Medical Image Analysis 93 (2024) 103095

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Separated collaborative learning for semi-supervised prostate segmentation with multi-site heterogeneous unlabeled MRI data

Zhe Xu [a], Donghuan Lu [b,*], Jie Luo [c], Yefeng Zheng [b], Raymond Kai-yu Tong [a,*]

[a] Department of Biomedical Engineering, The Chinese University of Hong Kong, Shatin, NT, Hong Kong, China
[b] Tencent Jarvis Research Center, Youtu Lab, Shenzhen, China
[c] Massachusetts General Hospital, Harvard Medical School, Boston, USA

ARTICLE INFO

Keywords:
Prostate segmentation
Semi-supervised learning
Data heterogeneity

ABSTRACT

Segmenting prostate from magnetic resonance imaging (MRI) is a critical procedure in prostate cancer staging and treatment planning. Considering the nature of labeled data scarcity for medical images, semi-supervised learning (SSL) becomes an appealing solution since it can simultaneously exploit limited labeled data and a large amount of unlabeled data. However, SSL relies on the assumption that the unlabeled images are abundant, which may not be satisfied when the local institute has limited image collection capabilities. An intuitive solution is to seek support from other centers to enrich the unlabeled image pool. However, this further introduces data heterogeneity, which can impede SSL that works under identical data distribution with certain model assumptions. Aiming at this under-explored yet valuable scenario, in this work, we propose a separated collaborative learning (SCL) framework for semi-supervised prostate segmentation with multi-site unlabeled MRI data. Specifically, on top of the teacher-student framework, SCL exploits multi-site unlabeled data by: (i) Local learning, which advocates local distribution fitting, including the pseudo label learning that reinforces confirmation of low-entropy easy regions and the cyclic propagated real label learning that leverages class prototypes to regularize the distribution of intra-class features; (ii) External multi-site learning, which aims to robustly mine informative clues from external data, mainly including the local-support category mutual dependence learning, which takes the spirit that mutual information can effectively measure the amount of information shared by two variables even from different domains, and the stability learning under strong adversarial perturbations to enhance robustness to heterogeneity. Extensive experiments on prostate MRI data from six different clinical centers show that our method can effectively generalize SSL on multi-site unlabeled data and significantly outperform other semi-supervised segmentation methods. Besides, we validate the extensibility of our method on the multi-class cardiac MRI segmentation task with data from four different clinical centers.

1. Introduction

Prostate disorder, e.g., cancer, benign hyperplasia and prostatitis, becomes a common problem in male. Especially, prostate cancer is the second leading cause of cancer deaths in American men (about 1 man in 41 will die of prostate cancer), behind only lung cancer.[1] Due to the difficulty in determining the prostate's anatomy, many diagnostic prostate biopsies, e.g., ultrasound-guided approaches, fail to detect the disorder (Bloch et al., 2015). With the superiority of high spatial resolution and soft-tissue contrast, currently, magnetic resonance imaging (MRI) guided methods have become the mainstream, wherein accurate prostate MRI segmentation is an essential pre-processing task for diagnosis and therapy planning of these diseases. Recently, deep learning (DL) based methods have greatly advanced automatic prostate segmentation (Jia et al., 2019; Liu et al., 2020b). However, their high performance usually relies on a large amount of labeled data with a similar distribution. With the nature of labeled data scarcity in medical imaging, such a perfect condition is usually violated in real-world clinical practice, hindering the training of an accurate prostate MRI segmentation model. More specifically, the labeled data scarcity can be further decomposed into two parts: ***label scarcity*** and ***image scarcity***.

Regarding ***label scarcity***, semi-supervised learning (SSL), which can simultaneously utilize a small set of labeled data and abundant unlabeled data with the identical distribution to support model training as shown in Fig. 1(a), has become an appealing approach (Tarvainen

* Corresponding authors.
E-mail addresses: jackxz@link.cuhk.edu.hk (Z. Xu), caleblu@tencent.com (D. Lu), kytong@cuhk.edu.hk (R.K.-y. Tong).

[1] https://www.cancer.org/cancer/prostate-cancer/about/key-statistics.html.

https://doi.org/10.1016/j.media.2024.103095
Received 1 June 2023; Received in revised form 11 September 2023; Accepted 24 January 2024
Available online 26 January 2024

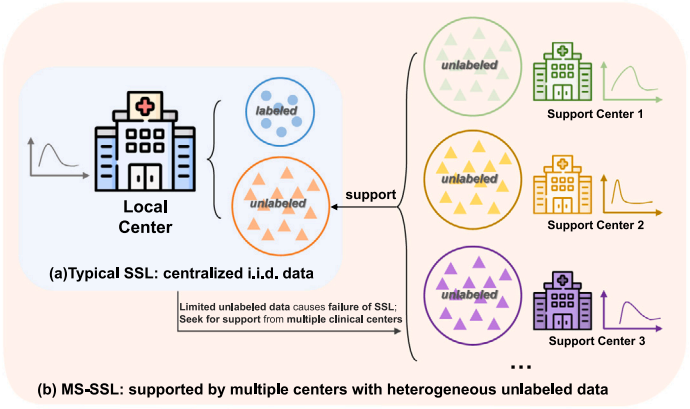

Fig. 1. Comparison between classical semi-supervised learning (SSL) and our focused multi-site semi-supervised learning (MS-SSL).

and Valpola, 2017; Cui et al., 2019; Zhang et al., 2022; Zheng et al., 2020; Xu et al., 2022b; Ouali et al., 2020; Luo et al., 2021; Xu et al., 2023). Despite high performance, the current SSL studies ignore the fact that the effectiveness of their unsupervised part also relies on the condition that the local unlabeled data is abundant. Unfortunately, such condition cannot be satisfied when the local institute has limited image collection capabilities or scarce local patient samples. That is, the unlabeled sample size of the local institute is also very limited, namely ***image scarcity***. As an exploratory study in the scenario of both label and image scarcity, we test some popular SSL methods and the results are presented in Fig. 2(a). As observed, if the amount of local unlabeled data is limited, most of existing methods still exhibit inferior performance when generalize to unseen test data, falling far behind the clinical requirements. What is worse, for some approaches, the limited amount of additional unlabeled data can even interfere with the supervised learning process, leading to obvious performance degradation. A possible reason is that a successful SSL desires for comprehensive data characteristics from unlabeled data as effective unsupervised information (Ben-David et al., 2008; Oliver et al., 2018). In other words, it inherently explores some statistical patterns from unlabeled data (similar to the law of large numbers Hsu and Robbins, 1947) to regularize the primary supervised training and alleviate overfitting to the limited labeled data. Yet, the restricted image collection ability of the local clinical center may lead to limited prostate variations of contrast, shape and texture, and even low image quality (Li and Zhou, 2014), hindering effective semi-supervised training.

Considering the high scanning costs and limited local patient samples, immediately enlarging the high-quality image pool in the local center is always infeasible. Thus, an intuitive way of tackling the *image scarcity* dilemma is to seek support from other clinical centers or public databases to enrich the unlabeled image pool. Note that some external sites with enough budgets may be willing to provide labels, but here we consider the general cooperation scenario with the lowest barriers of entry, i.e., only images are required. Although collecting multi-site images can provide more high-quality unlabeled data and comprehensive data characteristics to present the prostate of interest, this further poses a new notorious problem, i.e., data heterogeneity due to different scanners, scanning protocols and subject groups, as exampled in Table 1 and Fig. 5. Unfortunately, the existing SSL approaches typically rely on identical data distribution under certain model assumptions such as low-density assumption and manifold assumption (Van Engelen and Hoos, 2020). There is no specialized mechanism to tackle this problem within these methods. Different data distributions will violate the assumptions to some extent and result in decreased feature discriminability, thus with limited performance or even degradation, as shown in Fig. 2(b) and Section 4. In this regard, generalizing SSL on multiple sources of unlabeled data is not trivial but

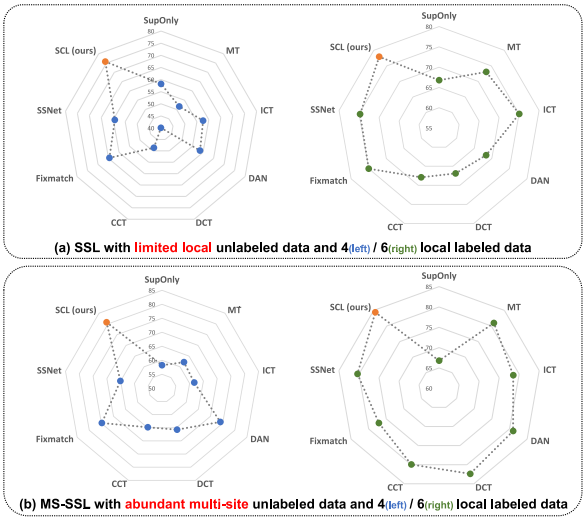

Fig. 2. (a) Prostate segmentation results of SSL with limited unlabeled data from the local center (C1 as listed in Table 1; 18 training scans in total; labeled/unlabeled: 4/14 scans (left) and 6/12 scans (right)); (b) Prostate segmentation results of MS-SSL with abundant unlabeled data from multiple clinical sites (C1 as local center, C2-C6 as external centers; labeled/unlabeled: 4/100 scans (left) and 6/98 scans (right)). Metric: Dice score (%).

of high clinical need. Proper mechanisms are called for this practical yet challenging SSL scenario.

Formally, we define this new SSL scenario as multi-site semi-supervised learning (MS-SSL), as depicted in Fig. 1. Being an under-explored setting itself, few efforts have been made, considering that this scenario seldom exists in classical computer vision because the unlabeled intra-domain natural images are usually easily accessible. To our best knowledge, in the medical imaging domain, the most relevant work is AHDC (Chen et al., 2021a). Yet, it simplifies the scenario into dual-domain semi-supervised learning, i.e., the additional unlabeled data is from a specific site rather than multiple arbitrary sites. Thus, AHDC intuitively devotes to image-to-image mapping to achieve dual-distribution alignment, which limits its application scope since simultaneously aligning multiple external domains to the local domain is challenging for a single image mapping network.

In this work, we propose a more generalized framework called Separated Collaborative Learning (SCL), as shown in Fig. 3, to achieve robust multi-site semi-supervised learning with application in prostate MRI segmentation. Specifically, SCL is built upon the popular teacher-student architecture (Cui et al., 2019). The design of separate learning

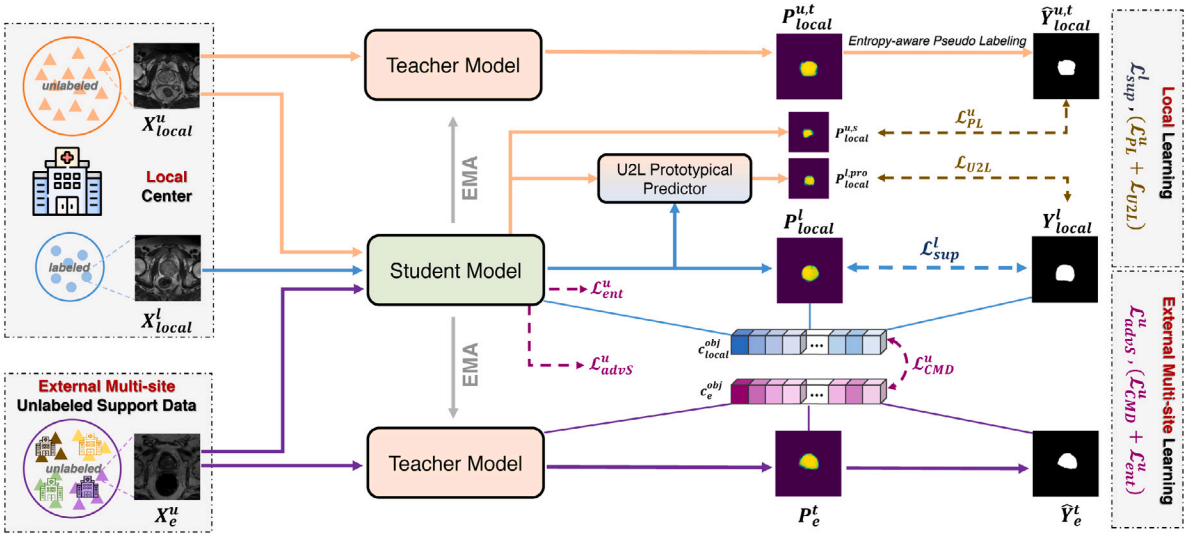

Fig. 3. Overview of our separated collaborative learning framework for multi-site semi-supervised prostate segmentation. The learning process is separated into local learning (responsible for accurate local distribution fitting) and external multi-site learning (robustly mining informative clues from external data to support local learning). The two teacher models are identical. U2L: unlabeled-to-labeled. EMA: exponential moving average.

is motivated by the observation that supervised learning plays an important role in distribution fitting (the main reason of why the recent CPS (Chen et al., 2021b) cannot handle MS-SSL, as detailed in Section 4.2). Enhancing such local distribution fitting (i.e., increasing the local labeled data) can also help existing SSL methods resist heterogeneity of multi-site data (as shown in Fig. 2(b)). Thus, the learning for local and external unlabeled data is separately performed in two tailored manners: **(i) Local learning**, advocating supervised-like label learning to fit local distribution better. It includes two key components: pseudo label learning, which helps reinforce confirmation of low-entropy easy regions; and cyclic propagated real label learning, which leverages class prototypes to perform non-parametric unlabeled-to-labeled prediction. The latter cyclic process serves to regularize the distribution of intra-class embeddings, benefiting former label propagation. Besides, a regional context loss is adopted for better perceiving the local mismatch in such a supervised-like process. **(ii) External multi-site learning**, which appreciates distribution-insensitive relationship modeling on region-of-interest between local labeled data and external unlabeled data from arbitrary sites and model behavior regularization. More specifically, it initially incorporates local-support category mutual dependence learning in combination with entropy minimization. This design is primarily motivated by the effectiveness of mutual information in measuring the amount of information shared by two variables (e.g., here are prostate prototypes) even from different domains (Shi and Sha, 2012). Subsequently, the stability learning under strong adversarial perturbations is introduced to enhance the model's robustness to image heterogeneity. The local and external multi-site learning complement each other in an interactive end-to-end learning scheme. Local learning is responsible for accurate local distribution fitting, while external learning robustly mines informative clues from external data so that we can achieve the goal of supporting the local (target) center to develop an accurate segmentation system.

Overall, the main contributions of this work are as follows:

- The typical SSL addresses label scarcity, ideally assuming an abundance of independent and identically distributed unlabeled local data, while neglecting a scenario of unlabeled image scarcity in local clinical centers. To overcome this, we, for the first time, present a new and practical scenario of multi-site semi-supervised learning (MS-SSL), which allows the enrichment of the unlabeled pool with heterogeneous unlabeled data from multiple arbitrary sites. This can support the development of semi-supervised segmentation models in local centers with scarce local patients or restricted image collection capabilities.
- Tailoring for this under-explored MS-SSL scenario, we propose a novel solution called Separated Collaborative Learning (SCL). Recognizing the different efficacy of local and external unlabeled data, we introduce two customized learning manners: local learning and external multi-site learning. These separated manners also work collaboratively to effectively exploit the informative clues present in multi-site unlabeled data.
- We propose a novel local-support category mutual dependence learning scheme. This scheme advocates mutual information-based distribution-insensitive relationship modeling on region-of-interests. Its goal is to facilitate effective and robust collaboration between local labeled data and heterogeneous external unlabeled data, thereby supporting local learning.
- Our method is extensively evaluated on public prostate MRI datasets from six different institutes with varying scanning protocols and patient demographics. The experimental results demonstrate the superiority of our approach in MS-SSL over the existing semi-supervised segmentation approaches. We also validate the extensibility of our method on the multi-class cardiac MRI segmentation task with data from four different clinical centers.

2. Related work

2.1. Semi-supervised medical image segmentation

Semi-supervised learning is actively studied because it can effectively alleviate the expertise-demanding annotation burden for medical images. It is nowadays widely recognized that most semi-supervised segmentation methods feature two schemes: pseudo labeling (Chen et al., 2021b; Zhang et al., 2022; Zheng et al., 2020; Rizve et al., 2021; Zhang et al., 2021) and consistency regularization (Samuli and Timo, 2017; Tarvainen and Valpola, 2017; Cui et al., 2019; Luo et al., 2021; Ouali et al., 2020; Yu et al., 2019; Xu et al., 2022b,a; Miyato et al., 2018). Pseudo labeling intuitively generates pseudo labels for unlabeled images and then mixes them with the original labeled data for further training. However, when simultaneously confronting extreme label scarcity and image scarcity, the data characteristics are limited, easily resulting in poor pseudo labels for biased training. Instead, consistency regularization is based on the smoothness assumption (Van Engelen and Hoos, 2020). For example, the Π model (Samuli and Timo, 2017) encourages consistency of multiple forward predictions under different input-level perturbations. The temporal ensembling strategy (Samuli

and Timo, 2017) improves the Π model by utilizing exponential moving average (EMA) predictions for unlabeled data as the consistency targets. As a computation-efficient improvement, the mean teacher (MT) model (Tarvainen and Valpola, 2017; Cui et al., 2019) introduces a self-ensembling teacher model updated by the EMA weights of the student model and exploits unlabeled data by encouraging consistency of the student's and teacher's predictions. Based on the MT model, several enhanced consistency-based methods were proposed, e.g., auxiliary branches for multi-task consistency (Luo et al., 2021), uncertainty to select reliable consistency targets (Yu et al., 2019), mutual information maximization to encourage invariant image representation under geometric transformations (Peng et al., 2020), contrastive learning for class separation (Wu et al., 2022), local distributional smoothness (Miyato et al., 2018) and cross-consistency training (Ouali et al., 2020). Besides the above two directions, other techniques have also been explored. For example, adversarial learning can be used to encourage the predicted segmentation of unlabeled data to be closer to that of labeled data (Zhang et al., 2017; Sedai et al., 2017). Entropy minimization (Grandvalet and Bengio, 2004) enforces low-entropy predictions of the unlabeled data. SSNet (Wu et al., 2022) encourages pixel-level smoothness and inter-class separation of the unlabeled data.

Despite good performance, the current SSL methods ignore the fact that their effectiveness also relies on the condition that the local unlabeled data is abundant. Unfortunately, such condition cannot be satisfied when the local institute has limited image collection capabilities, leading to the image scarcity problem. However, when the unlabeled image pool is enlarged with support from other clinical centers, the current methods have no specific mechanism to deal with the distribution shift problem, resulting in unstable performance. Regarding **using unlabeled data from external sites**, the most relevant work AHDC (Chen et al., 2021a) introduces image mapping networks to perform image-level dual-distribution alignment. However, it limits the application to the dual-domain scenario. We experimentally found that it struggles with multi-site data.

2.2. Domain adaptation and federated semi-supervised learning

While there are significant differences from our MS-SSL scenario, here, we also briefly discuss two research areas, domain adaptation and semi-supervised federated learning, to help readers in distinguishing between these contexts. Firstly, though **domain adaptation (DA)** can handle domain shift to certain extent, our scenario is inherently different, leading to essentially disparate methodological motivations. For example, unsupervised DA (UDA) (Xie et al., 2022; Zhao et al., 2022) relies on good knowledge transfer from abundant labeled source support data to unlabeled target data. Yet, MS-SSL can be regarded as having abundant unlabeled source support data and limited labeled target data, so the conditions of DA methods cannot be satisfied. Secondly, **federated semi-supervised learning (FSSL)** (Qiu et al., 2023; Yang et al., 2021) extends traditional federated supervised learning and focuses on label-efficient, privacy-preserving collaborative learning. In FSSL, some clients provide labeled data, while others with no annotation budget are also allowed to participate in optimizing the global model. While our MS-SSL shares a similar spirit of multi-site collaborative learning with FSSL, their objectives substantially differ. FSSL aims to train a superior global model through privacy-protected collaborative learning among clients, tailored to serve each individual client. In contrast, MS-SSL employs external partner institutes as support sites to address the image scarcity issue at the local center, enhancing local semi-supervised model training. Additionally, FSSL emphasizes data privacy and only allows model weight sharing, while MS-SSL involves centralized multi-site data, which is a practical paradigm within small-scale partner institutes. The decentralized nature of FSSL leads to fundamentally different and distinctive convergence challenges compared to MS-SSL. In this work, we aim to tailor a method for MS-SSL, enabling small clinical centers with limited image collection capabilities to perform cooperative and robust SSL. Considering these divergent focused scenarios, methodologically comparing MS-SSL with DA and FSSL is challenging.

3. Methodology

3.1. Preliminaries

We denote the local target dataset as $\mathcal{D}_{local}$ and the external unlabeled support datasets as $\mathcal{D}_e = \bigcup_{j=1}^{m} \mathcal{D}_{ej}$, where m is the number of support sites. The $\mathcal{D}_{local}$ contains both labeled data $\mathcal{D}^l_{local}$ and unlabeled data $\mathcal{D}^u_{local}$, wherein $\mathcal{D}^l_{local} = \{(X^l_{local(i)}, Y^l_{local(i)})\}_{i=1}^{n_l}$ with n_l labeled samples and $\mathcal{D}^u_{local} = \{X^u_{local(i)}\}_{i=n_l+1}^{n_l+n_u}$ with n_u unlabeled samples. $X^l_{local(i)}, X^u_{local(i)} \in \mathbb{R}^{H\times W\times D}$ represent the scans with height H, width W, and depth D, and $Y^l_{local(i)} \in \{0,1\}^{H\times W\times D}$ denotes the label of $X^l_{local(i)}$ (we focus on the binary segmentation). Due to large and greatly varying thickness of prostate MRI scans, our experiments are performed in 2D slices extracted from these scans. Therefore, for the upcoming content, we will refer to pixels instead of voxels. The external clinical centers only provide unlabeled data. Therefore, the jth external support dataset is further denoted as $\mathcal{D}_{ej} = \{X^u_{ej(i)}\}_{i=1}^{n_j}$ with n_j unlabeled samples. In summary, the goal of our MS-SSL is to boost the performance in the local center with the help of a local labeled set $\mathcal{D}^l_{local}$, a local unlabeled set $\mathcal{D}^u_{local}$ and multiple external unlabeled sets $\mathcal{D}_e = \bigcup_{j=1}^{m} \mathcal{D}_{ej}$.

3.2. Teacher-student framework

As shown in Fig. 3, we adopt the efficient teacher-student model (Tarvainen and Valpola, 2017) as our backbone framework. Here, we denote θ and $\tilde{\theta}$ as the weights of the student model and the teacher model, respectively. The teacher's weights $\tilde{\theta}_t$ at the training step t are updated by exponential moving average (EMA) of the student's weights θ_t, formulated as $\tilde{\theta}_t = \alpha\tilde{\theta}_{t-1} + (1-\alpha)\theta_t$, where α is the EMA decay rate and empirically set to 0.99 (Yu et al., 2019). By such, the teacher model performs self-ensembling by nature, which can help produce relatively stable pseudo labels (Tarvainen and Valpola, 2017) required by the calculation of losses $\mathcal{L}^u_{PL}$, $\mathcal{L}^u_{U2L}$ and $\mathcal{L}^u_{CMD}$ (as detailed later). The overall objective of our framework can be described as:

$$\min_{\theta}\left[\mathcal{L}^l_{sup}\left(\theta, \mathcal{D}^l_{local}\right) + \lambda\mathcal{L}^u(\theta, \tilde{\theta}, \mathcal{D}^u_{local}, \mathcal{D}_e)\right], \tag{1}$$

where $\mathcal{L}^l_{sup}$ is the supervised loss for $\mathcal{D}^l_{local}$; $\mathcal{L}^u$ is the unsupervised loss for $\mathcal{D}^u_{local}$ and $\mathcal{D}_e$; λ is a ramp-up weight scheduled by the time-dependent Gaussian function $\lambda(t) = \lambda_{max}\cdot e^{-5\left(1-\frac{t}{t_{max}}\right)^2}$ (Yu et al., 2019) to avoid the domination by meaningless guidance at the early training stage, where w_{max} and $\lambda_{\max}$ are the maximal weight and training step, respectively.

3.3. Local learning

3.3.1. Supervised learning for local labeled data

The entire supervised loss $\mathcal{L}^l_{sup}$ for local labeled data is a combination of the widely-used cross-entropy loss $\mathcal{L}_{CE}(P^l_{local}, Y^l_{local})$ and the Dice loss, where P^l_{local} represents the softmax prediction of X^l_{local}. Unlike the cross-entropy loss, which regards each pixel as an independent sample, Dice loss considers the information of neighboring pixels. Inspired by Liu et al. (2022), the Dice loss is extended to a regional manner to help the model better perceive the local mismatch for fine-grained information learning: divide the image equally into K non-overlapping regions of approximately equal size and individually estimate the Dice loss for each region. Then, their mean is the final loss, namely regional context loss $\mathcal{L}^l_{RC}$, formulated as $\mathcal{L}^l_{RC} = \frac{1}{K}\sum_{k=1}^{K}\mathcal{L}_{\text{Dice}}\left(P^{l,k}_{local}, Y^{l,k}_{local}\right)$. Thus, the supervised loss can be formulated as:

$$\mathcal{L}^l_{sup} = \frac{1}{2}\left[\mathcal{L}_{CE}(P^l_{local}, Y^l_{local}) + \mathcal{L}^l_{RC}(P^l_{local}, Y^l_{local})\right]. \tag{2}$$

3.3.2. Pseudo label learning for local unlabeled data

The EMA teacher model is adopted to generate pseudo labels for local unlabeled data because its self-ensembling nature can avoid sharp deterioration of the label quality. Specifically, denoting the prediction of the local unlabeled image X^u_{local} from the teacher model as $P^{u,t}_{local}$, the pseudo label corresponds to the class with the maximal posterior probability, i.e., $\hat{Y}^{u,t}_{local} = \arg\max_c(P^{u,t,c}_{local})$, where c denotes the semantic class. With limited local labeled data, it is hard for the model to produce confident pseudo labels. Besides, the high-disorder regions are usually ambiguous in nature. To avoid severe ambiguity propagation, we reinforce confirmation of low-entropy easy regions. Denoting the entropy of each pixel as $H^{(h,w)}_{P^{u,t}_{local}} = \frac{-1}{\log C}\sum_{c=1}^{C} P^{u,t(h,w,c)}_{local} \log P^{u,t(h,w,c)}_{local}$, if $H^{(h,w)}_{P^{u,t}_{local}} \leq \delta$, where δ is the ramp-down threshold ranging from $\frac{3}{4}$ to $\frac{1}{4}$ due to the fact that the model disorder will be decreased as training goes, this pixel will be included in the loss calculation. Observing that the cross-entropy loss is more sensitive to label noises and foreground-background imbalance (Wang et al., 2020) (many high-entropy pixels have been filtered), we only adopt the Dice loss with the same regional form as presented in Section 3.3.1 to impose pseudo supervision, formulated as:

$$\mathcal{L}^u_{PL} = \frac{1}{K}\sum_{k=1}^{K}\mathcal{L}_{\text{Dice}}\left(P^{u,s,k}_{local}, \hat{Y}^{u,t,k}_{local}\right). \tag{3}$$

3.3.3. Cyclic propagated real label learning for local unlabeled data

The above pseudo label learning helps reinforce confirmation of basic but easy knowledge (i.e., with the most similar patterns to labeled data). Inherently, pseudo labeling can be regarded as pixel-level label propagation from labeled data to unlabeled data, while the difficulty of pixel-level label propagation is caused by intra-class variation. In this regard, we introduce a proxy of cyclic propagated real label learning, i.e., unlabeled-to-labeled (U2L) prototypical predictor in Fig. 3, to regularize the distribution of intra-class features from local unlabeled data to facilitate pixel-level label propagation. We also provide the detailed illustration in Fig. 4 to help understand. This proxy is technically inspired by prototypical networks (Snell et al., 2017), which employs class prototypes (i.e., representative feature centroids that capture the essential characteristics of specific object classes within the image) to classify instead of a parameterized classifier, enabling explicit label propagation across images in few-shot learning. Since the cyclic process ultimately relies on real expert label supervision, it avoids error accumulation caused by pseudo labels. Intuitively, an accurate prototypical prediction requires both compact features and discriminative prototypes, thus improving such prediction can reduce intra-class variance. Specifically, we denote the feature map from the layer before the penultimate convolution in the student model as $F^{l,s}_{local}$ (for labeled data) and $F^{u,s}_{local}$ (for unlabeled data). Note that $F^{l,s}_{local}$ and $F^{u,s}_{local}$ are upsampled to the same size of images via bilinear interpolation. We use such features from the decoder because they are progressively refined and semantically more meaningful for the dense segmentation task. Assisted by the argmax pseudo label $\hat{Y}^{u,t}_{local}$ from the teacher, the object prototype from local unlabeled data can be estimated via masked average pooling:

$$c^{u,obj}_{local} = \frac{\sum_v \left[\hat{Y}^{u,t,obj}_{local(v)} \cdot P^{u,t,obj}_{local(v)} \cdot F^{u,s}_{local(v)}\right]}{\sum_v \left[\hat{Y}^{u,t,obj}_{local(v)} \cdot P^{u,t,obj}_{local(v)}\right]}, \tag{4}$$

where the probability $P^{u,t,obj}_{local(v)}$ weights each pixel v to the prototype generation. The background prototype $c^{u,bg}_{local}$ of unlabeled data is obtained in the same way. Then, we compare the features of labeled data $F^{l,s}_{local}$ with $c^{u,bg}_{local}$ and $c^{u,obj}_{local}$ to obtain the prototype-based predictions $P^{l,pro}_{local}$ for labeled data:

$$P^{l,pro}_{local} = \frac{\exp\left(sim(F^{l,s}_{local}, c^{u,i}_{local})/T\right)}{\sum_{i\in\{obj,bg\}}\exp\left(sim(F^{l,s}_{local}, c^{u,i}_{local})/T\right)}, \tag{5}$$

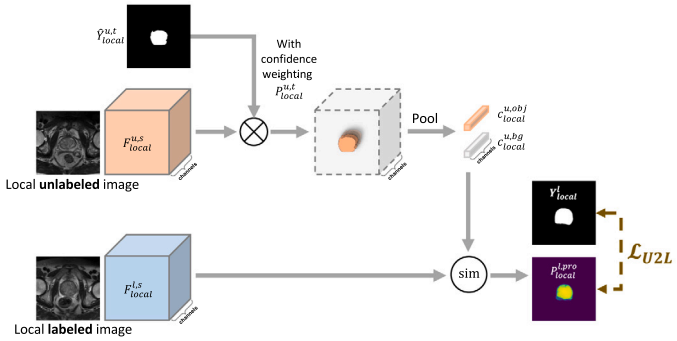

Fig. 4. Detailed illustration of the U2L prototypical predictor.

where we adopt cosine similarity for $sim(\cdot,\cdot)$ and empirically set the temperature T to 0.05 (Wang et al., 2019). As such, $P^{l,pro}_{local}$ can be supervised by the real expert label Y^l_{local} as:

$$\mathcal{L}_{U2L} = \frac{1}{K}\sum_{k=1}^{K}\mathcal{L}_{\text{Dice}}\left(P^{l,pro}_{local}, Y^l_{local}\right). \tag{6}$$

3.4. External multi-site learning

3.4.1. Maximizing local-support category mutual dependence

To exploit effective support from external data, we desire to maximize the similarity in feature space of prostate region between local and external data. Note that the background is excluded here to avoid distraction by irrelevant contexts. Due to data heterogeneity, we cannot assume the relationship between embeddings of the prostate of different sources is linear. Hence, inspired by that maximizing mutual information (MI), a concept in information theory, between aligned images and fixed images becomes a common paradigm in cross-modality image registration (Pluim et al., 2003; Viola and Wells, 1997), here, we advocate MI for distribution-insensitive relationship modeling on region-of-interests (ROIs) between local and external learning. MI can effectively measure the amount of information one distribution gives about another, even from different domains (Shi and Sha, 2012). Considering the variations in prostate structures across different centers and patients, we utilize feature prototypes of the prostate regions as ROIs. Here, same as the feature prototype generation in Section 3.3.3, we first obtain the object prototype from the external data c^{obj}_e on the fly, assisted by its argmax pseudo label $\hat{Y}^t_e$ and features F^t_e from the teacher model. Similarly, the object prototype from the local data c^{obj}_{local} can be obtained. Note that c^{obj}_{local} is extracted from the local labeled data because the expert labels can accurately mask the ROIs, avoiding error propagation caused by pseudo labels. EMA strategy is applied across training steps to alleviate the impact of unbalanced sampling of slices containing the prostate. Despite the improvement of approximate estimation for the lower or upper bound of MI, e.g., MINE (Belghazi et al., 2018; Meng et al., 2020), we empirically found that they are unstable during training. Since we use the features from the layer before the penultimate convolution in the decoder, after masked average pooling over all pixels of the same class, the class prototypes are not highly dimensional. Thus, we resort to classical Parzen windowing-based non-parametric explicit estimation (Pluim et al., 2003; Xu et al., 2008; Kwak and Choi, 2002) to estimate MI. Yet, our method is robust to different MI estimators (Section 4.4.4). For better readability, we give c^{obj}_{local} and c^{obj}_e two notations A and B, respectively. The MI between A and B is defined as:

$$\begin{aligned} I(A;B) &= \iint_{a,b} p_{(A,B)}(a,b)\log\left(\frac{p_{(A,B)}(a,b)}{p_A(a)p_B(b)}\right) \\ &\approx \sum_{a,b} p_{(A,B)}(a,b)\log\frac{p_{(A,B)}(a,b)}{p_A(a)p_B(b)}, \end{aligned} \tag{7}$$

where $p_{(A,B)}$ is the joint probability density function, and p_A and p_B are marginal probability density functions of A and B, respectively. These probabilities can be obtained by constructing a histogram of elements for each object prototype. Assuming each element should contribute continuously to a range of histogram bins instead of contributing only to the bin it falls into, we can use the widely-used Parzen windowing (Pluim et al., 2003) to estimate p to enable the approximation to be differentiable (Xu et al., 2008; Kwak and Choi, 2002; Guo, 2019). Similar to Guo (2019) and Xu et al. (2008), given a set S of n samples, each sample s contributes to $p(x)$ with a function of its distance to x: $p_S(x) = \frac{1}{n}\sum_{s\in S} W(x-s)$, where Gaussian function is widely used as the weighting function W (Pluim et al., 2003; Viola and Wells, 1997; Kwak and Choi, 2002). As such, we can calculate $p_A(x)$ and $p_B(x)$. Similarly, the joint distribution estimation can be achieved by: $p_{A,B}(x,y) = \frac{1}{n}\sum_{(a,b)\in(A,B)} W(x-a)W(y-b)$. Besides, the number of bins in the histogram of each prototype is denoted as k_b, which is empirically set to 8. $p_A(x)$ and $p_B(x)$ are estimated at k_b equally-spaced bin centers, and $p_{A,B}(x,y)$ at ${k_b}^2$ pairs of bin centers. Then, we use negative mutual information as the category mutual dependence (CMD) loss:

$$\mathcal{L}^u_{CMD} = -I(A;B) = -I(c^{obj}_{local}; c^{obj}_e). \tag{8}$$

Optimizing $\mathcal{L}^u_{CMD}$ encourages more shared information between local and external learning in a distribution-insensitive scheme, allowing the model to exploit informative clues of ROIs from other distribution-diverse sites. Additionally, since calculating c^{obj}_e requires the argmax pseudo label, we minimize the global predicted entropy of the student model to help reduce the ambiguity in generating $\hat{Y}^t_e$. For each pixel, its normalized entropy is:

$$H^{(h,w)}_{P^s_e} = \frac{-1}{\log C}\sum_{c=1}^{C} P_e^{s(h,w,c)} \log P_e^{s(h,w,c)}, \tag{9}$$

where C is class number. The entropy loss $\mathcal{L}^u_{ent}$ is defined as the mean normalized entropy of all pixels. $\mathcal{L}^u_{ent}$ is combined with $\mathcal{L}^u_{CMD}$ using equal weights, i.e., $\frac{1}{2}(\mathcal{L}^u_{CMD} + \mathcal{L}^u_{ent})$.

3.4.2. Adversarial stability learning

Encouraging prediction stability under perturbations, which takes the smoothness assumption (Van Engelen and Hoos, 2020), becomes an effective schema in typical SSL. This learning manner can still suit MS-SSL to make the model isotropically smooth locally around each input data point, which can enhance robustness to data heterogeneity. As suggested by Goodfellow et al. (2015), the adversarial direction is defined as the direction in the input space to which the label probability of the model is most sensitive, where smoothing the model in the adversarial (most anisotropic) direction can better promote local isotropy compared to isotropic smoothing. Thus, different from the common practice (e.g., MT Tarvainen and Valpola, 2017) that uses input-independent random Gaussian noise (i.e., isotropic smoothing), we aim to resolve the complex anisotropic sensitivity of each data point by introducing adversarial pixel-level perturbation (Goodfellow et al., 2015) which can naturally simulate the heterogeneity at each point to some extent. Specifically, since the ground truth is not accessible for external data, the adversarial perturbation is performed in a virtual manner, i.e., using pseudo labels. Inspired by Miyato et al. (2018), the pixel-level adversarial perturbation $r_{adv} \in R_{adv}$ can be approximated by:

$$r_{adv} \approx \epsilon \frac{g}{\|g\|_2}, g = \nabla_r \mathcal{D}\left[p(\hat{y}^s_e \mid x^u_e), p\left(y^s_e \mid x^u_e + r\right)\right], \tag{10}$$

where the gradient g denotes the fastest-changing direction of the metric $\mathcal{D}$, which is used as the direction of adversarial perturbation r_{adv} to perturb the original x^u_e and can be efficiently computed via back-propagation; ϵ is the magnitude; and r can be set as a random noise vector. We adopt the soft Dice loss as $\mathcal{D}$, and thus the adversarial stability loss $\mathcal{L}_{advS}$ is:

$$\mathcal{L}_{advS} = 1 - \frac{2\|p\left(\hat{Y}^s_e \mid X^u_e\right) \cap p\left(Y^s_e \mid X^u_e + R_{adv}\right)\|}{\|p\left(\hat{Y}^s_e \mid X^u_e\right)\| + \left\|p\left(Y^s_e \mid X^u_e + R_{adv}\right)\right\|}. \tag{11}$$

In this way, we encourage the model stability against the adversarial perturbation which can most greatly alter the output distribution and thus enhance robustness to data heterogeneity.

3.5. Final loss for unlabeled data

In summary, the final loss for the multi-site unlabeled data (i.e., $\mathcal{L}^u$ in Eq. (1)) can be formulated as:

$$\mathcal{L}^u = \underbrace{\frac{1}{2}\left(\mathcal{L}^u_{PL} + \mathcal{L}^u_{U2L}\right)}_{\text{Local learning}} + \underbrace{\frac{1}{2}\left(\mathcal{L}^u_{CMD} + \mathcal{L}^u_{ent}\right) + \mathcal{L}^u_{advS}}_{\text{External Multi-site Learning}}. \tag{12}$$

4. Experiments

4.1. Datasets and experimental setup

4.1.1. Datasets

To evaluate our method, prostate T2-weighted MR images from six different clinical centers (C1-6) are collected for a retrospective study. The prostate MR images from different centers usually present severe heterogeneity (manifested as different appearance characteristics) mainly due to the differences in scanners, field strengths, coil types, disease and other acquisition protocols (especially in-plane and through-plane resolution). Additionally, signal intensity values are not standardized. Similar to Liu et al. (2020a), the characteristics of the six data sources are summarized in Table 1. Besides, a central slice of the randomly picked scan from each center is shown in Fig. 5 to show the appearance differences.

More specifically, the samples of C1 and C2 are from the same NCI-ISBI 2013 challenge (Bloch et al., 2015) but from two different institutes, i.e., the Radboud University Nijmegen Medical Centre (RUNMC) in the Netherlands (C1, 30 samples) and Boston Medical Center (BMC) in the USA (C2, 30 samples). The 19 samples of C3 are collected from the Hospital Center Regional University of Dijon-Bourgogne (HCRUDB) included in the Initiative for Collaborative Computer Vision Benchmarking (I2CVB) dataset (Lemaître et al., 2015). The samples of C4, C5 and C6 are from the same Prostate MR Image Segmentation 2012 (PROMISE12) dataset (Litjens et al., 2014) but collected from three different centers, i.e., University College London (UCL) in the United Kingdom (C4, 13 samples), the Beth Israel Deaconess Medical Center (BIDMC) in the USA (C5, 12 samples) and Haukeland University Hospital (HK) in Norway (C6, 12 samples). Note that compared with C1 and C2, scans from the other four centers are acquired from patients with prostate cancer, either for detection or staging purpose. Yet, the specific stage of the patient and the location of prostate cancer are not disclosed. Besides the differences caused by scanning, such disease usually leads to an inherent semantic difference in the prostate area and further aggravates the distribution shift.

4.1.2. Pre-processing and data partitioning

For image pre-processing, following Liu et al. (2020a), we first center-crop the MR images from C3 with the roughly same view in the axial plane as images from other centers because the original scans of C3 were acquired from the whole body instead of the prostate area. Then, each image is resized to 384 × 384 pixels in the axial plane and the image intensity is normalized to zero mean and unit variance. Note that the inter-site heterogeneity of prostate MRI is hard to address by pre-processing techniques alone, which is extensively investigated in Liu et al. (2020b), because it comes from not only intensity variance but also other factors (e.g., coil used and different resolutions). For data partitioning, we regard C1 and C2 as two local (target) institutes. Specifically, when C1 is our local institute, we randomly separate data from C1 into 18 (60%), 3 (10%), and 9 (30%) cases as the training set, validation set and test set, respectively. Within C1, the training data will be further split into a small subset of labeled data and the remaining data will serve as an unlabeled set. Then, all the images from

Table 1
Details of the acquisition protocols and number of scans for the different centers. Each center supplied T2-weighted MR images of the prostate.

Center	Source	# Scans	Field strength (T)	Resolution (in-plane/through-plane in mm)	Coil	Scanner
C1	Radboud University Nijmegen Medical Centre (RUNMC) (Bloch et al., 2015)	30	3	0.6–0.625/3.6–4	Surface	Siemens
C2	Boston Medical Center (BMC) (Bloch et al., 2015)	30	1.5	0.4/3	Endorectal	Philips
C3	Hospital Center Regional University of Dijon-Bourgogne (HCRUDB) (Lemaître et al., 2015)	19	3	0.67–0.79/1.25	–	Siemens
C4	University College London (UCL) (Litjens et al., 2014)	13	1.5 and 3	0.325-0.625/3-3.6	–	Siemens
C5	Beth Israel Deaconess Medical Center (BIDMC) (Litjens et al., 2014)	12	3	0.25/2.2–3	Endorectal	GE
C6	Haukeland University Hospital (HK) (Litjens et al., 2014)	12	1.5	0.625/3.6	Endorectal	Siemens

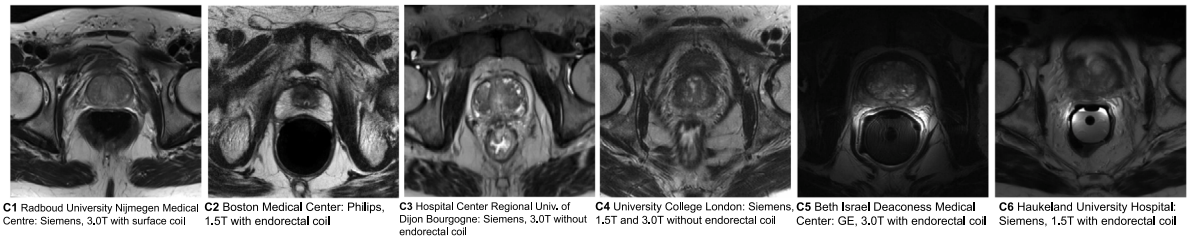

Fig. 5. Slices of the normalized prostate MR images from six different clinical centers to qualitatively show the appearance differences.

C2 to C6 serve as the external unlabeled data in conjunction with the unlabeled data from C1 to support the prostate segmentation in C1. Similarly, when C2 is our local institute, the data from C2 are split into 18 (60%), 3 (10%), and 9 (30%) cases as the training set, validation set and test set, respectively. The training data of C2 will be similarly split into a labeled subset and an unlabeled subset. Then, all the images from C1, C3, C4, C5 and C6 serve as the external unlabeled data in conjunction with the unlabeled data from C2 to support the prostate segmentation in C2.

4.1.3. Baseline approaches

We compare our method with the supervised-only (SupOnly) baselines and state-of-the-art semi-supervised medical image segmentation methods including: mean-teacher self-ensembling model (MT) (Cui et al., 2019), uncertainty-aware MT (UA-MT) (Yu et al., 2019), entropy minimization approach (EM) (Grandvalet and Bengio, 2004), interpolation consistency training (ICT) (Verma et al., 2022), deep adversarial network (DAN) (Zhang et al., 2017), cross-consistency training (CCT) (Ouali et al., 2020), deep co-training (DCT) (Qiao et al., 2018), FixMatch (Sohn et al., 2020), virtual adversarial training (VAT) (Miyato et al., 2018), cross pseudo supervision (CPS) (Chen et al., 2021b), SSNet (Wu et al., 2022) and AHDC (Chen et al., 2021a). All the methods are implemented with the same backbone and training protocols to ensure fairness.

4.1.4. Implementation and evaluation metrics

The framework is implemented on PyTorch with an NVIDIA GeForce RTX 3090 GPU. Considering the large variance in slice thickness among different clinical centers, we adopt the 2D architecture. Specifically, the commonly used 2D U-Net model (Ronneberger et al., 2015) (implemented in Xu et al., 2022a) is adopted as our backbone. The input patch size is set to 384 × 384 pixels, and the batch size is set to 36 including 12 labeled local slices, 12 unlabeled local slices and 12 unlabeled external slices. The maximum consistency weight λ_{max} is empirically set to 0.1 (Yu et al., 2019). The maximal training step $t_{\max}$ is set to 30,000. The magnitude ϵ is set to 6 (Wu et al., 2022). The number of regions K is set to 4, recommended by Liu et al. (2022). The channel number of the feature map from the layer before the penultimate convolution in the decoder is 16. The network is trained using the SGD optimizer (weight decay = 0.0001, momentum = 0.9). The learning rate is initialized as 0.01 and decayed by multiplication with $\left(1.0 - t/t_{\max}\right)^{0.9}$. Data augmentation, including randomly flip and rotation, is applied. For a fair comparison, no extra post-processing or ensemble methods are utilized. We adopt four common metrics for a comprehensive evaluation, including region-based metrics (i.e., Dice score and Jaccard) and surface-based metrics (i.e., average surface distance (ASD) and the 95-th Hausdorff distance (95HD)). The results are the average over three runs with different seeds.

4.2. Comparative experiments of C1 as the local center

Table 2 presents the quantitative results of different methods where C1 is our local target institute. The Supervised (joint) model indicates the model trained with all labeled data from the six sites. Compared to supervised-only baselines, our method achieves consistent improvements in terms of the four metrics, demonstrating that our method can effectively exploit multi-site unlabeled data. For example, the performance gain of our SCL reaches +16.09% and +18.11% Dice improvements under the settings of four and six local labeled scans, respectively. Further, it is observed that some of the recent SSL methods, e.g., MT (Tarvainen and Valpola, 2017), ICT (Verma et al., 2022) and CPS (Chen et al., 2021b), achieve worse results compared to the supervised-only baselines under the setting with four labeled scans. Especially, CPS (Chen et al., 2021b) fails to handle this setting and struggles even when two more labeled samples are further introduced. The main reason is that CPS is based on cross-modal pseudo labeling where all the unlabeled data is exploited in a supervised-like scheme. However, as mentioned in Section 1, supervised learning plays an

Table 2
Quantitative comparison when C1 serves as the local target center. "L." and "U." refer to labeled data and unlabeled data, respectively. Cross-subject standard deviations are shown in parentheses. The best mean results are shown in **bold**.

Method	# Scans used			Metrics			
	Local L. (C1)	Local U. (C1)	External U. (C2-6)	Dice (%) ↑	Jaccard (%) ↑	ASD (voxel) ↓	95HD (voxel) ↓
Supervised	4	0	0	64.68 (21.07)*	51.16 (21.72)*	2.07 (0.88)	7.30 (4.74)*
MT (Tarvainen and Valpola, 2017)	4	14	86	62.18 (23.70)*	49.12 (23.51)*	4.87 (2.71)*	9.39 (8.61)*
UA-MT (Yu et al., 2019)	4	14	86	68.55 (23.93)*	55.64 (23.84)*	3.56 (0.63)*	6.67 (6.27)*
EM (Grandvalet and Bengio, 2004)	4	14	86	64.44 (23.31)*	51.58 (23.71)*	1.88 (0.68)	6.26 (4.46)*
ICT (Verma et al., 2022)	4	14	86	61.70 (25.16)*	48.68 (22.63)*	2.53 (1.78)*	8.57 (9.69)*
DAN (Zhang et al., 2017)	4	14	86	74.05 (9.64)*	60.32 (12.37)*	5.26 (2.81)*	17.14 (11.15)*
CCT (Ouali et al., 2020)	4	14	86	64.83 (23.50)*	51.80 (22.47)*	2.72 (0.64)*	8.08 (8.73)*
DCT (Qiao et al., 2018)	4	14	86	65.73 (25.03)*	53.21 (23.22)*	1.84 (0.73)	8.48 (10.81)*
FixMatch (Sohn et al., 2020)	4	14	86	74.81 (17.04)*	62.16 (17.81)*	2.63 (1.84)*	6.20 (6.58)*
VAT (Miyato et al., 2018)	4	14	86	67.43 (19.87)*	53.94 (20.71)*	1.99 (0.83)*	8.54 (8.07)*
CPS (Chen et al., 2021b)	4	14	86	55.79 (25.58)*	42.85 (23.74)*	2.06 (0.85)	11.38 (10.69)*
SSNet (Wu et al., 2022)	4	14	86	65.10 (26.12)*	52.76 (23.65)*	2.50 (2.58)*	11.47 (19.51)*
AHDC (Chen et al., 2021a)	4	14	86	66.53 (25.74)*	54.21 (22.36)*	2.93 (2.87)	13.33 (17.21)*
SCL (ours)	4	14	86	**80.77** (9.25)	**68.68** (12.17)	**1.74** (0.77)	**5.14** (2.40)
Supervised	6	0	0	66.78 (23.26)*	54.24 (23.74)*	1.67 (0.81)*	6.67 (7.59)*
MT (Tarvainen and Valpola, 2017)	6	12	86	80.96 (11.15)*	70.14 (14.83)*	1.44 (0.53)	4.67 (2.68)*
UA-MT (Yu et al., 2019)	6	12	86	81.86 (13.82)*	70.77 (17.31)*	1.29 (0.99)	4.27 (3.73)
EM (Grandvalet and Bengio, 2004)	6	12	86	82.04 (10.55)*	71.43 (14.38)*	1.22 (0.79)	4.06 (2.96)
ICT (Verma et al., 2022)	6	12	86	78.52 (16.82)*	67.25 (19.02)*	1.29 (0.69)	4.88 (4.59)*
DAN (Zhang et al., 2017)	6	12	86	81.02 (8.45)*	70.69 (11.44)*	2.95 (2.03)*	9.69 (9.77)*
CCT (Ouali et al., 2020)	6	12	86	79.95 (17.27)*	69.40 (19.55)*	1.58 (1.51)*	4.47 (4.13)*
DCT (Qiao et al., 2018)	6	12	86	82.39 (12.68)*	71.74 (15.74)*	2.48 (1.37)*	5.53 (4.11)*
FixMatch (Sohn et al., 2020)	6	12	86	77.09 (18.45)*	65.69 (19.94)*	1.77 (1.05)*	5.66 (6.20)*
VAT (Miyato et al., 2018)	6	12	86	80.67 (12.96)*	69.35 (16.07)*	1.27 (0.68)	4.91 (4.56)*
CPS (Chen et al., 2021b)	6	12	86	65.02 (23.91)*	52.32 (23.65)*	1.70 (0.71)*	8.02 (7.76)*
SSNet (Wu et al., 2022)	6	12	86	80.37 (15.15)*	69.43 (17.88)*	**1.17** (0.63)	4.70 (4.75)*
AHDC (Chen et al., 2021a)	6	12	86	79.52 (14.32)*	68.02 (15.34)*	1.99 (0.78)*	7.87 (5.64)*
SCL (ours)	6	12	86	**84.89** (8.59)	**74.63** (11.85)	1.21 (0.51)	**3.59** (2.05)
Supervised (upper bound)	18	0	0	89.19 (4.33)	80.76 (6.71)	0.84 (0.28)	2.63 (0.81)
Supervised (joint)	18	0	86 (w/L)	88.33 (8.72)	80.02 (11.90)	0.86 (0.40)	2.92 (2.35)

* Indicates $p \leq 0.05$ from Wilcoxon signed rank test for pairwise comparison with our SCL.

important role in distribution fitting. Therefore, CPS is most sensitive to data heterogeneity because the model confuses about which distribution it should pay the most attention to. When we increase the amount of labeled data of C1, most of the SSL methods regain the ability to mine effective information from multi-site unlabeled data to support C1. This echoes our motivation described in Section 1 of why we perform separated learning and advocate label learning for local data. Despite improvement, these SSL methods have no proper mechanism to deal with multi-site data, thus still leading to limited segmentation performance. Besides, although AHDC (Chen et al., 2021a) generalizes SSL to dual-domain SSL, it requires that the external data is from a single specific site rather than multiple arbitrary sites. Since simultaneously aligning multiple external domains to the local domain is challenging for a single image mapping network, AHDC does not show an appealing advantage in MS-SSL. Notably, our SCL obtains better performance over the recent SSL methods. Fig. 6 presents the qualitative comparison between the proposed method and some top-performing approaches. Consistently, the predictions of our SCL fit more accurately with the ground truth, further verifying the effectiveness of our method.

4.3. Comparative experiments of C2 as the local center

Table 3 provides results of our method against several existing approaches where C2 serves as the local target institute. Note that we utilize six and eight C2 labeled scans here, considering radiologists' suggestion that a Dice score of 0.75 or higher is often considered as good prostate segmentation. Specifically, our method consistently outperforms the supervised baselines by a large margin under both local labeled data settings, e.g., +9.78% in Dice with six labeled data and +7.36% in Dice with eight labeled data, respectively. Meanwhile, we observe that most of the existing methods can exploit the useful clues from the multi-site unlabeled data in this experiment. However, CPS (Chen et al., 2021b) struggles again. It only achieves 43.41% and 65.34% Dice scores, respectively, under both labeled settings, which are far worse than the supervised baselines. This further verifies our claim that the supervised-like learning is critical to distribution fitting, and thus the learning manners for local and external unlabeled data should be separated. Interestingly, the dual-domain SSL method AHDC (Chen et al., 2021a) performs even worse than the supervised baseline under the setting with eight local labeled scans. The possible reason is that their bidirectional adversarial inference network (an image-to-image mapping component in Chen et al., 2021a) fails to simultaneously align the five external sites to the C2 domain, which impedes their following dual consistency learning scheme. Compared to AHDC, our method does not require unstable adversarial learning but considers the learning behavior of the model and domain-insensitive relationship modeling. Compared to other SSL methods, our SCL still yields better segmentation performance with significant differences in most metrics, further demonstrating the superiority and robustness of our approach.

4.4. Ablation study

Taking C1 as the local center as a case study, we further investigate our framework with the following experiments.

4.4.1. Effectiveness of each component

We conduct an ablation study under the setting with four local labeled scans to investigate how our framework works. The results are reported in Table 4. Firstly, when $\mathcal{L}_{RC}$ is replaced with the global Dice loss, the performance drops by 3.72% in Dice, showing that the regional form can better perceive the local mismatch and help the model learn fine-grained information. For local learning, if we only maintain $\mathcal{L}^u_{U2L}$, the overall results are worse than the variant without $\mathcal{L}^u_{PL}$. The potential reason is that the prototypical prediction is not good enough

Table 3
Comparison when C2 serves as the local target center and the other five centers construct the external unlabeled dataset $\mathcal{D}_e^u$. Standard deviations are in parentheses. The best results are in **bold**.

Method	# Scans used	Metrics			
	$\mathcal{D}_{local}^l/\mathcal{D}_{local}^u$, $\mathcal{D}_e^u$	Dice (%) ↑	Jaccard (%) ↑	ASD (voxel) ↓	95HD (voxel) ↓
Supervised	6/0, 0	71.19 (16.01)*	57.33 (16.80)*	6.81 (4.66)	21.13 (40.50)*
MT (Tarvainen and Valpola, 2017)	6/12, 86	77.38 (10.24)*	64.21 (13.19)*	11.50 (22.08)*	33.56 (61.07)*
UA-MT (Yu et al., 2019)	6/12, 86	78.31 (10.34)*	65.47 (13.26)*	9.30 (15.77)*	29.67 (49.70)*
EM (Grandvalet and Bengio, 2004)	6/12, 86	75.38 (10.49)*	61.62 (13.49)*	5.46 (4.39)	21.17 (29.83)*
ICT (Verma et al., 2022)	6/12, 86	77.67 (9.22)*	64.38 (11.81)*	5.92 (7.74)	22.31 (33.24)*
DAN (Zhang et al., 2017)	6/12, 86	75.93 (10.96)*	62.37 (13.38)*	17.18 (29.52)*	41.01 (65.83)*
CCT (Ouali et al., 2020)	6/12, 86	73.20 (15.20)*	59.79 (17.28)*	16.46 (30.92)*	33.56 (58.42)*
DCT (Qiao et al., 2018)	6/12, 86	78.70 (6.71)*	65.77 (9.16)*	5.88 (4.27)	19.02 (28.16)
FixMatch (Sohn et al., 2020)	6/12, 86	67.82 (14.80)*	53.16 (16.60)*	6.03 (3.83)*	19.88 (25.97)*
VAT (Miyato et al., 2018)	6/12, 86	78.83 (6.78)*	65.57 (9.16)*	7.97 (14.37)*	30.91 (53.76)*
CPS (Chen et al., 2021b)	6/12, 86	43.41 (23.39)*	30.32 (17.49)*	10.24 (17.81)	25.48 (34.82)*
SSNet (Wu et al., 2022)	6/12, 86	75.62 (10.99)*	62.03 (14.06)*	5.49 (8.85)	21.93 (45.64)*
AHDC (Chen et al., 2021a)	6/12, 86	74.65 (12.37)*	60.98 (14.33)*	6.02 (3.44)*	21.37 (27.23)*
SCL (ours)	6/12, 86	**80.97** (4.15)	**68.23** (5.77)	**5.01** (7.33)	**18.78** (32.34)
Supervised	8/0, 0	75.86 (10.24)*	62.20 (13.18)*	6.86 (6.22)*	18.95 (35.30)*
MT (Tarvainen and Valpola, 2017)	8/10, 86	77.90 (9.32)*	64.72 (11.99)*	4.39 (5.68)*	19.09 (38.06)*
UA-MT (Yu et al., 2019)	8/10, 86	77.44 (8.94)*	64.02 (11.41)*	2.97 (1.95)	13.03 (21.20)
EM (Grandvalet and Bengio, 2004)	8/10, 86	78.22 (7.46)*	64.84 (10.05)*	5.56 (7.87)*	26.58 (44.34)*
ICT (Verma et al., 2022)	8/10, 86	76.80 (11.84)*	63.65 (13.75)*	7.76 (12.29)*	30.09 (49.33)*
DAN (Zhang et al., 2017)	8/10, 86	76.18 (8.94)*	62.34 (11.37)*	4.10 (4.59)*	14.29 (21.53)*
CCT (Ouali et al., 2020)	8/10, 86	80.42 (7.60)*	67.31 (10.09)*	3.64 (2.25)*	14.23 (20.33)*
DCT (Qiao et al., 2018)	8/10, 86	82.07 (7.22)*	70.10 (9.35)*	3.31 (5.32)	15.14 (17.30)*
FixMatch (Sohn et al., 2020)	8/10, 86	66.17 (20.86)*	52.53 (20.05)*	14.19 (9.23)*	56.15 (40.39)*
VAT (Miyato et al., 2018)	8/10, 86	80.16 (10.42)*	68.44 (13.09)*	3.95 (1.37)*	14.27 (20.69)*
CPS (Chen et al., 2021b)	8/10, 86	65.34 (13.53)*	50.04 (15.19)*	6.98 (13.54)*	24.32 (50.18)*
SSNet (Wu et al., 2022)	8/10, 86	78.25 (9.90)*	65.27 (12.33)*	6.26 (11.77)*	23.20 (52.46)*
AHDC (Chen et al., 2021a)	8/10, 86	75.12 (11.23)*	61.97 (13.35)*	4.36 (3.51)*	17.77 (22.38)*
SCL (ours)	8/10, 86	**83.22** (6.27)	**71.73** (8.70)	**2.95** (3.77)	**12.02** (22.27)
Supervised (upper bound)	18/0, 0	85.01 (4.35)	74.15 (6.44)	2.76 (2.02)	7.02 (4.38)
Supervised (joint)	18/0, 86 (w/L)	85.93 (3.50)	75.32 (5.34)	1.12 (0.37)	3.22 (0.99)

* Indicates $p \leq 0.05$ from Wilcoxon signed rank test for pairwise comparison with our SCL.

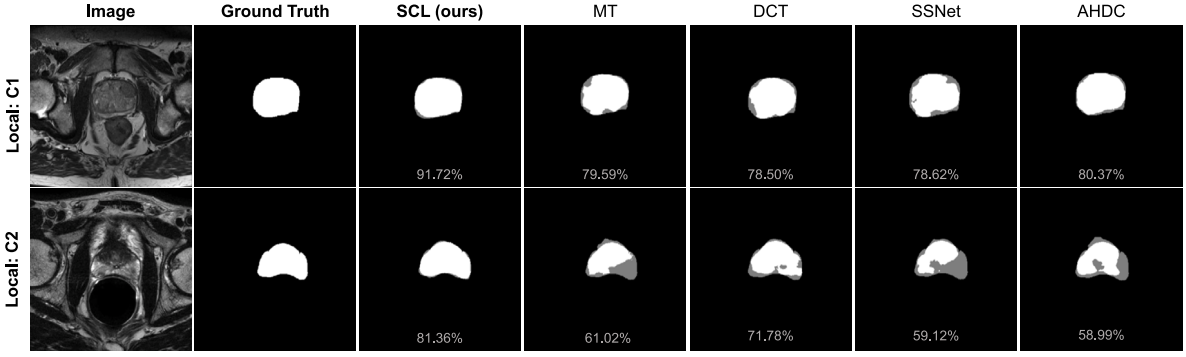

Fig. 6. 2D exemplar segmentation results with 4 (local: C1) and 6 (local: C2) local labeled scans for training. Gray color represents the inconsistency between the predicted result and the ground truth. The Dice score (%) of its corresponding 3D scan is shown at the bottom.

to produce high-quality cyclic predictions with very limited labeled data. Yet, $\mathcal{L}_{U2L}^u$ can be an effective regularization for the distribution of intra-class features and thus facilitate pixel-level label propagation. For external multi-site learning, we observe that the performance drops by 4.97% in Dice when we omit $\mathcal{L}_{CMD}^u$, showing that the local-support category mutual dependence plays an important role in robustly exploiting heterogeneous unlabeled data. Additionally, the global entropy minimization $\mathcal{L}_{ent}^u$ is complementally combined with $\mathcal{L}_{CMD}^u$ to reduce the ambiguity of the pseudo labels for prototype generation. If $\mathcal{L}_{ent}^u$ is removed, the performance slightly drops. Further, we observe that regularizing the model to be robust to adversarial perturbations ($\mathcal{L}_{advS}^u$) is an effective approach in MS-SSL. Besides exploiting the knowledge via stability learning, such stability under adversarial noises can enhance model robustness to heterogeneity. Replacing the adversarial perturbations to random noises (corresponding loss: $\mathcal{L}_{ran}^u$) (Yu et al., 2019), it can be observed that perturbing along the most "dangerous" adversarial direction is more effective in regularizing the learning behavior against data heterogeneity.

4.4.2. Impact of varying labeling budgets

Fig. 7 presents the Dice score under varying labeling budgets in C1. Compared to the supervised baseline, SCL can substantially improve the segmentation performance in the local center C1 due to the effective use of multi-site unlabeled data. Especially, when confronted with extremely scarce labeling budgets, the superiority of SCL is more prominent. SCL-Local denotes a special case that our SCL is applied to the local-only data, i.e., the unlabeled C1 data is used twice, one for local learning and the other for external learning. The results of other SSL methods under this scenario have been presented in Fig. 2(a). Excitingly, we observe that SCL-Local also achieves great performance improvement compared to the supervised baseline, showing that our

Table 4
Ablation study with 4 labeled scans from C1, wherein C1 as the local target center and C2 to C6 as the external support centers. Standard deviations are shown in parentheses. Best results are in **bold**.

Method	Metrics			
	Dice (%) ↑	Jaccard (%) ↑	ASD (voxel) ↓	95HD (voxel) ↓
SCL (ours)	**80.77** (9.25)	**68.68** (12.17)	1.74 (0.77)	**5.14** (2.40)
$-\mathcal{L}_{RC}$	77.05 (13.38)	64.91 (16.52)	3.28 (2.09)	9.52 (9.84)
$-\mathcal{L}^{u}_{PL}$	76.92 (14.29)	64.47 (17.11)	1.90 (0.75)	5.59 (2.40)
$-\mathcal{L}^{u}_{U2L}$	77.93 (12.16)	65.38 (15.60)	**1.69** (0.81)	5.32 (3.65)
$-(\mathcal{L}^{u}_{PL}+\mathcal{L}^{u}_{U2L})$	76.26 (14.28)	64.36 (17.07)	2.12 (1.29)	7.24 (4.54)
$-\mathcal{L}^{u}_{CMD}$	75.80 (17.07)	63.65 (19.16)	1.78 (0.84)	6.05 (4.17)
$-\mathcal{L}^{u}_{ent}$	79.11 (12.38)	67.02 (15.44)	1.85 (0.80)	5.82 (2.17)
$-\mathcal{L}^{u}_{advS}$	76.92 (10.12)	63.99 (12.26)	4.10 (2.01)	13.85 (7.99)
$-\mathcal{L}^{u}_{advS}+\mathcal{L}^{u}_{ran}$	78.11 (12.89)	65.74 (15.72)	1.88 (0.57)	6.77 (1.95)

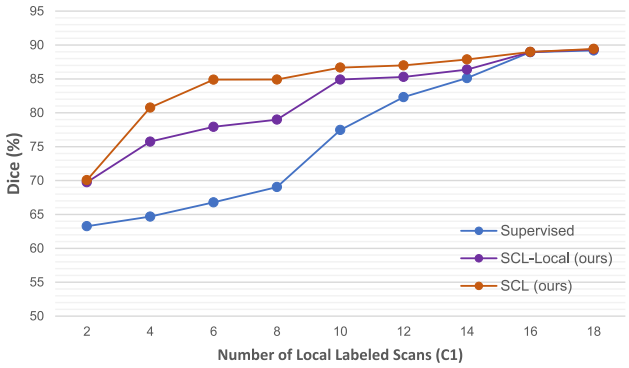

Fig. 7. Dice scores for local target center C1 under varying labeling budgets. SCL-Local denotes that our SCL is applied to local-only data, i.e., external unlabeled support scans are unavailable.

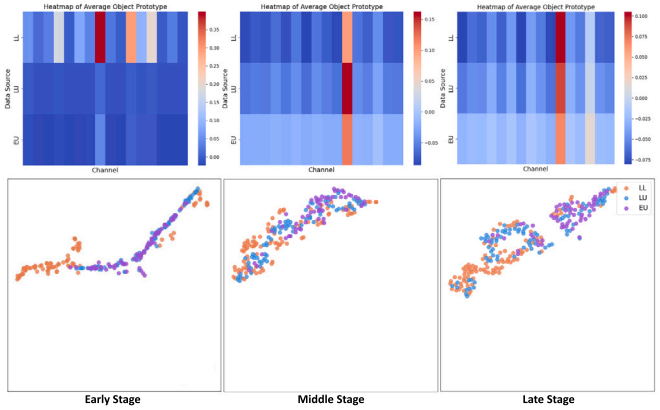

Fig. 8. Heatmap evolution of average object prototype profiles from the three subsets of data (upper row) and t-SNE visualization of object prototypes at different training stages colored by three subsets (bottom row).

SCL can robustly handle the situation of limited unlabeled data and thus benefit the typical SSL community as well. When the labeling budget is gradually increased, the performance of the supervised baseline gradually improves, but our SCL and SCL-Local can approach the saturated performance (i.e., upper bound) faster, revealing that our framework is an appealing alternative in both cases where external unlabeled data is available or not. However, since abundant external multi-site scans can greatly enrich data characteristics with better performance, applying our SCL in the MS-SSL setting is recommended.

4.4.3. Prototype evolution during training

Fig. 8 illustrates the progression of object (prostate) prototypes at different training stages using six local labeled scans. LL, LU and EU represent local labeled, local unlabeled and external unlabeled sets, respectively. The heatmap displays the statistically averaged prototype profiles of the three subsets of data. Each prototype vector has 16 channels. It is evident that at the early stage of training, the prototypes lack clarity and are not well-defined. As the training goes, the prototypes progressively specialize to capture the unique characteristics of the prostate. The t-SNE plot illustrates the distribution of object prototypes formed by randomly sampling an equal number of slices from LL, LU, and EU. Note that we employ ground-truth masks in LL prototype generation, whereas pseudo labels generated by the model are used for LU and EU prototype generation. Initially, the model's discriminative capabilities are limited, resulting in ill-defined prototypes. As training progresses, the intra-class variance within local unlabeled data gradually decreases through our local learning scheme, leading to a closer resemblance between LU and LL prototypes. We can also observe that although the average prototype profiles in the heatmap show that LL, LU, and EU prototypes become more specialized for prostate characteristics, there is still a distinction in prototype distribution between local and external data. By employing distribution-insensitive mutual information (MI) to model the relationship between c^{obj}_{local} and c^{obj}_{e}, our model can effectively harness valuable shared information from the abundant prostate characteristics in the external data while mitigating the impact of its heterogeneity.

4.4.4. Impact of different mutual information estimators

In contrast to non-parametric explicit estimation for mutual information, some studies proposed to use approximate estimation, i.e., turning it into a parameter optimization problem that optimizes the lower or upper bound of MI. Here, we further consider a popular approximate estimation method called Mutual Information Neural Estimation (MINE) (Belghazi et al., 2018), which approximates the lower bound of mutual information using the Donsker-Varadhan (DV) variational formula of the KL divergence between the joint distribution and the product of the marginals. The comparative results are reported in Table 5. It can be observed that our method is not very sensitive to the choice of MI estimation methods. However, we found that the efficiency of MINE is relatively low and it has high estimation variances. Especially, it easily leads to collapse during the late training stage here. In contrast, our adopted explicit estimation can more efficiently and stably estimate MI and achieve slightly better performance in this task.

5. Extended study

5.1. Efficacy of extensive data augmentations

In the above study, we applied the same weak augmentation techniques to augment images (Luo, 2021), ensuring a fair methodology comparison with recent semi-supervised medical image segmentation approaches. In addition to exploiting abundant unlabeled data to alleviate overfitting on limited labeled data, another intuitive solution is employing extensive data augmentation to simulate potential variation for general representation learning, as highlighted in the domain

Table 5
Results of using different mutual information estimation methods. C1 is the local target center. Standard deviations are shown in parentheses. Best results are in **bold**.

Method	Local L.	Metrics			
		Dice (%) ↑	Jaccard (%) ↑	ASD (voxel) ↓	95HD (voxel) ↓
SCL (Explicit)	4	**80.77** (9.25)	**68.68** (12.17)	**1.74** (0.77)	**5.14** (2.40)
SCL (MINE)	4	80.01 (8.82)	67.50 (11.36)	2.39 (1.13)	6.87 (4.05)
SCL (Explicit)	6	**84.89** (8.59)	**74.63** (11.85)	1.21 (0.51)	**3.59** (2.05)
SCL (MINE)	6	83.65 (0.10)	73.02 (13.34)	**1.17** (0.39)	3.87 (2.27)

Table 6
Results of introducing extensive data augmentations (Zhang et al., 2020). Standard deviations are in parentheses. Best results are in **bold**.

Local	Method	Local L.	Metrics			
			Dice (%) ↑	Jaccard (%) ↑	ASD (voxel) ↓	95HD (voxel) ↓
C1	Supervised	4	64.68 (21.07)	51.16 (21.72)	2.07 (0.88)	7.30 (4.74)
	Supervised (w/BigAug)		70.33 (14.01)	57.12 (15.52)	2.04 (0.67)	5.76 (2.05)
	SCL		80.77 (9.25)	68.68 (12.17)	**1.74** (0.77)	**5.14** (2.40)
	SCL (w/BigAug)		**84.66** (7.93)	**74.13** (10.84)	2.52 (2.34)	9.17 (13.88)
	Supervised	6	66.78 (23.26)	54.24 (23.74)	1.67 (0.81)	6.67 (7.59)
	Supervised (w/BigAug)		77.28 (11.88)	64.40 (14.82)	2.72 (1.24)	8.73 (7.56)
	SCL		84.89 (8.59)	74.63 (11.85)	**1.21** (0.51)	**3.59** (2.05)
	SCL (w/BigAug)		**86.25** (4.85)	**76.13** (7.28)	1.63 (0.80)	4.08 (1.80)
	Supervised	18	89.19 (4.33)	80.76 (6.71)	0.84 (0.28)	2.63 (0.81)
	Supervised (w/BigAug)		90.06 (2.78)	81.92 (4.60)	0.86 (0.31)	2.30 (0.72)
C2	Supervised	6	71.19 (16.01)	57.33 (16.80)	6.81 (4.66)	21.13 (40.50)
	Supervised (w/BigAug)		77.34 (10.73)	65.56 (13.20)	7.37 (5.23)	22.34 (35.78)
	SCL		80.97 (4.15)	68.23 (5.77)	**5.01** (7.33)	**18.78** (32.34)
	SCL (w/BigAug)		**81.61** (3.58)	**69.10** (5.19)	5.90 (6.09)	20.13 (27.79)
	Supervised	8	75.86 (10.24)	62.20 (13.18)	6.86 (6.22)	18.95 (35.30)
	Supervised (w/BigAug)		79.83 (5.06)	66.57 (7.26)	6.76 (4.12)	17.37 (24.36)
	SCL		83.22 (6.27)	71.73 (8.70)	**2.95** (3.77)	**12.02** (22.27)
	SCL (w/BigAug)		**83.76** (4.02)	**71.79** (5.93)	3.72 (3.70)	12.47 (21.57)
	Supervised	18	85.01 (4.35)	74.15 (6.44)	2.76 (2.02)	7.02 (4.38)
	Supervised (w/BigAug)		84.70 (2.89)	73.56 (4.32)	6.26 (5.63)	25.33 (20.21)

generalization work BigAug (Zhang et al., 2020). Thus, we further analyze the efficacy of extensive data augmentations. Following BigAug, we introduce stacked transformations including both weak and strong augmentations to our supervised baseline. These transformations encompass various aspects, such as image quality (sharpness, blurriness, and Gaussian noise), image appearance (brightness, contrast, and intensity), and spatial configuration (rotation, scaling, and elastic deformation). More details can be found in Zhang et al. (2020). As shown in Table 6, incorporating extensive augmentations can mitigate the overfitting issue observed with limited local labeled data. As the amount of local labeled data increases to the oracle, the performance gap between the results achieved with BigAug and the original supervised baseline using simple weak augmentation becomes less significant. Furthermore, integrating the stacked transformations into our SCL also yields overall performance improvements on the region-based metrics (i.e., Dice and Jaccard), particularly under scenarios of low labeling budget. This improvement can be primarily attributed to the fact that the supervised term with extensive data augmentations offers guidance towards more generalized representation, which complements the external multi-site learning with higher-quality and robust predictions. However, we also observe that excessive use of these strong augmentations led to a degradation on surface-based metrics (i.e., ASD and 95HD) to some extent. This may be attributed to the fact that certain types of strong augmentation can introduce biases into the training data that are not reflective of real-world data. These biases can have a negative impact on the model's predictions, particularly concerning surface-based metrics that demand precise delineation of object boundaries. In Zhang et al. (2020), the evaluation is limited to Dice score, which may overlook this issue.

5.2. Extensibility of SCL on multi-class segmentation

While our paper primarily focuses on whole prostate binary segmentation from T2-weighted MR images, in order to further assess the applicability and effectiveness of our SCL, we further extend our model to the multi-class segmentation task of cardiac magnetic resonance (CMR) images.

5.2.1. Materials and implementation

CMR images from four different clinical centers (C1-4) (Campello et al., 2021) are used for a retrospective evaluation. Table 7 summarizes the characteristics of the four data sources[2] and Fig. 9(a) presents randomly picked slices with one from each data source. The CMR images from C1-3 have undergone meticulous segmentation by proficient clinicians from their respective institutions, which includes delineating the contours for the left ventricle (LV), right ventricle (RV) blood pools, and left ventricular myocardium (MYO). The images from C4 are originally unlabeled. The implementation is consistent with our experiments in prostate segmentation, with the following exceptions: (i) the input patch size is set to 256×256 pixels, and (ii) the technical aspects related to prototypes are extended to accommodate multiple object classes. For the latter, technically, the proposed method can be easily extended by extracting the features for each to-be-segmented object category via masking different regions of interest (ROIs) and subsequently constructing their corresponding prototypes. In this extended study, we take C1 as the local target center and randomly divide their 75 scans into 50 for training, 5 for validation, and 20 for testing. All the images from C2 to C4 serve as the external unlabeled data.

5.2.2. Results

Table 8 presents the averaged results of the three object classes (LV, RV and MYO) with C1 as the local target center, wherein 5 (10%) or 10 (20%) local scans are annotated. The Supervised (joint) model denotes the model trained with all labeled data from C1, C2 and C3 (the images from C4 are originally unlabeled). As observed,

[2] https://www.ub.edu/mnms/.

Table 7
Details of the acquisition protocols for the cardiac MRI segmentation task (Campello et al., 2021) and number of scans for the different centers.

Center	Source	# Scans	Field strength (T)	In-plane resolution (mm)	Slice thickness (mm)	Scanner
C1	HVH	75	1.5	1.32	9.2	Siemens
C2	CSF	50	1.5	1.20	9.9	Philips
C3	UHE	25	1.5	1.45	9.9	Philips
C4	HUD	25	1.5	1.36	10	GE

Table 8
Comparison on the cardiac multi-class segmentation task. Standard deviations are in parentheses. The best results are in **bold**.

Method	# Scans used	Metrics			
	$\mathcal{D}^l_{local}/\mathcal{D}^u_{local}$, $\mathcal{D}^u_e$	Dice (%) ↑	Jaccard (%) ↑	ASD (voxel) ↓	95HD (voxel) ↓
Supervised	5/0, 0	65.70 (38.52)*	54.61 (38.97)*	5.70 (10.58)*	19.44 (33.50)*
BigAug (Zhang et al., 2020)	5/0, 0	68.33 (39.13)*	57.37 (41.12)*	2.43 (3.14)*	5.37 (12.13)
MT (Tarvainen and Valpola, 2017)	5/45, 100	61.43 (63.38)*	50.19 (57.82)*	5.66 (16.45)*	15.26 (27.29)*
UA-MT (Yu et al., 2019)	5/45, 100	65.87 (55.57)*	54.15 (51.61)*	2.78 (9.05)*	9.42 (24.78)*
EM (Grandvalet and Bengio, 2004)	5/45, 100	66.67 (56.34)*	54.94 (52.55)*	5.65 (15.28)	14.20 (28.03)*
ICT (Verma et al., 2022)	5/45, 100	64.52 (60.05)*	52.72 (55.17)*	3.41 (10.63)*	11.03 (24.57)*
DAN (Zhang et al., 2017)	5/45, 100	66.32 (46.67)*	54.04 (45.80)*	3.21 (6.55)*	13.48 (27.10)*
CCT (Ouali et al., 2020)	5/45, 100	68.82 (48.33)*	56.90 (47.84)*	3.02 (7.03)*	10.41 (22.21)*
DCT (Qiao et al., 2018)	5/45, 100	60.40 (63.74)*	48.88 (57.82)*	4.26 (13.06)*	13.88 (29.82)*
FixMatch (Sohn et al., 2020)	5/45, 100	62.26 (35.11)*	48.91 (36.32)*	9.52 (14.37)*	27.65 (41.49)*
VAT (Miyato et al., 2018)	5/45, 100	66.82 (54.89)*	55.25 (52.47)*	3.46 (12.43)*	11.46 (33.24)*
CPS (Chen et al., 2021b)	5/45, 100	39.05 (70.75)*	29.28 (57.50)*	5.17 (15.48)	14.87 (27.98)*
SSNet (Wu et al., 2022)	5/45, 100	66.08 (62.11)*	55.14 (57.85)*	2.21 (11.18)*	7.80 (23.95)*
AHDC (Chen et al., 2021a)	5/45, 100	67.31 (52.53)*	56.04 (51.29)*	4.38 (13.28)*	14.52 (38.25)*
SCL (ours)	5/45, 100	**72.96** (37.97)	**60.90** (41.04)	**1.03** (2.28)	**4.99** (10.43)
Supervised	10/0, 0	77.46 (30.61)*	66.25 (34.99)*	1.84 (3.71)	5.39 (12.67)
BigAug (Zhang et al., 2020)	10/0, 0	78.81 (24.53)*	68.27 (29.63)*	1.78 (3.02)	6.54 (13.12)*
MT (Tarvainen and Valpola, 2017)	10/40, 100	78.32 (30.74)*	66.91 (36.04)*	1.85 (4.06)	6.34 (10.17)*
UA-MT (Yu et al., 2019)	10/40, 100	78.36 (27.38)*	66.77 (32.64)*	**1.33** (3.23)	4.69 (11.08)
EM (Grandvalet and Bengio, 2004)	10/40, 100	78.71 (28.13)*	67.43 (33.11)*	1.89 (5.39)	5.77 (25.61)*
ICT (Verma et al., 2022)	10/40, 100	77.37 (25.01)*	66.36 (31.03)*	1.71 (1.83)	5.23 (18.66)
DAN (Zhang et al., 2017)	10/40, 100	77.31 (28.48)*	66.61 (33.97)*	1.72 (3.31)	6.47 (13.99)*
CCT (Ouali et al., 2020)	10/40, 100	78.10 (33.58)*	66.87 (37.96)*	1.62 (4.43)	4.77 (18.09)
DCT (Qiao et al., 2018)	10/40, 100	78.54 (25.73)*	66.82 (31.36)*	1.42 (3.64)	4.54 (15.44)
FixMatch (Sohn et al., 2020)	10/40, 100	68.05 (29.73)*	54.50 (32.75)*	9.95 (12.66)*	28.61 (38.67)*
VAT (Miyato et al., 2018)	10/40, 100	78.37 (20.59)*	66.76 (26.15)*	1.43 (2.99)	4.58 (12.62)
CPS (Chen et al., 2021b)	10/40, 100	61.33 (57.59)*	49.29 (52.63)*	3.02 (14.05)*	9.22 (40.10)*
SSNet (Wu et al., 2022)	10/40, 100	77.70 (31.33)*	66.64 (35.96)*	1.81 (2.91)	**3.77** (12.92)
AHDC (Chen et al., 2021a)	10/40, 100	78.98 (22.57)*	67.50 (30.39)*	1.73 (3.65)	6.56 (11.61)*
SCL (ours)	10/40, 100	**81.34** (18.79)	**69.91** (24.76)	1.43 (3.18)	4.67 (12.40)
Supervised (upper bound)	50/0, 0	86.89 (13.32)	77.62 (19.45)	0.75 (1.76)	2.66 (6.78)
Supervised (joint)	50/0, 75 (w/L)	86.02 (11.35)	77.01 (16.82)	0.77 (3.01)	3.16 (18.63)

* Indicates $p \leq 0.05$ from Wilcoxon signed rank test for pairwise comparison with our SCL.

compared to the supervised-only baselines, our SCL with {5, 10} local labeled scans achieves {11.05%, 5.01%} Dice improvement, showing its effectiveness in leveraging multi-site unlabeled data. Equipped with specialized mechanisms for learning informative representations from multi-site data and handling heterogeneity, our SCL achieves superior performance over the recent SSL methods, demonstrating the extensibility of our SCL on the multi-class problem. Fig. 9(b) further shows that the prediction of our method fit more accurately with the ground truth.

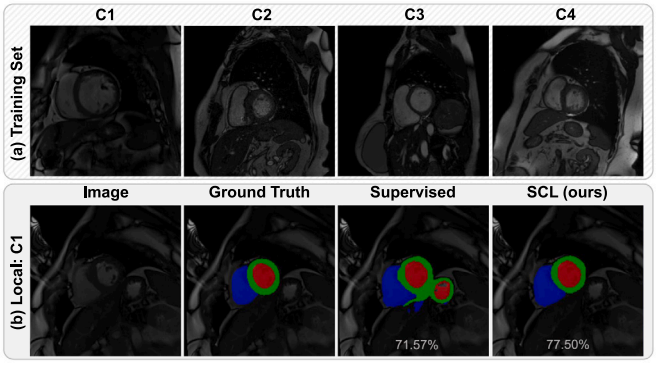

Fig. 9. (a) Slices of the normalized cardiac MR images from four different clinical centers. (b) 2D exemplar cardiac segmentation results (local: C1) with 5 local labeled scans for training. The Dice score (%) of its corresponding 3D scan is shown at the bottom.

6. Discussion

Impacts: The typical setting of SSL focuses on addressing label scarcity, assuming an abundance of independent and identically distributed (i.i.d.) unlabeled local data, while disregarding the scenario of image scarcity in low-resource local clinical centers. In such scenario, the availability of unlabeled images is also limited due to scarce local patients or restricted image collection capabilities. Our experimental findings in the context of prostate MRI segmentation reveal the significance of the quantity of unlabeled data in practical semi-supervised learning (as presented in Fig. 2). Furthermore, as observed in Tables 2 and 3, the Supervised (joint) model (trained with all labeled data from the six sites) does not significantly improve upon the Supervised (upper bound) model. This indicates that local centers still value their own data for personalized model development. If the local labeled data exhibits sufficient diversity, directly incorporating external heterogeneous labeled data does not provide significant benefits. However, in

cases where local centers lack the budget for precise expert annotations and the scale of local unlabeled data is insufficient to present comprehensive characteristics of the prostate of interest, these external heterogeneous unlabeled data can significantly increase the diversity of the unlabeled pool, thereby assisting in local semi-supervised learning. Hence, we introduce the new multi-site semi-supervised learning (MS-SSL) problem, which allows for the utilization of multi-site heterogeneous (non i.i.d.) unlabeled support data to enrich the unlabeled pool. This scenario introduces an additional challenge of data heterogeneity, violating the i.i.d. assumption of typical SSL (Van Engelen and Hoos, 2020), yet, it effectively addresses the potential clinical needs of low-resource local centers. We believe that this new problem setting has a high clinical impact and requires different learning strategies for local and external unlabeled data to effectively support the semi-supervised training of local personalized models.

Future Work: Although our proposed SCL demonstrates encouraging performance in the MS-SSL scenario, there are important aspects that require further investigation in future work. For example, it is beneficial to consider collaboration fairness by quantifying the contribution of each unlabeled support site and determining their relative importance. In this study, the external unlabeled data was directly mixed in advance and we treated each external site with equal priority, but future MS-SSL studies should consider assessing both the quantity and quality of the contributed unlabeled data from external collaborating sites. This evaluation will help determine the potential inclusion of each site's data in use-inspired studies. In the context of privacy-preserving collaborative learning, such as federated learning (FL), previous work (Tang et al., 2022) has explored client selection and contribution evaluation mechanisms. In FL, each client typically possesses labeled data, which can be utilized to evaluate the data quality of each client site (via the validation set) and quantify its contribution to the global model optimization. However, in MS-SSL, we do not use any label from external support sites. This poses a challenge in self-evaluating the contribution of each support site without any ground-truth annotations. Exploring this direction will provide valuable insights into the selection and utilization of external support sites, further enhancing the performance and applicability in real-world MS-SSL scenarios.

7. Conclusion

In this work, we proposed a separated collaborative learning (SCL) framework to achieve semi-supervised prostate MRI segmentation with multi-site heterogeneous unlabeled MRI data. Our key insight is to separate the training process of local and external unlabeled data according to the learning behavior of the model. Local learning advocates the supervised-like scheme, which is responsible for accurate local distribution fitting. In contrast, external multi-site learning appreciates mutual information-based distribution-insensitive relationship modeling on region-of-interest between local learning and external learning and model behavior regularization, which helps robustly mine informative clues from external data to support local learning. Extensive experiments on prostate MRI data from six different clinical centers showed that our method could effectively generalize SSL on multi-site unlabeled data and significantly outperformed other semi-supervised methods. We also demonstrated the extensibility of our method on the multi-class cardiac MRI segmentation task with data from four different clinical centers.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This research was done with Tencent Jarvis Research Center, Youtu Lab and supported by Hong Kong PhD Fellowship and General Research Fund (No. 14205419) from Research Grants Council of Hong Kong.

References

Belghazi, M.I., Baratin, A., Ozair, S., Bengio, Y., Courville, A., Hjelm, D., 2018. Mutual information neural estimation. In: International Conference on Machine Learning. PMLR, pp. 531–540.

Ben-David, S., Lu, T., Pál, D., 2008. Does unlabeled data provably help? Worst-case analysis of the sample complexity of semi-supervised learning. In: Annual Conference on Learning Theory. pp. 33–44.

Bloch, N., Madabhushi, A., Huisman, H., Freymann, J., Kirby, J., Grauer, M., Enquobahrie, A., Jaffe, C., Clarke, L., Farahani, K., 2015. NCI-ISBI 2013 Challenge: automated segmentation of prostate structures. Cancer Imaging Arch. 370 (6), 5.

Campello, V.M., Gkontra, P., Izquierdo, C., Martin-Isla, C., Sojoudi, A., Full, P.M., Maier-Hein, K., Zhang, Y., He, Z., Ma, J., et al., 2021. Multi-centre, multi-vendor and multi-disease cardiac segmentation: the M&Ms challenge. IEEE Trans. Med. Imaging 40 (12), 3543–3554.

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

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

Cui, W., Liu, Y., Li, Y., Guo, M., Li, Y., Li, X., Wang, T., Zeng, X., Ye, C., 2019. Semi-supervised brain lesion segmentation with an adapted mean teacher model. In: International Conference on Information Processing in Medical Imaging. Springer, pp. 554–565.

Goodfellow, I.J., Shlens, J., Szegedy, C., 2015. Explaining and harnessing adversarial examples. In: International Conference on Learning Representation.

Grandvalet, Y., Bengio, Y., 2004. Semi-supervised learning by entropy minimization. Adv. Neural Inf. Process. Syst. 17.

Guo, C.K., 2019. Multi-Modal Image Registration with Unsupervised Deep Learning (Ph.D. thesis). Massachusetts Institute of Technology.

Hsu, P.-L., Robbins, H., 1947. Complete convergence and the law of large numbers. Proc. Natl. Acad. Sci. 33 (2), 25–31.

Jia, H., Song, Y., Huang, H., Cai, W., Xia, Y., 2019. HD-Net: hybrid discriminative network for prostate segmentation in MR images. In: International Conference on Medical Image Computing and Computer Assisted Intervention. Springer, pp. 110–118.

Kwak, N., Choi, C.-H., 2002. Input feature selection by mutual information based on Parzen window. IEEE Trans. Pattern Anal. Mach. Intell. 24 (12), 1667–1671.

Lemaître, G., Martí, R., Freixenet, J., Vilanova, J.C., Walker, P.M., Meriaudeau, F., 2015. Computer-aided detection and diagnosis for prostate cancer based on mono and multi-parametric MRI: a review. Comput. Biol. Med. 60, 8–31.

Li, Y.-F., Zhou, Z.-H., 2014. Towards making unlabeled data never hurt. IEEE Trans. Pattern Anal. Mach. Intell. 37 (1), 175–188.

Litjens, G., Toth, R., van de Ven, W., Hoeks, C., Kerkstra, S., van Ginneken, B., Vincent, G., Guillard, G., Birbeck, N., Zhang, J., et al., 2014. Evaluation of prostate segmentation algorithms for MRI: the PROMISE12 challenge. Med. Image Anal. 18 (2), 359–373.

Liu, J., Desrosiers, C., Zhou, Y., 2022. Semi-supervised medical image segmentation using cross-model pseudo-supervision with shape awareness and local context constraints. In: International Conference on Medical Image Computing and Computer Assisted Intervention. Springer, pp. 140–150.

Liu, Q., Dou, Q., Heng, P.-A., 2020a. Shape-aware meta-learning for generalizing prostate MRI segmentation to unseen domains. In: International Conference on Medical Image Computing and Computer Assisted Intervention. Springer, pp. 475–485.

Liu, Q., Dou, Q., Yu, L., Heng, P.A., 2020b. MS-Net: multi-site network for improving prostate segmentation with heterogeneous MRI data. IEEE Trans. Med. Imaging 39 (9), 2713–2724.

Luo, X., 2021. SSL4MIS. URL: https://github.com/HiLab-git/SSL4MIS.

Luo, X., Chen, J., Song, T., Chen, Y., Wang, G., Zhang, S., 2021. Semi-supervised medical image segmentation through dual-task consistency. In: AAAI Conference on Artificial Intelligence.

Meng, Q., Matthew, J., Zimmer, V.A., Gomez, A., Lloyd, D.F., Rueckert, D., Kainz, B., 2020. Mutual information-based disentangled neural networks for classifying unseen categories in different domains: Application to fetal ultrasound imaging. IEEE Trans. Med. Imaging 40 (2), 722–734.

Miyato, T., Maeda, S.-i., Koyama, M., Ishii, S., 2018. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. IEEE Trans. Pattern Anal. Mach. Intell. 41 (8), 1979–1993.

Oliver, A., Odena, A., Raffel, C.A., Cubuk, E.D., Goodfellow, I., 2018. Realistic evaluation of deep semi-supervised learning algorithms. Adv. Neural Inf. Process. Syst. 31.

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

Peng, J., Pedersoli, M., Desrosiers, C., 2020. Mutual information deep regularization for semi-supervised segmentation. In: Medical Imaging with Deep Learning. PMLR, pp. 601–613.

Pluim, J.P., Maintz, J.A., Viergever, M.A., 2003. Mutual-information-based registration of medical images: a survey. IEEE Trans. Med. Imaging 22 (8), 986–1004.

Qiao, S., Shen, W., Zhang, Z., Wang, B., Yuille, A., 2018. Deep co-training for semi-supervised image recognition. In: Proceedings of the European Conference on Computer Vision. pp. 135–152.

Qiu, L., Cheng, J., Gao, H., Xiong, W., Ren, H., 2023. Federated semi-supervised learning for medical image segmentation via pseudo-label denoising. IEEE J. Biomed. Health Inf..

Rizve, M.N., Duarte, K., Rawat, Y.S., Shah, M., 2021. In defense of pseudo-labeling: An uncertainty-aware pseudo label selection framework for semi-supervised learning. arXiv preprint arXiv:2101.06329.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer Assisted Intervention. Springer, pp. 234–241.

Samuli, L., Timo, A., 2017. Temporal ensembling for semi-supervised learning. In: International Conference on Learning Representations.

Sedai, S., Mahapatra, D., Hewavitharanage, S., Maetschke, S., Garnavi, R., 2017. Semi-supervised segmentation of optic cup in retinal fundus images using variational autoencoder. In: International Conference on Medical Image Computing and Computer Assisted Intervention. Springer, pp. 75–82.

Shi, Y., Sha, F., 2012. Information-theoretical learning of discriminative clusters for unsupervised domain adaptation. In: International Conference on Machine Learning.

Snell, J., Swersky, K., Zemel, R., 2017. Prototypical networks for few-shot learning. In: Advances in Neural Information Processing Systems. Vol. 30, pp. 4080–4090.

Sohn, K., Berthelot, D., Li, C.-L., Zhang, Z., Carlini, N., Cubuk, E.D., Kurakin, A., Zhang, H., Raffel, C., 2020. FixMatch: Simplifying semi-supervised learning with consistency and confidence. arXiv preprint arXiv:2001.07685.

Tang, M., Ning, X., Wang, Y., Sun, J., Wang, Y., Li, H., Chen, Y., 2022. FedCor: Correlation-based active client selection strategy for heterogeneous federated learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10102–10111.

Tarvainen, A., Valpola, H., 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: Advances in Neural Information Processing Systems. pp. 1195–1204.

Van Engelen, J.E., Hoos, H.H., 2020. A survey on semi-supervised learning. Mach. Learn. 109 (2), 373–440.

Verma, V., Kawaguchi, K., Lamb, A., Kannala, J., Solin, A., Bengio, Y., Lopez-Paz, D., 2022. Interpolation consistency training for semi-supervised learning. Neural Netw. 145, 90–106.

Viola, P., Wells, III, W.M., 1997. Alignment by maximization of mutual information. Int. J. Comput. Vis. 24 (2), 137–154.

Wang, K., Liew, J.H., Zou, Y., Zhou, D., Feng, J., 2019. PANet: Few-shot image semantic segmentation with prototype alignment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9197–9206.

Wang, G., Liu, X., Li, C., Xu, Z., Ruan, J., Zhu, H., Meng, T., Li, K., Huang, N., Zhang, S., 2020. A noise-robust framework for automatic segmentation of COVID-19 pneumonia lesions from CT images. IEEE Trans. Med. Imaging 39 (8), 2653–2663.

Wu, Y., Wu, Z., Wu, Q., Ge, Z., Cai, J., 2022. Exploring smoothness and class-separation for semi-supervised medical image segmentation. In: International Conference on Medical Image Computing and Computer Assisted Intervention. Springer, pp. 34–43.

Xie, Q., Li, Y., He, N., Ning, M., Ma, K., Wang, G., Lian, Y., Zheng, Y., 2022. Unsupervised domain adaptation for medical image segmentation by disentanglement learning and self-training. IEEE Trans. Med. Imaging.

Xu, R., Chen, Y.-W., Tang, S.-Y., Morikawa, S., Kurumi, Y., 2008. Parzen-window based normalized mutual information for medical image registration. IEICE Trans. Inf. Syst. 91 (1), 132–144.

Xu, Z., Lu, D., Luo, J., Wang, Y., Yan, J., Ma, K., Zheng, Y., Tong, R.K.-Y., 2022a. Anti-interference from noisy labels: Mean-teacher-assisted confident learning for medical image segmentation. IEEE Trans. Med. Imaging 41 (11), 3062–3073.

Xu, Z., Wang, Y., Lu, D., Luo, X., Yan, J., Zheng, Y., Tong, R.K.-y., 2023. Ambiguity-selective consistency regularization for mean-teacher semi-supervised medical image segmentation. Med. Image Anal. 88, 102880.

Xu, Z., Wang, Y., Lu, D., Yu, L., Yan, J., Luo, J., Ma, K., Zheng, Y., Tong, R.K.-y., 2022b. All-around real label supervision: Cyclic prototype consistency learning for semi-supervised medical image segmentation. IEEE J. Biomed. Health Inf. 26 (7), 3174–3184.

Yang, D., Xu, Z., Li, W., Myronenko, A., Roth, H.R., Harmon, S., Xu, S., Turkbey, B., Turkbey, E., Wang, X., et al., 2021. Federated semi-supervised learning for COVID region segmentation in chest CT using multi-national data from China, Italy, Japan. Med. Image Anal. 70, 101992.

Yu, L., Wang, S., Li, X., Fu, C.-W., Heng, P.-A., 2019. Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation. In: International Conference on Medical Image Computing and Computer Assisted Intervention. Springer, pp. 605–613.

Zhang, B., Wang, Y., Hou, W., Wu, H., Wang, J., Okumura, M., Shinozaki, T., 2021. FlexMatch: Boosting semi-supervised learning with curriculum pseudo labeling. Adv. Neural Inf. Process. Syst. 34, 18408–18419.

Zhang, L., Wang, X., Yang, D., Sanford, T., Harmon, S., Turkbey, B., Wood, B.J., Roth, H., Myronenko, A., Xu, D., et al., 2020. Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation. IEEE Trans. Med. Imaging 39 (7), 2531–2540.

Zhang, Y., Yang, L., Chen, J., Fredericksen, M., Hughes, D.P., Chen, D.Z., 2017. Deep adversarial networks for biomedical image segmentation utilizing unannotated images. In: International Conference on Medical Image Computing and Computer Assisted Intervention. Springer.

Zhang, W., Zhu, L., Hallinan, J., Zhang, S., Makmur, A., Cai, Q., Ooi, B.C., 2022. BoostMIS: Boosting medical image semi-supervised learning with adaptive pseudo labeling and informative active annotation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20666–20676.

Zhao, Z., Zhou, F., Xu, K., Zeng, Z., Guan, C., Zhou, S.K., 2022. LE-UDA: Label-efficient unsupervised domain adaptation for medical image segmentation. IEEE Trans. Med. Imaging 42 (3), 633–646.

Zheng, H., Motch Perrine, S.M., Wang, C., Richtsmeier, J.T., Chen, D.Z., 2020. Cartilage segmentation in high-resolution 3D micro-CT images via uncertainty-guided self-training with very sparse annotation. In: International Conference on Medical Image Computing and Computer Assisted Intervention. Springer, pp. 802–812.