, SAM

gives the best segmentation performance, demonstrating each prompt type is necessary and curriculum combining them is effective.

Training Time. The time consumption primarily occurs during the finetuning process. Our model requires 9.5h, 2.7h, 21.1h training on TN3K, Kvasir, and QaTa-COV19. For comparison, the nnSAM model takes 15.2h, 12.5h, and 20.8h. In most cases, our training time is shorter than nnSAM but outperforms nnSAM on all three datasets, demonstrating that though our training process is somewhat complicated, the training time is acceptable.

4   Conclusion

In this paper, we present curriculum prompting for medical image segmentation using large foundation models, an efficient method to combine multiple prompts for better segmentation performance. We employ self-generated prompts that have progressively increasing granularity to systematically address segmentation challenges of varying difficulty levels. Compared to utilizing a singular type of prompt, our method introduces more prompt information while avoiding possible conflicts between different prompt types, and achieves state-of-the-art performance on three public medical datasets with different modalities and target lesions. We hope our study provides some inspiration about prompting vision foundation models for medical image segmentation.

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

References

1. Bengio, Y., Louradour, J., Collobert, R., Weston, J.: Curriculum learning. In: Proceedings of the 26th Annual International Conference on Machine Learning. p. 41–48. ICML '09, Association for Computing Machinery, New York, NY, USA (2009). https://doi.org/10.1145/1553374.1553380, https://doi.org/10.1145/1553374.1553380
2. Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., Zhang, Z., Cheng, D., Zhu, C., Cheng, T., Zhao, Q., Li, B., Lu, X., Zhu, R., Wu, Y., Dai, J., Wang, J., Shi, J., Ouyang, W., Loy, C.C., Lin, D.: MMDetection: Open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155 (2019)
3. Cheng, D., Qin, Z., Jiang, Z., Zhang, S., Lao, Q., Li, K.: Sam on medical images: A comprehensive study on three prompt modes. arXiv preprint arXiv:2305.00035 (2023)
4. Cheng, J., Tian, S., Yu, L., Gao, C., Kang, X., Ma, X., Wu, W., Liu, S., Lu, H.: Resganet: Residual group attention network for medical image classification and segmentation. Medical Image Analysis **76**, 102313 (2022)
5. 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)

6. Contributors, M.: Openmmlab pose estimation toolbox and benchmark. https://github.com/open-mmlab/mmpose (2020)

7. Degerli, A., Kiranyaz, S., Chowdhury, M.E., Gabbouj, M.: Osegnet: Operational segmentation network for covid-19 detection using chest x-ray images. In: 2022 IEEE International Conference on Image Processing (ICIP). pp. 2306–2310. IEEE (2022)

8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)

9. Fan, D.P., Ji, G.P., Zhou, T., Chen, G., Fu, H., Shen, J., Shao, L.: Pranet: Parallel reverse attention network for polyp segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 263–273. Springer (2020)

10. Gong, H., Chen, J., Chen, G., Li, H., Li, G., Chen, F.: Thyroid region prior guided attention for ultrasound segmentation of thyroid nodules. Computers in Biology and Medicine **155**, 106389 (2023)

11. 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)

12. 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)

13. 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)

14. Jha, D., Smedsrud, P.H., Riegler, M.A., Halvorsen, P., de Lange, T., Johansen, D., Johansen, H.D.: Kvasir-seg: A segmented polyp dataset. In: MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part II 26. pp. 451–462. Springer (2020)

15. Ke, L., Ye, M., Danelljan, M., Liu, Y., Tai, Y.W., Tang, C.K., Yu, F.: Segment anything in high quality. arXiv preprint arXiv:2306.01567 (2023)

16. 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)

17. Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N., et al.: Grounded language-image pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10965–10975 (2022)

18. Li, Y., Jing, B., Feng, X., Li, Z., He, Y., Wang, J., Zhang, Y.: nnsam: Plug-and-play segment anything model improves nnunet performance. arXiv preprint arXiv:2309.16967 (2023)

19. Li, Z., Li, Y., Li, Q., Wang, P., Guo, D., Lu, L., Jin, D., Zhang, Y., Hong, Q.: Lvit: language meets vision transformer in medical image segmentation. IEEE transactions on medical imaging (2023)

20. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499 (2023)

21. Lou, A., Guan, S., Ko, H., Loew, M.H.: Caranet: Context axial reverse attention network for segmentation of small medical objects. In: Medical Imaging 2022: Image Processing. vol. 12032, pp. 81–92. SPIE (2022)

22. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**, 1–9 (2024)
23. Mattjie, C., de Moura, L.V., Ravazio, R.C., Kupssinskü, L.S., Parraga, O., Delucis, M.M., Barros, R.C.: Zero-shot performance of the segment anything model (sam) in 2d medical imaging: A comprehensive evaluation and practical guidelines. arXiv preprint arXiv:2305.00109 (2023)
24. Putz, F., Grigo, J., Weissmann, T., Schubert, P., Hoefler, D., Gomaa, A., Tkhayat, H.B., Hagag, A., Lettmaier, S., Frey, B., et al.: The segment anything foundation model achieves favorable brain tumor autosegmentation accuracy on mri to support radiotherapy treatment planning. arXiv preprint arXiv:2304.07875 (2023)
25. Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., Chen, J., Huang, X., Chen, Y., Yan, F., Zeng, Z., Zhang, H., Li, F., Yang, J., Li, H., Jiang, Q., Zhang, L.: Grounded sam: Assembling open-world models for diverse visual tasks (2024)
26. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5693–5703 (2019)
27. Xu, Y., Zhang, J., Zhang, Q., Tao, D.: Vitpose: Simple vision transformer baselines for human pose estimation. Advances in Neural Information Processing Systems **35**, 38571–38584 (2022)
28. Zhang, C., Puspitasari, F.D., Zheng, S., Li, C., Qiao, Y., Kang, T., Shan, X., Zhang, C., Qin, C., Rameau, F., et al.: A survey on segment anything model (sam): Vision foundation model meets prompt engineering. arXiv preprint arXiv:2306.06211 (2023)