Semi-supervised TEE Segmentation via Interacting with SAM Equipped with Noise-Resilient Prompting

Sen Deng[1], Yidan Feng[1], Haoneng Lin[1], Yiting Fan[2*], Alex Pui-Wai Lee[3], Xiaowei Hu[4], Jing Qin[1*],

[1] Centre for Smart Health, School of Nursing, The Hong Kong Polytechnic University
[2] Department of cardiology, Shanghai Chest Hospital,
Shanghai Jiao Tong University School of Medicine, Shanghai, 200030, China
[3] Division of Cardiology, Department of Medicine and Therapeutics, The Chinese University of Hong Kong
[4] Shanghai Artificial Intelligence Laboratory
sendy.deng@connect.polyu.hk

Abstract

Semi-supervised learning (SSL) is a powerful tool to address the challenge of insufficient annotated data in medical segmentation problems. However, existing semi-supervised methods mainly rely on internal knowledge for pseudo labeling, which is biased due to the distribution mismatch between the highly imbalanced labeled and unlabeled data. Segmenting left atrial appendage (LAA) from transesophageal echocardiogram (TEE) images is a typical medical image segmentation task featured by scarcity of professional annotations and diverse data distributions, for which existing SSL models cannot achieve satisfactory performance. In this paper, we propose a novel strategy to mitigate the inherent challenge of distribution mismatch in SSL by, for the first time, incorporating a large foundation model (i.e. SAM in our implementation) into an SSL model to improve the quality of pseudo labels. We further propose a new self-reconstruction mechanism to generate both noise-resilient prompts to demonically improve SAM's generalization capability over TEE images and self-perturbations to stabilize the training process and reduce the impact of noisy labels. We conduct extensive experiments on an in-house TEE dataset; experimental results demonstrate that our method achieves better performance than state-of-the-art SSL models.

Introduction

Transesophageal Echocardiogram (TEE) is an increasingly utilized modality in ultrasonic cardiac imaging due to its unique advantage of proximity to cardiac structures (Fazlinezhad et al. 2020). The segmentation of left atrial appendage (LAA) from TEE images plays a crucial role in percutaneous LAA occlusion procedure, which is an effective treatment for preventing patients with atrial fibrillation from stroke (Nielsen-Kudsk et al. 2019). However, it is an extremely challenging task as: (1) compared with CT and MR images, the quality of TEE images is relatively low due to speckle noise, acoustic shadow, and motion artifacts, leading to severe boundary ambiguity and deficiency (Faletra et al. 2014); (2) there exist large variations in the size and morphology of the LAA among patients (Chen et al. 2020); and (3) it is laborious and expensive to obtain sufficient pixel-level annotated data from TEE experts for training.

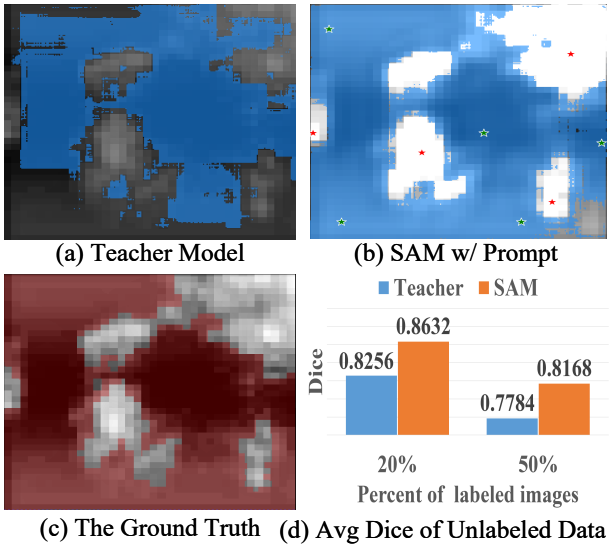

Figure 1: Comparison of the pseudo labels generated by (a) the teacher model and (b) SAM with the proposed noise-resilient prompts. (c) is the ground-truth mask for reference, and (d) shows the average dice of pseudo labels at different labeled ratios, which implies that SAM in our method generated more robust and reliable pseudo labels compared to the vanilla teacher model.

Many efforts have been dedicated to addressing this challenging task. Early methods for LAA segmentation (Song et al. 2016; Morais et al. 2019, 2018) mainly rely on manual configurations and adjustments, which are time-consuming and heavily dependent on individual expertise. More recently, deep learning-based methods have shown promising performance by incorporating shape priors (Zhu et al. 2023) or contextual information (Jin et al. 2018). However, these methods are fully-supervised and rely on large amount of labeled training data, which is hard to obtain in practice.

One promising solution to this problem is to leverage abundant unlabeled data through semi-supervised learning

*Corresponding author

(SSL), which have been actively explored in medical segmentation tasks (Luo et al. 2021a; You et al. 2023; Peiris et al. 2023; Zhu et al. 2021). Existing SSL methods typically leverage knowledge leart from a few labeled images through pseudo labeling, and then enhance representation learning through consistency regularization on unlabeled data (Luo et al. 2021b; Wang et al. 2022; You et al. 2022; Zhou et al. 2019). However, as pointed out in (Wang, Li, and Gool 2019), in practical scenarios, the distribution of labeled data often more or less deviates from the true distribution of the real-world dataset, which is caused by the small sample size and the randomness in sampling. This phenomenon greatly discounts the performance of most SSL models, which assume both labeled and unlabeled datasets follows same or similar distributions (Calderon-Ramirez, Yang, and Elizondo 2022). To address this issue and improve the performance of SSL models, some studies introduce adversarial learning to align the feature distributions (Wang, Li, and Gool 2019) between labeled and unlabeled datasets. Recently, some researchers propose random mixing strategies (Zhou et al. 2021; Bai et al. 2023) to build artificially aligned inputs to alleviate this inherent challenge of SSL. Such strategies, however, are not suitable for medical applications as they may disrupt original anatomical priors and confuse SSL models with mimic structures.

Different from existing SSL approaches, we, for the first time, propose to incorporate large foundation models with, rich yet general, external knowledge into SSL models to guide the pseudo label generation, reducing the adverse effect of distribution discrepancy and hence improving the quality of pseudo labels. As our task is segmentation, we choose the recently proposed segment anything model (SAM) (Kirillov et al. 2023) as our foundation model. To sufficiently take the advantage of SAM, we propose a novel strategy that iteratively harness SAM to enhance pseudo label generation for the SSL model while simultaneously improving SAM's generalization capability over the targeting datasets, i.e. the TEE images. The key idea of this strategy is to build a bidirectional interactive path between the SSL model and SAM, with one direction generating high-quality pseudo labels from 'external expert' (SAM) to the 'internal expert' (teacher model in SSL) and the other providing appropriate prompts by 'internal expert' for the 'external expert'. We further propose a novel noise-resilient prompting scheme based on a self-reconstruction mechanism (SRM) within the teacher-student model. The SRM produces the reconstructed images and candidate masks to generate noise-resilient prompts to enhance the reliability of SAM inference, and then the SAM reciprocally revises the pseudo labels for supervising the SSL learning. Moreover, to enhance the robustness against noisy pseudo labels, based on the prior work (Yang et al. 2022, 2023) that combines data augmentation with feature-level perturbation, we propose a self-perturbation for feature-level augmentation (SPFA) module based on the SRM. Different from the random drop-out strategy (Yang et al. 2023), our SPFA is able to stabilize the training process and hence achieve better performance on small datasets. Our contributions can be summarized as following:

- We propose a novel strategy to mitigate the inherent challenge of distribution mismatch in SSL by, for the first time, incorporating a large foundation model (i.e. SAM in our implementation) into an SSL model to improve the quality of pseudo labels.
- We further propose a new self-reconstruction mechanism to generate both noise-resilient prompts to demonically improve SAM's generalization capability over TEE images and self-perturbations to stabilize the training process and reduce the impact of noisy labels.
- We conduct extensive experiments on an in-house TEE dataset, and experimental results demonstrate that our method achieves better performance than state-of-the-art SSL models.

Related Work

LAA segmentation Segmentation of the left atrial appendage (LAA) anatomy in transesophageal echocardiography (TEE) images can greatly assist transcatheter LAA closure (LAAC) surgery. In the last decade, many algorithms have been developed to conduct automatic or semi-automatic extraction of LAA in TEE images (Song et al. 2016; Morais et al. 2019; Jin et al. 2018; Zhu et al. 2023). The semi-automatic methods were the mainstream approaches in the early stages, relying on manual setup and adjustments. In (Song et al. 2016), threshold segmentation was simply applied with human interactions for refinement. In (Morais et al. 2019), a double stage segmentation process based on B-spline Explicit Active Surfaces framework (BEAS) model was conducted to capture the LAA anatomical particularities. This method relied on manually defined centerline to generate preliminary results and involved multi-step refinement procedures. Benefiting from the development of deep learning techniques, completely automatic LAA segmentation methods have shown promising results over the time-consuming semi-automatic mehtods. In (Jin et al. 2018), a fully convolutional neural (FCN) network was fine-tuned to segment 2D LAA slice images on computed tomography (CT) images. Subsequently, to pile all predicted 2D slices to 3D volume, a dense 3D conditional random field was used to account for the 3D contextual information. In (Zhu et al. 2023), adversarial learning was introduced for mask reconstruction and LAA segmentation. Specifically, an adversarial latent space alignment loss incorporates the shape prior knowledge encoded by the reconstruction network into the segmentation network and improve the accuracy of LAA edge segmentation.

Semi-supervised learning The majority of existing SSL segmentation methods can be divided into two categories: consistency regularization (Tarvainen and Valpola 2017) and pseudo labeling (Lee et al. 2013). Pseudo labeling aims to generate labels for unlabeled data using the model's current predictions, and treating these generated labels as ground truth labels. To enhance the quality of pseudo labels, a novel local contrastive loss function was proposed in (Chaitanya et al. 2023) to enhance the extraction of pixel-level features. In (Seibold et al. 2022), annotated images were employed as the reference for establishing pixel correspondences between an unlabeled image and its semantic counterparts. In

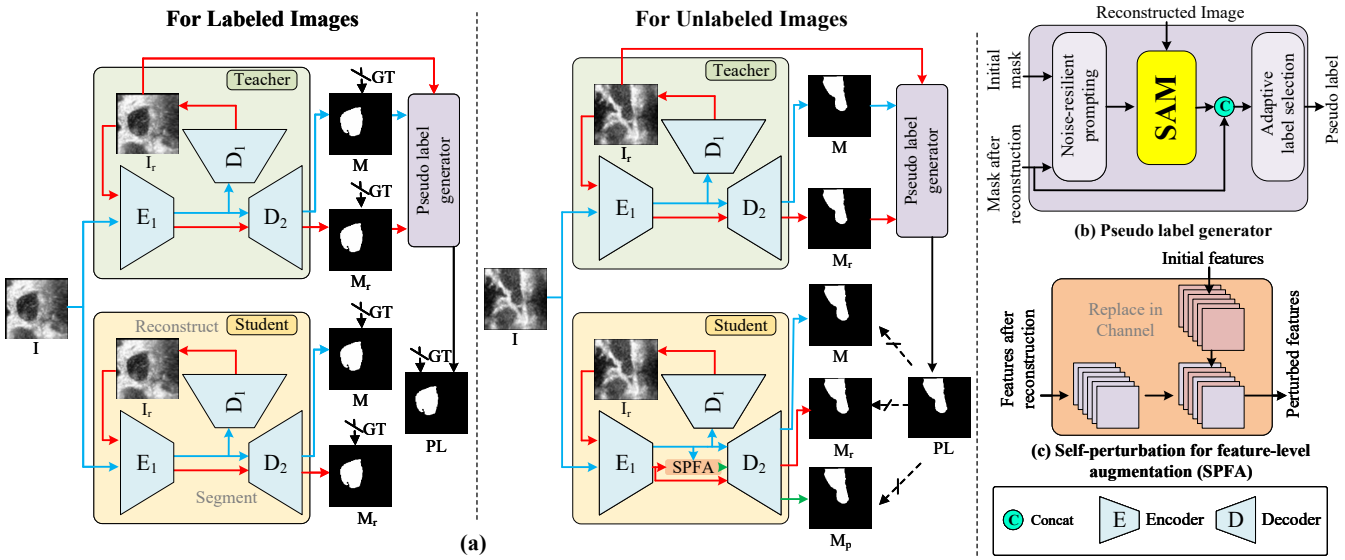

Figure 2: The overview of our proposed method during the training phase. The red, blue and green lines represent the path of initial, reconstructed and perturbed features, respectively. The teacher and student models share the same architecture, which is the self-reconstruction based segmentation network. The network accept the input noisy image I and obtain the reconstruction I_r within the inner loop, and then segment (I, I_r) to obtain the mask pair (M, M_r). PL denotes the pseudo label produced by the SAM-guided pseudo label generator, which is used as the supervision signal for mask predictions (M, M_r, M_p) from unlabeled images. Among them, M_p is the mask decoded from the perturbed feature generated by the proposed SPFA module. At test stage, the predicted mask after self-reconstruction M_r is the final output.

(Yao, Hu, and Li 2022), Fourier transformation was leveraged to acquire cross-domain knowledge, which can be interpreted as perturbation and regularization techniques that enhance the model's ability to learn more reliable pseudo labels. In (Wang et al. 2023), a dynamic competitive pseudo-label generation (DCPLG) module was presented to automatically select the best-performing sub-net as the pseudo label generator. Differently, the consistency regularization focuses on encouraging consistent predictions under perturbations to learn more robust and generalizable representations (Chen et al. 2021a; Luo et al. 2021c; Li et al. 2020). Specifically, various perturbations were employed to train the model, including perturbation of input data (Li et al. 2020; Tu et al. 2022), feature-level perturbations (Ouali, Hudelot, and Tami 2020), and perturbation between different networks (Chen et al. 2021b; Wu et al. 2022). The coherence across diverse tasks were also explored (Luo et al. 2021c). It was also argued that consistency regularization helps to prevent over-fitting over inaccurate pseudo labels (Yang et al. 2022). The authors in (Yang et al. 2022) proposed an advanced self-training framework equipped with strong data augmentation and selective re-training via prioritizing reliable unlabeled images. Their following work (Yang et al. 2023) further introduce feature-level perturbations as an effective supplement to the original data-level augmentation methods. Similarly, our approach combines pseudo-labeling and consistency regularization, aiming to enhance the pseudo-label quality while simultaneously reducing the impact of erroneous pseudo-labels.

Methodology

Preliminaries

To pave the way for clear comprehension, we begin by elucidating foundational concepts in the realm of our method.

Teacher-student model. The teacher-student model (Tarvainen and Valpola 2017) stands as a quintessential paradigm within semi-supervised learning. This model framework comprises a teacher network responsible for generating pseudo-labels for unlabeled images, in tandem with a student network that receives training leveraging both actual labels from labeled images and pseudo-labels from unlabeled images.

Noise2noise mapping. Traditional image denoising methods often demand substantial paired noisy/clean data for effective training. However, the procurement of such matched datasets can pose significant challenges. To alleviate this constraint, Noise2noise mapping (N2N) (Lehtinen et al. 2018) introduced an unsupervised image denoising approach. This technique operates under the assumption of zero-mean noise and trains a network to transform noisy images into other noisy observations sampled from the same distribution. An evolution of this concept, Neighbor2Neighbor (NB2NB) (Huang et al. 2021), took a stride towards further refinement. In NB2NB, the creation of multiple observations involves the sampling of a single input noisy image. For a given noisy input image I_n, NB2NB optimizes the reconstruction network by minimizing a compos-

ite loss function L_{rec}, expressed as:

$$L_{rec} = ||f(g_1(I_n)) - g_2(I_n)||_2^2, \quad (1)$$

where $g_1(\cdot)$ and $g_2(\cdot)$ respectively denote the sub-image sampler functions, while $f(\cdot)$ represents the reconstruction network. Simultaneously, an additional regularization term L_{reg} emphasizes the fundamental distinction in pixel values between ground-truth sub-sampled noisy image pairs:

$$L_{reg} = ||f(g_1(I_n)) - g_2(I_n) - (g_1(f(I_n)) - g_2(f(I_n)))||_2^2. \quad (2)$$

These preliminary concepts lay the groundwork for the subsequent unveiling of our comprehensive methodology.

The Overall Architecture of Our Method

We present the architecture of our method, depicted in Fig. 2. Our approach is fundamentally rooted in the teacher-student model, offering a unified structure. Both the teacher and student networks adopt a self-reconstruction based segmentation model, which takes the input image I and autonomously produces a pair of mask predictions: M and M_r (from images prior to and after reconstruction). In the teacher network, we introduce the SAM-guided pseudo label generator (PLG), building upon the predicted mask pair and the reconstructed image I_r. This PLG employs a two-fold process: firstly, it devises noise-resistant prompts for SAM prediction (M_{SAM}), and subsequently, it learns to amalgamate the various mask predictions into the ultimate pseudo label (PL). For images with available ground-truth labels (GT), these labels are employed to regulate all mask predictions. This regulation, in turn, influences the updates to both the self-reconstruction based segmentation network and the learnable parameters within the PLG. In the context of unlabeled images, the teacher model generates the pseudo label PL. This pseudo label acts as a supervisory signal for all outputs generated by the student model. This encompasses the mask prediction pair (M and M_r), along with an additional perturbed version (M_p) derived from self-perturbation for feature-level augmentation (SPFA). The conclusive segmentation outcome during the test phase emerges as the mask prediction post self-reconstruction (M_r) attained through the student model. This comprehensive approach melds various strategies to enhance segmentation performance and adapt to diverse scenarios.

Segmentation with Self-Reconstruction

TEE images often grapple with noisy backgrounds and ill-defined boundaries, which significantly impair segmentation performance and exacerbate the domain gap issue between TEE and natural images. The conventional strategy of pre-training a denoising model to mitigate these challenges might prove ineffective. This is due to the potential presence of noise, invisible to human perception yet detrimental to subsequent tasks (Li et al. 2023).

Hence, we introduce a novel approach by incorporating the self-reconstruction module into the segmentation network. Our architecture intertwines the segmentation module with the self-reconstruction module, sharing an encoder ($E_1(\cdot)$) while employing separate decoders ($D_1(\cdot)$ and $D_2(\cdot)$). The reconstructed image (I_r) is derived through $I_r = f_{rec}(I)$, where $f_{rec} = D_1(E_1(\cdot))$. This reconstructed image, in turn, undergoes iterative processing through the segmentation module ($f_{seg} = D_2(E_1(\cdot))$) to yield M_r. The outputs from the segmentation module are subjected to supervision via a fusion of focal loss (Lin et al. 2017) and dice loss (Milletari, Navab, and Ahmadi 2016):

$$L_{seg}(x, y) = \text{FocalLoss}(x, y) + 0.1 \cdot \text{DiceLoss}(x, y) \ . \quad (3)$$

Differing from NB2NB, we introduce an additional regularization term L_{rel} within the self-reconstruction module. This term emphasizes the relative disparity between M_r and M from a task-oriented standpoint. It is expressed as:

$$L_{rel} = \max(\text{Dice}(M, GT) - \text{Dice}(M_r, GT), 0) \ . \quad (4)$$

The holistic self-reconstruction loss is thereby consolidated as $L_{srec} = L_{rec} + L_{reg} + L_{rel}$. Through this integration of self-reconstruction, our method adapts to noisy environments while addressing the TEE image-to-natural image domain gap, ultimately enhancing the segmentation accuracy.

SAM-Powered Pseudo Label Generation

Our Pseudo Label Generator (PLG) harmonizes internal insights from the teacher model and external knowledge from SAM, culminating in more precise pseudo labels. Operating on the reconstructed image (I_r) alongside mask prediction pairs (M and M_r) from the teacher model, the PLG yields enhanced pseudo labels, which subsequently refine the teacher-student model. This intricate process unfolds in three sequential components: noise-resilient prompting, SAM integration, and Adaptive label selection.

Noise-resilient prompting The efficacy of the prompt holds paramount importance, profoundly influencing the final mask prediction produced by SAM. SAM employs diverse prompt types, encompassing masks, points, boxes, and text, sourced from interactive learning. Building on findings from (Gong et al. 2023), SAM's robustness to point position variations underscores the reliability of points prompts. For medical image segmentation, point prompts retain similarity in segmentation outcomes across different point positions. This characteristic strengthens our reliance on point prompts, particularly for TEE data. Given a labeled noisy input image (I) and corresponding mask prediction pairs (M_r, M), we initially identify high-confidence foreground/background points for each predicted mask. The criterion for this identification is as follows:

$$criterion = \begin{cases} foreground & if\ conf > \alpha \\ background & if\ conf > 1 - \alpha \end{cases} \quad (5)$$

where conf signifies the confidence of each point, and α is set at 0.85. The noise-resilient prompts emerge as a unanimous selection from $P(M, M_r) = Random(\text{criterion}(M) \cap \text{criterion}(M_r), k)$. Here, we randomly select k reliable points as prompts.

Adaptive label selection Our primary goal is to facilitate the model's adaptive selection of a more dependable pseudo

Algorithm 1: Pseudo code of our training strategy

Input: Labeled data$\{x, y\}$, Unlabeled data u
Parameter: Teacher T, Student S, total epoch number N_t
for *i = 1 to* N_t **do**
  $x_r^T = f_{rec}^T(x), x_r^S = f_{rec}^S(x)$
  $[m_r^T, m^T] = f_{seg}^T([x_r^T, x])$
  $[m_r^S, m^S] = f_{seg}^S([x_r^S, x])$
  Calculate loss: $L_{srec}, L_{seg}(m_{\cdot}^{\cdot}, y)$
  $Prompt = P(m_r^T, m^T)$
  $m_{SAM} = SAM(x_r^T, Prompt)$
  $pl = Con_{1x1}(m_r^T, m_{SAM}))$
  Calculate loss: L_{sel}
  $u_r^T = f_{rec}^T(u), u_r^S = f_{rec}^S(u)$
  $[\widetilde{m}_r^T, \widetilde{m}^T] = f_{seg}^T([u_r^T, u])$
  $[\widetilde{m}_r^S, \widetilde{m}^S] = f_{seg}^S([u_r^S, u])$
  $\widetilde{Prompt} = P(\widetilde{m}_r^T, \widetilde{m}^T)$
  $\widetilde{m}_{SAM} = SAM(u_r^T, \widetilde{Prompt})$
  $\widetilde{pl} = Con_{1x1}(Cat(\widetilde{m}_r^T, \widetilde{m}_{SAM}))$
  Calculate loss: $L_{pt}, L_{seg}(\widetilde{m}_{\cdot}^S, \widetilde{m}_{pseudo})$
  Back propagation to update parameters in $T\&S$

label from the teacher model's internal expertise and the external insights offered by SAM during training. This adaptability enhances the training process's effectiveness and results in improved performance. To achieve this, we employ a 1×1 convolution layer for soft selection. This layer is fine-tuned based on the following optimization criterion:

$$L_{sel} = L_{seg}(PL, GT) \,, \tag{6}$$

where $PL = Con_{1x1}(Cat(f_{seg}(I_c), SAM(I_c, P(I_n))))$ embodies the ultimate pseudo label. This formulation integrates the outputs of the segmentation module (f_{seg}) applied to I_c and the SAM module (SAM) operating on I_c and $P(I_n)$ (noise-resilient prompts). The Con_{1x1} signifies the 1×1 convolutional layer. This adaptive selection mechanism effectively guides the training process, resulting in more accurate pseudo labels.

Self-Perturbation for Feature-Level Augmentation

Taking into account the distinct attributes of TEE data, namely the limited dataset size and diminished image quality, we harness features extracted from the noisy input image to substitute a portion (p) of the features derived from the reconstructed image. This integration introduces domain-specific feature-level perturbations. To elaborate, when handling unlabeled data, we are presented with both the reconstructed image (I_r) and its corresponding noisy input counterpart (I). The ensuing mask prediction after feature-level perturbation is mathematically expressed as:

$$M_p = D_2(Replace(E_1(I), E_1(I_r))) \,. \tag{7}$$

In this equation, $Replace(\cdot)$ signifies the feature perturbation process, with a user-defined proportion of p. Subsequently, we aim to minimize the loss between the pseudo label (PL) and the perturbed variant of the mask prediction (M_p):

$$L_{pt} = L_{seg}(M_p, PL) \,. \tag{8}$$

This mechanism for feature-level augmentation, achieved through self-perturbation, injects adaptively tailored perturbations into the feature space. This approach bolsters model robustness and performance, particularly for TEE data marked by constrained dataset size and suboptimal image quality.

Training Strategy

We outline the training strategy in Algorithm 1, detailing the learning procedures for our proposed semi-supervised segmentation method. The algorithm also illustrates the dynamic interplay between SAM and the teacher-student model. For labeled images, the network update revolves around minimizing both the self-reconstruction loss and the segmentation losses:

$$L^{lb} = \beta_1 L_{srec} + \beta_2 L_{seg}(M_{\cdot}, GT) + \beta_3 L_{sel} \,. \tag{9}$$

When addressing unlabeled images, the pertinent loss function is formulated as:

$$L^{ulb} = \gamma_1 L_{seg}(M_{\cdot}, PL) + \gamma_2 L_{pt} \,. \tag{10}$$

The comprehensive loss that governs the training process is encapsulated in the total loss:

$$L^{total} = L^{lb} + L^{ulb} \,. \tag{11}$$

The combination of these distinct loss components optimizes the model's performance for both labeled and unlabeled images, achieving the goal of semi-supervised medical image segmentation.

Experimental Results

Implementation Details

Dataset The transesophageal echocardiography (TEE) imaging was collected using either an EPIQ7C or CVx ultrasound system (Philips Medical Systems, Andover, MA). The dataset consists of TEE data from 100 patients. We randomly divided the patients into three sets, with 70/30 for training and testing.

Metrics We employ two metrics to assess the segmentation performance in this paper, including the Dice similarity coefficient (Dice), and Average Surface Distance (ASD). Dice emphasizes the overlap of segmented and true regions, highlighting their similarity in terms of area, and ASD focuses on the spatial accuracy of the segmentation by measuring the distance between surface points, which is more sensitive to the boundary accuracy.

Network architecture and hyper-parameters The encoder E_1 and decoders D_1, D_2 adopt the same architecture borrowing from the classical segmentation model Unet (Ronneberger, Fischer, and Brox 2015). The weights of the losses for labeled images, β_1, β_2, and β_3, are experimentally set as 0.01, 1, and 1 respectively. For unlabeled images, the values of γ_1 and γ_2 are set as 0.01 and 0.1 respectively. The proportion of p in the SPFA algorithm and the selected points number k is set as 0.7 and 6 in our experiment.

Method	Data used		Metrics	
	Labeled	Unlabeled	Dice ↑	ASD ↓
Unet	100%	0	0.9019±0.0315	1.23±0.5158
Unet	50%	0	0.8457±0.0789	1.72±0.8821
ST++	50%	50%	0.8694±0.1059	1.48±1.4128
UM	50%	50%	0.8764±0.0271	1.40±0.9589
MCF	50%	50%	0.8490±0.1075	1.76±0.9014
BCP	50%	50%	0.8546±0.1302	1.51±1.5220
UCMT	50%	50%	0.8595±0.0864	1.68±0.3388
Ours	50%	50%	**0.8830±0.0367**	**1.28±0.5723**
Unet	20%	0	0.8290±0.0973	2.08±1.2337
ST++	20%	80%	0.8453±0.1862	1.77±0.0124
UM	20%	80%	0.8347±0.0748	1.83±1.6189
MCF	20%	80%	0.8303±0.0932	2.00±0.1621
BCP	20%	80%	0.8405±0.1434	1.75±1.0740
UCMT	20%	80%	0.8336±0.1110	1.99±0.8511
Ours	20%	80%	**0.8604±0.0518**	**1.65±0.6551**

Table 1: Comparisons with state-of-the-art semi-supervised segmentation methods on our TEE dataset, and the best results are highlighted.

Training details The hyper-parameters are consistent between the teacher model and student model. During training, each slice of TEE data is randomly cropped to obtain patches of size 192X192. The network is optimized using AdamW with a mini-batch size of 8 and trained for a total of 300 epochs. The learning rate is initialized as 0.004, which is divided by 2 every 60 epochs. All the experiments are conducted on 4 NVIDIA 3090 GPUs.

Comparison with State-of-the-art Methods

We compare our method with several SOTA methods on the TEE dataset, including two semi-supervised segmentation methods (originally for natural images): ST++ (Yang et al. 2022), UM (Yang et al. 2023), and three semi-supervised segmentation (originally for medical images): MCF (Wang et al. 2023), BCP (Bai et al. 2023), UCMT (Shen et al. 2023). The comparison was conducted using different labeled ratios: 50% and 20%. The quantitative evaluation results are presented in Table 1, which It can be observed that our method demonstrates the our method consistently outperformed other SOTA methods on both evaluation metrics. The visual comparisons are shown in Fig. 3, it can be observed that our segmentation result contains significantly fewer outliers and more faithfully reproduces the anatomy of the left atrial appendage.

Ablation Study and Analysis

In Tab. 2, we show quantitative results in order to validate the effectiveness of: SAM-powered pseudo label generator (GPL), self-reconstruction mechanism (SRM), Self-perturbation for feature-level augmentation (SPFA).

Analysis on SAM-powered GPL To validate the efficacy of SAM-powered GPL, we exclude it from our approach and simply employ the pseudo labels generated by the teacher model for training. As depicted in Table 2, the Dice and ASD is observed to decline from 0.8604 to 0.8366 and from 1.65 to 1.94, respectively, due to the removal of the GPL. This implies that the inclusion of the SAM with its extensive yet general knowledge significantly enhances the performance of semi-supervised segmentation.

Method	GPL	SPFA	Dice ↑	ASD ↓
Unet	×	×	0.8293	2.08
Unet	✓	×	0.8469	1.92
Unet	×	✓	0.8366	1.94
Ours	✓	✓	**0.8604**	**1.65**

Table 2: Quantitative comparison of the different components of our method. The experiment was conducted on our TEE dataset with of 20% labeled data.

Analysis on SRM Self-reconstruction mechanism (SRM) is the key component in our pipeline that enables the construction of SAM-powered GPL and SPFA. We validate the effectiveness of SRM by removing it and utilizing the SAM to directly generate pseudo labels for unlabeled data, which gives Dice/ASD of 0.8249/2.16, dropped by 0.03 and 0.51, respectively. The performance of directly using SAM to generate the pseudo labels is even worse than that of the corresponding fully-supervised Unet (0.8290/2.08 in Tab. 1). This observation suggests that, without the SRM produced noise-resilient prompts and the reconstructed image, the SAM predictions become unreliable due to the large domain gap, which will severely compromise the model's performance.

Additionally, in Fig. 4, we present the visualization of noisy/reconstructed images and their corresponding segmentation results, where the effects of self-reconstruction can be obviously observed. The image noise caused plenty disconnected outliers and the ambiguous boundaries greatly decimate the overall accuracy. However, thanks to the proposed SRM, the regularization term L_{rel} in SRM enforced the reconstructed image to have better task performance than the noisy one, which also encouraged the blurred edges to become more distinct during the reconstruction process.

Analysis on SPFA We have also excluded the SPFA from our method, and the corresponding results are presented in Table 2. After the removal of SPFA, the model's performance witnessed a decline due to the influence of noisy pseudo labels. Moreover, we also argue that inappropriate augmentation choices may impact the performance of segmentation. To investigate this, we compare two different data augmentation techniques with our SPFA approach: DA in (Yang et al. 2022), and supplemented with feature-level augmentation (FA dropout) in (Yang et al. 2023). The comparative results are shown in Table 3, SPFA appears to be the best partner of GPL (Tab. 2), and their synergistic effect has achieved the promising performance gain.

Conclusion

In this paper, we present a semi-supervised segmentation approach powered by the large generalized segmentation model SAM. Different from existing methods that inevitably suffered from the sample bias, our method introduces SAM

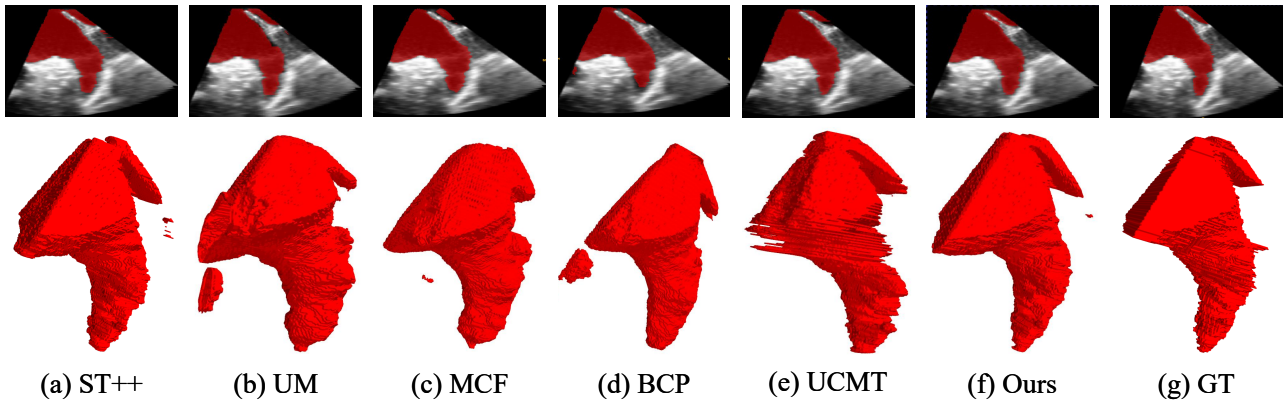

Figure 3: Comparisons of the segmentation results on our TEE data. From (a)-(g): The segmentation result of (a) ST++, (b) UniMatch, (c) MCF, (d) BCP, (e) UCMT, (f) ours and (g) the ground truth, respectively.

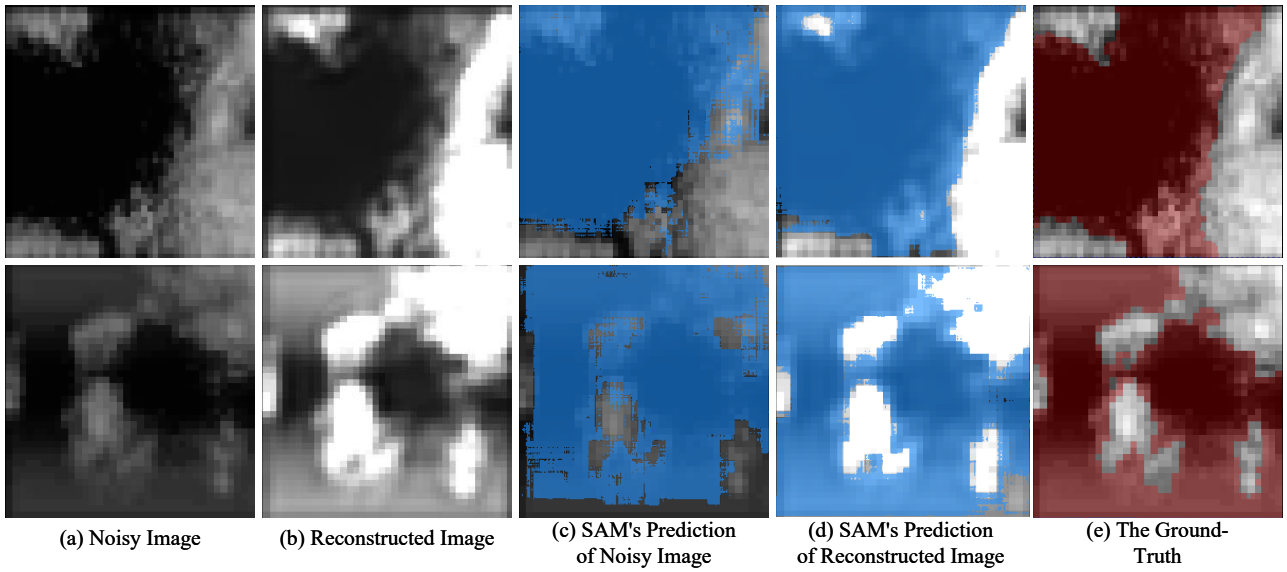

Figure 4: Visualization of noisy/reconstructed images and their corresponding segmentation results.

Metrics	only DA	FA dropout	Ours
Dice ↑	0.8385	0.8404	**0.8604**
ASD ↓	1.97	1.84	**1.65**

Table 3: Ablation results from vanilla SAM and different augmentation methods. The experiment was conducted on our TEE dataset with of 20% labeled data.

with general knowledge to revise the biased pseudo labels. For our TEE segmentation task, which represents an extremely challenging type for medical image segmentation, we propose the self-reconstruction mechanism, autonomously reconstructing the noisy input within the inner loop of our segmentation network, which enables the designs in SAM-powered PLG to improve the quality of pseudo labels, and SPFA to avoid over-fitting towards inaccurate pseudo labels. Our experiments, encompassing both qualitative and quantitative evaluations, showcase the superiority of our approach over existing semi-supervised methods. In essence, our proposed strategy advances LAA segmentation in TEE images by synergizing self-reconstruction learning and external knowledge integration within a semi-supervised framework, offering promise for robust segmentation in medical scenarios.

Acknowledgments

The work described in this paper is supported by a General Research Fund of Hong Kong Research Grants Council (project no. 15205919), an Innovation and Technology Fund of Hong Kong Innovation and Technology Commission (project no. GHP/050/20SZ), an Innovation and Technology Fund - Midstream Research Programme for Universities (project no. MRP/022/20X) and National Natural Sci-

ence Foundation of China (grant no. 82200557).

References

Bai, Y.; Chen, D.; Li, Q.; Shen, W.; and Wang, Y. 2023. Bidirectional Copy-Paste for Semi-Supervised Medical Image Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11514–11524.

Calderon-Ramirez, S.; Yang, S.; and Elizondo, D. 2022. Semisupervised deep learning for image classification with distribution mismatch: A survey. *IEEE Transactions on Artificial Intelligence*, 3(6): 1015–1029.

Chaitanya, K.; Erdil, E.; Karani, N.; and Konukoglu, E. 2023. Local contrastive loss with pseudo-label based self-training for semi-supervised medical image segmentation. *Medical Image Analysis*, 87: 102792.

Chen, H.-H.; Liu, C.-M.; Chang, S.-L.; Chang, P. Y.-C.; Chen, W.-S.; Pan, Y.-M.; Fang, S.-T.; Zhan, S.-Q.; Chuang, C.-M.; Lin, Y.-J.; et al. 2020. Automated extraction of left atrial volumes from two-dimensional computer tomography images using a deep learning technique. *International Journal of Cardiology*, 316: 272–278.

Chen, J.; Zhang, H.; Mohiaddin, R.; Wong, T.; Firmin, D.; Keegan, J.; and Yang, G. 2021a. Adaptive hierarchical dual consistency for semi-supervised left atrium segmentation on cross-domain data. *IEEE Transactions on Medical Imaging*, 41(2): 420–433.

Chen, X.; Yuan, Y.; Zeng, G.; and Wang, J. 2021b. Semi-supervised semantic segmentation with cross pseudo supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2613–2622.

Faletra, F. F.; Ramamurthi, A.; Dequarti, M. C.; Leo, L. A.; Moccetti, T.; and Pandian, N. 2014. Artifacts in three-dimensional transesophageal echocardiography. *Journal of the American Society of Echocardiography*, 27(5): 453–462.

Fazlinezhad, A.; Narayanasamy, H.; Wilansky, S.; and Naqvi, T. Z. 2020. Detection of LV apical thrombus by three-dimensional transesophageal echocardiography. *Echocardiography*, 37(1): 142–146.

Gong, S.; Zhong, Y.; Ma, W.; Li, J.; Wang, Z.; Zhang, J.; Heng, P.-A.; and Dou, Q. 2023. 3DSAM-adapter: Holistic Adaptation of SAM from 2D to 3D for Promptable Medical Image Segmentation. *arXiv preprint arXiv:2306.13465*.

Huang, T.; Li, S.; Jia, X.; Lu, H.; and Liu, J. 2021. Neighbor2Neighbor: Self-Supervised Denoising From Single Noisy Images. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, 14781–14790. Computer Vision Foundation / IEEE.

Jin, C.; Feng, J.; Wang, L.; Yu, H.; Liu, J.; Lu, J.; and Zhou, J. 2018. Left atrial appendage segmentation using fully convolutional neural networks and modified three-dimensional conditional random fields. *IEEE journal of biomedical and health informatics*, 22(6): 1906–1916.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; Dollár, P.; and Girshick, R. 2023. Segment Anything. *arXiv:2304.02643*.

Lee, D.-H.; et al. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, 896. Atlanta.

Lehtinen, J.; Munkberg, J.; Hasselgren, J.; Laine, S.; Karras, T.; Aittala, M.; and Aila, T. 2018. Noise2Noise: Learning Image Restoration without Clean Data. In Dy, J. G.; and Krause, A., eds., *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, 2971–2980. PMLR.

Li, C.; Zhou, H.; Liu, Y.; Yang, C.; Xie, Y.; Li, Z.; and Zhu, L. 2023. Detection-Friendly Dehazing: Object Detection in Real-World Hazy Scenes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(7): 8284–8295.

Li, X.; Yu, L.; Chen, H.; Fu, C.-W.; Xing, L.; and Heng, P.-A. 2020. Transformation-consistent self-ensembling model for semisupervised medical image segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 32(2): 523–534.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2980–2988.

Luo, X.; Chen, J.; Song, T.; and Wang, G. 2021a. Semi-supervised medical image segmentation through dual-task consistency. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 8801–8809.

Luo, X.; Chen, J.; Song, T.; and Wang, G. 2021b. Semi-supervised medical image segmentation through dual-task consistency. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 8801–8809.

Luo, X.; Chen, J.; Song, T.; and Wang, G. 2021c. Semi-supervised medical image segmentation through dual-task consistency. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 8801–8809.

Milletari, F.; Navab, N.; and Ahmadi, S.-A. 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, 565–571. Ieee.

Morais, P.; Queirós, S.; De Meester, P.; Budts, W.; Vilaça, J. L.; Tavares, J. M. R.; and D'hooge, J. 2018. Fast segmentation of the left atrial appendage in 3-D transesophageal echocardiographic images. *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, 65(12): 2332–2342.

Morais, P.; Vilaça, J. L.; Queirós, S.; De Meester, P.; Budts, W.; Tavares, J. M. R.; and D'hooge, J. 2019. Semiautomatic estimation of device size for left atrial appendage occlusion in 3-D TEE images. *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, 66(5): 922–929.

Nielsen-Kudsk, J. E.; Berti, S.; De Backer, O.; Aguirre, D.; Fassini, G.; Cruz-Gonzalez, I.; Grassi, G.; and Tondo, C. 2019. Use of intracardiac compared with transesophageal echocardiography for left atrial appendage occlusion in the

amulet observational study. *JACC: Cardiovascular Interventions*, 12(11): 1030–1039.

Ouali, Y.; Hudelot, C.; and Tami, M. 2020. Semi-supervised semantic segmentation with cross-consistency training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12674–12684.

Peiris, H.; Hayat, M.; Chen, Z.; Egan, G.; and Harandi, M. 2023. Uncertainty-guided dual-views for semi-supervised volumetric medical image segmentation. *Nature Machine Intelligence*, 1–15.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, 234–241. Springer.

Seibold, C. M.; Reiß, S.; Kleesiek, J.; and Stiefelhagen, R. 2022. Reference-guided pseudo-label generation for medical semantic segmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 2171–2179.

Shen, Z.; Cao, P.; Yang, H.; Liu, X.; Yang, J.; and Zaiane, O. R. 2023. Co-training with High-Confidence Pseudo Labels for Semi-supervised Medical Image Segmentation. *IJCAI*.

Song, H.; Zhou, Q.; Deng, Q.; Chen, J.; Zhang, L.; Tan, T.; and Guo, R. 2016. Morphologic assessment of the left atrial appendage in patients with atrial fibrillation by gray values–inverted volume-rendered imaging of three-dimensional transesophageal echocardiography: A comparative study with computed tomography. *Journal of the American Society of Echocardiography*, 29(11): 1100–1108.

Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30.

Tu, P.; Huang, Y.; Zheng, F.; He, Z.; Cao, L.; and Shao, L. 2022. Guidedmix-net: Semi-supervised semantic segmentation by using labeled images as reference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 2379–2387.

Wang, Q.; Li, W.; and Gool, L. V. 2019. Semi-supervised learning by augmented distribution alignment. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1466–1475.

Wang, X.; Zhao, K.; Zhang, R.; Ding, S.; Wang, Y.; and Shen, W. 2022. Contrastmask: Contrastive learning to segment every thing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11604–11613.

Wang, Y.; Xiao, B.; Bi, X.; Li, W.; and Gao, X. 2023. MCF: Mutual Correction Framework for Semi-Supervised Medical Image Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15651–15660.

Wu, Y.; Ge, Z.; Zhang, D.; Xu, M.; Zhang, L.; Xia, Y.; and Cai, J. 2022. Mutual consistency learning for semi-supervised medical image segmentation. *Medical Image Analysis*, 81: 102530.

Yang, L.; Qi, L.; Feng, L.; Zhang, W.; and Shi, Y. 2023. Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7236–7246.

Yang, L.; Zhuo, W.; Qi, L.; Shi, Y.; and Gao, Y. 2022. St++: Make self-training work better for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4268–4277.

Yao, H.; Hu, X.; and Li, X. 2022. Enhancing pseudo label quality for semi-supervised domain-generalized medical image segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 3099–3107.

You, C.; Dai, W.; Min, Y.; Staib, L.; and Duncan, J. S. 2023. Bootstrapping semi-supervised medical image segmentation with anatomical-aware contrastive distillation. In *International Conference on Information Processing in Medical Imaging*, 641–653. Springer.

You, C.; Zhou, Y.; Zhao, R.; Staib, L.; and Duncan, J. S. 2022. Simcvd: Simple contrastive voxel-wise representation distillation for semi-supervised medical image segmentation. *IEEE Transactions on Medical Imaging*, 41(9): 2228–2237.

Zhou, Y.; Wang, Y.; Tang, P.; Bai, S.; Shen, W.; Fishman, E.; and Yuille, A. 2019. Semi-supervised 3D abdominal multi-organ segmentation via deep multi-planar co-training. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 121–140. IEEE.

Zhou, Y.; Xu, H.; Zhang, W.; Gao, B.; and Heng, P.-A. 2021. C3-semiseg: Contrastive semi-supervised segmentation via cross-set learning and dynamic class-balancing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7036–7045.

Zhu, X.; Zhang, S.; Hao, H.; and Zhao, Y. 2023. Adversarial-based latent space alignment network for left atrial appendage segmentation in transesophageal echocardiography images. *Frontiers in Cardiovascular Medicine*, 10: 1153053.

Zhu, Y.; Zhang, Z.; Wu, C.; Zhang, Z.; He, T.; Zhang, H.; Manmatha, R.; Li, M.; and Smola, A. J. 2021. Improving semantic segmentation via efficient self-training. *IEEE transactions on pattern analysis and machine intelligence*.