Category-Level Regularized Unlabeled-to-Labeled Learning for Semi-supervised Prostate Segmentation with Multi-site Unlabeled Data

Zhe Xu[1], Donghuan Lu[2](✉), Jiangpeng Yan[3], Jinghan Sun[4], Jie Luo[5], Dong Wei[2], Sarah Frisken[5], Quanzheng Li[5], Yefeng Zheng[2], and Raymond Kai-yu Tong[1](✉)

[1] Department of Biomedical Engineering, The Chinese University of Hong Kong, Hong Kong, China
jackxz@link.cuhk.edu.hk, kytong@cuhk.edu.hk

[2] Tencent Healthcare Co., Jarvis Lab, Shenzhen, China
calebIu@tencent.com

[3] Department of Automation, Tsinghua University, Beijing, China

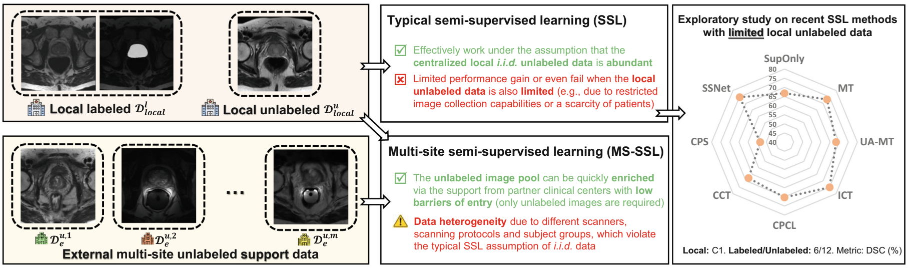
[4] Xiamen University, Xiamen, China

[5] Harvard Medical School, Boston, MA, USA

Abstract. Segmenting prostate from MRI is crucial for diagnosis and treatment planning of prostate cancer. Given the scarcity of labeled data in medical imaging, semi-supervised learning (SSL) presents an attractive option as it can utilize both limited labeled data and abundant unlabeled data. However, if the local center has limited image collection capability, there may also not be enough unlabeled data for semi-supervised learning to be effective. To overcome this issue, other partner centers can be consulted to help enrich the pool of unlabeled images, but this can result in data heterogeneity, which could hinder SSL that functions under the assumption of consistent data distribution. Tailoring for this important yet under-explored scenario, this work presents a novel Category-level regularized Unlabeled-to-Labeled (CU2L) learning framework for semi-supervised prostate segmentation with multi-site unlabeled MRI data. Specifically, CU2L is built upon the teacher-student architecture with the following tailored learning processes: (i) local pseudo-label learning for reinforcing confirmation of the data distribution of the local center; (ii) category-level regularized non-parametric unlabeled-to-labeled learning for robustly mining shared information by using the limited expert labels to regularize the intra-class features across centers to be discriminative and generalized; (iii) stability learning under perturbations to further enhance robustness to heterogeneity. Our method is evaluated on prostate MRI data from six different clinical centers and shows superior performance compared to other semi-supervised methods.

Keywords: Prostate segmentation · Semi-supervised · Heterogeneity

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14223, pp. 3–13, 2023.
https://doi.org/10.1007/978-3-031-43901-8_1


Local labeled $\mathcal{D}^{l}_{local}$
Local unlabeled $\mathcal{D}^{u}_{local}$
$\mathcal{D}^{u,1}_{e}$
$\mathcal{D}^{u,2}_{e}$
...
$\mathcal{D}^{u,m}_{e}$
External multi-site unlabeled support data
SupOnly
MT
UA-MT
ICT
CPCL
CCT
CPS
SSNet
80
75
70
65
60
55
50
45
40

Typical semi-supervised learning (SSL)

☑ Effectively work under the assumption that the **centralized local *i.i.d.* unlabeled data** is **abundant**

☒ Limited performance gain or even fail when the **local unlabeled data** is also **limited** (e.g., due to restricted image collection capabilities or a scarcity of patients)

Multi-site semi-supervised learning (MS-SSL)

☑ The **unlabeled image pool** can be quickly **enriched** via the support from partner clinical centers with **low barriers of entry** (only unlabeled images are required)

⚠ **Data heterogeneity** due to different scanners, scanning protocols and subject groups, which violate the typical SSL assumption of *i.i.d.* data

Exploratory study on recent SSL methods with limited local unlabeled data

Local: C1. **Labeled/Unlabeled:** 6/12. Metric: DSC (%)

Fig. 1. Comparison between typical semi-supervised learning (SSL) and our focused multi-site semi-supervised learning (MS-SSL), and an exploratory validation on recent SSL methods with limited local unlabeled data.

Table 1. Details of the acquisition protocols and number of scans for the six 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	RUNMC [1]	30	3	0.6–0.625/3.6–4	Surface	Siemens
C2	BMC [1]	30	1.5	0.4/3	Endorectal	Philips
C3	HCRUDB [4]	19	3	0.67–0.79/1.25	–	Siemens
C4	UCL [5]	13	1.5 and 3	0.325–0.625/3–3.6	–	Siemens
C5	BIDMC [5]	12	3	0.25/2.2–3	Endorectal	GE
C6	HK [5]	12	1.5	0.625/3.6	Endorectal	Siemens

1 Introduction

Prostate segmentation from magnetic resonance imaging (MRI) is a crucial step for diagnosis and treatment planning of prostate cancer. Recently, deep learning-based approaches have greatly improved the accuracy and efficiency of automatic prostate MRI segmentation [7,8]. Yet, their success usually requires a large amount of labeled medical data, which is expensive and expertise-demanding in practice. In this regard, semi-supervised learning (SSL) has emerged as an attractive option as it can leverage both limited labeled data and abundant unlabeled data [3,9–11,15,16,21–26,28]. Nevertheless, the effectiveness of SSL is heavily dependent on the ***quantity*** and ***quality*** of the unlabeled data.

Regarding ***quantity***, the abundance of unlabeled data serves as a way to regularize the model and alleviate overfitting to the limited labeled data. Unfortunately, such "abundance" may be unobtainable in practice, i.e., the local unlabeled pool is also limited due to restricted image collection capabilities or scarce patient samples. As a specific case shown in Table 1, there are only limited prostate scans available per center. Taking C1 as a case study, if the amount of local unlabeled data is limited, existing SSL methods may still suffer from inferior performance when generalizing to unseen test data (Fig. 1). To efficiently enrich the unlabeled pool, seeking support from other centers is a viable solution, as illustrated in Fig. 1. Yet, due to differences in imaging protocols and variations in patient demographics, this solution usually introduces data heterogeneity, leading to a ***quality*** problem. Such heterogeneity may impede the performance of SSL which typically assumes that the distributions of labeled data and unlabeled data are independent and identically distributed (i.i.d.) [16]. Thus, proper mechanisms are called for this practical but challenging SSL scenario.

Here, we define this new SSL scenario as multi-site semi-supervised learning (MS-SSL), allowing to enrich the unlabeled pool with multi-site heterogeneous images. Being an under-explored scenario, few efforts have been made. To our best knowledge, the most relevant work is AHDC [2]. However, it only deals with additional unlabeled data from a specific source rather than multiple arbitrary sources. Thus, it intuitively utilizes image-level mapping to minimize dual-distribution discrepancy. Yet, their adversarial min-max optimization often leads to instability and it is difficult to align multiple external sources with the local source using a single image mapping network.

In this work, we propose a more generalized framework called Category-level regularized Unlabeled-to-Labeled (CU2L) learning, as depicted in Fig. 2, to achieve robust MS-SSL for prostate MRI segmentation. Specifically, CU2L is built upon the teacher-student architecture with customized learning strategies for local and external unlabeled data: (i) recognizing the importance of supervised learning in data distribution fitting (which leads to the failure of CPS [3] in MS-SSL as elaborated in Sec. 3), the local unlabeled data is involved into pseudo-label supervised-like learning to reinforce fitting of the local data distribution; (ii) considering that intra-class variance hinders effective MS-SSL, we introduce a non-parametric unlabeled-to-labeled learning scheme, which takes advantage of the scarce expert labels to explicitly constrain the prototype-propagated predictions, to help the model exploit discriminative and domain-insensitive features from heterogeneous multi-site data to support the local center. Yet, observing that such scheme is challenging when significant shifts and various distributions are present, we further propose category-level regularization, which advocates prototype alignment, to regularize the distribution of intra-class features from arbitrary external data to be closer to the local distribution; (iii) based on the fact that perturbations (e.g., Gaussian noises [15]) can be regarded as a simulation of heterogeneity, perturbed stability learning is incorporated to enhance the robustness of the model. Our method is evaluated on prostate MRI data from six different clinical centers and shows promising performance on tackling MS-SSL compared to other semi-supervised methods.

2 Methods

2.1 Problem Formulation and Basic Architecture

In our scenario of MS-SSL, we have access to a local target dataset $\mathcal{D}_{local}$ (consisted of a labeled sub-set $\mathcal{D}^l_{local}$ and an unlabeled sub-set $\mathcal{D}^u_{local}$) and the external unlabeled support datasets $\mathcal{D}^u_e = \bigcup_{j=1}^{m} \mathcal{D}^{u,j}_e$, where m is the number of support centers. Specifically, $\mathcal{D}^l_{local} = \{(X^l_{local(i)}, Y^l_{local(i)})\}_{i=1}^{n_l}$ with n_l labeled scans and $\mathcal{D}^u_{local} = \{X^u_{local(i)}\}_{i=n_l+1}^{n_l+n_u}$ with n_u unlabeled scans.

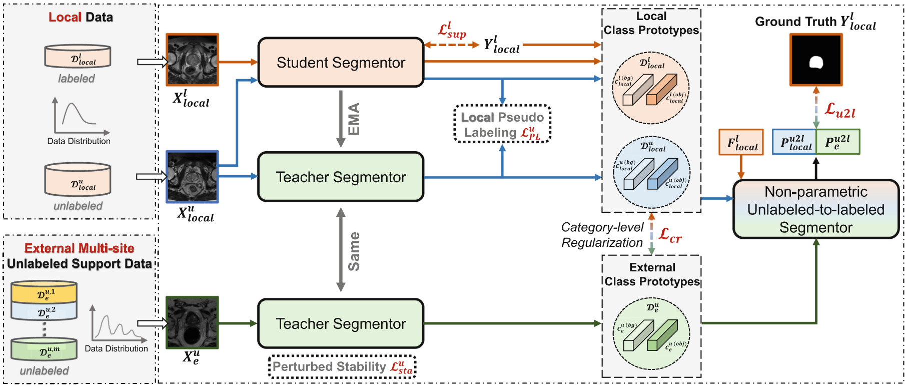


Fig. 2. Illustration of the proposed Category-level Regularized Unlabeled-to-labeled Learning (CU2L) framework. EMA: exponential moving average.

$X^l_{local(i)}, X^u_{local(i)} \in \mathbb{R}^{H\times W\times D}$ denote 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 binary segmentation). Similarly, the j-th external unlabeled support dataset is denoted as $\mathcal{D}^{u,j}_e = \{X^{u,j}_{e(i)}\}^{n_j}_{i=1}$ with n_j unlabeled samples. Considering the large variance on slice thickness among different centers [7,8], our experiments are performed in 2D. Thus, we refer to pixels in the subsequent content. As shown in Fig. 2, our framework is built upon the popular teacher-student framework. Specifically, the student f^s_θ is an in-training model optimized by loss back-propagation as usual while the teacher model $f^t_{\tilde{\theta}}$ is slowly updated with a momentum term that averages previous weights with the current weights, where θ denotes the student's weights and $\tilde{\theta}$ the teacher's weights. $\tilde{\theta}$ is updated by $\tilde{\theta}_t = \alpha\tilde{\theta}_{t-1} + (1-\alpha)\theta_t$ at iteration t, where α is the exponential moving average (EMA) coefficient and empirically set to 0.99 [26]. Compared to the student, the teacher performs self-ensembling by nature which helps smooth out the noise and avoid sudden changes of predictions [15]. Thus, the teacher model is suitable for handling the heterogeneous external images and producing relatively stable pseudo labels (will be used later). As such, our task of MS-SSL can be formulated as optimizing the following loss:

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

where $\mathcal{L}^l_{sup}$ is the supervised guidance from local labeled data and $\mathcal{L}^u$ denotes the additional guidance from the unlabeled data. λ is a trade-off weight scheduled by the time-dependent ramp-up Gaussian function [15] $\lambda(t) = w_{max} \cdot e^{-5(1-t/t_{max})^2}$, where w_{max} and $t_{\max}$ are the maximal weight and iteration, respectively. The key challenge of MS-SSL is the proper design of $\mathcal{L}^u$ for robustly exploiting multi-site unlabeled data $\{\mathcal{D}^u_{local}, \mathcal{D}^u_e\}$ to support the local center.

2.2 Pseudo Labeling for Local Distribution Fitting

As mentioned above, supervised-like learning is advocated for local unlabeled data to help the model fit local distribution better. Owning the self-ensembling property, the teacher model provides relatively stable pseudo labels for the student model. Given the predicted probability map $P_{local}^{u,t}$ of X_{local}^{u} from the teacher model, the pseudo label $\hat{Y}_{local}^{u,t}$ corresponds to the class with the maximal posterior probability. Yet, with limited local labeled data for training, it is difficult to generate high-quality pseudo labels. Thus, for each pixel, if $\max_c(p_{local}^{u,t}) \geq \delta$, where c denotes the c-th class and δ is a ramp-up threshold ranging from 0.75 to 0.9 as training goes, this pixel will be included in loss calculation. Considering that the cross-entropy loss has been found very sensitive to label noises [18], we adopt the partial Dice loss $\mathcal{L}_{\text{Dice}}$ [27] to perform pseudo label learning, formulated as: $\mathcal{L}_{PL}^{u} = \frac{1}{K}\sum_{k=1}^{K}\mathcal{L}_{\text{Dice}}\left(P_{local}^{u,s,k}, \hat{Y}_{local}^{u,t,k}\right)$, where $P_{local}^{u,s}$ denotes the prediction of X_{local}^{u} from the student model. The Dice loss is calculated for each of the K equally-sized regions of the image, and the final loss is obtained by taking their mean. Such a regional form [6] can help the model better perceive the local discrepancies for fine-grained learning.

2.3 Category-Level Regularized Unlabeled-to-Labeled Learning

Unlabeled-to-Labeled Learning. Inherently, the challenge of MS-SSL stems from intra-class variation, which results from different imaging protocols, disease progress and patient demographics. Inspired by prototypical networks [13,19,25] that compare class prototypes with pixel features to perform segmentation, here, we introduce a non-parametric unlabeled-to-labeled (U2L) learning scheme that utilizes expert labels to explicitly constrain the prototype-propagated predictions. Such design is based on two considerations: (i) a good prototype-propagated prediction requires both compact feature and discriminative prototypes, thus enhancing this prediction can encourage the model to learn in a variation-insensitive manner and focus on the most informative clues; (ii) using expert labels as final guidance can prevent error propagation from pseudo labels. Specifically, we denote the feature map of the external unlabeled image X_e^u before the penultimate convolution in the teacher model as $F_e^{u,t}$. Note that $F_e^{u,t}$ has been upsampled to the same size of X_e^u via bilinear interpolation but with L channels. With the argmax pseudo label $\hat{Y}_e^{u,t}$ and the predicted probability map $P_e^{u,t}$, the object prototype from the external unlabeled data can be computed via confidence-weighted masked average pooling: $c_e^{u(obj)} = \frac{\sum_v\left[\hat{Y}_{e(v)}^{u,t,obj} \cdot P_{e(v)}^{u,t,obj} \cdot F_{e(v)}^{u,t}\right]}{\sum_v\left[\hat{Y}_{e(v)}^{u,t,obj} \cdot P_{e(v)}^{u,t,obj}\right]}$. Likewise, the background prototype $c_e^{u(bg)}$ can also be obtained. Considering the possible unbalanced sampling of prostate-containing slices, EMA strategy across training steps (with a decay rate of 0.9) is applied for prototype update. Then, as shown in Fig. 2, given the feature map F_{local}^{l} of the local labeled image X_{local}^{l} from the in-training student model, we can compare $\{c_e^{u(obj)}, c_e^{u(bg)}\}$ with the features F_{local}^{l} pixel-by-pixel and obtain the prototype-propagated prediction

P_e^{u2l} for X_{local}^l, formulated as: $P_e^{u2l} = \frac{\exp\left(\mathrm{sim}\left(F_{local}^l, c_e^{u(i)}\right)/T\right)}{\sum_{i\in\{obj,bg\}}\exp\left(\mathrm{sim}\left(F_{local}^l, c_e^{u(i)}\right)/T\right)}$, where we use cosine similarity for $\mathrm{sim}(\cdot,\cdot)$ and empirically set the temperature T to 0.05 [19]. Note that a similar procedure can also be applied to the local unlabeled data X_{local}^u, and thus we can obtain another prototype-propagated unlabeled-to-labeled prediction P_{local}^{u2l} for X_{local}^l. As such, given the accurate expert label Y_{local}^l, the unlabeled-to-labeled supervision can be computed as:

$$\mathcal{L}_{u2l} = \frac{1}{K}\left(\sum_{k=1}^{K}\mathcal{L}_{\mathrm{Dice}}\left(P_e^{u2l,k}, Y_{local}^{l,k}\right) + \sum_{k=1}^{K}\mathcal{L}_{\mathrm{Dice}}\left(P_{local}^{u2l,k}, Y_{local}^{l,k}\right)\right). \tag{2}$$

Category-Level Regularization. Being a challenging scheme itself, the above U2L learning can only handle minor intra-class variation. Thus, proper mechanisms are needed to alleviate the negative impact of significant shift and multiple distributions. Specifically, we introduce category-level regularization, which advocates class prototype alignment between local and external data, to regularize the distribution of intra-class features from arbitrary external data to be closer to the local one, thus reducing the difficulty of U2L learning. In U2L, we have obtained prototypes from local unlabeled data $\{c_{local}^{u(obj)}, c_{local}^{u(bg)}\}$ and external unlabeled data $\{c_e^{u(obj)}, c_e^{u(bg)}\}$. Similarly, the prototypes of object $c_{local}^{l(obj)}$ and background $c_{local}^{l(bg)}$ of the local labeled data can be obtained but using expert labels and student's features. Then, the category-level regularization is formulated as:

$$\mathcal{L}_{cr} = \frac{3}{4}\left[d(c_{local}^{l(obj)}, c_e^{u(obj)}) + d(c_{local}^{u(obj)}, c_e^{u(obj)})\right] + \frac{1}{4}\left[d(c_{local}^{l(bg)}, c_e^{u(bg)}) + d(c_{local}^{u(bg)}, c_e^{u(bg)})\right], \tag{3}$$

where mean squared error is adopted as the distance function $d(\cdot,\cdot)$. The weight of background prototype alignment is smaller due to less relevant contexts.

Stability Under Perturbations. Although originally designed for typical SSL, encouraging stability under perturbations [26] can also benefit MS-SSL, considering that the perturbations can be regarded as a simulation of heterogeneity and enforcing such perturbed stability can regularize the model behavior for better generalizability. Specifically, for the same unlabeled input $X^u \in \{\mathcal{D}_{local}^u \cup \mathcal{D}_e^u\}$ with different perturbations ξ and ξ' (using the same Gaussian noises as in [26]), we encourage consistent pre-softmax predictions between the teacher and student models, formulated as $\mathcal{L}_{sta}^u = d\left(f_{\tilde{\theta}}^t(X^u+\xi), f_\theta^s(X^u+\xi')\right)$, where mean squared error is also adopted as the distance function $d(\cdot,\cdot)$.

Overall, the final loss for the multi-site unlabeled data is summarized as:

$$\mathcal{L}^u = \mathcal{L}_{PL}^u(\mathcal{D}_{local}^u) + [\mathcal{L}_{u2l}(\mathcal{D}_{local}, \mathcal{D}_e^u) + \mathcal{L}_{cr}(\mathcal{D}_{local}, \mathcal{D}_e^u)] + \mathcal{L}_{sta}^u(\mathcal{D}_{local}^u, \mathcal{D}_e^u). \tag{4}$$

3 Experiments and Results

Materials. We utilize prostate T2-weighted MR images from six different clinical centers (C1–6) [1,4,5] to perform a retrospective evaluation. Table 1 summa-

Table 2. Quantitative comparison. $*$ indicates $p \leq 0.05$ from the Wilcoxon signed rank test for pairwise comparison with our method. Standard deviations are shown in parentheses. The best mean results are shown in **bold**.

Method	# Scans Used			Local Site: C1		Local Site: C2	
	$\mathcal{D}^l_{local}$	$\mathcal{D}^u_{local}$	$\mathcal{D}^u_e$	DSC (%)	Jaccard (%)	DSC (%)	Jaccard (%)
Supervised	6	0	0	66.78 (23.26)*	54.24 (23.74)*	71.19 (16.01)*	57.33 (16.80)*
MT [15]	6	12	86	80.96 (11.15)*	70.14 (14.83)*	77.38 (10.24)*	64.21 (13.19)*
UA-MT [26]	6	12	86	81.86 (13.82)*	70.77 (17.31)*	78.31 (10.34)*	65.47 (13.26)*
ICT [17]	6	12	86	78.52 (16.82)*	67.25 (19.02)*	77.67 (9.22)*	64.38 (11.81)*
CCT [11]	6	12	86	79.95 (17.27)*	69.40 (19.55)*	73.20 (15.20)*	59.79 (17.28)*
FixMatch [14]	6	12	86	77.09 (18.45)*	65.69 (19.94)*	67.82 (14.80)*	53.16 (16.60)*
CPCL [25]	6	12	86	81.40 (14.42)*	71.27 (18.02)*	75.92 (13.59)*	62.96 (16.27)*
CPS [3]	6	12	86	65.02 (23.91)*	52.32 (23.65)*	43.41 (23.39)*	30.32 (17.49)*
SSNet [20]	6	12	86	80.37 (15.15)*	69.43 (17.88)*	75.62 (10.99)*	62.03 (14.06)*
AHDC [2]	6	12	86	79.52 (14.32)*	68.02 (15.34)*	74.65 (12.37)*	60.98 (14.33)*
CU2L (ours)	6	12	86	**85.93** (9.18)	**76.36** (12.87)	**80.29** (10.77)	**68.01** (13.92)
Supervised	8	0	0	69.04 (25.07)*	57.52 (25.34)*	75.86 (10.24)*	62.20 (13.18)*
MT [15]	8	10	86	80.18 (15.47)*	69.22 (17.85)*	77.90 (9.32)*	64.72 (11.99)*
UA-MT [26]	8	10	86	82.42 (12.45)*	71.75 (15.70)*	77.44 (8.94)*	64.02 (11.41)*
ICT [17]	8	10	86	79.88 (13.61)*	68.42 (16.96)*	76.80 (11.84)*	63.65 (13.75)*
CCT [11]	8	10	86	79.29 (14.21)*	67.71 (17.29)*	80.42 (7.60)*	67.31 (10.09)*
FixMatch [14]	8	10	86	80.46 (13.46)*	69.14 (16.29)*	66.17 (20.86)*	52.53 (20.05)*
CPCL [25]	8	10	86	81.34 (14.01)*	70.56 (17.13)*	77.49 (11.20)*	64.53 (13.93)*
CPS [3]	8	10	86	76.17 (15.58)*	63.74 (17.77)*	65.34 (13.53)*	50.04 (15.19)*
SSNet [20]	8	10	86	77.22 (15.09)*	65.19 (18.81)*	78.25 (9.90)*	65.27 (12.33)*
AHDC [2]	8	10	86	78.53 (16.23)*	66.01 (18.45)*	75.12 (11.23)*	61.97 (13.35)*
CU2L (ours)	8	10	86	**86.46** (6.72)	**76.74** (9.97)	**82.30** (9.93)	**70.71** (12.94)
Supervised (upper bound)	18	0	0	89.19 (4.33)	80.76 (6.71)	85.01 (4.35)	74.15 (6.44)

rizes the characteristics of the six data sources, following [7,8], where [7,8] also reveal the severity of inter-center heterogeneity here through extensive experiments. The heterogeneity comes from the differences in scanners, field strengths, coil types, disease and in-plane/through-plane resolution. Compared to C1 and C2, scans from C3 to C6 are taken from patients with prostate cancer, either for detection or staging purposes, which can cause inherent semantic differences in the prostate region to further aggravate heterogeneity. Following [7,8], we crop each scan to preserve the slices with the prostate region only and then resize and normalize it to 384 × 384 px in the axial plane with zero mean and unit variance. We take C1 or C2 as the local target center and randomly divide their 30 scans into 18, 3, and 9 samples as training, validation, and test sets, respectively.

Implementation and Evaluation Metrics. The framework is implemented on PyTorch using an NVIDIA GeForce RTX 3090 GPU. Considering the large variance in slice thickness among different centers, we adopt the 2D architecture. Specifically, 2D U-Net [12] is adopted as our backbone. The input patch size is set to 384 × 384, and the batch size is set to 36 including 12 labeled local slices, 12 unlabeled local slices and 12 unlabeled external slices. The supervised loss $\mathcal{L}^l_{sup}$

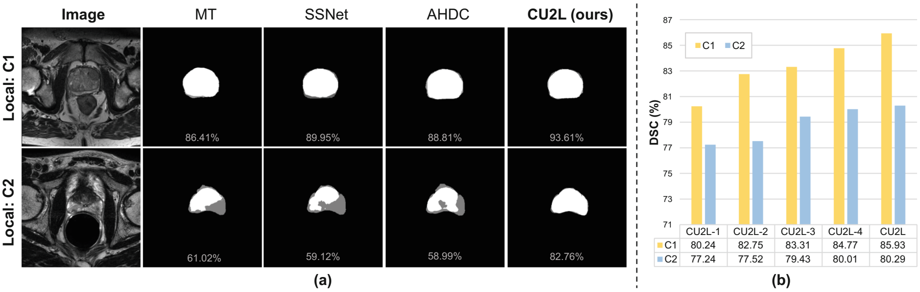


Fig. 3. (a) Exemplar results under the setting with 6 labeled scans. Grey color indicates the mismatch between the prediction and the ground truth. (b) Ablation results. (Color figure online)

consists of the cross-entropy loss and the K-regional Dice loss [6]. The maximum consistency weight w_{max} is set to 0.1 [20,26]. $t_{\max}$ is set to 20,000. K is empirically set to 2. The network is trained using the SGD optimizer and the learning rate is initialized as 0.01 and decayed by multiplication with $(1.0 - t/t_{\max})^{0.9}$. Data augmentation is applied, including random flip and rotation. We adopt the Dice similarity coefficient (DSC) and Jaccard as the evaluation metrics and the results are the average over three runs with different seeds.

Comparison Study. Table 2 presents the quantitative results with either C1 or C2 as the local target center, wherein only 6 or 8 local scans are annotated. Besides the supervised-only baselines, we include recent top-performing SSL methods [2,3,11,14,15,17,20,25,26] for comparison. All methods are implemented with the same backbone and training protocols to ensure fairness. As observed, compared to the supervised-only baselines, our CU2L with $\{6, 8\}$ local labeled scans achieves $\{19.15\%, 17.42\%\}$ and $\{9.1\%, 6.44\%\}$ DSC improvements in {C1, C2}, showing its effectiveness in leveraging multi-site unlabeled data. Despite the violation of the assumption of i.i.d. data, existing SSL methods can still benefit from the external unlabeled data to some extent compared to the results using local data only as shown in Fig. 1, revealing that the quantity of unlabeled data has a significant impact. However, due to the lack of proper mechanisms for learning from heterogeneous data, limited improvement can be achieved by them, especially for CPS [3] and FixMatch [14] in C2. Particularly, CPS relies on cross-modal pseudo labeling which exploits all the unlabeled data in a supervised-like fashion. We attribute its degradation to the fact that supervised learning is crucial for distribution fitting, which supports our motivation of performing pseudo-label learning on local unlabeled data only. As a result, its models struggle to determine which distribution to prioritize. Meanwhile, the most relevant AHDC [2] is mediocre in MS-SSL, mainly due to the instability of adversarial training and the difficulty of aligning multiple distributions to the local distribution via a single image-mapping network. In contrast, with specialized mechanisms for simultaneously learning informative representations from multi-site data and handling heterogeneity, our CU2L obtains the best performance over the recent SSL methods. Figure 3(a) further shows that the predictions of our method fit more accurately with the ground truth.

Ablation Study. To evaluate the effectiveness of each component, we conduct an ablation study under the setting with 6 local labeled scans, as shown in Fig. 2(b). Firstly, when we remove $\mathcal{L}_{PL}^{u}$ (CU2L-1), the performance drops by $\{5.69\%$ (C1), 3.05%(C2)$\}$ in DSC, showing that reinforcing confirmation on local distribution is critical. CU2L-2 represents the removal of both $\mathcal{L}_{u2l}$ and $\mathcal{L}_{cr}$, and it can be observed that such an unlabeled-to-labeled learning approach combined with class-level regularization is crucial for exploring multi-site data. If we remove $\mathcal{L}_{cr}$ which accompanies with $\mathcal{L}_{u2l}$ (CU2L-3), the performance degrades, which justifies the necessity of this regularization to reduce the difficulty of unlabeled-to-labeled learning process. CU2L-4 denotes the removal of $\mathcal{L}_{sta}^{u}$. As observed, such a typical stability loss [15] can further improve the performance by introducing hand-crafted noises to enhance the robustness to real-world heterogeneity.

4 Conclusion

In this work, we presented a novel Category-level regularized Unlabeled-to-Labeled (CU2L) learning framework for semi-supervised prostate segmentation with multi-site unlabeled MRI data. CU2L robustly exploits multi-site unlabeled data via three tailored schemes: local pseudo-label learning for better local distribution fitting, category-level regularized unlabeled-to-labeled learning for exploiting the external data in a distribution-insensitive manner and stability learning for further enhancing robustness to heterogeneity. We evaluated our method on prostate MRI data from six different clinical centers and demonstrated its superior performance compared to other semi-supervised methods.

Acknowledgement. This research was done with Tencent Jarvis Lab and Tencent Healthcare (Shenzhen) Co., LTD and supported by General Research Fund from Research Grant Council of Hong Kong (No. 14205419) and the National Key R&D Program of China (No. 2020AAA0109500 and No. 2020AAA0109501).

References

1. Bloch, N., et al.: NCI-ISBI 2013 challenge: automated segmentation of prostate structures. The Cancer Imaging Arch. **370**(6), 5 (2015)
2. Chen, J., et al.: Adaptive hierarchical dual consistency for semi-supervised left atrium segmentation on cross-domain data. IEEE Trans. Med. Imaging **41**(2), 420–433 (2021)
3. Chen, X., Yuan, Y., Zeng, G., Wang, J.: Semi-supervised semantic segmentation with cross pseudo supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2613–2622 (2021)

4. Lemaître, G., Martí, R., Freixenet, J., Vilanova, J.C., Walker, P.M., Meriaudeau, F.: Computer-aided detection and diagnosis for prostate cancer based on mono and multi-parametric MRI: a review. Comput. Biol. Med. **60**, 8–31 (2015)
5. Litjens, G., et al.: Evaluation of prostate segmentation algorithms for MRI: the PROMISE12 challenge. Med. Image Anal. **18**(2), 359–373 (2014)
6. Liu, J., Desrosiers, C., Zhou, Y.: Semi-supervised medical image segmentation using cross-model pseudo-supervision with shape awareness and local context constraints. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13438, pp. 140–150. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16452-1_14
7. Liu, Q., Dou, Q., Heng, P.-A.: Shape-aware meta-learning for generalizing prostate MRI segmentation to unseen domains. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12262, pp. 475–485. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59713-9_46
8. Liu, Q., Dou, Q., Yu, L., Heng, P.A.: MS-net: multi-site network for improving prostate segmentation with heterogeneous MRI data. IEEE Trans. Med. Imaging **39**(9), 2713–2724 (2020)
9. Luo, X., Chen, J., Song, T., Chen, Y., Wang, G., Zhang, S.: Semi-supervised medical image segmentation through dual-task consistency. In: AAAI Conference on Artificial Intelligence (2021)
10. Luo, X., et al.: Efficient semi-supervised gross target volume of nasopharyngeal carcinoma segmentation via uncertainty rectified pyramid consistency. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12902, pp. 318–329. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87196-3_30
11. Ouali, Y., Hudelot, C., Tami, M.: Semi-supervised semantic segmentation with cross-consistency training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12674–12684 (2020)
12. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
13. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. In: Advances in Neural Information Processing Systems, vol. 30, pp. 4080–4090 (2017)
14. Sohn, K., et al.: FixMatch: simplifying semi-supervised learning with consistency and confidence. arXiv preprint arXiv:2001.07685 (2020)
15. Tarvainen, A., Valpola, H.: 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 (2017)
16. Van Engelen, J.E., Hoos, H.H.: A survey on semi-supervised learning. Mach. Learn. **109**(2), 373–440 (2020)
17. Verma, V., et al.: Interpolation consistency training for semi-supervised learning. Neural Netw. **145**, 90–106 (2022)
18. Wang, G., et al.: A noise-robust framework for automatic segmentation of COVID-19 pneumonia lesions from CT images. IEEE Trans. Med. Imaging **39**(8), 2653–2663 (2020)
19. Wang, K., Liew, J.H., Zou, Y., Zhou, D., Feng, J.: PANet: few-shot image semantic segmentation with prototype alignment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9197–9206 (2019)

20. Wu, Y., Wu, Z., Wu, Q., Ge, Z., Cai, J.: Exploring smoothness and class-separation for semi-supervised medical image segmentation. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13435, pp. 34–43. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16443-9_4
21. Wu, Y., Xu, M., Ge, Z., Cai, J., Zhang, L.: Semi-supervised left atrium segmentation with mutual consistency training. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12902, pp. 297–306. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87196-3_28
22. Xu, Z., et al.: Anti-interference from noisy labels: Mean-teacher-assisted confident learning for medical image segmentation. IEEE Trans. Med. Imaging **41**, 3062–3073 (2022)
23. Xu, Z., et al.: Noisy labels are treasure: mean-teacher-assisted confident learning for hepatic vessel segmentation. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12901, pp. 3–13. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87193-2_1
24. Xu, Z., et al.: Ambiguity-selective consistency regularization for mean-teacher semi-supervised medical image segmentation. Med. Image Anal. **88**, 102880 (2023)
25. Xu, Z., et al.: All-around real label supervision: cyclic prototype consistency learning for semi-supervised medical image segmentation. IEEE J. Biomed. Health Inform. **26**(7), 3174–3184 (2022)
26. Yu, L., Wang, S., Li, X., Fu, C.-W., Heng, P.-A.: Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11765, pp. 605–613. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32245-8_67
27. Zhai, S., Wang, G., Luo, X., Yue, Q., Li, K., Zhang, S.: PA-seg: learning from point annotations for 3D medical image segmentation using contextual regularization and cross knowledge distillation. IEEE Trans. Med. Imaging **42**(8), 2235–2246 (2023). https://doi.org/10.1109/TMI.2023.3245068
28. Zhang, W., et al.: 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 (2022)