RadSAM: Segmenting 3D radiological images with a 2D promptable model

Julien Khlaut[1,2,3], Elodie Ferreres[1], Daniel Tordjman[1], Helene Philippe[1,2,4], Tom Boeken[2,3], Pierre Manceron[1], and Corentin Dancette[1,5]

[1] Raidium, France firstname.lastname@raidium.eu
[2] Université de Paris Cité,
[3] AP-HP, Hôpital Européen Georges Pompidou, Department of Vascular and Oncological Interventional Radiology, HEKA INRIA, INSERM PARCC U 970, Paris
[4] AP-HP. Nord, Department of Radiology, FHU MOSAIC, Beaujon Hospital, Clichy
[5] corresponding author: corentin.dancette@raidium.eu

Abstract. Medical image segmentation is a crucial and time-consuming task in clinical care, where precision is extremely important. The Segment Anything Model (SAM) offers a promising approach, providing an interactive interface based on visual prompting and edition. However, this model and adaptations for medical images are built for 2D images, whereas a whole medical domain is based on 3D images, such as CT and MRI. This requires one prompt per slice, making the segmentation process tedious. We propose RadSAM, a novel method for segmenting 3D objects with a 2D model from a single prompt, based on an iterative inference pipeline to reconstruct the 3D mask slice-by-slice. We introduce a benchmark to evaluate the model's ability to segment 3D objects in CT images from a single prompt and evaluate the models' out-of-domain transfer and edition capabilities. We demonstrate the effectiveness of our approach against state-of-the-art 2D and 3D models using the AMOS abdominal organ segmentation dataset.

Keywords: Computer Vision · Interactive Segmentation · Computed Tomography

1 Introduction

Image segmentation is an essential task for medical imaging [12]: it allows specialists to compute multiple metrics regarding anatomical and pathological objects useful for clinical care. The manual segmentation process is long, error-prone, subjective, and led to the adoption of simplified criteria such as RECIST [2] to avoid segmenting whole 3D structures.

Deep-learning-based methods for organs and tumors have been studied extensively; most of them are trained on specific datasets using U-net architectures [5, 3, 10, 19]. However, standard semantic segmentation models are subject to limitations in their clinical usage: if the model outputs a wrong segmentation map, it requires manual correction. Recently, the Segment Anything Model (SAM) [7]

was introduced and proposed a prompted segmentation framework, allowing users to interactively guide the model with a spatial prompt to obtain an initial segmentation and correct the model to refine the mask. This framework is promising for medical image segmentation, as it allows the precise segmentation of diverse structures more consistently and quickly than manual segmentation. However, the original Segment Anything model (SAM) was shown to be unreliable for medical data [9, 18]. Thus, Ma et al. proposed MedSAM [8], a fine-tuned SAM for medical segmentation. SAM-Med 2D [1], Medical SAM Adapter [16] and SA-Med2D [17] feature similar approaches. Yet, they lack important SAM features. More importantly, in radiology, high-quality imaging is produced in 3D, with Magnetic resonance imaging (MRI) and Computed Tomography (CT). Using a 2D model requires one prompt for each 2D slice to segment full 3D objects, with potentially noncoherent segmentation for consecutive slices. SAM-Med-3D [14], propose a model that directly segments 3D medical images from the full volume. However, such a 3D model has high training memory constraints, making it hard to train and deploy. SAM-2 [13] was recently introduced for interactive video segmentation, requiring an iterative pipeline during training.

We introduce RadSAM (for Radiological SAM). This promptable segmentation model can segment 3D structures in CT images from a single prompt, using the memory footprint of a 2D model. In addition to points and box prompts, we train a 2D segmentation model with a mask prompt as the first input. The model learns to reconstruct the ground-truth mask from a degraded input. We then introduce an iterative inference method to leverage the novel mask prompt and forward instructions from one slice to the next with minimal information loss. This allows us to segment 3D radiological objects from a single prompt and to add 3D editing capabilities. This is different from SAM2 mainly in term of training: SAM2 requires feeding the model with a whole volume during training, as it forwards a memory tensor iteratively and backpropagates the loss through all the slices. This requires more memory than our training pipeline, where we only train in 2D by perturbing the ground truth mask as input. Our main contribution is to segment 3D objects without specifically training the model in 3D. We benchmark our promptable segmentation model on 3D imaging using two CT organ segmentation datasets: AMOS [6] and TotalSegmentator [15]. We evaluate its capacity to segment from multiple prompt types, transferability to other datasets, and edition performances.

2 Benchmark

2.1 Datasets

To produce our benchmark, we consider only models trained on the AMOS dataset [6]. It comprises a large and diverse collection of clinical CT scans for abdominal organ segmentation and some MRIs. We only use the CT part containing 500 CT scans and voxel-level annotations of 15 abdominal organs. In addition to AMOS, we use another dataset for additional experiments: TotalSegmentator

(TS) [15]: This dataset consists of 1204 CT scans with detailed annotations of 104 anatomical structures (27 organs, 59 bones, 10 muscles, and 8 vessels).

2.2 Input prompts

The benchmark evaluates the model's response to different prompting strategies, as detailed in Figure 1a. We consider 2 approaches. The first one is slice-level prompting, which consists of predicting a mask for each slice of interest by giving a prompt on each slice. The second one is volume-level prompting; this consists of a unique 2D prompt on one slice and 2 boundary annotations indicating the maximum and minimum slices. We also study the models' ability to respond to edition prompts. For the 2D slice-level prompting, we add a point to each of the slices, whereas for the volume-level prompting with boundaries, we add a point for the whole volume.

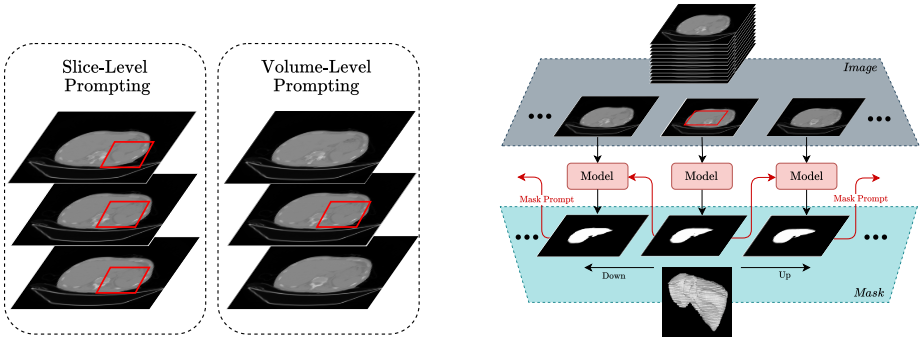
3 RadSAM

3.1 Architecture

We use the SAM [7] architecture and weights as a starting point to train RadSAM. The model takes as input a 2D image $v \in \mathbb{R}^{3\times H\times W}$ and a combination of one or multiple prompts, including a bounding-box $p_{\text{box}} \in \mathbb{R}^4$, one or multiple points $p_{\text{point}} \in \mathbb{R}^{N\times 2}$ along with their positive/negative labels $p_l \in [0,1]^N$, or a mask $p_m \in \mathbb{R}^{H'\times W'}$, and returns one or multiple masks $m \in R^{H\times W}$. We always generate 4 masks: one primary and three secondary masks, used for oracle prediction. For all evaluations, we use the primary mask unless specified otherwise. The training objective we use is the sum between the soft dice loss [11] $\mathcal{L}_d$, commonly used for medical imaging, and the binary cross-entropy loss $\mathcal{L}_{\text{BCE}}$.

3.2 Prompts generation

We generate three prompts for each object to train and evaluate our model: point, bounding box, and mask. The prompts are randomly created from the ground truth mask. For the mask prompting, a novelty compared to SAM, we generate a noisy version of the ground-truth mask with random transforms (rotation: 5.0 degrees, scaling: 10%, translation: 15%, erosion: 5 steps, and dilation: 5 steps). The scale of those perturbations is very important because it prevents the model from collapsing and reproducing the input mask exactly. It also enables the model to segment a structure with a mask prompt from a neighbor slice. This novel prompt type enables iterative generation (Sec. 3.4) or manually drawing a mask. To unify the input prompts, we encode all input masks with binary values instead of logits like in SAM.



(a) Prompting strategies. Slice-level prompting means one prompt for each slice. Volume-level prompting consists of a single 2D prompt, and annotations for the top and bottom slices.

(b) RadSAM iterative segmentation pipeline. The model is applied slice by slice, with the mask output passed as a prompt to segment the next slice.

Fig. 1: RadSAM method for prompting strategies and iterative segmentation

3.3 Training and Inference Details

The training was done with a constant learning rate of 10^{-5}, a batch size of 2 and was trained on 16 NVIDIA V100s for 7 epochs with the Adam optimizer. In term of inference speed RadSAM can process 23.42 images per second on a RTX 4090. To infer a volume composed of N slices, it needs N/23.42 seconds.

3.4 Iterative Segmentation for volume-level prompting

We propose an inference pipeline where the user provides a single 2D prompt with top and bottom slice boundaries. We display the inference pipeline in Figure 1b. The user provides an initial prompt on a single slice i and the model returns a mask m_i for this slice. We then create a new prompt for the slices $i+1$ and $i-1$ based on m_i. For RadSAM, we simply use the mask m_i as the new prompt. We generate a bounding box around this mask for models that do not support the initial mask prompting, such as SAM or MedSAM.

3.5 Edition in iterative segmentation

Allowing editing through the iterative pipeline is more challenging than with a 2D mode, but it is important as errors can accumulate over the steps. We propose a method to integrate them to perform editions of the whole 3D mask with a single edition point. These correction points are analogous to those used in 2D scenarios and are placed where the mask is either incorrectly present or absent. They are sampled uniformly among the errors.

The propagation strategy is similar to that used for the first segmentation. When the user adds a correction point to the slice i, we feed the previous mask

Table 1: Dice scores on AMOS for RadSAM, SAM, MedSAM, SAM-Med3D and nnU-Net. MedSAM does not support point prompting and edition, and nnU-Net does not support prompting. n represents the number of slices. The 3 represents the initial prompt and the two boundaries for volume-level prompts. We display the 95% confidence intervals for all results we evaluate. Best results without edition and with edition are respectively in bold and underlined. |P| and |E| represent the number of initial prompts and the number of edits.

	Volume-level prompt				Slice-level prompt			
Model	|P|	|E|	Bbox	Point	|P|	|E|	Bbox	Point
nnU-Net	∅	∅	88.87		∅	∅	88.87	
SAM-Med3D	1	-	-	79.94				
	1	9	-	83.99				
SAM	3	–	50.93 ±1.40	38.78 ±1.73	n	–	70.20 ±0.90	33.07 ±1.60
	3	20	60.60 ±1.36	47.16 ±1.74	n	$10n$	85.43 ±0.35	84.31 ±0.45
MedSAM	3	–	53.19 ±1.47	–	n	–	81.36 ±0.70	–
RadSAM	3	–	**84.99** ±0.74	**84.05** ±0.86	n	–	**91.08** ±0.35	**85.09** ±0.70
	3	20	<u>91.11</u> ±0.39	<u>90.63</u> ±0.46	n	$10n$	<u>96.73</u> ±0.13	<u>96.78</u> ±0.13

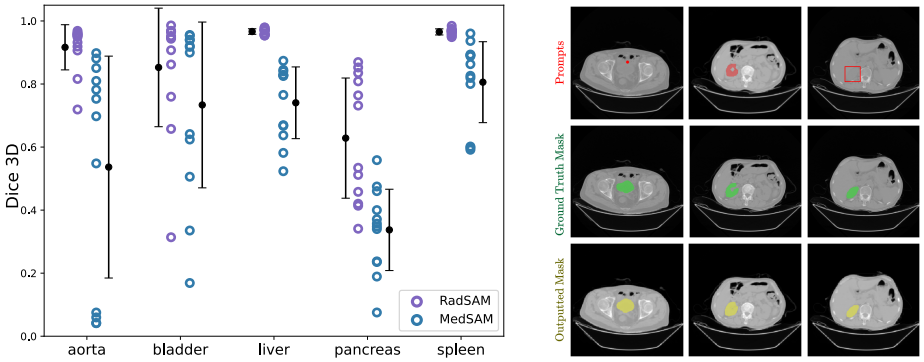
m_i along with the correction point to obtain a corrected mask m_i'. We then start the propagation again with this new corrected mask. If the model encounters a slice with previous correction points, we can re-use them in addition to the previous mask to give the model more information.

4 Results

4.1 Main Results

We compare RadSAM with two publicly available 2D prompted segmentation models: MedSAM [8], trained on multiple medical datasets including AMOS [6], and SAM [7], trained on large datasets of natural images in Table 1. We also compare with SAM-Med3D [14], a 3D segmentation model trained with volume-level point prompting and editing, but without bounding boxes support. Unless specified otherwise, we evaluate all models on the AMOS validation set, composed of 100 CT volumes. All evaluations report the 3D Dice metric, computed on the whole organ, even in the slice-level prompting evaluations. We also report scores of a 3D nnU-Net [4], a semantic segmentation model without prompting. RadSAM is trained for 7 epochs on the AMOS training set with random translations, rotations, shear, zoom, gaussian noise, and clipping of HU values.

Volume-Level Prompting We first evaluate models with volume-level prompting, with slice boundaries on the top and bottom of the organ. RadSAM reaches a dice of 84.99 with a single bounding box prompt, largely beating MedSAM,



(a) Quantitative evaluation results using the iterative pipeline on some classes of AMOS. We represented the standard deviation

(b) Qualitative results from various prompts.

Fig. 2: Detailed quantitative and qualitative evaluation results.

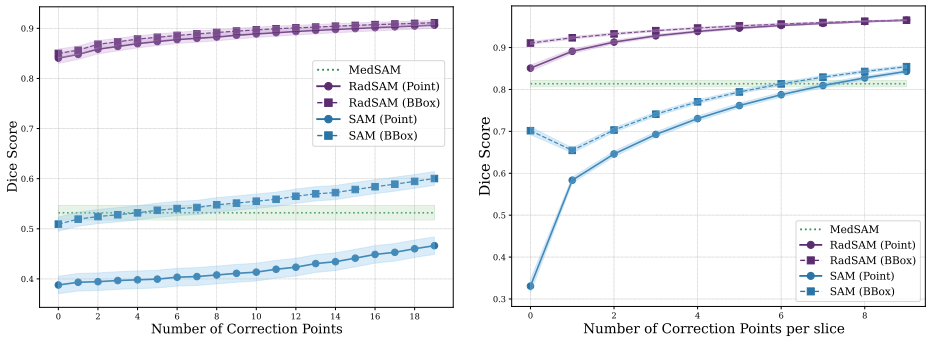
which gets a dice of 53.19. With a point prompt, which can be much more ambiguous, our model obtains 84.05, losing only 0.94 points compared to the box prompt. With 20 edition points, RadSAM gains 6.12 points, beating the semantic segmentation model nnU-Net. SAM and MedSAM perform very poorly on this setup, accumulating errors over the iterations. Finally, RadSAM also shows a gain of 4.11 points over SAM-Med3D using only one point and a score of 88.61 using a point prompt and 9 edition points, yielding a 4.62 points improvement over SAM-Med3D. This shows that a 2D-based model can beat a full 3D model. Figure 2a shows organ-level scores along with their variances. RadSAM provides more accurate 3D masks and a much smaller variance interval on most anatomical structures.

Slice-Level Prompting We give one prompt per slice where the object is present, i.e. in total n. The dice scores are still computed on the final 3D object. Additionally, we assessed the model's performance with 10 editing points added to each slice. RadSAM obtains a dice score of 91.08, beating the scores of nnU-Net, MedSAM and SAM by a significant margin.

Figure 2b shows a qualitative example of our model's predictions for each type of prompt: point, box, and mask. We display the prompt, the ground-truth mask, and the model's prediction.

4.2 Edition Capabilities

We report the editing capabilities of the three models with volume-level and slice-level promptings in Figure 3. We show important gains with editing: in volume-level prompting, the model gains 6.12 dice points when going from 0 to 20 edition points. In slice-level prompting, the model gains 5.65 points with 10 points per slice. SAM, while not specifically trained on medical images, beats



(a) Volume-level prompting (b) Slice-level prompting

Fig. 3: Dice scores with various numbers of edition points on AMOS. MedSAM is represented as a straight line, as it does not support edition.

Table 2: Generalization performances of a model trained on AMOS, evaluated on TotalSegmentator (TS). Known are the classes from TS that are present in AMOS, and unknown are all the other classes from TS.

(a) Volume-level prompting

	Known	Unknown
SAM	29.71 ±2.72	**33.86** ±0.65
MedSAM	44.46 ±3.28	24.85 ±1.87
RadSAM	**60.98** ±3.22	30.57 ±0.68

(b) Slice-level prompting

	Known	Unknown
SAM	50.77 ±2.30	54.53 ±0.46
MedSAM	67.26 ±2.08	50.33 ±0.47
RadSAM	**78.72** ±1.74	**62.94** ±0.41

MedSAM on both setups with sufficient edition prompts. This highlights the importance of integrating the edition mode.

4.3 Transfer Learning

Medical models must be robust to images from various domains, as images coming from different hospitals or machines can have different distributions. We use the TotalSegmentator dataset to assess our model's performance on out-of-domain images. We split its classes into two subsets: "known" classes present in AMOS and "unknown" classes, which are not. We show the results in Table 2. RadSAM demonstrates superior generalization, outperforming MedSAM on out-of-domain known classes, with 60.98 dice (+16.52 with respect to MedSAM) with volume-level prompting and 78.72 (+11.46 points) with slice-level prompting. As expected, SAM obtains a lower score on those classes. For unknown classes, we observe that while slice-level prompting gives average performances (62.94 dice, losing around 20 points to the known classes), the volume-level prompting model's performances degrade much more, losing around 30 points.

Table 3: Ablation studies.

(a) Impact of the iterative prompt (mask or bbox) on the 3D dice. Results in gray show unsupported prompts.

Iterative prompt	Mask		Bbox	
Initial prompt	Bbox	Point	Bbox	Point
MedSAM	12.34	1.17	58.26	3.22
RadSAM	84.99	84.05	66.56	59.54

(b) Comparing training dataset of RadSAM, on TS (all classes) and AMOS using slice-level prompting.

	TS		AMOS	
Training strategy	Bbox	Point	Bbox	Point
TS	90.87	88.78	82.51	59.38
TS → AMOS	76.65	62.47	91.40	86.20
TS + AMOS	90.20	87.67	90.05	83.27

(c) Model size evaluation of RadSAM. Both models are trained for 7 epochs.

	Bbox	Point
ViT-B	91.08	85.09
ViT-L	91.59	86.15

(d) Oracle evaluation of RadSAM: the best mask among the 3 predictions is selected.

	Bbox	Point
Dice 3D	91.08	85.09
Dice oracle	92.52	90.25

5 Ablation Study

We perform ablations of some critical parameters of our approach.

Impact of iterative prompt in the volume-level pipeline: We evaluate the choice of using the mask prompt for the iterative inference pipeline. We compare this prompt to using a bounding box around the mask of the previous slice. Our results, in Table 3a, show that the mask prompt drastically increases the dice score, going from 66.56 to 84.99 (+18.43) with an initial bounding box prompt, with similar gains from the initial point prompt.

Scaling the model size: All our previous experiments used the ViT-B architecture to reduce computing costs. We evaluate the performance of ViT-L in Table 3c and show that scaling significantly increases performance: the model gains around 0.5 to 1 dice point for each prompt type.

Varying training datasets: We compare the performance of our model under three different training scenarios in Table 3b: (1) training only on TotalSegmentator, (2) training on TotalSegmentator followed by fine-tuning on AMOS, and (3) joint training on both datasets. The results reveal that fine-tuning on AMOS significantly boosts performance but this improvement comes at the cost of reduced performance on TotalSegmentor. In contrast, training on both datasets simultaneously achieves high performance.

Oracle evaluation: Table 3d compares the main predicted mask and the oracle mask (the best among the three outputted masks). The oracle obtains significantly better performance at a low user cost: one additional click to select the desired mask.

6 Conclusion

We propose a simple method for 3D prompted segmentation using volume-level prompts from a 2D segmentation model. Naive approaches with existing models, like using bounding boxes to forward the prompt from slice to slice, perform very poorly. However, adding a novel prompt type, the mask, as the iterative prompt drastically improves those results. RadSAM demonstrates performances close to the state-of-the-art segmentation model nnU-Net, surpassing it with edition points, as well as beating native 3D models. Additionally, we show that RadSAM generalizes well when evaluated on other datasets with the same classes. This work lays the foundation for more effective clinical utilization of segmentation models. Their interactivity through prompting and editing is essential to give users control over the output mask for most medical tasks where decisions impact clinical care.

Acknowledgments. This work was granted access to the HPC resources of IDRIS under the allocation 2024-AD011013489R2 made by GENCI.

Disclosure of Interests. The authors are affiliated with Raidium.

References

1. Cheng, J., Ye, J., Deng, Z., Chen, J., Li, T., Wang, H., Su, Y., Huang, Z., Chen, J., Jiang, L., Sun, H., He, J., Zhang, S., Zhu, M., Qiao, Y.: SAM-Med2D (Aug 2023). https://doi.org/10.48550/arXiv.2308.16184, http://arxiv.org/abs/2308.16184, arXiv:2308.16184
2. Eisenhauer, E.A., Therasse, P., Bogaerts, J., Schwartz, L.H., Sargent, D.J., Ford, R., Dancey, J.E., Arbuck, S.G., Gwyther, S., Mooney, M., Rubinstein, L., Shankar, L.K., Dodd, L.E., Kaplan, R.M., Lacombe, D., Verweij, J.: New response evaluation criteria in solid tumours: revised recist guideline (version 1.1). European journal of cancer **45 2**, 228–47 (2009), https://api.semanticscholar.org/CorpusID:8748071
3. 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)
4. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (Feb 2021). https://doi.org/10.1038/s41592-020-01008-z, https://www.nature.com/articles/s41592-020-01008-z, publisher: Nature Publishing Group

5. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)
6. Ji, Y., Bai, H., Yang, J., Ge, C., Zhu, Y., Zhang, R., Li, Z., Zhang, L., Ma, W., Wan, X., Luo, P.: AMOS: A Large-Scale Abdominal Multi-Organ Benchmark for Versatile Medical Image Segmentation (Sep 2022). https://doi.org/10.48550/arXiv.2206.08023, http://arxiv.org/abs/2206.08023, arXiv:2206.08023
7. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment Anything (Apr 2023). https://doi.org/10.48550/arXiv.2304.02643, http://arxiv.org/abs/2304.02643, arXiv:2304.02643
8. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment Anything in Medical Images (Apr 2024). https://doi.org/10.48550/arXiv.2304.12306, http://arxiv.org/abs/2304.12306, arXiv:2304.12306
9. Mazurowski, M.A., Dong, H., Gu, H., Yang, J., Konz, N., Zhang, Y.: Segment anything model for medical image analysis: An experimental study. Medical Image Analysis **89**, 102918 (Oct 2023). https://doi.org/10.1016/j.media.2023.102918, https://www.sciencedirect.com/science/article/pii/S1361841523001780
10. Milletarì, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. 2016 Fourth International Conference on 3D Vision (3DV) pp. 565–571 (2016), https://api.semanticscholar.org/CorpusID:206429151
11. 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)
12. Pham, D.L., Xu, C., Prince, J.L.: Current methods in medical image segmentation. Annual review of biomedical engineering **2**(1), 315–337 (2000)
13. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
14. Wang, H., Guo, S., Ye, J., Deng, Z., Cheng, J., Li, T., Chen, J., Su, Y., Huang, Z., Shen, Y., Fu, B., Zhang, S., He, J., Qiao, Y.: Sam-med3d (2023)
15. Wasserthal, J., Breit, H.C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D., Cyriac, J., Yang, S., Bach, M., Segeroth, M.: TotalSegmentator: robust segmentation of 104 anatomical structures in CT images (Jun 2023). https://doi.org/10.48550/arXiv.2208.05868, http://arxiv.org/abs/2208.05868, arXiv:2208.05868
16. 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 (2023)
17. Ye, J., Cheng, J., Chen, J., Deng, Z., Li, T., Wang, H., Su, Y., Huang, Z., Chen, J., Jiang, L., et al.: Sa-med2d-20m dataset: Segment anything in 2d medical imaging with 20 million masks. arXiv preprint arXiv:2311.11969 (2023)
18. Zhang, Y., Shen, Z., Jiao, R.: Segment anything model for medical image segmentation: Current applications and future directions. Computers in Biology and Medicine **171**, 108238 (Mar 2024). https://doi.org/10.1016/j.compbiomed.2024.108238, https://www.sciencedirect.com/science/article/pii/S0010482524003226
19. Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: Unet++: A nested u-net architecture for medical image segmentation. Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support : 4th International

Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, held in conjunction with MICCAI 2018, Granada, Spain, S... **11045**, 3–11 (2018), https://api.semanticscholar.org/CorpusID:50786304