ECT-3DMedSAM: Efficient Cross Teaching Using Segment Anything Model for Semi-Supervised 3D Medical Image Segmentation

Zhewen Huang[1] zhuang5@gradcenter.cuny.edu
Sara R. Guariglia[2] Sararose.X.Guariglia@opwdd.ny.gov
Jiaqi Yang[1] jyang2@gradcenter.cuny.edu
Chia-Ling Tsai[3] chialing.tsai@qc.cuny.edu

[1] *Computer Science Department, The Graduate Center of the City University of NY, New York, NY*

[2] *Environmental Health & Neurodevelopment Laboratory, New York State Institute for Basic Research in Developmental Disabilities, Staten Island, NY*

[3] *Computer Science Department, Queens College of the City University of NY, Flushing, NY*

Editors: Under Review for MIDL 2026

Abstract

The advent of foundation models has established new benchmarks in volumetric medical image segmentation by leveraging large-scale pre-training and temporal context. However, effectively adapting these data-hungry models to downstream tasks with limited annotations remains a critical bottleneck. Standard Semi-Supervised Medical Image Segmentation (SSMIS) approaches typically rely on conventional CNNs, which lack the semantic generalization capabilities required for complex 3D anatomical structures. In this paper, we propose a novel cross-teaching framework tailored for the efficient adaptation of the 3D foundation model (MedSAM-2). We introduce a parameter-efficient design that shares frozen image and prompt encoders between two parallel, Low-Rank Adaptation (LoRA) learnable mask decoders. Furthermore, we replace the memory-intensive attention mechanism with a simplified temporal propagation module for reducing the memory consumption while maintaining critical local volumetric coherence. Our model processes the same input volume through weakly and strongly augmentations to create a synergistic learning loop where the two decoders mutually supervise each other. We validate our method across three distinct datasets and modalities. Experimental results demonstrate that our framework effectively bridges the domain gap across different modalities and improve segmentation accuracy, especially in boundaries precision compared to existing baselines.

Keywords: Semi-supervised learning, 3D image segmentation, Transformer, Segment Anything Model, Multi-modalities

1. Introduction

Medical image segmentation is a critical step in various clinical applications, ranging from anatomical structure analysis to disease diagnosis and treatment planning. However, training robust deep learning models typically requires large-scale datasets with precise pixel-level annotations, which are expensive and time-consuming to obtain due to the need for expert knowledge (Ma et al., 2024). To mitigate the burden of manual annotation, Semi-Supervised Medical Image Segmentation (SSMIS) has emerged as a practical solution, enabling models to learn from limited labeled data alongside abundant unlabeled data.

Many influential SSMIS approaches rely on consistency learning, which encourages the model to produce invariant predictions under various perturbations. These methods can be broadly categorized based on the type of perturbation used. Input perturbation methods (Shu et al., 2023; Yang et al., 2022; Zou et al., 2021) encourage the model to maintain consistency across different perturbations of the same image input for the same model. Network perturbation methods (Liu et al., 2022; Chen et al., 2021) utilize and combine the strengths of different network architectures or initializations to generate diverse predictions for mutual supervision. Furthermore, hybrid frameworks (Luo et al., 2022; Yang et al., 2024; Chi et al., 2024) combine input and network perturbation, aiming at precise decision boundary in a larger variety of cases using mixed and multiple perturbations.

Recent progress in large-scale vision foundation models trained on broad and diverse data has opened new possibilities for reducing annotation costs in image analysis. The Segment Anything Model (SAM) (Kirillov et al., 2023) demonstrates remarkable zero-shot generalization capabilities. Following this success, recent works have introduced 3D-native foundation models in medical field (Wang et al., 2025; Ma et al., 2025a), which employs hierarchical vision transformers and memory attention modules to explicitly model volumetric dependencies. These pre-trained vision foundation models have demonstrated impressive segmentation performance and generalization capabilities for downstream tasks.

Despite these advancements, current semi-supervised methods face significant challenges when applied to complex 3D medical data with domain shifts. First, while hybrid frameworks combine multiple perturbations, they typically rely on conventional CNN based architectures trained from scratch. These models lack the extensive prior knowledge of foundation models, making them prone to overfitting when labeled data is extremely scarce or when facing significant domain disparities between labeled and unlabeled sets. Second, directly leveraging large pretrained foundation models like MedSAM-2 (Ma et al., 2025a) is non-trivial. The extensive prior knowledge inherent in these models often leads to overconfidence in incorrect predictions on unseen domains, leading to error accumulation that hinders the effective utilization of unlabeled data (Ma et al., 2025b). These large foundation models are also time and resource consuming to finetune.

To address these challenges, we propose a dual-stream, semi-supervised cross-teaching framework tailored for the efficient adaptation of MedSAM-2 to semi-supervised medical image segmentation. Unlike previous approaches that rely on architectural heterogeneity or computationally expensive model duplication, we introduce a parameter-efficient design that shares frozen image and prompt encoders between two parallel, learnable mask decoders.

This setup processes the same volumetric input through distinct augmentation strategies, a weakly augmented stream and a strongly augmented stream (Chen et al., 2021), while extracting unified features from the shared backbone. By decoupling the decoding process, we compel the two streams to learn robust, view-invariant representations through mutual supervision. Rather than a unilateral teacher-student dynamic, our framework enforces a bidirectional consistency constraint where the predictions of the weak stream and the strong stream are jointly optimized to minimize discrepancy. This mutual alignment effectively mitigates foundation model overconfidence and prevents error accumulation. Our main contributions are summarized as follows:

- We propose a dual-stream cross-teaching framework, ECT-3DMedSAM, tailored for adapting foundation models to semi-supervised 3D medical segmentation. This approach effectively minimizes the overconfidence often observed in foundation models and bridges the domain gap between pre-training data and downstream medical tasks.
- We introduce a parameter-efficient fine-tuning strategy that integrates Low-Rank Adaptation (LoRA) with a simplified temporal propagation mechanism. We also freeze the massive image and prompt encoders and injecting trainable LoRA layers solely into the mask decoder. This design reduces computational overhead while effectively leveraging the promptable nature of foundation models, enabling the effective training of foundation models on limited labeled data.
- We conduct comprehensive experiments on three diverse public datasets in multi-modal scenarios. The results demonstrate that our method outperforms existing state-of-the-art semi-supervised learning methods and foundation models.

2. Methodology

2.1. Problem Setting

We address the task of semi-supervised volumetric medical image segmentation, where the goal is to leverage a small set of annotated data alongside a larger set of unannotated data to train a robust segmentation model. Let $\mathcal{D}_L = \{(X_i^{lab}, Y_i)\}_{i=1}^{N_L}$ denote the labeled dataset consisting of N_L volumes and their corresponding ground truth masks, and $\mathcal{D}_U = \{X_i^{unlab}\}_{i=1}^{N_U}$ denote the unlabeled dataset with N_U volumes, where typically $N_U \gg N_L$. Each input volume $X \in \mathbb{R}^{D\times H\times W}$ represents a 3D medical scan. For any input volume X (whether from $\mathcal{D}_L$ or $\mathcal{D}_U$), we apply two distinct augmentations to generate a weak stream $X_{weak} = \mathcal{A}_{weak}(X)$ and a strong stream $X_{strong} = \mathcal{A}_{strong}(X)$. Let P as the volumetric prediction.

2.2. Overview

We propose a dual-stream semi-supervised cross-teaching framework to adapt the generalist segmentation capabilities of MedSAM-2 to specific medical domains using limited labeled data. The overall framework is illustrated in Figure 1. To balance computational efficiency with effective consistency learning, we utilize shared, frozen image and prompt encoders to extract unified volumetric features from the input. These features are then processed by two parallel, LoRA (Hu et al., 2021) learnable mask decoders: one dedicated to the weakly-augmented stream and the other to the strongly-augmented stream for learning robustness. These two decoder streams mutually supervise each other, enforcing the model to produce consistent predictions regardless of input perturbations.

2.3. Adapted MedSAM-2 Architecture

While MedSAM-2 demonstrates powerful segmentation capabilities, directly fine-tuning the entire architecture is computationally prohibitive and high data demanding when using two augmented streams with both labeled and unlabeled branches. Therefore, we introduce two critical architectural modifications to tailor it for efficient semi-supervised 3D segmentation.

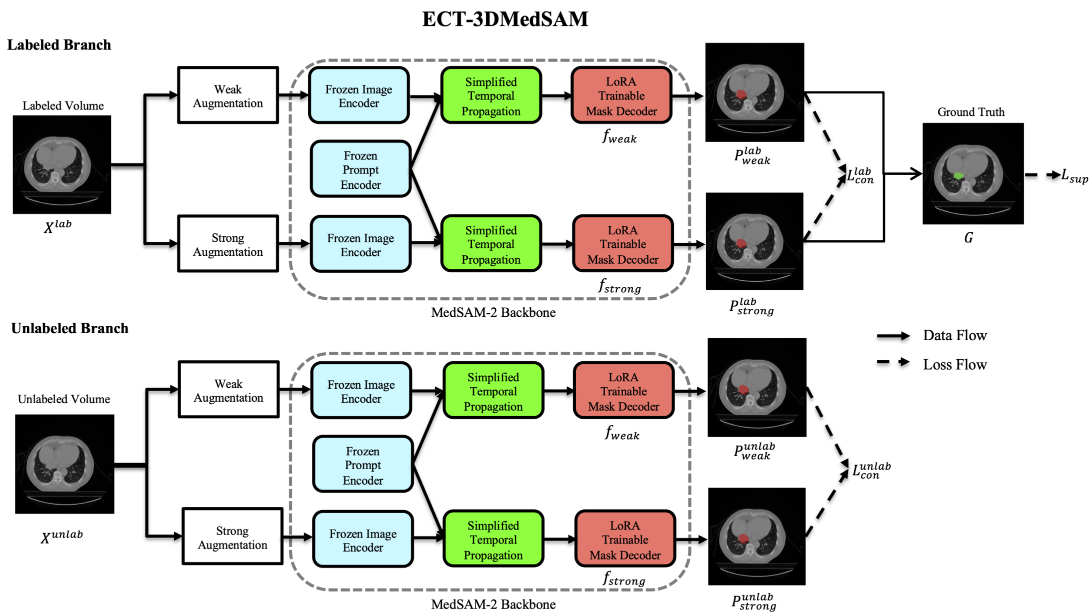

Figure 1: Overview of ECT-3DMedSAM. The architecture consists of two parallel data streams trained jointly: Labeled Branch (top) and Unlabeled Branch (bottom). Note that the weak decoder (f_{weak}) and the strong decoder (f_{strong}) are shared across both branches, which are mathematically identical. Both decoders receive gradient updates from the loss on the labeled branch and the unlabeled branch, enabling synergistic learning from both data sources.

Parameter-Efficient Fine-Tuning via LoRA. Instead of full fine-tuning, we adapt LoRA to the mask decoders to adapt the model to the target domain. We freeze the heavy image and prompt encoders to preserve the rich, pre-trained feature representations. We apply LoRA comprehensively across the mask decoder. Rather than limiting adaptation to the attention mechanism, we inject low-rank decomposition matrices into all linear and transposed convolutional layers. This ensures that the adaptation signal propagates through the entire decoding process from the transformer's semantic processing to the final spatial upscaling and mask generation. This strategy reduces the number of trainable parameters while allowing the decoder to learn task-specific segmentation boundaries.

Simplified Temporal Propagation. The original MedSAM-2 employs a sophisticated memory attention module with an 8-frame memory bank to handle long-range temporal occlusions in videos. However, in most of the tasks based on volumetric medical images, the spatial continuity between adjacent slices is actually more critical than long-range dependencies, which may introduce irrelevant noise from distant anatomical sections. To optimize computational efficiency and focus on local volumetric coherence, we replace the multi-frame memory bank with a simpler propagation mechanism. Specifically, the mask prediction M_t at slice t is conditioned solely on the image embedding E_t and the prediction from the immediate previous slice M_{t-1}. This simplified propagation further reduces memory and computing resource consumption.

2.4. Consistency Learning and Loss Functions

Supervised Loss ($\mathcal{L}_{sup}$). For the labeled subset, we apply supervision to both the weakly and strongly augmented predictions. To prioritize geometric overlap in the semi-supervised setting, we employ a composite loss function with a 1:20 weighting ratio between Focal and Dice loss:

$$\mathcal{L}_{sup} = \sum_{s\in\{weak,strong\}} (1.0 \cdot \mathcal{L}_{Focal}(P_s^{lab}, G) + 20.0 \cdot \mathcal{L}_{Dice}(P_s^{lab}, G)) \tag{1}$$

where P_s represents the prediction from stream s, and G is the ground truth.

Consistency Losses ($\mathcal{L}_{con}$). We enforce consistency between the weak and strong streams for both labeled and unlabeled data. This bidirectional constraint ensures that the model learns robust features invariant to the intensity of augmentation. We utilize Mean Squared Error (MSE) to minimize the discrepancy:

$$\mathcal{L}_{con}^{lab} = ||P_{weak}^{lab} - P_{strong}^{lab}||_2^2, \quad \mathcal{L}_{con}^{unlab} = ||P_{weak}^{unlab} - P_{strong}^{unlab}||_2^2 \tag{2}$$

Total Loss. The final objective function integrates these components, with both consistency terms weighted by a time-dependent ramp-up factor $\lambda(t)$ to ensure stable convergence:

$$\mathcal{L}_{total} = \mathcal{L}_{sup} + \lambda(t) \cdot (\mathcal{L}_{con}^{lab} + \mathcal{L}_{con}^{unlab}) \tag{3}$$

3. Experiments

3.1. Datasets and Metrics

In our experiments, we use three volumetric datasets across different modalities (CT and MRI) to evaluate performance. For the semi-supervised learning setting, we partition the training set such that only 20% of the cases are used as labeled data, while the remaining cases serve as unlabeled data.

NSCLC-Radiomics Dataset (Aerts et al., 2014) comprises CT scans from 422 patients with non-small cell lung cancer (NSCLC). In our experimental setup, we utilize the first 200 cases for training and the remaining 222 cases for test.

The PROMISE12 Dataset (Litjens et al., 2014) consists of T2-weighted MRI scans from patients with various prostatic diseases acquired at multiple locations to introduce domain variability. We use 30 cases for training and 7 cases for test. Detailed information of this dataset is provided in the appendix.

The M&Ms Dataset (Campello et al., 2021) contains Cardiac Magnetic Resonance (CMR) images acquired from four distinct scanner vendors (Siemens, Philips, GE, and Canon) across six clinical centers. The dataset provides ground truth masks strictly for the Left Ventricle (LV), Right Ventricle (RV), and Left Ventricle Myocardium (MYO) at the End-Diastolic (ED) and End-Systolic (ES) phases. We utilize a split of 256 cases for training and 64 cases for test.

The evaluation metrics include the Dice Similarity Coefficient (DSC), Jaccard index, 95% Hausdorff Distance (95HD), and Average Surface Distance (ASD).

3.2. Implementation Details

Our framework is implemented using PyTorch and trained on NVIDIA L40S GPUs. The LoRA parameters are configured with a rank of $r = 16$ and $\alpha = 32$. The model is trained for 50 epochs using the Adam optimizer with an initial learning rate of 1×10^{-4} and a weight decay of 1×10^{-4}. We employ a polynomial learning rate scheduler with a power of 0.9 to decay the learning rate during training.

The total batch size is set to 4, composed of a balanced mix of 2 labeled and 2 unlabeled volumes per iteration. The consistency loss weight λ is set to 0.1, utilizing a Gaussian ramp-up function over the beginning epochs to stabilize early training. All data augmentations are applied at the volumetric level rather than slice-by-slice. For weak data augmentations, random rotation and flipping are used. For strong data augmentations, in addition to the same geometric transfer in weak data augmentations, color jitter, blur and cutout are used. All input volumes are resized to 512×512 spatial resolution for processing. For each class, one prompt point is randomly generated from the mask in the middle frame in labeled branch. The unlabeled branch uses center point for prompts, and it shares identical encoder and decoder parameters with the labeled branch. In test, we use the mask decoder trained with weak augmentation volumes. All the baselines are finetuned using the training set for fair comparison.

3.3. Comparison with State-of-the-Art Methods

Table 1 demonstrates the performance comparison between our model, SSMIS approaches, and state-of-the-art 3D medical foundation models on the NSCLC dataset for cancer segmentation. As observed, our proposed method achieves the best DSC of 72.31%. Notably, our method shows superior boundary precision, decreasing 63% in 95HD and 64.5% in ASD compared to MedSAM-2, proving its efficacy in delineating precise margins.

Table 1: Comparison of different methods on NSCLC dataset (the best performance is marked as **bold**, and the second-best is underlined). **Avg.** means the average performance of all volumes in the dataset. **# of Params** is the number of trainable parameters.

Method	# of Params	DSC ↑	Jaccard ↑	95HD ↓	ASD ↓
		Avg.			
UNet (Ronneberger et al., 2015)	31.05M	9.39	5.22	228.93	136.14
BCP (Bai et al., 2023)	18.92M	25.40	18.38	120.86	90.77
SAM-Med3D (Wang et al., 2025)	100.51M	30.85	21.51	45.49	20.21
MedSAM-2 (Ma et al., 2025a)	38.96M	<u>71.00</u>	**59.36**	<u>17.81</u>	<u>4.93</u>
ECT-3DMedSAM (Ours)	**0.94M**	**72.31**	<u>57.55</u>	**6.59**	**1.75**

Table 2 illustrates the performance comparison on the PROMISE12 dataset for prostate segmentation across three distinct institutions. Our method shows the most balanced performance across the three institutions, achieving the highest average DSC of 74.17%. It also outperforms all comparison methods on the challenging UCL domain with a DSC of 79.66%, indicating superior generalization capability. Furthermore, our method achieves best in boundary metrics, reducing the 95HD from 37.34 to 17.68 and the ASD from 11.92 to 4.69 comparing to MedSAM-2, confirming its reliability in edges in multi-domain medical image segmentation. Figure 2 visualizes the superiority of our method, which minimize the overconfidence on unseen domains in other foundation models.

Table 2: Comparison of different methods on PROMISE12 dataset

Method	Segmentation DSC ↑			DSC ↑	Jaccard ↑	95HD ↓	ASD ↓
	BIDMC	HK	UCL	Avg.			
UNet	<u>66.71</u>	54.41	5.86	37.12	27.08	113.22	50.37
BCP	**72.58**	39.45	28.87	44.38	33.92	45.10	17.08
SAM-Med3D	26.82	45.22	26.14	31.78	20.02	98.86	34.18
MedSAM-2	45.93	**90.96**	<u>77.15</u>	<u>72.18</u>	**63.57**	<u>37.34</u>	<u>11.92</u>
ECT-3DMedSAM (Ours)	55.88	<u>84.24</u>	**79.66**	**74.17**	<u>60.68</u>	**17.68**	**4.69**

Table 3 presents the quantitative results on the M&Ms dataset, which poses significant challenges due to domain shifts across four different scanner vendors. First, we observe that our proposed model adapts MedSAM-2 to the specific domain, doubling the segmentation accuracy of MedSAM-2 to an average DSC of 30.21% and demonstrating the effectiveness of our cross-teaching framework. Second, our model achieves the best performance in distance metrics, reducing the 95HD to 18.40 and ASD to 5.73. This demonstrates that our method produces topologically coherent segmentation with precise boundaries. Finally, although our model underperforms BCP a little in overlap metrics, our method shows superior robustness on the most challenging domain, Vendor C. While UNet and BCP drop to 13.71% and 16.93% DSC, respectively on Vendor C, our method achieves 35.08%, indicating a stronger capability to generalize to unseen or difficult domains compared to models trained from scratch.

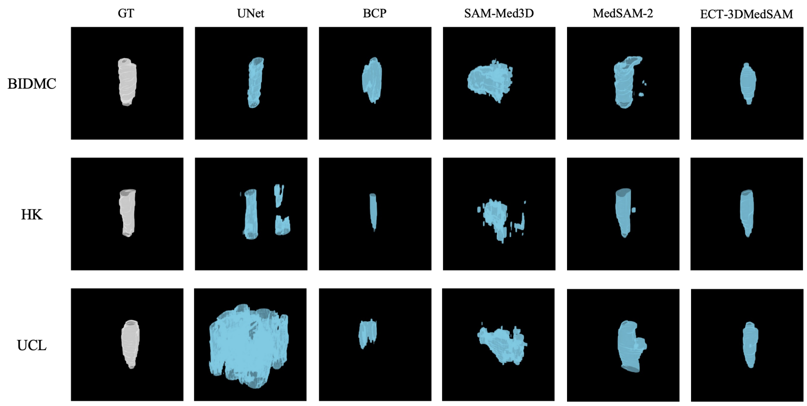

Figure 2: Qualitative comparison results on each institution in the PROMISE12 dataset

Table 3: Comparison of different methods on M&Ms dataset

Method	**(LV / MYO / RV Segmentation) DSC ↑**				**DSC ↑**	**Jaccard ↑**	**95HD ↓**	**ASD ↓**
	Vendor A	**Vendor B**	**Vendor C**	**Vendor D**	**Avg.**			
UNet	29.94 / 20.18 / 34.60	**23.66** / 11.22 / 33.58	18.07 / 8.60 / 14.46	56.19 / **46.00** / 57.93	29.66	21.53	91.59	82.74
BCP	**37.93** / **39.44** / 55.59	18.72 / **40.34** / 55.96	14.93 / 16.16 / 19.71	25.33 / 28.33 / 19.16	**34.37**	**24.47**	88.57	69.54
SAM-Med3D	7.26 / 3.42 / 7.40	1.58 / 1.64 / 3.62	8.74 / 1.16 / 12.61	5.94 / 3.60 / 8.04	4.77	2.69	76.70	30.30
MedSAM-2	3.18 / 19.41 / 0.16	17.34 / 16.34 / 3.90	22.37 / **30.53** / 12.25	**73.27** / 17.64 / 3.41	15.45	12.54	44.82	17.37
ECT-3DMedSAM (Ours)	9.63 / 4.84 / **74.28**	19.12 / 6.26 / **64.20**	**28.40** / 11.33 / **65.51**	14.49 / 5.34 / **62.49**	30.21	21.81	**18.40**	**5.73**

3.4. Ablation studies

We conduct ablation studies in order to verify the effectiveness of each module in our method. All experiments are conducted on the PROMISE12 dataset.

Table 4 shows the effectiveness of each module in our model. The baseline MedSAM-2 baseline suffers from significant boundary errors due to overconfidence. Incorporating the dual-stream mechanism substantially refines these boundaries, halving the 95HD from 41.40 to 20.21. Furthermore, the addition of the unlabeled data stream proves critical for generalization, increasing the DSC to 73.96%. Exchanging sophisticated memory attention based propagation approach with simplified temporal propagation achieves the best results with a peak DSC of 74.17% and the lowest ASD of 4.69.

Table 4: Ablation experiments on the PROMISE12 dataset. **Base** is MedSAM-2 with frozen image and prompt encoders and the LoRA trainable decoder. **DS** is adding dual-stream framework. **CT** is adding the unlabeled branch for cross-teaching. **SP** is using the simplified temporal propagation.

Base	DS	CT	SP	DSC↑	Jaccard↑	95HD↓	ASD↓
✓				64.97	54.81	41.40	13.09
✓	✓			71.57	58.58	20.21	5.05
✓	✓	✓		73.96	60.50	18.89	5.22
✓	✓	✓	✓	**74.17**	**60.68**	**17.48**	**4.69**

4. Conclusion

In this paper, we proposed a dual-stream MedSAM-2 based dual-stream cross-teaching framework to address the challenges of semi-supervised 3D medical image segmentation under limited supervision. By synergizing the generalist capabilities of the MedSAM-2 foundation model with a consistency-based cross-teaching paradigm, our approach effectively mitigates the overconfidence inherent in large-scale foundation models while preventing overfitting to the limited source domain.

Unlike methods that rely on computationally expensive full fine-tuning or heavy memory banks, we also introduced a parameter-efficient architecture sharing frozen image and prompt encoders between two parallel, LoRA-adapted mask decoders. Furthermore, by replacing the standard memory attention module with a streamlined simple propagation, we reduced the memory consumption while maintaining the volumetric coherence essential for medical scans. These architectural optimizations enable the effective training of foundation models on limited labeled data, overcoming the resource constraints that typically hinder their adaptation to downstream medical tasks.

Our extensive experiments across diverse modalities demonstrate that our method consistently outperforms both conventional SSMIS architectures and 3D medical foundation models, especially in boundary metrics. These results confirm that identifying robust, consistency-invariant features via dual-stream cross-teaching is a powerful strategy for adapting foundation models to complex, multi-modalities medical semi-supervised segmentation tasks.

Acknowledgments

This work was supported by PSC-CUNY Award 65406-00 53 and New York State Institute for Basic Research (Office for People with Developmental Disabilities).

References

Hugo J. W. L. Aerts et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature Communications*, 5(1):4006, June 2014. ISSN 2041-1723. doi: 10.1038/ncomms5006. URL https://www.nature.com/articles/ncomms5006. Publisher: Nature Publishing Group.

Yunhao Bai, Duowen Chen, Qingli Li, Wei Shen, and Yan Wang. Bidirectional Copy-Paste for Semi-Supervised Medical Image Segmentation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11514–11524, Vancouver, BC, Canada, June 2023. IEEE. ISBN 979-8-3503-0129-8. doi: 10.1109/CVPR52729.2023.01108. URL https://ieeexplore.ieee.org/document/10204177/.

Víctor M. Campello et al. Multi-Centre, Multi-Vendor and Multi-Disease Cardiac Segmentation: The M&Ms Challenge. *IEEE Transactions on Medical Imaging*, 40(12):3543–3554, December 2021. ISSN 1558-254X. doi: 10.1109/TMI.2021.3090082. URL https://ieeexplore.ieee.org/document/9458279.

Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-Supervised Semantic Segmentation with Cross Pseudo Supervision. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2613–2622, Nashville, TN, USA, June 2021. IEEE. ISBN 978-1-6654-4509-2. doi: 10.1109/CVPR46437.2021.00264. URL https://ieeexplore.ieee.org/document/9577639/.

Hanyang Chi, Jian Pang, Bingfeng Zhang, and Weifeng Liu. Adaptive Bidirectional Displacement for Semi-Supervised Medical Image Segmentation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4070–4080, Seattle, WA, USA, June 2024. IEEE. ISBN 979-8-3503-5300-6. doi: 10.1109/CVPR52733.2024.00390. URL https://ieeexplore.ieee.org/document/10658193/.

Edward J. Hu et al. LoRA: Low-Rank Adaptation of Large Language Models, October 2021. URL http://arxiv.org/abs/2106.09685. arXiv:2106.09685 [cs].

Alexander Kirillov et al. Segment Anything. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3992–4003, Paris, France, October 2023. IEEE. ISBN 979-8-3503-0718-4. doi: 10.1109/ICCV51070.2023.00371. URL https://ieeexplore.ieee.org/document/10378323/.

Geert Litjens et al. Evaluation of prostate segmentation algorithms for MRI: The PROMISE12 challenge. *Medical Image Analysis*, 18(2):359–373, February 2014. ISSN 1361-8415. doi: 10.1016/j.media.2013.12.002. URL https://www.sciencedirect.com/science/article/pii/S1361841513001734.

Yuyuan Liu et al. Perturbed and Strict Mean Teachers for Semi-supervised Semantic Segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4248–4257, New Orleans, LA, USA, June 2022. IEEE. ISBN 978-1-6654-6946-3. doi: 10.1109/CVPR52688.2022.00422. URL https://ieeexplore.ieee.org/document/9879201/.

Xiangde Luo, Minhao Hu, Tao Song, Guotai Wang, and Shaoting Zhang. Semi-Supervised Medical Image Segmentation via Cross Teaching between CNN and Transformer. In *Proceedings of The 5th International Conference on Medical Imaging with Deep Learning*, pages 820–833. PMLR, December 2022. URL https://proceedings.mlr.press/v172/luo22b.html. ISSN: 2640-3498.

Jun Ma et al. Segment anything in medical images. *Nature Communications*, 15(1):654, January 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-44824-z. URL https://www.nature.com/articles/s41467-024-44824-z. Publisher: Nature Publishing Group.

Jun Ma et al. MedSAM2: Segment Anything in 3D Medical Images and Videos, April 2025a. URL http://arxiv.org/abs/2504.03600. arXiv:2504.03600 [eess].

Qinghe Ma et al. Steady Progress Beats Stagnation: Mutual Aid of Foundation and Conventional Models in Mixed Domain Semi-Supervised Medical Image Segmentation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5175–5185, June 2025b. doi: 10.1109/CVPR52734.2025.00488. URL https://ieeexplore.ieee.org/document/11094453. ISSN: 2575-7075.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24574-4. doi: 10.1007/978-3-319-24574-4_28.

Yucheng Shu, Hengbo Li, Bin Xiao, Xiuli Bi, and Weisheng Li. Cross-Mix Monitoring for Medical Image Segmentation With Limited Supervision. *Trans. Multi.*, 25:1700–1712, January 2023. ISSN 1520-9210. doi: 10.1109/TMM.2022.3154159. URL https://doi.org/10.1109/TMM.2022.3154159.

Haoyu Wang et al. SAM-Med3D: A Vision Foundation Model for General-Purpose Segmentation on Volumetric Medical Images. *IEEE Transactions on Neural Networks and Learning Systems*, 36(10):17599–17612, October 2025. ISSN 2162-2388. doi: 10.1109/TNNLS.2025.3586694. URL https://ieeexplore.ieee.org/document/11105528.

Lihe Yang, Wei Zhuo, Lei Qi, Yinghuan Shi, and Yang Gao. ST++: Make Self-trainingWork Better for Semi-supervised Semantic Segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4258–4267, New Orleans, LA, USA, June 2022. IEEE. ISBN 978-1-6654-6946-3. doi: 10.1109/CVPR52688.2022.00423. URL https://ieeexplore.ieee.org/document/9880151/.

Yuan Yang, Lin Zhang, and Lei Ren. Semi-supervised medical image segmentation via cross teaching between MobileNet and MobileViT. *Image Vision Comput.*, 150(C), October

2024. ISSN 0262-8856. doi: 10.1016/j.imavis.2024.105196. URL https://doi.org/10.1016/j.imavis.2024.105196.

Yuliang Zou et al. Pseudoseg: Designing pseudo labels for semantic segmentation. In *International Conference on Learning Representations*, 2021.

Appendix A. Detailed information of the PROMISE12 Dataset

Table 5 shows detailed information of the PROMISE12 dataset used in the experiment.

Table 5: Detailed information of the PROMISE12 Dataset

Institution	**Case number**	**Field strength (T)**	**Endorectal coil**	**Manufactor**
HK	12	1.5	Endorectal	Siemens
BIDMC	12	3	Endorectal	GE
UCL	13	1.5 and 3	No	Siemens