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

Qinghe Ma[1], Jian Zhang[1], Zekun Li[1], Lei Qi[2], Qian Yu[3], Yinghuan Shi[1,*]
[1]Nanjing University [2]Southeast University [3]Shandong Women's University

Abstract

*Large pretrained visual foundation models exhibit impressive general capabilities. However, the extensive prior knowledge inherent in these models can sometimes be a double-edged sword when adapting them to downstream tasks in specific domains. In the context of semi-supervised medical image segmentation with domain shift, foundation models like MedSAM tend to make overconfident predictions, some of which are incorrect. The error accumulation hinders the effective utilization of unlabeled data and limits further improvements. In this paper, we introduce a Synergistic training framework for Foundation and Conventional models (**SynFoC**) to address the issue. We observe that a conventional model trained from scratch has the ability to correct the high-confidence mispredictions of the foundation model, while the foundation model can supervise it with high-quality pseudo-labels in the early training stages. Furthermore, to enhance the collaborative training effectiveness of both models and promote reliable convergence towards optimization, the consensus-divergence consistency regularization is proposed. We demonstrate the superiority of our method across four public multi-domain datasets. In particular, our method improves the Dice score by 10.31% on the Prostate dataset. Our code is available at* `https://github.com/MQinghe/SynFoC`.

1. Introduction

Semi-supervised medical image segmentation (SSMIS) [2, 3, 14, 15, 35, 49, 55, 56] offers an effective way to tackle problems with limited annotations [17, 40]. In recent years, domain shifts in real-world clinical scenarios have garnered significant attention, where labeled and unlabeled data are drawn from different distributions due to variations in equipment parameters, patient populations, and disease severity [18, 53, 57]. When large amounts of medical data are collected, it is challenging to exam the distribution to which they belong. Consequently, many related settings have been explored [7, 30, 32, 38, 51, 61], with Mixed Domain Semi-Supervised Medical Image Segmentation (MiDSS) being a more general framework [32]. In this setting, a limited number of labeled samples are sourced from a single domain, while a substantial number of unlabeled samples originate from multiple mixed domains.

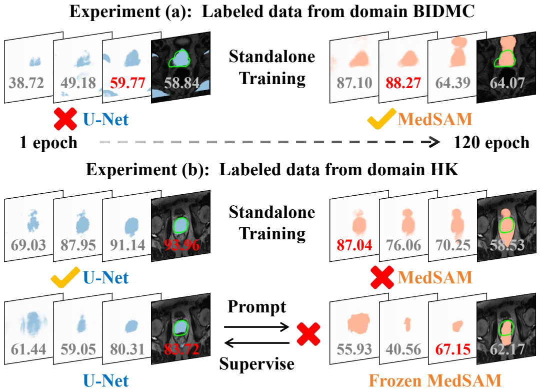

Figure 1. Illustration of pseudo-labels generation across training stages for various methods. In experiment (a) and (b) on prostate dataset, 20 labeled data come from BIDMC and HK, respectively. We report the Dice Coefficient between each pseudo-label and the ground truth, which is represented by the green contour. Standalone training of U-Net and MedSAM, as well as the guidance from MedSAM to U-Net, fail to effectively address MiDSS.

For SSMIS, existing methods usually train a conventional model (*e.g*., U-Net [37]) using labeled and unlabeled data from scratch. However, the domain shift renders the training of conventional models particularly vulnerable to noisy pseudo-labels [8, 32]. Recently, large pretrained visual foundation models [24, 31, 44, 63] have demonstrated impressive segmentation performance and generalization capabilities for downstream tasks. Considering the limitations of conventional models, we wonder whether foundation models can serve as off-the-shelf tools for addressing these problems. In other words, can they be effectively adapted to specific domains by leveraging a small amount of labeled data alongside mixed-domain unlabeled data?

To answer the question, we conduct experiments to explore how does the single foundation model or conventional

[*] Corresponding author: Yinghuan Shi (syh@nju.edu.cn). Qinghe Ma, Jian Zhang, Zekun Li and Yinghuan Shi are with the State Key Laboratory for Novel Software Technology and National Institute of Healthcare Data Science, Nanjing University, China.

model perform in MiDSS. MedSAM [31] is served as the foundation model and U-Net as the conventional model. As illustrated in Fig. 1(a), when labeled data are drawn from the domain BIDMC, U-Net overfits to the labeled data, resulting in poor performance. In comparison, MedSAM exhibits superior segmentation capability in the early stages of training due to its inherent extensive prior knowledge. Similarly, as shown in Fig. 1(b), when labeled data comes from the domain HK, it fails to rectify high-confidence wrong predictions, hindering further performance improvement. In contrast, U-Net actively correct high-uncertainty mispredictions, achieving a higher performance ceiling. Fig. 1(a) and Fig. 1(b) show that neither the conventional nor the foundation model is universally effective. The conventional model tends to overfit to the labeled data when there are significant domain shifts between labeled and unlabeled data, while the foundation model, not limited to MedSAM struggles to correct high-confidence mispredictions due to large-scale pretraining, leading to error accumulation. Additionally, many existing studies train a conventional model guided by the pseudo-labels from frozen foundation model. As shown in Fig. 1(b), pseudo-labels from U-Net provide bounding box prompts to MedSAM, which, in turn, offers additional supervisory signals for U-Net. However, under this training scheme, the performance of the conventional model is often limited by the foundation model.

Unlike previous studies where foundation models dominate [9, 42, 59, 60], considering the complementary characteristics of both model, we believe that conventional models also play a critical role in further boosting the performance of foundation models. In this paper, we propose a mixed-domain semi-supervised medical image segmentation **Syn**ergistic training framework where **Fo**undation (*e.g.*, MedSAM) and **C**onventional (*e.g.*, U-Net) models are synergistically trained (**SynFoC**). We dynamically adjusts the dominance of each model during training: MedSAM leads in the early stages to ensure training quality of U-Net, while U-Net takes the lead in later stages to correct high-confidence errors, unlocking performance potential of MedSAM. Beyond that, to boost their representational abilities jointly, we employ consistency regularization [28, 56] to enhances information sharing. Yet, for regions with consistent predictions, encouraging higher confidence offers a more reliable optimization direction. Thus, we propose region-specific regularization to promote training effectiveness.

Our main contributions are summarized as follows:

- We identify that error accumulation from overconfident predictions in the foundation model hinders performance improvement when transferred to downstream tasks.
- We introduce the Self-Mutual Confidence evaluation module (SMC), which determines the integration ratio of the pseudo-labels from both models by evaluating both self-stability and mutual consistency.

Setting	Limited Annotations	Domain Shift	Unknown Domain Labels
SSMIS	✓	✗	-
UDA	✗	✓	✗
LE-UDA	✓	✓	✗
MiDSS	✓	✓	✓

Table 1. Various settings and challenges in clinical scenarios.

- We design the Consensus-Divergence Consistency Regularization (CDCR) to encourage both models to make high-confidence and reliable predictions, while aligning their representation capabilities.
- Extensive experiments are conducted on four public multi-domain datasets, demonstrating that our method outperforms other state-of-the-art approaches[1]. With only 20 labeled data from Prostate dataset, our method obtains an improvement of over 10% Dice than other methods.

2. Related Work

Medical Image Segmentation with Limited Annotation. Existing semi-supervised medical image segmentation (SSMIS) methods can be categorized into pseudo-label [26, 36, 52] and consistency regularization based methods [5, 11, 19, 23, 47]. Pseudo-label based methods generate pseudo-labels for unlabeled data to update the network iteratively. Wang *et al*. [46] propose to evaluate the performance of different networks to select more reliable pseudo-labels dynamically. Consistency regularization based methods aim to generate consistent predictions for unlabeled data under perturbations of the input, feature, or network. Chen *et al*. [11] encourage models with different initializations to produce the same predictions. However, SSMIS methods typically follow the assumption that labeled and unlabeled data are sampled from the same distribution [50, 54], which can be impractical in real-world applications.

Medical Image Segmentation with Domain Shift. Domain shift is frequently encountered in real-world clinical scenarios due to differences in equipment parameters, patient populations, and disease severity. Many related problem settings, such as unsupervised domain adaptation (UDA) [7, 38, 51], label-efficient UDA (LE-UDA) [30, 61], and MiDSS (Mixed Domain SSMIS) [32], have been proposed. As shown in Tab. 1, MiDSS represents a more general scenario that confronts challenges from limited annotations, domain shifts, and unknown domain labels, where labeled data are limited and come from a single domain, while unlabeled data are mixed from multiple domains. Other scenarios, such as SSMIS, UDA, and LE-UDA [30, 61], can be regarded as specific cases of this broader challenge. Ma *et al*. [32] propose to generate symmetric intermediate samples to fully utilize the intermediate domain information. However, conventional models often struggle to effectively transfer domain knowledge and are prone to over-

[1] Our method also achieves competitive performance on SSMIS and UDA setting, as shown in the supplementary materials.

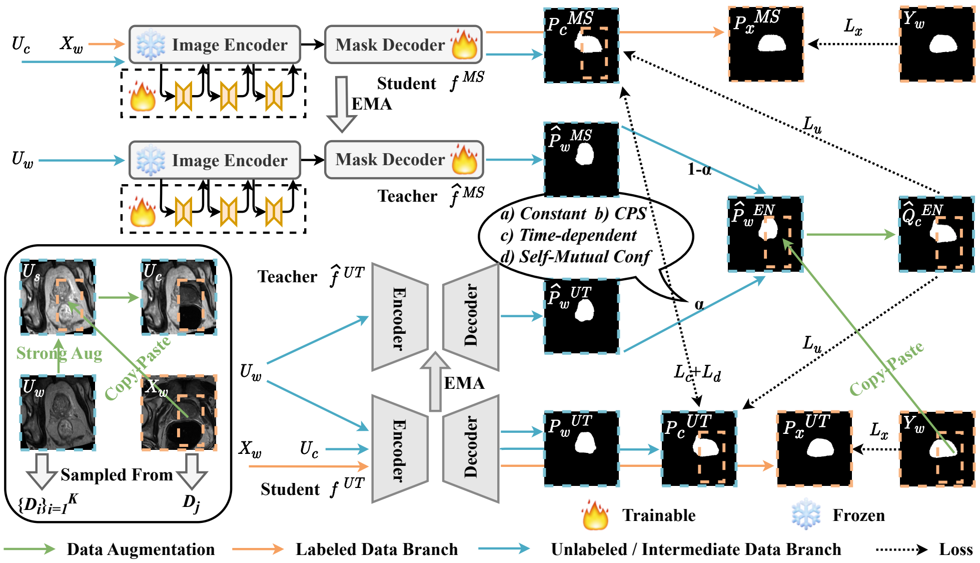

Figure 2. The overall framework of our SynFoC. For U-Net and MedSAM, the teacher model generates pseudo-labels for intermediate samples to guiding the student model. To reduce computational costs, we applies the LoRA module to MedSAM. We design various pseudo-label integration strategies to combine the predictions of both models, aiming to achieve higher-quality pseudo-labels. Additionally, we introduce consensus-divergence consistency regularization to enhance the efficiency of the synergistic training.

fitting when there is a significant domain gap between labeled and unlabeled data. To address this, we leverage the powerful feature extraction and generalization capabilities of foundation models to accelerate the early-stage training of conventional models while ensuring training quality.

Foundation Models in Medical Image Segmentation. Pre-trained on massive datasets, foundation models like the Segment Anything Model (SAM) [25] demonstrates exceptional generalization capability across various downstream tasks with prompts such as points and bounding boxes. Several efforts have been made to adapt SAM for medical images. MedSAM [31] is fine-tuned on 1.57 million image-mask pairs, and SAM-Med2D [12] on 4.6 million images and 19.7 million masks from both public and private datasets. Nonetheless, fully fine-tuning SAM is resource-intensive, and precise medical prompts generation requires expert knowledge, making the process time-consuming. To overcome these challenges, many works focus on prompt-free, efficient fine-tuning of foundation models [10, 13, 22, 27, 48, 58]. SAM-adaptor and SAMed, for instance, freeze the pre-trained image encoder and employ adapter [20] or low-rank-based [21] strategies.

Recently, many works have explored leveraging SAM for SSMIS tasks. For instance, In SemiSAM [59], the segmentation model generates prompt information for SAM, which in turn produces predictions that offer additional supervisory signals to the conventional model. Even so, static SAM fails to achieve optimal performance on specific datasets. CPC-SAM [34] automatically generate prompts and supervision across two decoder branches, enabling effective learning from both labeled and unlabeled data. These methods typically treat foundation models as the dominant, even discarding the conventional models. However, we believe that conventional models play a key role in further boosting the performance of foundation models.

3. Method

3.1. Problem Formulation and Preliminary

Let $\mathcal{L} = \{(X_i, Y_i)\}_{i=1}^N$, and $\mathcal{U} = \{U_i\}_{i=1}^M$ denote the labeled and unlabeled data sets, where N and M are their respective sizes, with $M \geq N$. $X_i, U_i \in \mathbb{R}^{W\times H\times L}$ represents the input image, and $Y_i \in \{0, 1, \dots, C\}^{W\times H}$ denotes the ground truth, where C represents the number of semantic classes, with 0 indicating the background. The training data originates from K different data centers $\mathcal{D} = \{D_i\}_{i=1}^K$, where the labeled data $\mathcal{L}$ is sampled from a single domain D_j and the unlabeled data $\mathcal{U}$ from multiple domains.

We first introduce the training paradigm in the MiDSS scenario. For any network structure, we define a teacher model $\hat{f}$ and a student model f in the Mean Teacher architecture. Weak and strong augmentations are applied to unlabeled data U to generate U_w and U_s, respectively. The teacher model predicts $\hat{P}_w$ on U_w and obtains the pseudo-label $\hat{Q}_w = \arg\max(\hat{P}_w)$. To bridge the domain gap, we generate an intermediate sample U_c by pasting part of the weakly-augmented labeled data X_w onto U_s. The pseudo-label $\hat{Q}_c$ for U_c is obtained from Y_w and $\hat{Q}_w$ similarly:

$$\begin{aligned} U_c &= X_w \odot M + U_s \odot (\mathbf{1} - M), \\ \hat{Q}_c &= Y_w \odot M + \hat{Q}_w \odot (\mathbf{1} - M), \end{aligned} \tag{1}$$

where $M \in \{0, 1\}^{W\times H}$ is a one-centered mask that indicates the region for Copy-Paste. $\mathbf{1}$ represents an all-ones

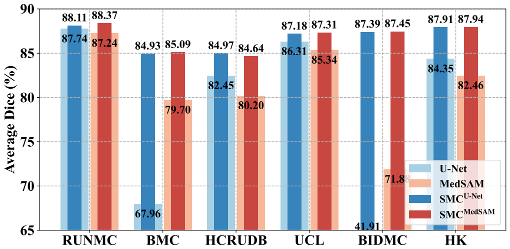

Figure 3. The performance comparison of U-Net and MedSAM under standalone and SMC-based synergistic training on Prostate.

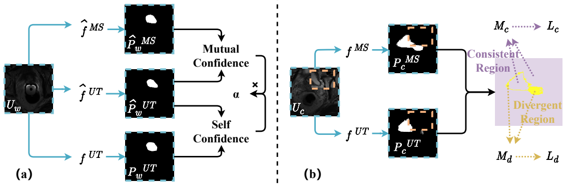

Figure 4. Illustration of self-confidence and mutual confidence evaluation and Consensus-Divergence consistency regularization.

matrix, and $\odot$ denotes the element-wise multiplication, respectively. $\hat{Q}_c$ will be used as the supervision to supervise the student model prediction P_c of U_c. $\hat{f}$ is updated by the Exponential Moving Average (EMA) of f.

3.2. Overview

The overall framework of our SynFoC is illustrated in Fig. 2. Based on the training paradigm mentioned above, we incorporate the foundation model MedSAM with the lightweight model U-Net for synergistic training. The supervised training of both models is independent, with the supervised loss L_x calculated based on the difference between the prediction P_x of X_w and Y_w. For unsupervised loss L_u, we determine the pseudo-label ensemble ratio by self and mutual confidence, providing higher-quality supervision for U_c. Finally, we apply specific consistency regularizations on the consistent and divergent regions of the predictions from the two student models, calculating L_c and L_d accordingly. The overall loss L_{total} is defined as:

$$L_{total} = L_x + \lambda(L_u + L_c + L_d), \tag{2}$$

where $\lambda(t) = e^{-5(1-t/t_{max})}$ is a time-dependent Gaussian warming-up function. t represents the current training step, and t_{max} is the maximum number of steps.

3.3. Synergistic Training of Foundation and Conventional Models

In the MiDSS scenario, the insufficient discriminative segmentation information from labeled data and the inefficiency of domain knowledge transfer lead to U-Net overfitting on the labeled data, as illustrated by the performance of BIDMC in Fig. 3. MedSAM, with its powerful feature extraction and generalization capabilities, effectively address these shortcomings. To reduce computational cost, following SAMed [58] and apply Low-rank (LoRA) to the frozen image encoder, training it together with the mask decoder. The EMA update also involves only these two modules. However, MedSAM tends to make high-confidence predictions even in the early stages of training due to its extensive prior knowledge. As demonstrated by the performance of HCRUDB in Fig. 3, the high-confidence mispredictions are difficult to rectify, hindering performance improvement.

To enhance the stability of the early training for U-Net and further improve the performance of MedSAM, we propose a synergistic training strategy, which dynamically adjusts the instance-wise ensemble ratio α for predictions generated by both models. We denote MedSAM as f^{MS} and U-Net as f^{UT}, with corresponding teacher models $\hat{f}^{MS}$ and $\hat{f}^{UT}$. The ensemble pseudo-label $\hat{P}_w^{EN}$ is defined as:

$$\hat{P}_w^{EN} = \alpha \hat{P}_w^{UT} + (1-\alpha)\hat{P}_w^{MS}, \tag{3}$$

where $\hat{P}_w^{UT}$ and $\hat{P}_w^{MS}$ represent the pseudo-labels generated by f^{UT} and f^{MS} for U_w, respectively. α can be determined in various ways. A straightforward way is to set it as a constant, such as $\alpha = 0.5$, or to adopt a CPS strategy where $\alpha = 1$ for MedSAM and $\alpha = 0$ for U-Net. Considering the complementary characteristics between both models during training, a time-dependent linear change can be applied, *i.e.*, $\alpha = t/t_{max}$. Furthermore, we incorporate instance-wise pseudo-label uncertainty and determine the ratio α by taking into account both self-confidence and mutual confidence (SMC), as observed in Fig. 4(a). According to Fig. 3, both models demonstrate significant performance improvement under SMC-based synergistic training.

Self-Confidence. It is difficult for teacher and student models to achieve consensus for hard unlabeled data [16, 45]. High-quality pseudo-labels require models to demonstrate consistent and stable predictions. We evaluate the Self-confidence Φ^{self} of the U-Net by measuring the consistency between P_w^{UT} and $\hat{P}_w^{UT}$ from the f^{UT} and $\hat{f}^{UT}$:

$$\Phi^{self} = \frac{1}{C}\sum_{i=1}^{C}\frac{2\times(|\mathbb{1}(Q_w^{UT}=i)\cap\mathbb{1}(\hat{Q}_w^{UT}=i)|)}{|\mathbb{1}(Q_w^{UT}=i)|+|\mathbb{1}(\hat{Q}_w^{UT}=i)|}, \tag{4}$$

where $\mathbb{1}(\cdot)$ represents the indicator function, $\mathbb{1}(Q_w^{UT}=i)$ is the binary mask for pixels predicted as class i, and $|A|$ denotes the number of pixels where value is 1. The intersection $|\mathbb{1}(Q_w^{UT}=i)\cap\mathbb{1}(\hat{Q}_w^{UT}=i)|$ counts the number of pixels predicted as class i by both models.

Mutual Confidence. When optimization converges, regardless of the performance, the model always exhibits high stability, making it difficult to ensure the quality of the pseudo-labels. Given robust feature extraction capabilities of MedSAM, it can quickly pinpoint segmentation target areas and generate reasonably accurate results. To further assess the reliability of the predictions generated by U-Net, we measure the consistency between $\hat{P}_w^{UT}$ and $\hat{P}_w^{MS}$ produced by U-Net and MedSAM to determine the mutual confidence Φ^{mut}, which is defined as follows:

$$\Phi^{mut} = \frac{1}{C}\sum_{i=1}^{C}\frac{2\times(|\mathbb{1}(\hat{Q}_w^{MS}=i)\cap\mathbb{1}(\hat{Q}_w^{UT}=i)|)}{|\mathbb{1}(\hat{Q}_w^{MS}=i)|+|\mathbb{1}(\hat{Q}_w^{UT}=i)|}. \tag{5}$$

We assess the reliability of the predictions generated by U-Net from above two perspectives. We consider the prediction of U-Net is reliable when both Φ^{self} and Φ^{mut} approach 1, leading to large ensemble ratio α, and vice versa:

$$\alpha = \Phi^{self} \times \Phi^{mut}. \tag{6}$$

According to Eq. (3), we obtain the ensembled probability map $\hat{P}_w^{EN}$ and the pseudo-label $\hat{Q}_w^{EN}$ of U_w. Next, we generate the intermediate sample U_c along with its pseudo-label $\hat{Q}_c^{EN}$ by Eq. (1). $\hat{Q}_c^{EN}$ guides the predictions P_c^{UT} and P_c^{MS} on U_c, with the unsupervised loss L_u defined as:

$$\begin{aligned} L_u = L_{ce}(\hat{Q}_c^{EN}, P_c^{UT}, W_c^{EN}) + L_{dice}(\hat{Q}_c^{EN}, P_c^{UT}, W_c^{EN}) + \\ L_{ce}(\hat{Q}_c^{EN}, P_c^{MS}, W_c^{EN}) + L_{dice}(\hat{Q}_c^{EN}, P_c^{MS}, W_c^{EN}), \end{aligned} \tag{7}$$

where $W_c^{EN} = \mathbb{1}(\max(\hat{P}_w^{EN}) \geq \tau)$ represents the high-confidence regions in $\hat{Q}_c^{EN}$, and $\tau = 0.95$ is a predefined confidence threshold used to filter out noisy labels. L_{ce} and L_{dice} denote the cross-entropy loss and dice loss, which are formulated in the supplementary materials.

3.4. Region-Specific Consistency Regularization

We hypothesize that the regions where both models make consistent predictions are more likely to be accurate, while the divergent regions reflect the differences in their representational capabilities. To enhance the synergistic training efficiency of both models while aligning their representational capabilities, we propose the consensus-divergence consistency regularization (CDCR). Referring to Fig. 4(b), by comparing the predictions of the two models, we obtain consistent and divergent regions, denoted as M_c and M_d, respectively, and apply different constraints to each region:

$$M_c = \mathbb{1}(Q_s^{UT} = Q_s^{MS}), M_d = \mathbf{1} - M_c. \tag{8}$$

Consensus Consistency Regularization We encourage the models to generate high-confidence predictions, characterized by low-entropy probability distributions, in the areas where predictions are reliable. Specifically, we minimize the Shannon entropy for predictions in the M_c:

$$L_c = -\frac{1}{S}\sum(P_c^{UT}\log P_c^{UT} + P_c^{MS}\log P_c^{MS}) \odot M_c, \tag{9}$$

where $S = W \times H \times C$.

Divergence Consistency Regularization We aim to reduce the prediction discrepancies to promote consistent improvement in representational capabilities. Therefore, we minimize the mean squared error (MSE) within the M_d:

$$L_d = -\frac{1}{S}\sum \text{MSE}(P_c^{UT}, P_c^{MS}) \odot M_d. \tag{10}$$

	U-Net alone	MedSAM alone	SynFoC
Training (h)	4.09	7.54	8.61
Testing (s)	11.66	21.04	21.04

Table 2. The training (h) and testing (s) time on Prostate dataset.

3.5. Remarks

To maintain the decent image resolution of predicted segmentation prediction, the input image of MedSAM is upsampled from $W \times H \times L$ to $512 \times 512 \times L$ [58]. The output resolution of the segmentation logits for each class is 128×128, which differs from that of U-Net $(W \times H)$. Interpolation is used to align the resolutions when matrix calculations involve different resolutions. During the testing phase, We retain only the student MedSAM model and resize the segmentation results to $W \times H$ for inference. We present the training and testing times in Tab. 2. On Prostate dataset, our method takes approximately 1 extra hour compared to training MedSAM alone. We believe this additional time is worthwhile due to the significant improvements in training stability and model performance. During the testing phase, we require about 10 extra seconds compared to U-Net, which is acceptable. SynFoC is a general method that achieve competitive performance in traditional SSMIS and UDA settings, facilitating breakthroughs for foundation model, not limited to MedSAM, in downstream tasks. Further details are in the supplementary materials.

4. Experiments

4.1. Datasets and Evaluation Metrics

The Prostate dataset [29] is a well-organized multi-site dataset for prostate MRI segmentation, consisting of T2-weighted MRI data collected from six different data sources across three public datasets. For each domain, the data is split into training and validation sets in a 4:1 ratio. Each 2D slice is resized to a resolution of 384×384 pixels.

The Fundus dataset [43] consists of fundus images from four different medical centers, used for the segmentation of the optic cup and optic disc. The data from each domain has been pre-split into training and test sets, with an 800×800 region of interest (ROI) cropped from each image. We resize the images to 256×256 for processing.

The M&Ms dataset [6] is collected from four different magnetic resonance scanner vendors, with annotations available only for the end-systole and end-diastole phases. We split annotated data of each vendor into training and test sets at a 4:1 ratio for the segmentation tasks of the left ventricle (LV), left ventricle myocardium (MYO), and right ventricle (RV). Each slice is resized to 288×288.

Methods	Venue	#L	(Prostate Segmentation) DSC ↑						DSC ↑	Jaccard ↑	95HD ↓	ASD ↓
			RUNMC	BMC	HCRUDB	UCL	BIDMC	HK	Avg.			
SupOnly	-	20	22.11	21.81	19.60	13.87	18.16	26.98	20.42	15.63	118.15	79.70
UA-MT [56]	MICCAI'19	20	19.09	13.66	16.07	37.30	15.23	11.22	18.76	13.44	127.59	85.76
FixMatch [39]	NeurIPS'20	20	81.69	65.27	53.70	70.40	10.20	81.22	60.41	52.53	49.47	28.83
SS-Net [50]	MICCAI'22	20	14.92	11.64	14.49	34.31	15.45	12.52	17.22	12.65	119.73	81.38
BCP [3]	CVPR'23	20	64.79	62.46	50.49	55.08	63.31	57.64	58.96	48.74	56.81	27.77
CauSSL [33]	ICCV23	20	20.36	31.11	15.68	27.27	26.17	26.66	24.54	18.03	116.15	70.57
ABD [14]	CVPR'24	20	53.10	62.28	9.17	59.22	51.92	22.19	42.98	32.35	75.73	47.17
SymGD [32]	CVPR'24	20	88.34	83.26	83.99	85.45	42.03	78.02	76.85	67.88	39.02	21.08
SAMed [58]	arXiv'23	20	63.67	65.62	65.25	68.65	48.83	75.31	64.56	53.49	35.85	15.83
SemiSAM† [59]	arXiv'23	20	78.07	67.16	76.49	69.25	65.97	69.76	71.12	59.69	28.19	12.61
H-SAM [13]	CVPR'24	20	56.29	51.23	41.59	60.50	51.94	51.86	52.24	42.37	69.31	36.15
CPC-SAM [34]	MICCAI'24	20	77.23	66.46	66.98	76.79	79.71	75.42	73.77	62.85	31.08	13.16
SynFoC$^{\text{U-Net}}$	This paper	20	88.09	85.22	84.97	87.74	87.63	87.74	86.90	79.11	11.25	4.82
SynFoC$^{\text{MedSAM}}$	This paper	20	**88.54**	**85.74**	**84.89**	**87.51**	**87.92**	**88.34**	**87.16**	**79.30**	**10.26**	**4.41**

Table 3. Comparison of different methods on Prostate dataset. #L represents the number of labeled data. The best performance is marked as **bold**, and the second-best is underlined. † denotes that we reproduce the results of SemiSAM.

Methods	Venue	#L	(Optic Cup / Optic Disc Segmentation) DSC ↑				DSC ↑	Jaccard ↑	95HD ↓	ASD ↓
			Domain 1	Domain 2	Domain 3	Domain 4	Avg.			
SupOnly	-	20	59.54 / 73.89	71.28 / 74.23	50.87 / 64.29	35.61 / 63.30	61.63	52.65	48.28	28.86
UA-MT [56]	MICCAI'19	20	59.35 / 78.46	63.08 / 74.45	35.24 / 47.73	36.18 / 55.43	56.24	47.00	48.64	31.35
FixMatch [39]	NeurIPS'20	20	81.18 / 91.29	72.04 / 87.60	80.41 / 92.95	74.58 / 87.07	83.39	73.48	11.77	5.60
SS-Net [50]	MICCAI'22	20	59.42 / 78.15	67.32 / 85.05	45.69 / 69.91	38.76 / 61.13	63.18	53.49	44.90	25.73
BCP [3]	CVPR'23	20	71.65 / 91.10	77.19 / 92.00	72.63 / 90.77	77.67 / 91.42	83.05	73.66	11.05	5.80
CauSSL [33]	ICCV'23	20	63.38 / 80.60	67.52 / 80.72	49.53 / 63.88	39.43 / 49.43	61.81	51.80	41.25	23.94
ABD [14]	CVPR'24	20	73.92 / 79.71	65.19 / 90.96	77.61 / 86.11	74.79 / 86.72	79.38	69.28	13.99	8.14
SymGD [32]	CVPR'24	20	83.71 / 92.96	80.47 / 89.93	84.18 / 92.97	83.71 / 93.38	87.66	79.10	8.21	3.89
SAMed [58]	arXiv'23	20	71.00 / 93.53	81.77 / 90.04	82.07 / 92.25	71.62 / 93.14	84.47	75.69	9.83	5.25
SemiSAM† [59]	arXiv'23	20	83.70 / 93.21	72.40 / 87.72	81.39 / 92.11	79.17 / 91.10	85.10	75.50	9.48	4.60
H-SAM [13]	CVPR'24	20	76.97 / 93.01	79.01 / 90.47	76.85 / 91.86	81.03 / 92.42	85.20	76.26	9.67	4.99
CPC-SAM [34]	MICCAI'24	20	75.99 / **94.34**	80.10 / **93.08**	83.19 / 92.81	83.43 / 93.20	87.02	78.65	8.50	4.24
SynFoC$^{\text{U-Net}}$	This paper	20	84.26 / 92.78	78.51 / 90.12	85.33 / 93.05	82.05 / 93.88	87.50	78.93	7.74	3.79
SynFoC$^{\text{MedSAM}}$	This paper	20	**85.44** / 93.29	**82.50** / 90.52	**85.51 / 93.56**	**83.78 / 94.23**	**88.60**	**80.50**	**6.56**	**3.47**

Table 4. Comparison of different methods on Fundus dataset.

The BUSI dataset [1] includes breast ultrasound images categorized into three classes based on breast cancer: normal, benign, and malignant. Since there is no segmentation target for the normal class, we divide the dataset into two domains based on tumor type: benign and malignant. For each domain, the data is split into training and test sets at a 4:1 ratio. Each image is resized to 256 × 256.

We also evaluate our SynFoC in SSMIS and UDA on the **ACDC** [4] and **MS-CMRSeg** [62] datasets, respectively. The evaluation metrics include the Dice Similarity Coefficient (DSC), Jaccard Index, 95% Hausdorff Distance (95HD), and Average Surface Distance (ASD).

4.2. Implementation Details

Our method is implemented by Pytorch and trained on an NVIDIA GeForce RTX 3090 GPU. For U-Net, we use Stochastic Gradient Descent (SGD) with a momentum of 0.9 and a weight decay of 0.0001, with an initial learning rate of 0.03. For MedSAM, we adopt the ViT_B version, where the input is resized to 512 × 512, and the LoRA rank is set to 4. MedSAM is optimized with the AdamW optimizer with an initial learning rate of 0.0001, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a weight decay of 0.1. We experiment with the following numbers of labeled data for each dataset: 20 for prostate and fundus, 5 for M&MS, and 64 (1/8) or 129 (1/4) for BUSI. Except for M&MS, the batch size is set to 8, with an equal split between labeled and unlabeled samples. The batch size for M&MS is reduced to 4 due to the extremely small number of labeled data (only 5). The maximum number of training iterations is set to 60,000 for Prostate and M&Ms datasets, and 30,000 for Fundus and BUSI datasets. For each dataset, taking Prostate dataset as an example, we use 20 labeled data from a specific domain, such as RUNMC, as the labeled data, while the remaining data serves as the unlabeled data. The performance is evaluated on multiple domain test sets, and the average performance across all six domains is reported as the experimental results, corresponding to the values in the column for RUNMC in Tab. 3. We compare our approach with various methods based on the conventional model (UA-MT [56], FixMatch [39, 41], SS-Net [50], BCP [3], CauSSL [33], ABD [14], SymGD [32]) and those based on the foundation model (SAMed [58], SemiSAM [59], H-SAM [13], CPC-SAM [34]). To ensure a fair comparison, we select U-Net

Method	Venue	#L	(LV / MYO / RV Segmentation) DSC ↑				DSC ↑	Jaccard ↑	95HD ↓	ASD ↓
			Vendor A	Vendor B	Vendor C	Vendor D	Avg.			
SupOnly	-	5	33.65 / 19.07 / 27.38	47.98 / 47.79 / 29.89	23.55 / 12.89 / 20.52	43.82 / 34.15 / 37.29	31.50	24.04	65.13	40.41
UA-MT [56]	MICCAI'19	5	15.64 / 10.57 / 10.38	40.07 / 35.84 / 11.64	17.68 / 13.30 / 12.06	32.32 / 19.76 / 17.90	19.76	14.53	78.23	54.74
FixMatch [39]	NeurIPS'20	5	80.57 / 66.29 / 65.13	87.88 / 79.77 / 77.01	83.37 / 75.47 / 71.89	89.13 / 78.83 / 78.34	77.81	67.47	9.09	**4.85**
SS-Net [50]	MICCAI'22	5	9.90 / 6.89 / 4.77	32.68 / 32.30 / 15.26	7.15 / 6.13 / 4.39	23.20 / 16.24 / 5.28	13.68	10.06	84.29	64.06
BCP [3]	CVPR'23	5	49.99 / 18.12 / 19.55	84.41 / 69.04 / 68.75	57.25 / 40.28 / 42.80	69.10 / 56.43 / 58.83	52.88	43.97	37.10	22.67
CauSSL [33]	ICCV'23	5	33.83 / 18.92 / 17.43	35.00 / 32.70 / 21.42	12.38 / 16.48 / 16.13	28.35 / 28.21 / 22.89	23.65	17.12	69.80	41.75
ABD [14]	CVPR'24	5	38.74 / 24.05 / 1.56	29.47 / 24.20 / 17.36	14.62 / 7.87 / 10.85	39.69 / 31.30 / 35.08	22.90	17.28	63.48	51.38
SymGD [32]	CVPR'24	5	62.35 / 70.53 / 68.15	**89.24** / **81.35** / **81.37**	82.15 / 78.32 / 77.40	89.23 / 79.85 / 78.00	78.16	67.98	12.91	6.50
SAMed [58]	ArXiv'23	5	66.99 / 51.21 / 28.20	77.67 / 60.16 / 48.80	72.98 / 49.72 / 37.68	77.72 / 55.85 / 42.36	55.78	43.61	35.27	15.38
SemiSAM† [59]	ArXiv'23	5	28.88 / 21.73 / 21.52	87.16 / 78.00 / 73.00	82.46 / 72.49 / 67.07	83.28 / 72.94 / 67.30	62.99	53.79	21.41	10.99
H-SAM [13]	CVPR'24	5	50.85 / 31.56 / 32.94	58.01 / 38.82 / 39.82	59.17 / 39.18 / 45.95	69.64 / 48.47 / 48.90	46.94	36.07	36.44	20.40
CPC-SAM [34]	MICCAI'24	5	**87.05** / 74.31 / 72.00	83.65 / 72.31 / 70.88	85.02 / 73.73 / 74.03	86.28 / 74.21 / 68.19	76.81	65.70	12.37	5.13
SynFoC$^{\text{U-Net}}$	U-Net	5	85.81 / 76.80 / 73.71	86.70 / 78.23 / 74.29	85.88 / 76.61 / 74.77	87.90 / 78.32 / 75.54	79.55	69.86	9.33	5.20
SynFoC$^{\text{MedSAM}}$	MedSAM	5	85.65 / **76.40** / **76.18**	88.03 / 78.42 / 75.40	**87.09** / **78.63** / **77.71**	**89.51** / **79.90** / 77.13	**80.84**	**70.94**	**8.15**	**3.65**

Table 5. Comparison of different methods on M&Ms dataset.

Method	Venue	(Cancer) DSC ↑		DSC ↑	Jaccard ↑	95HD ↓	ASD ↓
		Benign	Malignant	Avg.			
		64 (1/8) labels					
SupOnly	-	55.38	63.51	59.45	48.94	61.10	25.44
UA-MT [56]	MICCAI'19	53.51	62.68	58.10	47.96	55.50	26.56
FixMatch [39]	NeurIPS'20	59.49	69.80	64.65	54.60	46.48	20.68
SS-Net [50]	MICCAI'22	56.11	63.36	59.74	49.50	51.30	22.66
BCP [3]	CVPR'23	60.49	65.20	62.85	52.68	50.80	18.62
CauSSL [33]	ICCV'23	49.54	59.31	54.43	44.26	57.09	29.05
ABD [14]	CVPR'24	50.45	62.71	56.58	47.03	49.40	23.27
SymGD [32]	CVPR'24	60.04	72.78	66.41	56.45	40.26	18.20
SAMed [58]	Arxiv'23	66.89	63.52	65.21	54.13	46.47	18.39
SemiSAM† [59]	Arxiv'23	60.65	66.35	63.50	53.25	50.43	23.46
H-SAM [13]	CVPR'24	67.76	63.87	65.82	54.99	41.58	17.25
CPC-SAM [34]	MICCAI'24	**71.87**	65.86	68.87	57.46	40.77	16.29
SynFoC$^{\text{U-Net}}$	This paper	59.56	72.69	66.13	56.41	43.52	20.67
SynFoC$^{\text{MedSAM}}$	This paper	70.16	**73.05**	**71.61**	**61.31**	**34.77**	**15.05**
		129 (1/4) labels					
SupOnly	-	59.57	66.05	62.81	52.31	54.26	23.29
UA-MT [56]	MICCAI'19	56.54	64.92	60.73	50.69	50.29	23.15
FixMatch [39]	NeurIPS'20	61.28	71.13	66.21	55.77	49.73	21.28
SS-Net [50]	MICCAI'22	56.94	64.18	60.56	51.09	49.69	21.32
BCP [3]	CVPR'23	61.96	67.21	64.59	53.68	55.82	22.87
CauSSL [33]	ICCV'23	58.97	62.57	60.77	50.47	48.05	21.46
ABD [14]	CVPR'24	56.62	63.85	60.24	49.34	48.98	20.83
SymGD [32]	CVPR'24	61.68	72.09	66.89	56.79	44.23	17.82
SAMed [58]	Arxiv'23	70.79	68.09	69.44	58.40	38.49	14.67
SemiSAM† [59]	Arxiv'23	61.61	68.25	64.93	55.12	48.10	19.08
H-SAM [13]	CVPR'24	70.86	65.55	68.21	57.26	38.83	18.12
CPC-SAM [34]	MICCAI'24	**74.01**	71.13	72.57	61.80	34.04	13.30
SynFoC$^{\text{U-Net}}$	This paper	61.28	71.45	66.37	56.45	49.89	22.58
SynFoC$^{\text{MedSAM}}$	This paper	73.74	**75.75**	**74.75**	**64.90**	**31.29**	**12.45**

Table 6. Comparison of different methods on BUSI dataset.

as the conventional model and MedSAM as the foundation model. Among the methods, SupOnly indicates the performance of the model trained solely on labeled data.

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

The Prostate dataset. Tab. 3 presents the performance of different methods on Prostate dataset using 20 labeled data. Our method outperforms other state-of-the-art methods by a large margin, achieving an improvement of 10.31% in DSC. For the test data sampled from the same and different domains as the labeled data, we provide visual comparisons in Fig. 5 to validate the superiority of our method.

The Fundus dataset. As shown in Tab. 4, we conduct experiments on Fundus dataset using 20 labeled data. The segmentation performance for the optic cup and disc is separated by a slash. Through the synergistic training of U-Net and MedSAM, both models complement shortcomings of each other and enhance their performance mutually. Our method surpasses all other approaches across four metrics, achieving state-of-the-art results. Visual comparisons on Fundus dataset are presented in the supplementary materials, and the same applies to M&Ms and BUSI datasets.

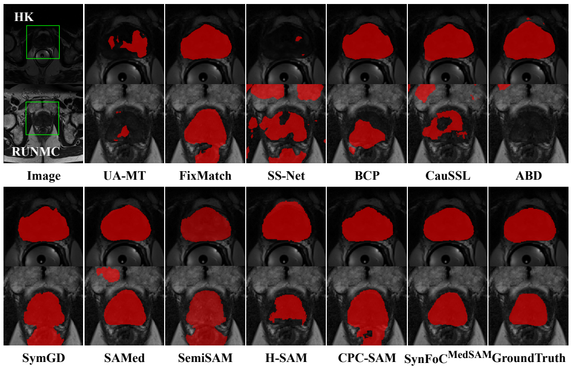

Figure 5. Visual comparison of different methods on Prostate dataset. The test samples are drawn from the labeled domain (HK) and another domain (RUNMC), respectively.

The M&Ms dataset. In Tab. 5, we evaluate the performance of various methods on M&Ms dataset using only 5 labeled data. The segmentation results for the LV, MYO, and RV are separated by slashes. Despite the extremely limited number of labeled data, our method still achieves a 2.14% improvement in DSC compared to other methods.

The BUSI dataset. We further examine our method on BUSI dataset in Tab. 6. When 12.5% and 25% of labeled data are available, our SynFoC outperforms other methods by 2.74% and 2.18% in DSC, respectively.

4.4. Ablation Studies

We conduct ablation studies to verify the effectiveness of each module in our method. All experiments are conducted on Prostate dataset with 20 labeled data.

Effectiveness of each module. As shown in Tab. 7, **Base** refers to the SSMIS method described in Sec. 3.1, where intermediate samples are generated to promote the

Base	SMC	CDCR	CR	DSC ↑						DSC ↑	Jaccard ↑	95HD ↓	ASD ↓
				RUNMC	BMC	HCRUDB	UCL	BIDMC	HK	Avg.	Avg.	Avg.	Avg.
$\checkmark_{\text{U-Net}}$				87.74	67.96	82.45	86.31	41.91	84.35	75.12	65.76	54.67	29.08
$\checkmark_{\text{MedSAM}}$				87.24	79.70	80.20	85.34	71.88	82.46	81.14	71.91	22.78	9.13
✓	✓			88.37	85.09	84.64	87.31	87.45	87.94	86.80	78.85	10.78	4.64
✓			✓	87.99	81.84	82.99	86.18	71.92	86.34	82.88	73.81	17.09	7.21
✓		✓		88.20	84.86	83.36	86.75	87.15	86.82	86.19	77.94	11.91	5.03
✓	✓	✓		**88.54**	**85.74**	**84.89**	**87.51**	**87.92**	**88.34**	**87.16**	**79.30**	**10.26**	**4.41**

Table 7. Ablation experiments on Prostate dataset.

Strategy	DSC ↑	Jaccard ↑	95HD ↓	ASD ↓
$\text{Constant}_{0.5}$	85.93	77.75	11.78	5.06
CPS	84.30	75.36	12.59	5.75
Linear	86.19	78.12	11.21	4.85
Self-only	86.23	78.16	11.15	4.85
Mutual-only	86.29	78.18	11.09	4.76
SMC	**86.80**	**78.85**	**10.78**	**4.64**

Table 8. Ablation study of different synergistic strategies.

training process. $\checkmark_{\text{U-Net}}$ and $\checkmark_{\text{MedSAM}}$ represent the **Base** method employing U-Net and MedSAM as the backbone models, respectively. Compared to the foundation model, the well-trained conventional model achieves a higher performance ceiling. However, when there is a significant discrepancy between labeled data and unlabeled data, the conventional model suffers from error accumulation, leading to training failure, whereas the foundation model continues to demonstrate strong segmentation performance. **SMC** denotes the synergistic training of U-Net and MedSAM, where the pseudo-label ensemble weight is determined based on self-mutual confidence. **CDCR** refers to the introduction of consensus-divergence consistency regularization between the two models. Both methods significantly enhance the training effectiveness of the models. In contrast to **CR**, which directly minimizes the MSE loss between P_c^{UT} and P_c^{MS}, **CDCR** reduces the uncertainty in the consistent predictions of the two models, accelerating a more reliable convergence. Incorporating both **SMC** and **CDCR** improves performance, yielding the best results.

Different synergistic strategies. α can be determined through various approaches, such as **constant**$_{0.5}$, **CPS**, **linear**, and **SMC** mentioned in Sec. 3.3. We also evaluate the performance of employing self-confidence and mutual confidence individually. As shown in Tab. 8, our **SMC** assesses the reliability of U-Net pseudo-labels at the instance level from multiple perspectives, dynamically adjusting the ensemble ratio to achieve optimal performance.

Alpha Variation and Pseudo-Label Quality. In Fig. 6, self-confidence clearly reflects the stability of U-Net, as observed in the fluctuations during epochs 38 to 40, while mutual confidence roughly indicates the quality of the pseudo-labels from U-Net. The α more precisely captures the relationship between the quality of U-Net's pseudo-labels and its stability. The inset further illustrates that the quality of ensemble-generated pseudo-labels consistently surpasses that of the individual model outputs.

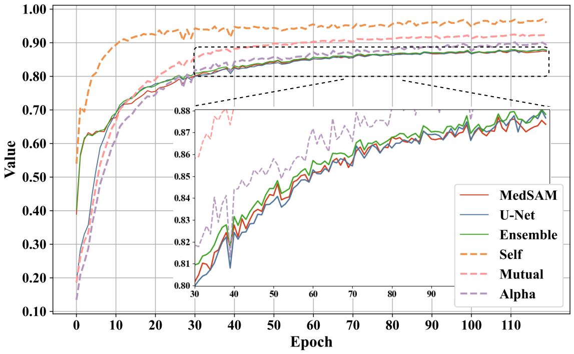

Figure 6. The variation of parameters alongside the quality of pseudo-labels, with 20 labeled data from the BMC domain. The dashed lines represent the curves for self-confidence, mutual confidence, and α, respectively. The solid lines indicate the quality of pseudo-labels generated by MedSAM, U-Net, and the ensemble pseudo-labels, measured by the DSC against the ground truth.

5. Conclusion

In this paper, we propose a novel synergistic training framework for foundation and conventional models that complements shortcomings of each other. We precisely measure the quality of the pseudo-labels from U-Net by self-mutual confidence, designing the instance-wise pseudo-label ensemble ratio to generate higher-quality pseudo-labels. Furthermore, we introduce the consensus-divergence consistency regularization to ensure consistent improvement in the representation capabilities of both models and promote reliable convergence during training. Extensive experiments conducted on four public multi-domain datasets validates the effectiveness of our method.

Acknowledgements. This work was supported by NSFC Project (62222604, 62206052, 624B2063), China Postdoctoral Science Foundation (2024M750424), Fundamental Research Funds for the Central Universities (020214380120, 020214380128), State Key Laboratory Fund (ZZKT2024A14), Postdoctoral Fellowship Program of CPSF (GZC20240252), Jiangsu Funding Program for Excellent Postdoctoral Talent (2024ZB242), Jiangsu Science and Technology Major Project (BG2024031), Shandong Natural Science Foundation (ZR2023MF037), and Open Project Program of the State Key Laboratory of CAD&CG (A2320), Zhejiang University.

References

[1] Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in brief*, 28:104863, 2020. 6

[2] Wenjia Bai, Ozan Oktay, Matthew Sinclair, Hideaki Suzuki, Martin Rajchl, Giacomo Tarroni, Ben Glocker, Andrew King, Paul M Matthews, and Daniel Rueckert. Semi-supervised learning for network-based cardiac mr image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, pages 253–260. Springer, 2017. 1

[3] 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. 1, 6, 7

[4] 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. 6

[5] Heng Cai, Lei Qi, Qian Yu, Yinghuan Shi, and Yang Gao. 3d medical image segmentation with sparse annotation via cross-teaching between 3d and 2d networks. In *Medical Image Computing and Computer-Assisted Intervention*, pages 614–624. Springer, 2023. 2

[6] Victor M Campello, Polyxeni Gkontra, Cristian Izquierdo, Carlos Martin-Isla, Alireza Sojoudi, Peter M Full, Klaus Maier-Hein, Yao Zhang, Zhiqiang He, Jun Ma, et al. Multi-centre, multi-vendor and multi-disease cardiac segmentation: the m&ms challenge. *IEEE Transactions on Medical Imaging*, 40(12):3543–3554, 2021. 5

[7] 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

[8] Lin Chen, Zhixiang Wei, Xin Jin, Huaian Chen, Miao Zheng, Kai Chen, and Yi Jin. Deliberated domain bridging for domain adaptive semantic segmentation. *Advances in Neural Information Processing Systems*, 35:15105–15118, 2022. 1

[9] Shiyun Chen, Li Lin, Pujin Cheng, and Xiaoying Tang. Aslseg: Adapting sam in the loop for semi-supervised liver tumor segmentation. *arXiv preprint arXiv:2312.07969*, 2023. 2

[10] Tianrun Chen, Lanyun Zhu, Chaotao Ding, Runlong Cao, Yan Wang, Zejian Li, Lingyun Sun, Papa Mao, and Ying Zang. Sam fails to segment anything?–sam-adapter: Adapting sam in underperformed scenes: Camouflage, shadow, medical image segmentation, and more. *arXiv preprint arXiv:2304.09148*, 2023. 3

[11] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2613–2622, 2021. 2

[12] Junlong Cheng, Jin Ye, Zhongying Deng, Jianpin Chen, Tianbin Li, Haoyu Wang, Yanzhou Su, Ziyan Huang, Jilong Chen, Lei Jiang, et al. Sam-med2d. *arXiv preprint arXiv:2308.16184*, 2023. 3

[13] Zhiheng Cheng, Qingyue Wei, Hongru Zhu, Yan Wang, Liangqiong Qu, Wei Shao, and Yuyin Zhou. Unleashing the potential of sam for medical adaptation via hierarchical decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3511–3522, 2024. 3, 6, 7

[14] 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. 1, 6, 7

[15] Deng-Ping Fan, Tao Zhou, Ge-Peng Ji, Yi Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Inf-net: Automatic covid-19 lung infection segmentation from ct images. *IEEE Transactions on Medical Imaging*, 39(8):2626–2637, 2020. 1

[16] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33:21271–21284, 2020. 4

[17] Katharina Grünberg, Oscar Jimenez-del Toro, Andras Jakab, Georg Langs, Tomàs Salas Fernandez, Marianne Winterstein, Marc-André Weber, and Markus Krenn. Annotating medical image data. *Cloud-Based Benchmarking of Medical Image Analysis*, pages 45–67, 2017. 1

[18] Lin Lawrence Guo, Stephen R Pfohl, Jason Fries, Alistair EW Johnson, Jose Posada, Catherine Aftandilian, Nigam Shah, and Lillian Sung. Evaluation of domain generalization and adaptation on improving model robustness to temporal dataset shift in clinical medicine. *Scientific reports*, 12(1): 2726, 2022. 1

[19] Along He, Tao Li, Yanlin Wu, Ke Zou, and Huazhu Fu. Frcnet frequency and region consistency for semi-supervised medical image segmentation. *arXiv preprint arXiv:2405.16573*, 2024. 2

[20] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019. 3

[21] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 3

[22] Xinrong Hu, Xiaowei Xu, and Yiyu Shi. How to efficiently adapt large segmentation model (sam) to medical images. *arXiv preprint arXiv:2306.13731*, 2023. 3

[23] Siyao Jiang, Huisi Wu, Junyang Chen, Qin Zhang, and Jing Qin. Ph-net: Semi-supervised breast lesion segmentation via patch-wise hardness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11418–11427, 2024. 2

[24] Aliasghar Khani, Saeid Asgari Taghanaki, Aditya Sanghi, Ali Mahdavi Amiri, and Ghassan Hamarneh. Slime: Segment like me. *arXiv preprint arXiv:2309.03179*, 2023. 1

[25] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 3

[26] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, International Conference on Machine Learning*, page 896. Atlanta, 2013. 2

[27] Shumeng Li, Lei Qi, Qian Yu, Jing Huo, Yinghuan Shi, and Yang Gao. Stitching, fine-tuning, re-training: A sam-enabled framework for semi-supervised 3d medical image segmentation. *IEEE Transactions on Medical Imaging*, 2025. 3

[28] Xiaomeng Li, Lequan Yu, Hao Chen, Chi-Wing Fu, Lei Xing, and Pheng-Ann Heng. Transformation-consistent self-ensembling model for semisupervised medical image segmentation. *IEEE transactions on neural networks and learning systems*, 32(2):523–534, 2020. 2

[29] Quande Liu, Qi Dou, and Pheng-Ann Heng. Shape-aware meta-learning for generalizing prostate mri segmentation to unseen domains. In *Medical Image Computing and Computer-Assisted Intervention*, pages 475–485. Springer, 2020. 5

[30] Zelun Luo, Yuliang Zou, Judy Hoffman, and Li F Fei-Fei. Label efficient learning of transferable representations acrosss domains and tasks. *Advances in Neural Information Processing Systems*, 30, 2017. 1, 2

[31] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024. 1, 2, 3

[32] Qinghe Ma, Jian Zhang, Lei Qi, Qian Yu, Yinghuan Shi, and Yang Gao. Constructing and exploring intermediate domains in mixed domain semi-supervised medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11642–11651, 2024. 1, 2, 6, 7

[33] Juzheng Miao, Cheng Chen, Furui Liu, Hao Wei, and Pheng-Ann Heng. Caussl: Causality-inspired semi-supervised learning for medical image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21426–21437, 2023. 6, 7

[34] Juzheng Miao, Cheng Chen, Keli Zhang, Jie Chuai, Quanzheng Li, and Pheng-Ann Heng. Cross prompting consistency with segment anything model for semi-supervised medical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, pages 167–177. Springer, 2024. 3, 6, 7

[35] Muyang Qiu, Jian Zhang, Lei Qi, Qian Yu, Yinghuan Shi, and Yang Gao. The devil is in the statistics: Mitigating and exploiting statistics difference for generalizable semi-supervised medical image segmentation. In *European Conference on Computer Vision*, pages 74–91. Springer, 2024. 1

[36] Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. *arXiv preprint arXiv:2101.06329*, 2021. 2

[37] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015. 1

[38] Hyungseob Shin, Hyeongyu Kim, Sewon Kim, Yohan Jun, Taejoon Eo, and Dosik Hwang. Sdc-uda: volumetric unsupervised domain adaptation framework for slice-direction continuous cross-modality medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7412–7421, 2023. 1, 2

[39] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in Neural Information Processing Systems*, 33:596–608, 2020. 6, 7

[40] Aiham Taleb, Winfried Loetzsch, Noel Danz, Julius Severin, Thomas Gaertner, Benjamin Bergner, and Christoph Lippert. 3d self-supervised methods for medical imaging. *Advances in Neural Information Processing Systems*, 33:18158–18172, 2020. 1

[41] Pratima Upretee and Bishesh Khanal. Fixmatchseg: Fixing fixmatch for semi-supervised semantic segmentation. *arXiv preprint arXiv:2208.00400*, 2022. 6

[42] Haoran Wang, Lian Huai, Wenbin Li, Lei Qi, Xingqun Jiang, and Yinghuan Shi. Weakmedsam: Weakly-supervised medical image segmentation via sam with sub-class exploration and prompt affinity mining. *arXiv preprint arXiv:2503.04106*, 2025. 2

[43] Shujun Wang, Lequan Yu, Kang Li, Xin Yang, Chi-Wing Fu, and Pheng-Ann Heng. Dofe: Domain-oriented feature embedding for generalizable fundus image segmentation on unseen datasets. *IEEE Transactions on Medical Imaging*, 39 (12):4237–4248, 2020. 5

[44] Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. Seggpt: Towards segmenting everything in context. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1130–1140, 2023. 1

[45] Yikai Wang, Chengming Xu, Chen Liu, Li Zhang, and Yanwei Fu. Instance credibility inference for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 4

[46] Yongchao Wang, Bin Xiao, Xiuli Bi, Weisheng Li, and Xinbo Gao. Mcf: Mutual correction framework for semi-supervised medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15651–15660, 2023. 2

[47] You Wang, Lei Qi, Qian Yu, Yinghuan Shi, and Yang Gao. Enhancing weakly supervised medical segmentation via heterogeneous co-training with box-wise augmentation

and pseudo-label filtering. In *International Conference on Intelligence Science*, pages 331–345. Springer, 2024. 2

[48] Junde Wu, Wei Ji, Yuanpei Liu, Huazhu Fu, Min Xu, Yanwu Xu, and Yueming Jin. Medical sam adapter: Adapting segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.12620*, 2023. 3

[49] Yicheng Wu, Zongyuan Ge, Donghao Zhang, Minfeng Xu, Lei Zhang, Yong Xia, and Jianfei Cai. Mutual consistency learning for semi-supervised medical image segmentation. *Medical Image Analysis*, 81:102530, 2022. 1

[50] 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. 2, 6, 7

[51] Qingsong Xie, Yuexiang Li, Nanjun He, Munan Ning, Kai Ma, Guoxing Wang, Yong Lian, and Yefeng Zheng. Unsupervised domain adaptation for medical image segmentation by disentanglement learning and self-training. *IEEE Transactions on Medical Imaging*, 43(1):4–14, 2022. 1, 2

[52] Lihe Yang, Wei Zhuo, Lei Qi, Yinghuan Shi, and Yang Gao. ST++: Make self-training work better for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4268–4277, 2022. 2

[53] Yuzhe Yang, Haoran Zhang, Judy W Gichoya, Dina Katabi, and Marzyeh Ghassemi. The limits of fair medical imaging ai in real-world generalization. *Nature Medicine*, pages 1–11, 2024. 1

[54] Chenyu You, Yuan Zhou, Ruihan Zhao, Lawrence Staib, and James S Duncan. Simcvd: Simple contrastive voxel-wise representation distillation for semi-supervised medical image segmentation. *IEEE Transactions on Medical Imaging*, 41(9):2228–2237, 2022. 2

[55] Chenyu You, Weicheng Dai, Yifei Min, Fenglin Liu, David Clifton, S Kevin Zhou, Lawrence Staib, and James Duncan. Rethinking semi-supervised medical image segmentation: A variance-reduction perspective. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[56] Lequan Yu, Shujun Wang, Xiaomeng Li, Chi-Wing Fu, and Pheng-Ann Heng. Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, pages 605–613. Springer, 2019. 1, 2, 6, 7

[57] Angela Zhang, Lei Xing, James Zou, and Joseph C Wu. Shifting machine learning for healthcare from development to deployment and from models to data. *Nature Biomedical Engineering*, 6(12):1330–1345, 2022. 1

[58] Kaidong Zhang and Dong Liu. Customized segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.13785*, 2023. 3, 4, 5, 6, 7

[59] Yichi Zhang, Yuan Cheng, and Yuan Qi. Semisam: Exploring sam for enhancing semi-supervised medical image segmentation with extremely limited annotations. *arXiv preprint arXiv:2312.06316*, 2023. 2, 3, 6, 7

[60] Yizhe Zhang, Tao Zhou, Shuo Wang, Ye Wu, Pengfei Gu, and Danny Z Chen. Samdsk: Combining segment anything model with domain-specific knowledge for semi-supervised learning in medical image segmentation. *arXiv preprint arXiv:2308.13759*, 2023. 2

[61] Ziyuan Zhao, Fangcheng Zhou, Kaixin Xu, Zeng Zeng, Cuntai Guan, and S Kevin Zhou. Le-uda: Label-efficient unsupervised domain adaptation for medical image segmentation. *IEEE Transactions on Medical Imaging*, 42(3):633–646, 2022. 1, 2

[62] 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. 6

[63] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *Advances in Neural Information Processing Systems*, 36, 2024. 1