SAM2 as a Key Bridge Between 2D Slices and 3D Volumes for Sparsely-supervised Medical Image Segmentation

Peng Yu, Duowen Chen, Yunhao Bai, and Yan Wang(✉)

Shanghai Key Laboratory of Multidimensional Information Processing, East China Normal University, Shanghai, China
51275904040@stu.ecnu.edu.cn, duowen_chen@hotmail.com, 51215904056@stu.ecnu.edu.cn, ywang@cee.ecnu.edu.cn

Abstract. Deep learning-based 3D medical image segmentation is limited by the high cost of voxel-level annotation. A practical alternative is to label only three orthogonal central slices per volume, yet such sparse 2D supervision provides insufficient constraints to reconstruct topologically consistent 3D structures. Existing methods struggle to bridge this dimensional gap, often failing to segment 2D slices far from the labeled slices, and cannot transfer morphological priors from 2D labeled slices into 3D structures. To overcome these limitations, we propose a novel framework that pioneers the use of the Segment Anything Model 2 (SAM2) as a dimensional bridge. SAM2 has the ability to propagate 2D annotations throughout the 3D volume, but it suffers from propagation drift, i.e., the progressive accumulation of segmentation errors during long-range propagation. To mitigate the drift, we not only propose a cross-view consensus mechanism for adaptively fusing these 2D slice-wise predictions and extending sparse 2D annotations to more robust dense annotations, but also introduce a dynamic 3D refinement module to capture inherent volumetric continuity. Then a confidence-weighted fusion method is designed to obtain the final segmentation results. Experiments on three public benchmarks demonstrate that our method achieves consistent gains (6.7% in Dice on LA2018 dataset) over state-of-the-art baselines. Code is publicly available at https://github.com/DeepMedLab-ECNU/SAM2-bridge.

Keywords: Sparse annotation · SAM2 · Multi-view propagation · Pseudo-label refinement.

1 Introduction

Accurate 3D medical image segmentation is crucial for clinical diagnosis and treatment planning [19,9,26,27]. However, training high-performance deep learning models typically relies on **dense voxel-level annotations**, spanning generic semantic segmentation [2,6,21], disease-specific tasks [8,14], and recent advances in foundation models [12], all of which are prohibitively expensive and time-consuming to acquire in clinical practice. To alleviate this burden, semi-supervised

learning under sparse annotation has emerged as a highly promising alternative [4,5,17,1,11,15,25]. A particularly clinically friendly setting is to annotate only three orthogonal central slices (axial, coronal, and sagittal) for each labeled 3D scan [23,20]. Under this extremely sparse supervision, our primary goal is to reconstruct the complete anatomical structure of 3D volumes relying solely on the prior information from these 2D slices.

However, bridging the gap between three-slice supervision and dense 3D prediction presents significant challenges. Existing methods primarily follow two distinct trajectories. The first category directly trains models under the extremely sparse supervision of three-view central slices, incorporating mechanisms such as cross-view consistency [23,20,5]. However, due to severely limited supervision signals, performance often degrades on slices far from the annotated centers, struggling to propagate 2D morphological priors into the 3D volumetric space. Furthermore, when single-view predictions are unreliable, these consistency constraints risk degenerating into erroneous alignments. The second category relies on image registration [4,17], attempting to propagate labels by calculating inter-slice deformations. However, this propagation is strictly limited to the annotated data and cannot be extended to other unlabeled data. Since registration-based labels cannot be "created from nothing," the potential of the unlabeled training set remains largely unexploited, limiting the model's performance.

To this end, we pioneer the introduction of SAM2 [16] for sparse-to-dense 3D medical segmentation and propose a unified segmentation framework comprising three core modules: SAM2-guided multi-view propagation, cross-view consensus-weighted fusion, and dynamic pseudo-label refinement. In the 2D stage, to strengthen the morphological prior and stabilize the memory-based propagation process, we stack the annotated central slices across cases to construct view-specific global morphological anchors. Subsequently, by inserting these anchors as prompts into the center of the target sequence, we are the first to leverage SAM2's powerful inter-frame memory mechanism under the sparse-supervision setting, effectively diffusing the sparse 2D slice annotation across the entire volumetric space. Nevertheless, during long-range frame-by-frame propagation, as predictions move further away from the central anchors, errors inevitably accumulate and magnify. While existing multi-view methods attempt to correct these errors using complementary views, this alignment often reinforces mistakes if the single-view predictions are inaccurate. To address this issue, we introduce a 2D Cross-view Consensus mechanism. By measuring the consensus across the other two orthogonal views, we dynamically assess the slice-level reliability of the current view, enabling an adaptive weighted fusion of the 2D slice-wise predictions from all three planes.

Finally, while this fusion strategy successfully generates dense pseudo-labels, it still stems from slice-wise aggregation of 2D predictions. This slice-wise aggregation neglects the inherent 3D spatial continuity of voxels, which can lead to local fragmentation near object boundaries. In the 3D stage, to bridge this dimensional gap, we introduce a dynamic soft pseudo-label refinement module formulated within a 3D Teacher-Student architecture. This module adaptively

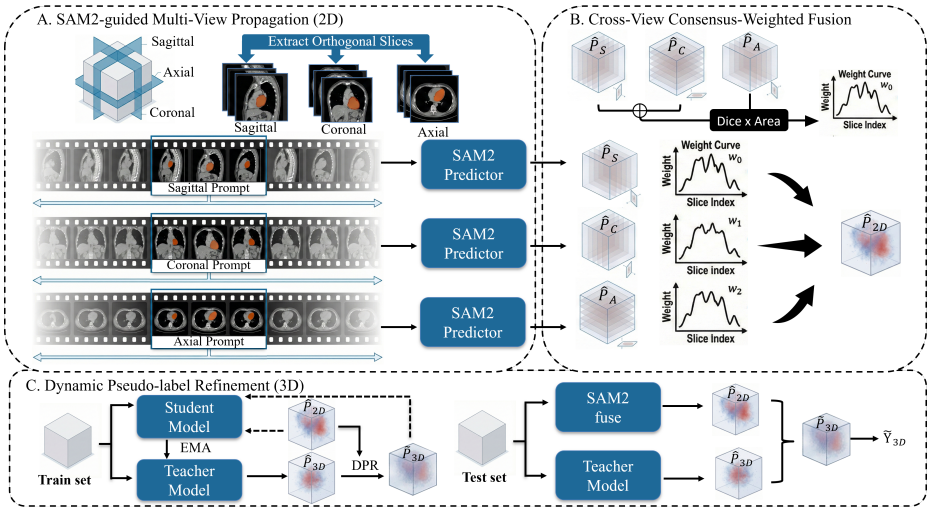

Fig. 1. Overview of our proposed SAM2-guided multi-view propagation and dynamic soft pseudo-label refinement framework.

balances 2D priors with 3D predictions, thereby progressively restoring structural continuity.

In summary, our main contributions are as follows:

- To the best of our knowledge, we present the first application of SAM2 in sparse-supervised 3D medical segmentation. By introducing a propagation mechanism driven by cross-case morphological anchors, we effectively generate dense pseudo-labels from minimal planar annotations.
- We design a fusion strategy based on consensus weights to adaptively combine predictions at the slice level, effectively reducing propagation drift and mitigating single-view bias among different views.
- We introduce a dynamic soft pseudo-label refinement stage within a Teacher-Student framework, adaptively balancing 2D priors and 3D context to enhance topological coherence.

2 Method

2.1 SAM2-guided Multi-View Propagation

In sparsely annotated 3D medical image segmentation, we have a small labeled set $\mathcal{D}_{\text{labeled}} = \{(X_i, \{Y_i^a, Y_i^c, Y_i^s\})\}_{i=1}^{N}$ and a large unlabeled set $\mathcal{D}_{\text{unlabeled}} = \{X_j\}_{j=1}^{M}$ ($M \gg N$). For the labeled data, supervision signals are strictly limited to the 2D masks of the three central orthogonal slices (axial, coronal, and sagittal).

To reconstruct continuous 3D volumes from such discrete supervision, we propose a SAM2-guided multi-view propagation module. For any view $v \in \{a, c, s\}$, we concatenate the central slices and corresponding masks of all labeled samples in that view to construct a global topological anchor sequence $\mathcal{S}^v = \{(X_i^v, Y_i^v)\}_{i=1}^N$. This cross-subject concatenation breaks the limitation of a single sample, greatly enriching the morphological priors within SAM2's memory bank.

When generating pseudo-labels, we unroll the target volume along view v and embed $\mathcal{S}^v$ at the center of the sequence as the starting point to execute bidirectional propagation. Relying on SAM2's powerful inter-frame memory mechanism, the model automatically tracks topological evolution across adjacent slices, ultimately outputting voxel-wise probability predictions $\hat{P}^a, \hat{P}^c, \hat{P}^s$ and initial pseudo-labels $\hat{Y}^a, \hat{Y}^c, \hat{Y}^s$ across the three orthogonal dimensions.

2.2 Cross-View Consensus-Weighted Fusion

Although SAM2-based single-view propagation can effectively capture intra-dimension anatomical continuity, as the propagation gradually moves away from the topological anchors, the model inevitably accumulates errors and perception biases (e.g., boundary over-segmentation or overconfidence in distal slices). To this end, we propose a slice-level consistency-weighted fusion strategy to dynamically quantify the reliability of each slice.

First, we construct a continuous reference consensus through the predictions of the other two orthogonal views to evaluate the reliability of a specific view. Taking the axial plane (a) as an example, its consensus probability map $\tilde{P}^{c,s}$ is jointly generated by averaging the predicted probability volumes of the coronal (c) and sagittal (s) planes:

$$\tilde{P}^{c,s} = \frac{\hat{P}^c + \hat{P}^s}{2}. \tag{1}$$

Here, we discard the traditional hard thresholding to preserve fine-grained spatial uncertainty and probability distribution information during the consensus construction phase.

Subsequently, we quantify the weights at the fine-grained 2D slice level. To suppress tiny noise patches that might exhibit deceptively high Dice scores despite low prediction probabilities or small areas, we introduce the intersection activation area as a scaling factor. For the k-th slice in the axial plane, its dynamic weight $w^{a,k}$ is defined based on the target prediction probability map $\hat{P}^{a,k}$ and the consensus probability map $\tilde{P}^{c,s,k}$:

$$w^{a,k} = \mathrm{Dice}(\hat{P}^{a,k}, \tilde{P}^{c,s,k}) \cdot \langle \hat{P}^{a,k}, \tilde{P}^{c,s,k} \rangle, \tag{2}$$

where $\mathrm{Dice}(\cdot,\cdot)$ denotes the soft Dice computed in probability space, and $\langle \cdot,\cdot \rangle$ quantifies the effective intersection area between the two probability maps. Computed directly in the probability space, this design avoids quantization errors caused by hard threshold truncation. Furthermore, it strictly ensures that a high weight is assigned only when the single-view prediction highly aligns with the

cross-view consensus in both morphological contour and probability confidence, while possessing a sufficient valid activation area.

Finally, we expand the 1D slice weights $\{w^a, w^c, w^s\}$ along their respective axes into 3D weight matrices W^a, W^c, W^s, and perform adaptive weighted fusion on the initial probability volumes of the three views:

$$P^{2D} = \frac{W^a \odot \hat{P}^a + W^c \odot \hat{P}^c + W^s \odot \hat{P}^s}{W^a + W^c + W^s}. \tag{3}$$

We denote this fused 2D-derived volumetric probability as P^{2D}, which serves as the soft pseudo supervision in the subsequent refinement stage. A hard pseudo-label can be obtained by thresholding P^{2D} when needed.

2.3 Dynamic Pseudo-label Refinement

While the aforementioned multi-view fusion strategy successfully generates dense pseudo-labels, they still originate from slice-wise predictions and late-stage fusion, lacking native 3D receptive field constraints. Consequently, the volumetric results may exhibit local topological discontinuities and fragmented noise. To inject 3D spatial context and dynamically rectify these labels during training, we introduce a Teacher-Student (T-S) framework [18] based on **3D V-Net** [13].

The student network is denoted as f_θ, and the teacher network $f_{\bar{\theta}}$ is updated via the Exponential Moving Average (EMA) of the student's weights: $\bar{\theta} \leftarrow \alpha\bar{\theta} + (1-\alpha)\theta$. We utilize voxel-wise confidence from both the 2D fusion results and the 3D teacher predictions (denoted as P_T^{3D}) to guide the refinement. To adaptively balance the two sources, we define the confidence C as the degree of deviation from the uncertainty point (0.5):

$$C^{2D} = |2 \cdot (P^{2D} - 0.5)|, \quad C_T^{3D} = |2 \cdot (P_T^{3D} - 0.5)|. \tag{4}$$

During training, instead of performing hard thresholding, we construct dynamic soft pseudo-labels $\tilde{Y}^{3D}$ through dual-confidence weighting in the probability space. In the initial stage, the model primarily anchors on regions with $C^{2D} \geq \gamma$ (γ=0.95) to establish stable anatomical representations. As training progresses and the teacher network acquires stronger volumetric perception, the refined target is generated as:

$$\tilde{Y}^{3D} = \frac{C^{2D} \odot P^{2D} + C_T^{3D} \odot P_T^{3D}}{C^{2D} + C_T^{3D} + \epsilon}. \tag{5}$$

This formulation ensures an adaptive supervision source: when the 2D prior is ambiguous and the teacher is more stable (larger C_T^{3D}), $\tilde{Y}^{3D}$ shifts towards the teacher's output to suppress 2D artifacts. Conversely, if the teacher remains uncertain, $\tilde{Y}^{3D}$ retains more high-confidence 2D information to avoid premature interference from unreliable 3D corrections.

To synchronize the refinement intensity with the model's maturity, we introduce a progressive weight that increases with the training progress η (the ratio

of current to total iterations):

$$W_{dyn} = \eta \cdot W_{max} \cdot C_T^{3D}, \tag{6}$$

where W_{max} is a predefined upper bound. For uncertain voxels ($C^{2D} < \gamma$), W_{dyn} scales the 3D rectification: it suppresses refinement early in training (η small) and increases it as the teacher becomes confident. This yields a progressive refinement process that starts from high-confidence regions and gradually expands to the remaining voxels, producing spatially coherent 3D segmentations.

To keep training and inference consistent, we apply the same confidence-weighted fusion at test time to combine the 2D multi-view probabilities and the 3D teacher predictions for the final segmentation.

3 Experiments

Dataset. We evaluated our method on three public 3D medical segmentation benchmarks: LA2018 [22] (LGE-MRI), KiTS19 [10] (abdominal CT), and LiTS [3] (contrast-enhanced abdominal CT). We follow the standard train/test splits: 80/20 for LA2018, 190/20 for KiTS19, and 100/31 for LiTS. Under the sparse-supervision setting, each labeled volume provides only three *central orthogonal* 2D masks, with the remainder treated as unlabeled.

Implementation. The framework is built in `PyTorch`. SAM2-Small extracts tri-view probability maps via bidirectional propagation, followed by consensus-weighted fusion into soft pseudo-labels. A 3D V-Net is then trained for 6,000 iterations using a teacher–student scheme ($\alpha = 0.99$) and SGD (LR=0.01). We use a sliding-window strategy for inference and report Dice, Jaccard, 95HD, and ASD for evaluation.

Comparison with SOTA methods. We compare our method with representative semi-supervised 3D medical segmentation baselines, including MT [18], UA-MT [24], CPS [7], BCP [2], MF-ConS [20], Desco [4], and SGTC [23]. Following the protocols of Desco [4] and SGTC [23], we evaluate all methods under the identical sparse-supervision setting with $N_L = 5$ labeled volumes.

As shown in Tab. 1, our approach achieves the best performance across all three datasets. On LA2018, our method attains a Dice score of 87.6%, outperforming the runner-up SGTC by 6.7 points, while reducing the HD from 25.3 to 10.4 (59% relative decrease). On the KiTS19 and LiTS datasets, our method demonstrates consistent stability, achieving Dice scores of 93.8% and 94.1%, respectively, along with the lowest HD (3.1 and 5.2, respectively).

Furthermore, visual results on the LiTS dataset (Fig. 2) illustrate the differences among methods in handling complex boundaries. Compared to the baselines, our method generates smoother prediction contours, effectively reduces inter-slice discontinuities and isolated artifacts, and maintains a strong visual agreement with the ground truth.

Ablation Study. We assess the contribution of each core module on the LA2018 dataset, with results summarized in Table 2. Lacking cross-dimensional constraints and suffering from error drift, single-view propagation (ID 1–3) delivers

Table 1. Comparison with state-of-the-art methods under sparse supervision.

Method	N_L	LA		KiTS19		LiTS	
		Dice ↑	HD ↓	Dice ↑	HD ↓	Dice ↑	HD ↓
MT [18]	5	63.2 ± 13.8	43.7 ± 10.3	77.2 ± 13.3	56.3 ± 24.9	75.8 ± 13.5	59.1 ± 26.9
UA-MT [24]	5	60.4 ± 11.1	46.3 ± 9.4	70.4 ± 25.8	28.6 ± 18.4	67.3 ± 24.8	31.5 ± 19.2
CPS [7]	5	64.3 ± 9.0	42.1 ± 7.6	76.9 ± 11.1	57.4 ± 14.2	76.8 ± 9.9	59.0 ± 13.1
BCP [2]	5	66.1 ± 13.3	40.9 ± 9.7	82.8 ± 9.4	27.1 ± 20.5	82.2 ± 9.1	30.6 ± 21.0
MF-ConS [20]	5	75.3 ± 10.5	25.8 ± 9.1	86.2 ± 8.4	15.9 ± 18.1	86.2 ± 7.7	16.5 ± 18.7
Desco [4]	5	77.4 ± 5.8	24.9 ± 10.6	87.0 ± 2.8	11.6 ± 2.2	89.3 ± 1.4	10.1 ± 2.4
SGTC [23]	5	80.9 ± 6.3	25.3 ± 16.3	89.5 ± 6.8	8.1 ± 7.3	90.7 ± 6.4	12.6 ± 14.0
Ours	**5**	**87.6 ± 2.9**	**10.4 ± 3.2**	**93.8 ± 3.8**	**3.1 ± 4.3**	**94.1 ± 1.7**	**5.2 ± 6.4**

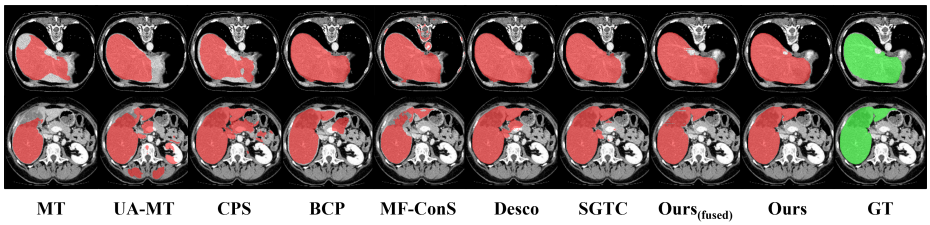

Fig. 2. Visual comparison on the LiTS dataset under sparse supervision.

limited performance with severe boundary errors. However, integrating these discrete views via consensus-weighted fusion (ID 4) triggers a consistent performance surge, elevating the Dice score to 84.82% and drastically cutting the HD to 15.96 voxels. This validates that exploiting cross-view structural complementarity successfully neutralizes view-specific perception biases.

Moreover, the comparison between ID 5 and ID 6 underscores the critical role of our Dynamic Pseudo-labeling Refinement (DPR). Although a standard 3D Teacher-Student module (ID 5) captures volumetric context, its reliance on static hard pseudo-labels renders it vulnerable to 2D propagation artifacts. Conversely, the complete framework equipped with DPR (ID 6) yields the optimal performance (87.63% Dice, 10.42 HD). This proves that our dual-confidence weighting dynamically harmonizes 2D priors with 3D spatial contexts, effectively suppressing initial label noise and restoring intrinsic anatomical continuity.

Qualitative Ablation Analysis. Visual examples of the ablation study are presented in Fig. 3(a). As observed from left to right, single-view propagations (Prop-S, Prop-C, and Prop-A) typically yield fragmented masks with severe boundary inaccuracies. The cross-view Fusion module successfully integrates these discrete views, substantially improving structural completeness. Finally, compared to the ground truth shown in the last column, our proposed dynamic pseudo-label refinement strategy effectively refines boundary details and suppresses artifacts, resulting in highly accurate and anatomically plausible segmentations.

Table 2. Ablation study of our proposed components on the LA dataset.

ID	Prop-S	Prop-C	Prop-A	Fusion	MT	DPR	Dice (%) ↑	Jac. (%) ↑	HD (vx) ↓	ASD (vx) ↓
1	✓						76.66	62.70	31.69	9.78
2		✓					76.75	62.58	23.08	6.26
3			✓				66.11	51.15	37.33	12.19
4	✓	✓	✓	✓			84.82	73.01	15.96	4.20
5	✓	✓	✓	✓	✓		85.90	75.32	13.54	3.21
6	✓	✓	✓	✓	✓	✓	**87.63**	**77.96**	**10.42**	**2.98**

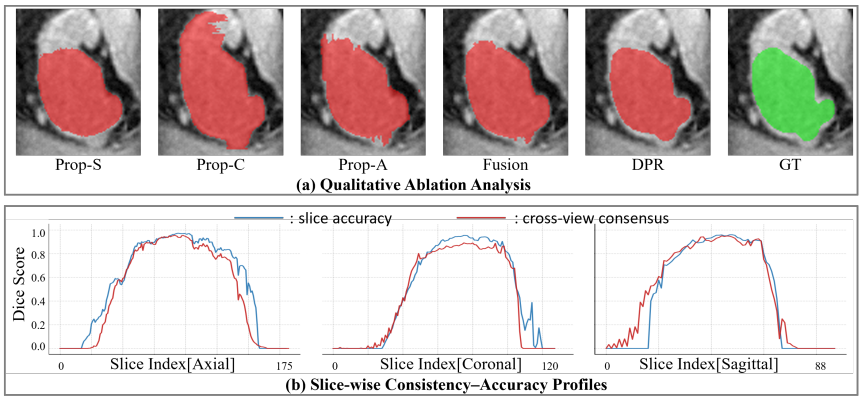

Fig. 3. (a) Qualitative ablation visualization; (b) Slice-wise curves of accuracy and cross-view consensus in the axial/coronal/sagittal planes.

Correlation Analysis of Cross-view Consistency. To validate our cross-view consensus, we calculated the Spearman correlation (ρ) between actual slice-wise accuracy and cross-view consensus. As visualized for a single representative case in Fig. 3(b), these two metrics exhibit highly synchronized trends. Quantitatively, the strong positive correlations evaluated across all cases (Axial $\rho = 0.805$, Coronal $\rho = 0.850$, Sagittal $\rho = 0.818$) confirm that consensus reliably reflects prediction quality, fully justifying our slice-weighted fusion.

4 Conclusion

In this paper, we propose a SAM2-based framework for sparsely supervised 3D medical segmentation, where only the labeled volumes are annotated with three orthogonal central slices. By integrating global morphological anchors and a cross-view consensus mechanism, we effectively propagate these sparse 2D annotations into dense volumetric supervision while mitigating drift. The framework is further enhanced by a dynamic 3D refinement stage using a confidence-weighted Teacher-Student architecture to capture spatial continuity. Extensive evaluations across the LA2018, KiTS19, and LiTS datasets demonstrate that our framework achieves consistent improvements.

Acknowledgements. This work was supported by the National Natural Science Foundation of China (Grant No. 62471182), Science and Technology Commission of Shanghai Municipality Basic Research Program (Grant No. 25JD1401300), Shanghai Rising-Star Program (Grant No. 24QA2702100), and the Science and Technology Commission of Shanghai Municipality (Grant No. 22DZ2229004).

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

References

1. Assefa, M., Naseer, M., Ganapathi, I.I., Ali, S.S., Seghier, M.L., Werghi, N.: Dycon: Dynamic uncertainty-aware consistency and contrastive learning for semi-supervised medical image segmentation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 30850–30860 (2025)
2. Bai, Y., Chen, D., Li, Q., Shen, W., Wang, Y.: Bidirectional copy-paste for semi-supervised medical image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11514–11524 (2023)
3. 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)
4. Cai, H., Li, S., Qi, L., Yu, Q., Shi, Y., Gao, Y.: Orthogonal annotation benefits barely-supervised medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3302–3311 (2023)
5. Cai, H., Qi, L., Yu, Q., Shi, Y., Gao, Y.: 3d medical image segmentation with sparse annotation via cross-teaching between 3d and 2d networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 614–624. Springer (2023)
6. Chen, D., Bai, Y., Shen, W., Li, Q., Yu, L., Wang, Y.: Magicnet: Semi-supervised multi-organ segmentation via magic-cube partition and recovery. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 23869–23878 (2023)
7. Chen, X., Yuan, Y., Zeng, G., Wang, J.: Semi-supervised semantic segmentation with cross pseudo supervision. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2613–2622 (2021)
8. Fan, D.P., Zhou, T., Ji, G.P., Zhou, Y., Chen, G., Fu, H., Shen, J., Shao, L.: Inf-net: Automatic covid-19 lung infection segmentation from ct images. IEEE transactions on medical imaging **39**(8), 2626–2637 (2020)
9. Fu, Y., Lei, Y., Wang, T., Curran, W.J., Liu, T., Yang, X.: A review of deep learning based methods for medical image multi-organ segmentation. Physica Medica **85**, 107–122 (2021)
10. Heller, N., Sathianathen, N., Kalapara, A., Walczak, E., Moore, K., Kaluzniak, H., Rosenberg, J., Blake, P., Rengel, Z., Oestreich, M., et al.: The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes. arXiv preprint arXiv:1904.00445 (2019)
11. Hu, M., Yin, J., Ma, Z., Ma, J., Zhu, F., Wu, B., Wen, Y., Wu, M., Hu, C., Hu, B., et al.: beta-fft: Nonlinear interpolation and differentiated training strategies for semi-supervised medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 30839–30849 (2025)

12. Konwer, A., Yang, Z., Bas, E., Xiao, C., Prasanna, P., Bhatia, P., Kass-Hout, T.: Enhancing sam with efficient prompting and preference optimization for semi-supervised medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20990–21000 (2025)
13. 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)
14. Qi, W., Wu, J., Chan, S.C.: Gradient-aware for class-imbalanced semi-supervised medical image segmentation. In: European Conference on Computer Vision. pp. 473–490. Springer (2024)
15. Qiu, M., Zhang, J., Qi, L., Yu, Q., Shi, Y., Gao, Y.: The devil is in the statistics: Mitigating and exploiting statistics difference for generalizable semi-supervised medical image segmentation. In: European Conference on Computer Vision. pp. 74–91. Springer (2024)
16. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K.V., Carion, N., Wu, C.Y., Girshick, R., Dollár, P., Feichtenhofer, C.: Sam 2: Segment anything in images and videos. In: International Conference on Learning Representations (ICLR) (2025), oral
17. Su, H., Liu, Z., Yao, L., Li, S., Lin, H., Chen, G., Chen, X., Lei, H., Lei, B.: Sparsely annotated medical image segmentation via cross-sam of 3d and 2d networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 530–540. Springer (2025)
18. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in neural information processing systems **30** (2017)
19. Wang, Y., Zhou, Y., Shen, W., Park, S., Fishman, E.K., Yuille, A.L.: Abdominal multi-organ segmentation with organ-attention networks and statistical fusion. Medical image analysis **55**, 88–102 (2019)
20. Wu, X., Xu, Z., Tong, R.K.y.: Few slices suffice: Multi-faceted consistency learning with active cross-annotation for barely-supervised 3d medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 286–296. Springer (2024)
21. Wu, Y., Ge, Z., Zhang, D., Xu, M., Zhang, L., Xia, Y., Cai, J.: Mutual consistency learning for semi-supervised medical image segmentation. Medical image analysis **81**, 102530 (2022)
22. Xiong, Z., Xia, Q., Hu, Z., Huang, N., Bian, C., Zheng, Y., Vesal, S., Ravikumar, N., Maier, A., Yang, X., et al.: A global benchmark of algorithms for segmenting the left atrium from late gadolinium-enhanced cardiac magnetic resonance imaging. Medical image analysis **67**, 101832 (2021)
23. Yan, K., Cai, Q., Zhang, F., Cao, Z., Liu, Z.: Sgtc: Semantic-guided triplet co-training for sparsely annotated semi-supervised medical image segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 9112–9120 (2025)
24. Yu, L., Wang, S., Li, X., Fu, C.W., Heng, P.A.: Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 605–613. Springer (2019)
25. Zhao, H., Meng, H., Yang, D., Xie, X., Wu, X., Li, Q., Niu, J.: Guidednet: semi-supervised multi-organ segmentation via labeled data guide unlabeled data. In:

Proceedings of the 32nd ACM international conference on multimedia. pp. 886–895 (2024)
26. Zhou, Y., Li, Z., Bai, S., Wang, C., Chen, X., Han, M., Fishman, E., Yuille, A.L.: Prior-aware neural network for partially-supervised multi-organ segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10672–10681 (2019)
27. Zhou, Y., Wang, Y., Tang, P., Bai, S., Shen, W., Fishman, E., Yuille, A.: Semi-supervised 3d abdominal multi-organ segmentation via deep multi-planar co-training. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 121–140. IEEE (2019)