Steady Progress Beats Stagnation: Mutual Aid of Foundation and Conventional Models in Mixed Domain Semi-Supervised Medical Image Segmentation

Supplementary Material

1. Loss Function Formulations

We provide the specific definitions of L_{ce} and L_{dice} mentioned in Eq. (7) (Sec. 3.3):

$$L_{ce}(y, p, w) = -\frac{1}{H \times W} \sum_{i=1}^{H \times W} w_i y_i \log p_i, \quad (1)$$

$$L_{dice}(y, p, w) = 1 - \frac{2 \times \sum_{i=1}^{H \times W} w_i p_i y_i}{\sum_{i=1}^{H \times W} w_i (p_i^2 + y_i^2)}, \quad (2)$$

where y_i, p_i, and w_i is i^{th} pixel of y, p, and w, respectively.

2. Visual Results Across Multiple Datasets

We provide visual results of different methods on Fundus, M&Ms, and BUSI datasets. As shown in Figs. 1 to 3, our SynFoC achieves optimal segmentation results on both test samples from the same and different domains as labeled data, with minimal error compared to the ground truth.

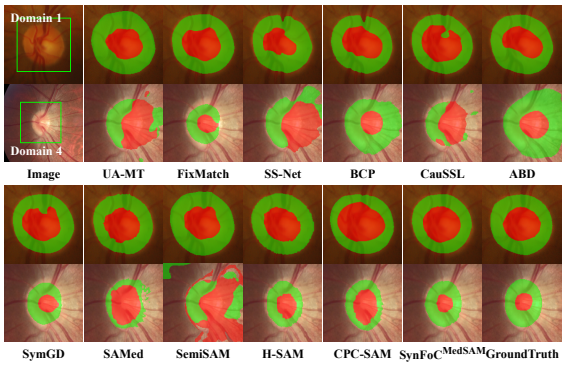


Figure 1. Visual comparison of different methods on Fundus dataset. The test samples are drawn from the labeled domain 1 and another domain 4, respectively.

3. Reproduction of SemiSAM Method

We compare all other methods by their official code implementations, whereas we reproduce SemiSAM since its public code has not been released. SemiSAM, based on the SSMIS framework, utilize predictions from conventional model to generate prompts for frozen foundation model. In turn, foundation model generates predictions based on the prompts to provide additional supervision for conventional model. To address domain shift issue, we replace the UAMT used in the original paper with the training method

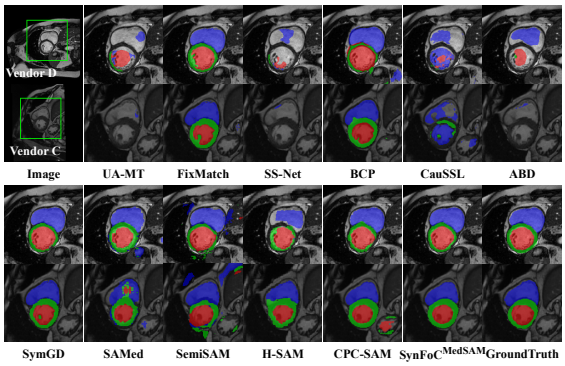


Figure 2. Visual comparison of different methods on M&Ms dataset. The test samples are drawn from the labeled Vendor D and another Vendor C, respectively.

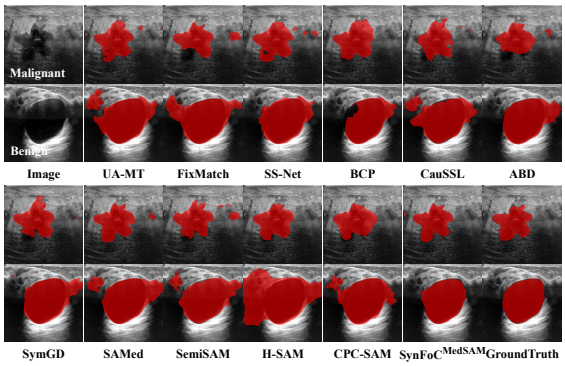


Figure 3. Visual comparison of different methods on BUSI dataset. The test samples are drawn from the labeled domain Malignant and another domain Benign, respectively.

described in Sec. 3.1. We standardize the use of MedSAM as the foundation model. Since MedSAM is fine-tuned on large-scale medical data with bounding boxes based on SAM, we provide bounding box prompts from the conventional model to the frozen MedSAM in SemiSAM.

4. Comparison with More Methods

As shown in Tab. 1, we conduct further comparisons on Prostate and Fundus datasets to demonstrate the superiority of our method. The methods include UDA approaches (SIFA [3] and UDA-VAE++ [5]), MedSAM with precise bounding box prompts, and fully fine-tuned MedSAM (instead of LoRA-based strategy). It can be observed that UDA methods struggle to achieve satisfactory performance when the number of labeled data is limited. Despite being pro-

Method	DSC ↑	Jaccard ↑	95HD ↓	ASD ↓
Prostate 20 labels				
SIFA [3]	59.33	45.40	53.90	24.29
UDA-VAE++ [5]	64.36	50.27	33.14	15.11
MedSAM$^{\text{Bounding box}}$	77.39	63.67	13.22	6.27
MedSAM$^{\text{Full Fine-tuning}}$	81.78	72.21	30.78	13.73
SynFoC	**87.16**	**79.30**	**10.26**	**4.41**
Fundus 20 labels				
SIFA [3]	67.78	54.77	20.16	10.93
UDA-VAE++ [5]	73.51	61.40	17.60	9.86
MedSAM$^{\text{Bounding box}}$	77.82	64.87	15.21	6.62
MedSAM$^{\text{Full Fine-tuning}}$	85.99	77.18	9.04	4.48
SynFoC	**88.60**	**80.50**	**6.56**	**3.47**

Table 1. Comparison of different methods on Prostate and Fundus datasets.

vided with precise bounding box prompts, MedSAM still falls short in specific datasets. Even with full fine-tuning, MedSAM struggles to correct high-confidence mispredictions, while significantly increasing training costs.

5. More Ablation Studies

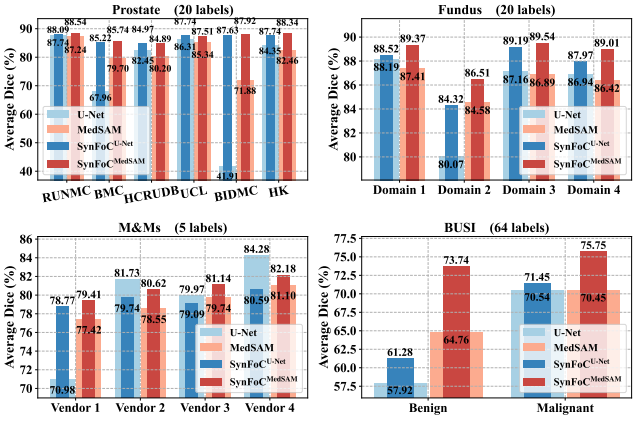
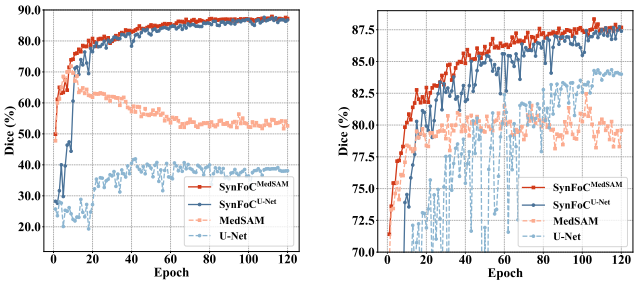


Figure 4. The performance comparison of U-Net and MedSAM under standalone training and our SynFoC across four datasets.

Comparison with Standalone U-Net and MedSAM Across Four Datasets. We present the advantages of our method over standalone U-Net and MedSAM on four datasets in Fig. 4. Unlike in Fig. 3 (Sec. 3.3), where we compare standalone U-Net and MedSAM with SMC-based synergistic training, here we compare them with our overall method, SynFoC. Our method effectively mitigates U-Net's overfitting and further advances MedSAM's performance in downstream tasks across all four datasets, demonstrating superior capabilities in handling significant domain gaps in training data.

In Fig. 5, we present the performance curves of standalone U-Net and MedSAM, as well as U-Net and MedSAM trained within our SynFoC framework, across two experiments on Prostate dataset (labeled data sourced from BIDMC and HK, respectively). Both U-Net and MedSAM demonstrate significant performance improvements when trained with SynFoC.



(a) Labeled data from BIDMC (b) Labeled data from HK

Figure 5. The experimental results on Prostate with 20 labeled data from BIDMC and HK. Each subplot displays the performance curves of individually trained MedSAM and U-Net , as well as the performance curves of MedSAM and U-Net under SynFoC.

U-Net	MedSAM	DSC ↑	Jaccard ↑	95HD ↓	ASD ↓
SGD	SGD	76.56	65.85	24.51	10.70
Adam	Adam	87.01	79.10	10.61	4.43
Adam	SGD	-	-	-	-
SGD	Adam	**87.16**	**79.30**	**10.26**	**4.41**

Table 2. Ablation study of different optimizer choices.

Different optimizer choices. We explore the impact of different optimizer choices for U-Net and MedSAM on Prostate dataset. As shown in Tab. 2, the best performance is achieved when U-Net and MedSAM are optimized with SGD and Adam, respectively.

τ	DSC ↑	Jaccard ↑	95HD ↓	ASD ↓
0.85	86.71	78.73	10.87	4.58
0.90	87.08	79.19	10.85	4.54
0.95	**87.16**	**79.30**	**10.26**	**4.41**
0.98	86.68	78.78	11.55	4.98
0.99	86.62	78.71	11.12	4.90

Table 3. Ablation study of different confidence threshold τ.

Discussion on τ. On Prostate dataset, we investigate the effect of varying the threshold τ on our method. In Tab. 3, a setting of 0.95 yields the optimal performance, and the results remain stable across other threshold values.

6. The Performance on SSMIS and UDA settings

Our SynFoC offers a general solution to address domain shifts and limited labeled data. As shown in Tabs. 5 and 6, we conduct experiments on the ACDC dataset [2] (containing 100 patients' scans) and the MSCMRSeg dataset [8]

Method	DSC ↑ RUNMC	BMC	HCRUDB	UCL	BIDMC	HK	DSC ↑ Avg.	Jaccard ↑ Avg.	95HD ↓ Avg.	ASD ↓ Avg.
Standalone$^{\text{U-Net}}$	87.74	67.96	82.45	86.31	41.91	84.35	75.12	65.76	54.67	29.08
Standalone$^{\text{SAM}}$	81.15	75.26	81.93	84.78	75.09	81.26	79.91	70.35	21.59	9.22
SynFoC$^{\text{U-Net}}$	88.05	84.42	84.11	88.00	86.84	87.31	86.46	78.84	11.66	5.03
SynFoC$^{\text{SAM}}$	**88.41**	**84.94**	**84.51**	**88.71**	**86.71**	**87.63**	**86.82**	**79.21**	**10.57**	**4.65**

Table 4. Ablation experiments on Prostate dataset.

Method	Scans used L	U	Metrics DSC ↑	ASD ↓
SupOnly	3(5%)	0	47.83	12.62
	7(%10)	0	79.41	2.70
	70(All)	0	91.44	0.99
SS-Net [7]	3(5%)	67(95%)	65.83	2.28
BCP [1]			87.59	0.67
ABD [4]			**88.96**	**0.52**
SynFoC			88.32	0.70
SS-Net [7]	7(10%)	63(90%)	86.78	1.40
BCP [1]			88.84	1.17
ABD [4]			**89.81**	**0.49**
SynFoC			89.68	1.17

Table 5. Comparison of different methods on ACDC dataset.

Method	DSC ↑ MYO	LV	RV	Avg.
Adaptation from **CT** to **MRI**				
NoAdapt	14.50	34.51	31.10	26.70
SIFA [3]	67.69	83.31	**79.04**	76.68
UDA-VAE [6]	68.42	84.41	72.59	75.14
UDA-VAE++ [5]	70.75	**88.64**	75.82	78.40
SynFoC	**71.47**	86.90	78.81	**79.06**
Adaptation from **MRI** to **CT**				
NoAdapt	12.32	30.24	37.25	26.60
SIFA [3]	60.89	79.32	82.39	74.20
UDA-VAE [6]	58.58	79.43	80.43	72.81
UDA-VAE++ [5]	68.74	85.08	81.42	78.41
SynFoC	**78.26**	**88.25**	**82.97**	**83.16**

Table 6. Comparison of different methods on MSCMRSeg dataset.

(containing 35 labeled CT images and 45 labeled LGE-MRI images), demonstrating its competitive performance in traditional SSMIS and UDA settings.

7. Common Challenges of Foundation Models

As shown in Tab. 4, SAM also struggles to correct high-confidence mispredictions. Due to large-scale pretraining, the issue of error accumulation is a common challenge for foundation models in downstream tasks. Our SynFoC method is not limited to the combination of U-Net and MedSAM. Through the Synergistic training of conventional and foundation models, we achieve significant performance improvements for both models.

8. Limitations and Future Works

As shown in Fig. 6, in the experiments on M&Ms dataset, by deeply analyzing the results, we found that SynFoC and most existing methods struggle with extremely small targets. Visual analysis of error cases reveals that tiny size and low boundary contrast often lead to over- or under-segmentation. Additionally, our method focuses on 2D medical image segmentation and lacks exploration in 3D medical image segmentation. Future work could enhance the precise segmentation of extremely small targets and extend the framework to 3D medical images.

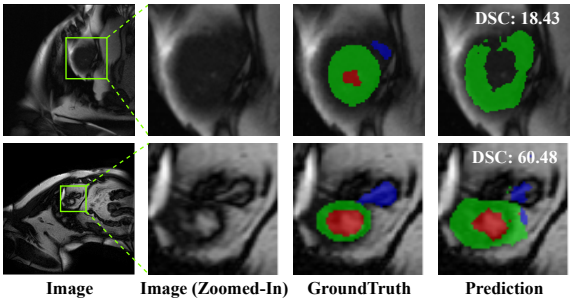


Figure 6. Visual results of error cases on M&Ms dataset.

References

[1] Yunhao Bai, Duowen Chen, Qingli Li, Wei Shen, and Yan Wang. Bidirectional copy-paste for semi-supervised medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11514–11524, 2023. 3

[2] Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE Transactions on Medical Imaging*, 37(11): 2514–2525, 2018. 2

[3] Cheng Chen, Qi Dou, Hao Chen, Jing Qin, and Pheng Ann Heng. Unsupervised bidirectional cross-modality adaptation via deeply synergistic image and feature alignment for medical image segmentation. *IEEE Transactions on Medical Imaging*, 39(7):2494–2505, 2020. 1, 2, 3

[4] Hanyang Chi, Jian Pang, Bingfeng Zhang, and Weifeng Liu. Adaptive bidirectional displacement for semi-supervised medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4070–4080, 2024. 3

[5] Changjie Lu, Shen Zheng, and Gaurav Gupta. Unsupervised domain adaptation for cardiac segmentation: Towards structure mutual information maximization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2588–2597, 2022. 1, 2, 3

[6] Fuping Wu and Xiahai Zhuang. Unsupervised domain adaptation with variational approximation for cardiac segmentation. *IEEE Transactions on Medical Imaging*, 40(12):3555–3567, 2021. 3

[7] Yicheng Wu, Zhonghua Wu, Qianyi Wu, Zongyuan Ge, and Jianfei Cai. Exploring smoothness and class-separation for semi-supervised medical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, pages 34–43. Springer, 2022. 3

[8] Xiahai Zhuang. Multivariate mixture model for myocardial segmentation combining multi-source images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(12): 2933–2946, 2018. 2