Knowledge Bridges the Intent Gap: Contextual Fusion in Medical Fine-Grained Segmentation

Hengyuan Zhang[1,2*], Peng Qiao[1,2*†], Wenyu Li[1,2], Yan Jia[1,2], and Yong Dou[1,2†]

[1] College of Computer Science and Technology, National University of Defense Technology
[2] National Key Laboratory of Parallel and Distributed Computing, National University of Defense Technology
{zhanghengyuan,pengqiao,wenyu18,jia.yan20,yongdou}@nudt.edu.cn

Abstract. Segment Anything Model (SAM) has been widely used in common medical image segmentation for its great zero-shot generalization by providing points or box as prompt. However, we find that SAM and its variants do not cope well with complex fine-grained segmentation tasks such as kidney anatomical structure segmentation due to the discrepancy between the model's interpretation of the task and the actual intent conveyed by the prompts. This paper introduces a new approach called Knowledge SAM (KSAM). By providing a pair of example image and corresponding fine-grained segmentation mask as the knowledge prompt, model can utilize the contextual information to better understand the meaning of the unseen fine-grained segmentation task. To accommodate knowledge prompts, we design two modules specifically designed for knowledge prompt feature fusion. KSAM outperforms the SAM models based on different prompts across both our proposed kidney anatomical structure dataset and REFUGE. Notably, our approach demonstrates competitive performance while offering better extensibility on new tasks compared with prompt-free methods.

Keywords: SAM · fine-grained segmentation · knowledge prompt.

1 Introduction

Segment Anything Model (SAM) has gained widespread adoption as a visual foundation model for various segmentation tasks, including abdominal organ segmentation[2,23] in medical images and instance segmentation[22]. However, SAM's prompt methods including point and box prompt are designed for regular targets in natural images, which make SAM not suited for medical segmentation tasks with complex target shapes. Specifically, in fine-grained segmentation tasks such as kidney anatomical structure segmentation [4,21,9,26,10,17], simple

[*] Both authors contributed equally to this research.
[†] Corresponding author.

prompts make it difficult for the model to fully comprehend the task and accurately identify the target objects for segmentation. Additionally, labeled datasets for such fine-grained segmentation tasks are scarce. Consequently, developing effective strategies to fine-tune SAM and design better prompt methods that allow the model to fully understand the segmentation task remain a challenging yet worthwhile research problem. Vanilla SAM usually have the problems: **1)** SAM inherently relies on high quality sparse prompts, which are difficult to generate automatically in practical applications. Moreover, even minor variations in the prompts can significantly influence the model's performance. **2)** Fine-tuning typically requires a substantial amount of data.

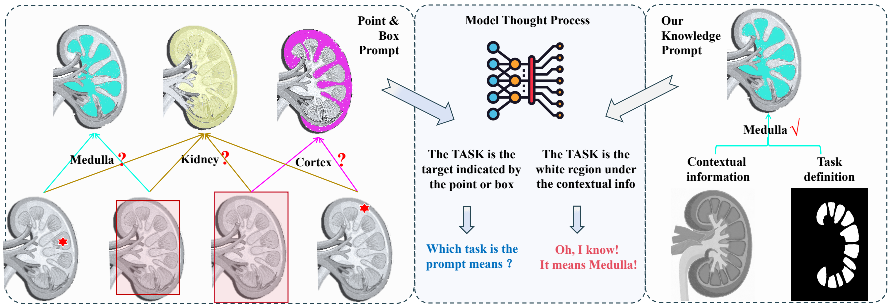
Point & Box Prompt
Medulla ?
Kidney ?
Cortex ?
Model Thought Process
The TASK is the target indicated by the point or box
The TASK is the white region under the contextual info
Which task is the prompt means ?
Oh, I know! It means Medulla!
Our Knowledge Prompt
Medulla √
Contextual information
Task definition

Fig. 1: **Right.**Model can generate eight comprehensions which are all correct with four simple prompts. **Left.**Knowledge prompt can provide a definate segmentation task. **Mid.**Model thought process according to prompt.

Given the difficulty of obtaining high-quality prompts, prompt-free variants of SAM [7,8,18,28] have been proposed. However, these approaches rigidly solidify the segmentation capabilities of model to predefined tasks. Other works such as [29,11,27,16], prompt the model by providing a similar image with an existing mask. These methods select the most related point by calculating similarity pixel-by-pixel with the target image, allowing the model to segment without sparse prompts. However, the accuracy of similarity-based point selection can be limited, potentially leading to significant false prompting issues [25].

To solve Problem.2), some efficient fine-tuning methods [23,5,28,19] reduce the number of training parameters. Nevertheless, the demand for a large amount of training data still exists. For some complex tasks, it is difficult to obtain enough data, which can lead to overfitting. At the same time, we find that point and box prompt can only provide position information, which may not be well-suited for the segmentation of medical images. Target organs usually have similar characteristics with surrounding tissues in medical images. SAM is highly sensitive to object edges, tending to segment the entire organ when given a point prompt, even if the origin goal is to segment a specific structure within the organ. For example, in fine-grained kidney structure segmentation, the model's comprehension of the task can vary significantly when using point prompts or box prompts, as shown in Fig. 1. This happens when the region indicated by point or box has multiple semantic labels.

To address the comprehension ambiguity problem, we propose a novel prompt method called knowledge prompt, which leverages richer contextual information to enhance the model's understanding of the segmentation task. A knowledge prompt is an image pair comprising a template and its corresponding mask, where the mask specifies the target region need to be segmented. The template provides rich contextual semantics about the target's surroundings as prior knowledge. Guided by the mask, KSAM accurately segments the target region by maintaining contextual awareness of adjacent anatomical structures. Compared with point or box prompts which only provide positional information, knowledge prompts can incorporate richer semantic details and fully express the segmentation task definition. We annotate a kidney anatomical structures dataset (KAS) shows like Fig. 2 to verify the effectiveness of knowledge prompt on fine-grained segmentation tasks. On KAS and REFUGE dataset, we competitively achieve the best results compared with various SAM variants with different prompt approaches, achieving 89.3% and 89.9% DSC respectively. Our main contributions are summarized as:

- An unbiased knowledge prompt method is employed, effectively addressing the challenges of obtaining high quality prompts and prompt ambiguity.
- We propose two modules to achieve knowledge fusion, making model better understand segmentation task definition and complete the segmentation without uncertainty according to the provided knowledge prompt.
- We provide superior extensibility and competitive performance compared with prompt-free methods.

2 Method

2.1 Knowledge-based Prompt

Due to the specificity and complexity of medical images, it is necessary to provide richer prompt information than simple points or boxes to accomplish fine-grained segmentation tasks. As shown in Fig. 3, we demonstrate the predictions using point prompts on kidney anatomical structure segmentation tasks. The soft prediction results show that the model produces completely opposite segmentation outcomes. The model inaccurately predicts lower confidence in regions that should have high-confidence features.

Therefore, we propose knowledge prompt(KP) to solve the problem in fine-grained segmentation task. Such approach contains enough medical prior knowledge to guide the model. The pipline of KSAM and related modules are shown in Fig. 4. Given the image to be segmented $X \in \mathbb{R}^{3\times H\times W}$, the knowledge prompt $KP = \{KP_{instance}, KP_{task}\}$. where $KP_{instance} \in \mathbb{R}^{3\times H\times W}$, $KP_{task} \in \mathbb{R}^{H\times W}$, model $F(\cdot)$, with output $Pred = F(X, KP_{instance}, KP_{task})$, $Pred \in \{0,1\}^{H\times W}$, each value of the prediction mask corresponds to whether the pixel matches the segmentation task under knowledge prompt KP.

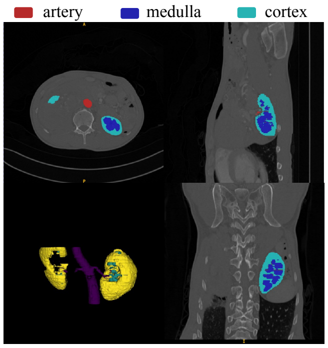

Fig. 2: Overview of the KAS, including planes in the Axial, Sagittal and Coronal and a 3D reconstruction view.

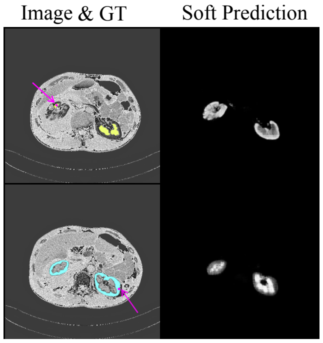

Fig. 3: Using point prompt, soft predictions from MedSA [23].Points pointed by arrows indicate the target object.

We fuse input image with the knowledge prompt through the proposed Knowledge Merge module to enhence the origin input features. $KP_{instance}$ can be seen as the contextual of the segmentation task that KP_{task} indicates, which containing plenty of information about the target object. The enhanced feature $mergedE$ with contextual information of the target object can help decoder to achieve better segmentation. Prompt Merge module compromising four blocks learns prompt embedding corresponding to KP which represents the defination of the segmentation task. Finally, $mergedE$ together with the fused prompt embedding $promptE$ are decoded by Mask Decoder to obtain the final segmentation mask. Since vanilla SAM do not support such prompt, we try to reuse the convolutional module in prompt encoder and add a lightweight adapter to support RGB images.

$$DE_0, DE_1 = PrEnc(PrAdaptor(KP_{instance}, KP_{task})) \tag{1}$$

$$Output = MaskDec(mergedE, promptE) \tag{2}$$

where $PrAdaptor$ and $PrEnc$ denote lightweight adapter and the original prompt encoder of the SAM respectively. $MaskDec$ represents the mask decoder of vanilla SAM and $Output$ is the prediction.

2.2 Knowledge Merge Module

We use Knowledge Merge module(KM) to enhance the original input The fusion module adopts an architecture similar to VNet [15] shown in Fig. 4. The module is consist of convolutional blocks and handles two branches. DE_0, DE_1 come from Eq. (1) and IE, SE are features generated from $X, KP_{Instance}$ through the image encoder respectively. DE_0, DE_1, SE are accomplished by the operation $\odot$, which finally choose summation operation. Image feature are continuously fused with the prompt features in depth and compressed into reduced dimensions.

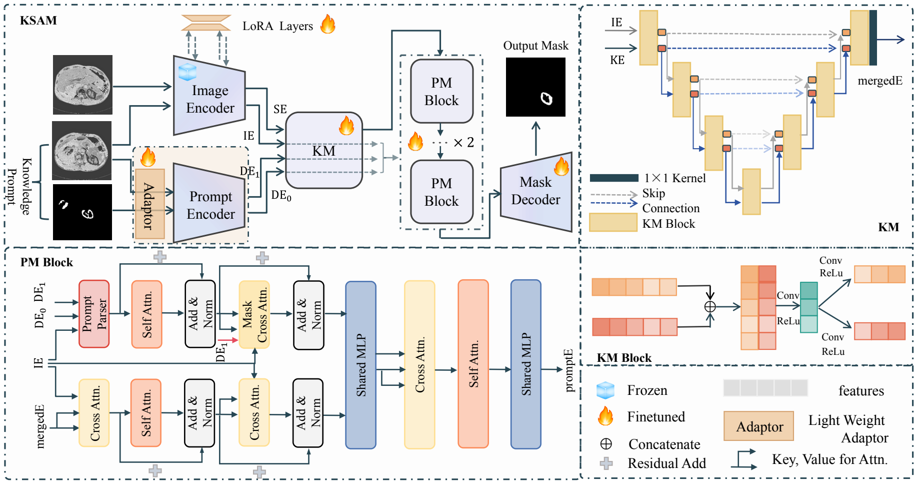

Fig. 4: The overview of KSAM and detailed structres of proposed modules.

Each convolutional block accepts two vectors as input and then concatenates them along channel dimension. Followed by convolutional operation and nonlinear activation function, the vector involved with the knowledge prompt features are reverted to two features and then be used as input to the subsequent convolutional block.

2.3 Prompt Merge Module

Since we remove sparse prompt, we design Prompt Merge Module(PM) to compute prompt embedding which will be used in decoding stage. PM is consisted with attention blocks with each block receive four features as input. PM block have two branches and have symmetrical structure. We reference the Prompt Parser structure [24] instead of cross attention in prompt branch for the fusion of images, knowledge prompts and task definition.

Besides that, we use a learnable mask attention in [7] to enhance the prompt features. The mask further helps the model to retain and enhance the features relevant to the segmentation task. Finally, the two features are transformed by a shared MLP to obtain the prompt embedding.

$$X_{maskAtt} = \text{sigmoid}(DE_1) \otimes Att \tag{3}$$

$\otimes$ denotes the pixel-by-pixel multiplication of the matrix and Att represents attention similarity. By using the sigmoid function, task features DE_1 is mapped to $(0, 1)$. The feature of the background will be mapped to a small non-zero value. After multiplication, both the target foreground features can be effectively preserved according to the mask, while the background part of the features are added to increase the robustness.

2.4 Decoder change And Loss Function

The decoder maintains the structure of the vanilla SAM and decodes the feature vectors $mergedE$ and $promptE$ to get the final segmentation result. The loss we used is the weighted sum of the Binary Cross Entropy loss and the Dice loss.

3 Experiments

3.1 Dataset

KAS The contrast-enhanced CT images come from the partner hospitals and have obtained authorization.We annotate the areas of artery and more complex structures including cortex and medulla using 3D Slicer under the guidance of experienced medical specialists. Given the difficulty and rarity of fine-grained segmentation tasks, the amount of available training data is limited. To further increase task complexity, we use CT images from a patient with severe damage to the left kidney. These images are significantly different from the normal kidney form. This choice allows us to assess whether the model can learn the essential features of the organ's structure, rather than relying on fixed morphological features. In the end, we collected a total of 75 samples.

REFUGE is a fundus dataset containing 400 images available for training. Since a pair of fundus image consists of optic cup as well as disc part, where the cup is surrounded by the disc, it is difficult to distinguish between the cup and the disc in a clear sense when using traditional point prompt or box prompt. The processing settings for the data follow[23].

3.2 Implementation Details

The implementation of the experiments is based on Pytorch and all models except OnePrompt are trained on a single NVIDIA V100 GPU. The LoRA fine-tuning method used is referenced to [23] with the rank of LoRA set to 4. We use a ViT-B backbone to perform all experiments, and the input images as well as the knowledge prompt have the same resolution of 512×512. The optimizer used is Adam with the setting of $\beta_1 = 0.9, \beta_2 = 0.999$. The warmup mechanism is added while training. The maximum epoch and learning rate are set to 300 and 0.0001. The metric for evaluation is Dice Similarity Coefficient (DSC) [12].

3.3 Results

As shown in Tab. 1, we compare KSAM with several prevalent variants, encompassing point-based, box-based, prompt-free, instance-based, and traditional methods. The outputs of each methods is shown in Fig. 5.

On KAS Dataset, the use of point and box prompt is usually more effective in segmenting artery due to the fact that artery is an independent structure

Table 1: Results of different Prompt models. Evaluated on KAS Dataset and REFUGE by DSC. The 1st, 2nd and 3rd-best performances are highlighted by red, orange and yellow respectively. († represents model is fully finetuned.)

Types	Methods	KAS				REFUGE		
		Artery	Medulla	Cortex	Avg↑	Disc	Cup	Avg↑
Point/Box Prompt	MedSA [23]	0.795	0.580	0.620	0.665	0.807	0.644	0.726
	MedSAM [13](bbox 20pixel †)	0.929	0.815	0.878	0.874	0.951	0.818	0.884
	MedSAM [13](bbox 30pixel †)	0.923	0.656	0.856	0.812	0.951	0.462	0.707
	MedSAM2 [14](bbox 20pixel †)	0.927	0.879	0.692	0.833	0.958	0.832	0.895
	MedSAM2 [14](bbox 30pixel †)	0.926	0.873	0.771	0.857	0.956	0.636	0.796
	SAM-Med2D [6](one point)	0.865	0.791	0.702	0.786	0.774	0.660	0.717
	SAM-Med2D [6](bbox 20pixel)	0.904	0.805	0.718	0.809	0.886	0.681	0.784
prompt-free Methods	HSAM [7]	0.921	0.891	0.878	0.897	0.721	0.701	0.711
	AutoSAM [8]	0.930	0.876	0.880	0.895	0.608	0.663	0.636
	SAMed [28]	0.914	0.874	0.875	0.888	0.586	0.538	0.562
Traditional Methods	PANet [20]	0.086	0.348	0.295	0.243	0.340	0.535	0.438
	TransUnet [3]	0.888	0.865	0.845	0.866	0.540	0.486	0.513
	SwinUnet [1]	0.829	0.867	0.822	0.840	0.732	0.726	0.729
Instance prompt	PerSAM [29]	0.071	0.079	0.169	0.106	0.399	0.316	0.358
	OnePrompt [24](SegLab)	0.131	0.442	0.488	0.354	0.800	0.420	0.610
	ours	0.922	0.882	0.881	0.893	0.931	0.866	0.899

with a clearer outline structure, and is less prone to the problem of unclear task designation. However, it performs bad on medulla and cortex, which have complex structures, and is unable to correctly understand the structure that the cue is trying to decompose. Although MedSAM shows a similar performance reaching an 87.8% DSC, it adopts a fully finetuned training approach while pretraining on more than 1 million medical data images yet still lower than our method. For instance prompt methods, mainstream methods usually obtain point prompts according to similarity and reuse the original prompt encoder. Obviously there is no guarantee that the most related point is always right. On REFUGE Dataset, we outperform all the other methods reaching an 89.9% DSC for the special structure of organs.

All the other methods perform poorly in the cup category. Although MedSAM as well as MedSAM2 using fully finetune achieve close reults, when the prompt provided are disturbed, the performance is severely degraded up to a performance gap of almost 18%.

Although the prompt-free method HSAM achieves 0.4% higher performance than ours, HSAM is unscalable when facing an unseen task, as its ability is fixed during the training stage.

We first fine-tune KSAM on REFUGE and then refine the weights obtained from REFUGE KAS. The same parameter settings are used in both stages, and no additional fine-tuning tricks are employed. We evaluate the performance on the REFUGE and KAS separately every 10 epochs using the weights refine-tuned on KAS. The results obtained are shown as Fig. 6.

We achieve an average DSC of 67.15% at 25 epoch which has already surpassed MedSA that need 300 epochs. When fine-tuned by 15 epochs, we still achieve 73.9% DSC on the REFUGE, which outperform all baselines except for

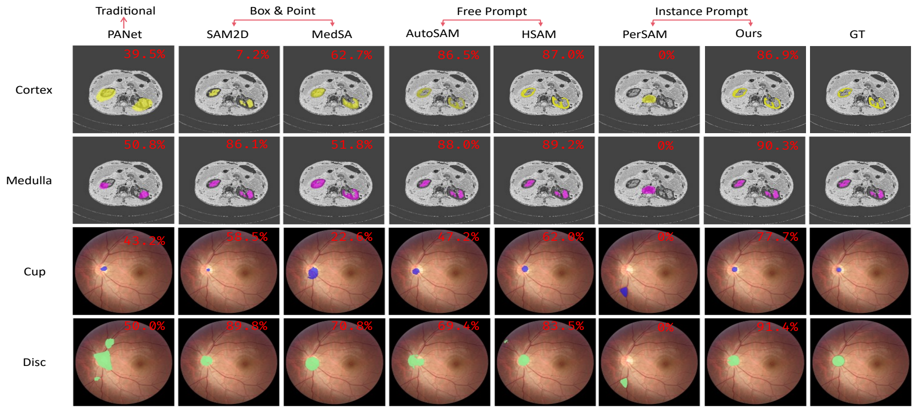

Fig. 5: Comparison of output results. Each column represents segmentation results for different tasks on KAS and REFUGE.

MedSAM and SAM-Med2D which using fully finetune and pretrained on millions of medical images. For Traditional Methods as well as Free Prompt Methods, they have no capability in this experimental setup. The tasks have already been fixed at the beginning, so they will completely forget what have learned before.

3.4 Ablation Study

As shown in Tab. 2, we conduct ablation experiments on KM and PM. Due to KP serving as a essential structure which can not be simply eliminated, it is used in all experiments. The direct use of KP results in poor performance because vanilla SAM does not support this prompt method, and no additional components are adapted. Adding PM improves performance by 8.1%, enhancing prompt understanding. Adding KM further boosts performance by 36.8%, significantly enhancing the encoded features through knowledge prompt integration.

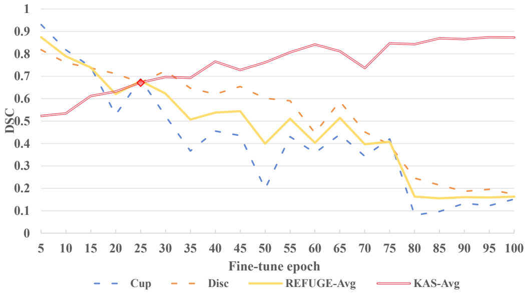

Fig. 6: Scalability test on datasets.

Table 2: Ablation of the modules proposed.

ID	KP	KM	PM	DSC ↑
1	✓	-	-	0.304
2	✓	-	✓	0.385
3	✓	✓	-	0.672
4	✓	✓	✓	0.893

The best result is achieved when all modules are active, enabling the model to incorporate knowledge features and discriminate between prompt tasks. It should be noted that experiment 2 is similar with OnePrompt when just use PM and it is logical to have similar performance(38.5% vs. 35.4%).

4 Conclusion

We use knowledge prompts to address the issue of misinterpretation in complex fine-grained tasks. KSAM achieves higher performance with partial fine-tuning, even surpassing some models that are fully fine-tuned. The performance of scalability far exceeds the prompt free approach while maintaining the same level of segmentation capability. These results demonstrate that our method offers a novel approach to help visual prompt models better understand what to segment.

Acknowledgments. Thanks to everyone for their efforts in this work.

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: Proceedings of the European Conference on Computer Vision Workshops(ECCVW) (2022)
2. Chen, F., Tang, J., Wang, P., Wang, T., Li, S., Deng, T.: Deap-3dsam: Decoder enhanced and auto prompt sam for 3d medical image segmentation. In: 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 1852–1859 (2024). https://doi.org/10.1109/BIBM62325.2024.10822764
3. 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)
4. Chen, Z., Xiao, C., Liu, Y., Hassan, H., Li, D., Liu, J., Li, H., Xie, W., Zhong, W., Huang, B.: Comprehensive 3d analysis of the renal system and stones: Segmenting and registering non-contrast and contrast computed tomography images. Information Systems Frontiers pp. 1–15 (2024)
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. Cheng, J., Ye, J., Deng, Z., Chen, J., Li, T., Wang, H., Su, Y., Huang, Z., Chen, J., Sun, L.J.H., He, J., Zhang, S., Zhu, M., Qiao, Y.: Sam-med2d (2023)
7. 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. pp. 3511–3522 (2024)
8. Hu, X., Xu, X., Shi, Y.: How to efficiently adapt large segmentation model (sam) to medical images. arXiv preprint arXiv:2306.13731 (2023)
9. Jin, C., Shi, F., Xiang, D., Zhang, L., Chen, X.: Fast segmentation of kidney components using random forests and ferns. Medical physics **44**(12), 6353–6363 (2017)

10. Korfiatis, P., Denic, A., Edwards, M.E., Gregory, A.V., Wright, D.E., Mullan, A., Augustine, J., Rule, A.D., Kline, T.L.: Automated segmentation of kidney cortex and medulla in ct images: a multisite evaluation study. Journal of the American Society of Nephrology **33**(2), 420–430 (2022)
11. Liu, Y., Zhu, M., Li, H., Chen, H., Wang, X., Shen, C.: Matcher: Segment anything with one shot using all-purpose feature matching. arXiv preprint arXiv:2305.13310 (2023)
12. Luo, X., Liao, W., Xiao, J., Chen, J., Song, T., Zhang, X., Li, K., Metaxas, D.N., Wang, G., Zhang, S.: Word: A large scale dataset, benchmark and clinical applicable study for abdominal organ segmentation from ct image. Medical Image Analysis **82**, 102642 (2022)
13. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024)
14. Ma, J., Kim, S., Li, F., Baharoon, M., Asakereh, R., Lyu, H., Wang, B.: Segment anything in medical images and videos: Benchmark and deployment. arXiv preprint arXiv:2408.03322 (2024)
15. 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)
16. Tang, L., Jiang, P.T., Xiao, H., Li, B.: Towards training-free open-world segmentation via image prompt foundation models. International Journal of Computer Vision pp. 1–15 (2024)
17. Tang, Y., Gao, R., Lee, H.H., Xu, Z., Savoie, B.V., Bao, S., Huo, Y., Fogo, A.B., Harris, R., de Caestecker, M.P., et al.: Renal cortex, medulla and pelvicaliceal system segmentation on arterial phase ct images with random patch-based networks. In: Medical Imaging 2021: Image Processing. vol. 11596, pp. 379–386. SPIE (2021)
18. Tianrun, C., et al.: Sam fails to segment anything? sam-adapter: Adapting sam in underperformed scenes: Camouflage shadow and more. arXiv preprint arXiv:2304.09148 (2023)
19. Wang, H., Guo, S., Ye, J., Deng, Z., Cheng, J., Li, T., Chen, J., Su, Y., Huang, Z., Shen, Y., et al.: Sam-med3d: towards general-purpose segmentation models for volumetric medical images. arXiv preprint (2023)
20. Wang, K., Liew, J.H., Zou, Y., Zhou, D., Feng, J.: Panet: Few-shot image semantic segmentation with prototype alignment. In: proceedings of the IEEE/CVF international conference on computer vision. pp. 9197–9206 (2019)
21. Wang, S., He, Y., Kong, Y., Zhu, X., Zhang, S., Shao, P., Dillenseger, J.L., Coatrieux, J.L., Li, S., Yang, G.: Cpnet: cycle prototype network for weakly-supervised 3d renal compartments segmentation on ct images. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24. pp. 592–602. Springer (2021)
22. Wang, X., Xie, L., Qiao, P., Dou, Y., Liu, S., Li, W., Yang, K.: Sam-nerf: Nerf-based 3d instance segmentation with segment anything model. In: International Conference on Artificial Neural Networks. pp. 434–448. Springer (2024)
23. 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)
24. Wu, J., Xu, M.: One-prompt to segment all medical images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11302–11312 (2024)

25. Xu, Z., Chen, Q.: Nubbledrop: A simple way to improve matching strategy for prompted one-shot segmentation. arXiv preprint arXiv:2405.11476 (2024)
26. Yang, X., Le Minh, H., Cheng, K.T.T., Sung, K.H., Liu, W.: Renal compartment segmentation in dce-mri images. Medical image analysis **32**, 269–280 (2016)
27. Yin, F., Li, J., Wei, Y., Zhang, W., Xu, C.: Sams: One-shot learning for the segment anything model using similar images. In: 2024 International Joint Conference on Neural Networks (IJCNN). pp. 1–8. IEEE (2024)
28. Zhang, K., Liu, D.: Customized segment anything model for medical image segmentation. arXiv preprint arXiv:2304.13785 (2023)
29. Zhang, R., Jiang, Z., Guo, Z., Yan, S., Pan, J., Ma, X., Dong, H., Gao, P., Li, H.: Personalize segment anything model with one shot. arXiv preprint arXiv:2305.03048 (2023)