Towards Expert-Amateur Collaboration: Prototypical Label Isolation Learning for Left Atrium Segmentation with Mixed-Quality Labels

Zhe Xu[1], Jiangpeng Yan[3], Donghuan Lu[2(✉)], Yixin Wang[4], Jie Luo[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
caleblu@tencent.com
[3] Department of Automation, Tsinghua University, Beijing, China
[4] Department of Bioengineering, Stanford University, Stanford, CA, USA
[5] Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA

Abstract. Deep learning-based medical image segmentation usually requires abundant high-quality labeled data from experts, yet, it is often infeasible in clinical practice. Without sufficient expert-examined labels, the supervised approaches often struggle with inferior performance. Unfortunately, directly introducing additional data with low-quality cheap annotations (e.g., crowdsourcing from non-experts) may confuse the training. To address this, we propose a Prototypical Label Isolation Learning (PLIL) framework to robustly learn left atrium segmentation from scarce high-quality labeled data and massive low-quality labeled data, which enables effective expert-amateur collaboration. Particularly, PLIL is built upon the popular teacher-student framework. Considering the structural characteristics that the semantic regions of the same class are often highly correlated and the higher noise tolerance in the high-level feature space, the self-ensembling teacher model isolates clean and noisy labeled voxels by exploiting their relative feature distances to the class prototypes via multi-scale voting. Then, the student follows the teacher's instruction for adaptive learning, wherein the clean voxels are introduced as supervised signals and the noisy ones are regularized via perturbed stability learning, considering their large intra-class variation. Comprehensive experiments on the left atrium segmentation benchmark demonstrate the superior performance of our approach.

Keywords: Image Segmentation · Class Prototype · Label Noises

1 Introduction

Segmenting the left atrium (LA) from magnetic resonance images (MRI) is critical in treating atrial fibrillation. Recent success of deep learning (DL)-

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14226, pp. 99–109, 2023.
https://doi.org/10.1007/978-3-031-43990-2_10

based methods usually requires a large amount of high-quality (HQ) labeled data (termed as Set-HQ). However, since labeling medical images is expertise-demanding and laborious, acquiring massive HQ labeled data from experts is expensive and not always feasible. Without sufficient HQ labels, the DL approaches often struggle with inferior performance. Despite the recent success of semi-supervised learning (SSL) that leverages abundant unlabeled data [3,11,20,21], it is still difficult for SSL to accurately propagate label information at the voxel level especially when the HQ labeled data is extremely scarce. Thus, an intuitive cost-efficient alternative is to collect additional labels via cheaper ways, e.g., crowdsourcing from non-experts, as depicted in Fig. 1. Unfortunately, the quality of cheap labels is always unsatisfactory. Directly introducing additional data with low-quality (LQ) noisy labels (termed as Set-LQ) may mislead the model training, easily causing performance degradation [10,18]. Such a pervasive dilemma poses a challenging yet practical scenario: how to robustly learn segmentation from scarce HQ labeled data and abundant LQ noisy labeled data?

The existing works on mining LQ labeled data for medical image segmentation can be categorized by two distinct application scenarios: (i) **HQ-agnostic**, e.g., Set-HQ and Set-LQ are mixed as one dataset [6,9,24,26–28]. TriNet [24] uses a tri-network that integrates predictions from two peer networks to supervise the third network; PNL [28] introduces an image-level label quality evaluation module to identify clean labels to tune the network. (ii) **HQ-aware**, e.g., recruiting experts to obtain a reasonable amount of HQ labeled data and thus Set-HQ and Set-LQ are separate. Such scenario extends SSL [3,11,20,21] to further exploit the potentially useful information of LQ labels, which will be more beneficial when the HQ labeled data is extremely scarce (as detailed in Sect. 3). Luo et al. [10] proposed to implicitly decouple the learning processes for Set-HQ and Set-LQ using two separate decoders; KDEM [5] extends [10] with knowledge distillation and entropy minimization regularization. However, this implicit decoupling strategy is experimentally hard-to-control. Thus, MTCL [18] estimates the joint distribution matrix between observed and latent true labels to explicitly characterize mislabeled locations for smooth label refurbishment. However, MTCL is based on the class-conditional noise (CCN) assumption that the noise is independent of input features given the true label, which may be impractical [2]. Considering the clinical practice, we advocate the HQ-aware scenario because: (a) HQ/LQ labeled data can be separated since the sources of medical annotation are usually recorded and acquiring a reasonable amount of HQ labels from radiologists is feasible; (b) the separation may implicitly embed rewarding prior knowledge on discriminating HQ/LQ labeled data into training.

Tailoring for the HQ-aware scenario, in this work, we propose the Prototypical Label Isolation Learning (PLIL) framework for left atrium segmentation, enabling effective expert-amateur collaboration. Specifically, PLIL is built upon the popular teacher-student framework. Besides the prime supervised signals from HQ labeled data, PLIL robustly exploits the additional LQ labeled data via two steps: (i) Considering the structural characteristics that semantic regions of the same class are often highly correlated and the higher noise tolerance in

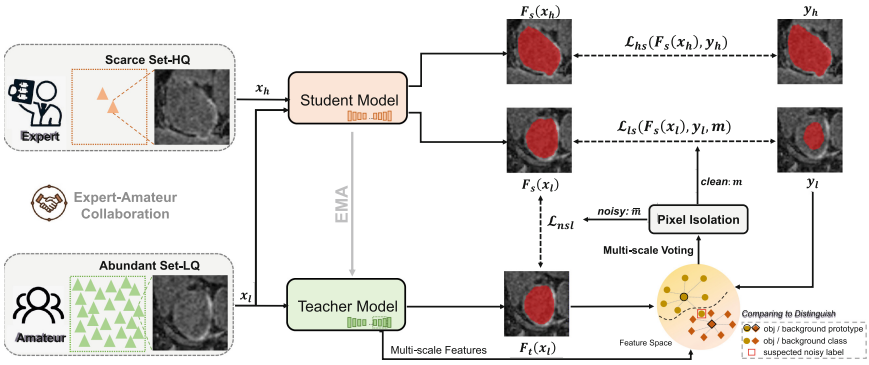

Fig. 1. Overview of our prototypical label isolation learning (PLIL) framework for robustly learning segmentation with scarce HQ labeled data and abundant LQ labeled data. m is the estimated clean-label selection mask; $\bar{m}$ is the noisy-label selection mask.

the high-level feature space [13,23], the self-ensembling teacher model isolates clean and noisy labeled voxels by exploiting their relative feature distances to the class prototypes via multi-scale voting. Besides the advantage of explicit spatial isolation, this strategy takes the input features into account, which is more realistic compared to [18] as the mislabeled voxels often present difficult and ambiguous regions in the image. (ii) Synergistically, the student follows the teacher's instruction for adaptive learning, wherein the clean voxels are further introduced as supervised signals and the noisy ones are especially regularized via perturbed stability learning, considering their vulnerable large intra-class variation in general. Comprehensive experiments on left atrium segmentation under extreme budget settings demonstrate the superior performance of our approach. The ablation study further verifies the effectiveness of each component.

2 Methods

2.1 Problem Formulation

Our PLIL framework is depicted in Fig. 1. Following the HQ-aware scenario, we have access to scarce expert-examined HQ labeled data $\mathcal{S}_h = \left\{\left(x_{h(i)}, y_{h(i)}\right)\right\}_{i=1}^{M}$ that only contains M samples, and abundant non-expert LQ noisy labeled data $\mathcal{S}_l = \left\{\left(x_{l(i)}, y_{l(i)}\right)\right\}_{i=M+1}^{N}$ that consists of $N-M$ (usually $\gg M$) samples, where $x_{h(i)}, x_{l(i)} \in \mathbb{R}^{\Omega_i}$ denote the images and $y_{h(i)}, y_{l(i)} \in \{0,1\}^{\Omega_i \times C}$ are the given HQ or LQ label (C denotes the class number). Our goal is to learn segmentation with scarce Set-HQ and abundant Set-LQ by optimizing the following loss function:

$$\mathcal{L} = \mathcal{L}_{HQ} + \lambda \mathcal{L}_{LQ}, \tag{1}$$

where $\mathcal{L}_{HQ}$ and $\mathcal{L}_{LQ}$ denote the guidance from HQ and LQ labeled data, respectively. λ is a trade-off weight for $\mathcal{L}_{LQ}$, scheduled by the time-dependent ramp-up

Gaussian function [4] $\lambda(t) = e^{-5\left(1-\frac{t}{t_{max}}\right)^2}$, where t is the current iteration and $t_{\max}$ is the maximal iteration. Since our method heavily relies on the manipulation in the feature space, such weighting schedule can reduce the interference of LQ labeled data to the feature space learning at the early training stage. The HQ labeled data provides prime HQ supervised guidance $\mathcal{L}_{hs}$, i.e., $\mathcal{L}_{HQ} = \mathcal{L}_{hs}$. Following [21], we adopt the cross-entropy loss $\mathcal{L}_{ce}$ and Dice loss $\mathcal{L}_{dice}$ with equal weights for $\mathcal{L}_{hs}$. To further exploit Set-LQ while alleviating confirmation bias [10], we aim to spatially isolate the *clean* and *noisy* labeled voxels and make better use of the suspected *noisy* labeled voxels rather than discarding them.

2.2 Prototypical Label Isolation for Adaptive Learning

Teacher-Student Architecture. Our framework is built upon the popular teacher-student architecture [15], where the student model F_s is updated by back-propagation and the teacher F_t is updated by the exponential moving average (EMA) weights of the student θ across training steps. Denoting the weights of the teacher model at step t as $\tilde{\theta}_t$, $\tilde{\theta}_t$ is updated by: $\tilde{\theta}_t = \alpha\tilde{\theta}_{t-1} + (1-\alpha)\theta_t$, where α is the EMA decay rate and empirically set to 0.99 [15]. As such, the teacher model owns the self-ensembling property [4], which can avoid sharp deterioration of the feature quality and thus suits our following prototypical label isolation strategy that appreciates high-quality and smooth embedding space.

Multi-scale Voting-Based Prototypical Label Isolation. Considering the structural characteristics that the targeted segmentation regions of the same class are often highly correlated and the higher noise tolerance in the high-level feature space [1,13,14,23], our label isolation strategy is inherently motivated by the assumption that for a *clean* labeled voxel, its features should lie closer to its corresponding class prototype (class-wise feature centroid); otherwise, a potential *noisy* labeled voxel is suspected. Specifically, we determine whether a voxel-wise label is a *clean* one by exploiting the relative feature distances to the class prototypes. Considering that different layers perceive the entire image with different perspectives, a multi-scale voting mechanism is introduced. Technically, given a medical scan x_l of Set-LQ and its noisy label y_l, we denote the last i-th feature map from the teacher model F_t as e_i^{temp}, which is then upsampled to $e_i \in \mathbb{R}^{H\times W\times Z\times L_i}$ (H, W and D denote height, width and depth of x_l, respectively, and L_i is the channel number) to be consistent with the size of segmentation mask via trilinear interpolation. Then, we resort to the pseudo label from the teacher model as the "mask" for the target class, which will be utilized to extract the class features. Denoting the teacher's prediction of x_l as $F_t(x_l)$, the pseudo label corresponds to the class with the maximal posterior probability. Since the HQ labeled data is scarce which makes it hard to obtain confident prediction, Monte Carlo dropout [7] based model uncertainty is leveraged to calibrate the pseudo label. For the LQ labeled image x_l, K stochastic forward inferences through F_t are performed with random dropout. Then, the normalized predictive entropy of the mean of the K softmax predictions is regarded as

the uncertainty map u [21]. When the uncertainty u_v at voxel v is smaller than a threshold η, i.e., $u_v < \eta$, this voxel will be used as the final pseudo mask $\hat{F}_t(x_l)$. As such, at the i-th scale, the object prototype q_i^{obj} can be obtained via the masked average pooling [20,25] as: $q_i^{obj} = \frac{\sum_v \hat{F}_{t(v)}^{obj} \cdot p_{t(v)}^{obj} \cdot e_{i(v)}}{\sum_v \hat{F}_{t(v)}^{obj} \cdot p_{t(v)}^{obj}}$, where the predicted probabilities of object $p_{t(v)}^{obj}$ from the teacher model weight the contribution of voxel v to prototype generation. Similarly, the background prototype q_i^{bg} can be also obtained. Then, the relative feature distances $d_{i(v)}^{obj}$ and $d_{i(v)}^{bg}$ between the feature vector of voxel v and the prototypes are defined as:

$$d_{i(v)}^{obj} = \left\| e_{i(v)} - q_i^{obj} \right\|_2 \quad \text{and} \quad d_{i(v)}^{bg} = \left\| e_{i(v)} - q_i^{bg} \right\|_2. \tag{2}$$

Intuitively, if the given label $y_{l(v)}$ at voxel v is object (background) yet its feature vector e_v lies closer to the background (object) prototype than the object (background) prototype, this voxel will be isolated to the *noisy* group. Otherwise, it will be selected as the *clean* labeled one. Formally, the i-th scale determines the *clean-label* selection mask m_i for image x_l as:

$$m_{i(v)} = \mathbb{1}[y_{l(v)} = 1] \cdot \mathbb{1}[d_{i(v)}^{obj} < d_{i(v)}^{bg}] + \mathbb{1}[y_{l(v)} = 0] \cdot \mathbb{1}[d_{i(v)}^{obj} > d_{i(v)}^{bg}]. \tag{3}$$

We select the last three scales of features from the teacher model to perform multi-scale voting. Thus, for the final *clean-label* selection mask, $m_v = 1$ if $\sum_i^3 m_{i(v)} \geq 2$. The *noisy-label* selection mask $\bar{m}$ is the negation of m.

Adaptive Learning Scheme for Isolated Voxels. As shown in Fig. 1, the additional supervised loss for Set-LQ ($\mathcal{L}_{ls}$) is applied to the isolated *clean* labeled voxels, which takes the form of m-masked cross-entropy loss and Dice loss as:

$$\mathcal{L}_{ls} = \sum_v (m_v \cdot \mathcal{L}_{ce,v} + m_v \cdot \mathcal{L}_{dice,v}). \tag{4}$$

For the noisy group, since it is extremely difficult to perfectly find out the noisy labels, we do not advocate label refinement as in [18] to avoid additional error propagation. Instead, we regularize the model behavior on these ambiguous noisy voxels via perturbed stability learning [15], i.e., encouraging consistent pre-softmax predictions between the student and teacher model for the same input with different perturbations ξ and ξ', formulated as:

$$\mathcal{L}_{nsl} = \frac{\sum_v \bar{m}_v \left\| F_{t(v)}(x_l + \xi) - F_{s(v)}(x_l + \xi') \right\|^2}{\sum_v \bar{m}_v}. \tag{5}$$

The design of $\bar{m}$-masked stability loss is motivated by the fact that the estimated noisy group correlates with the voxels with large intra-class variation, wherein these voxels often exhibit difficult and ambiguous nature, which potentially have serious instability problem. Besides, compared to [15], such a noise-selective stability learning avoids the distraction by the redundant easy

regions, considering this loss takes the form of mean squared error (MSE) with the average nature. As such, the LQ loss $\mathcal{L}_{LQ}$ in Eq. 1 can be formulated as $\mathcal{L}_{LQ} = \mathcal{L}_{ls} + \beta\mathcal{L}_{nsl}$, where β is a tradeoff weight for the two learning manners. By combining $\mathcal{L}_{HQ}$ and $\mathcal{L}_{LQ}$, the model can not only receive HQ supervision from the scarce Set-HQ but also adaptively exploit different kinds of productive information in Set-LQ towards effective expert-amateur collaboration.

3 Experiments and Results

Materials. The left atrium (LA) segmentation dataset [17] provides 100 3D gadolinium-enhanced magnetic resonance images (GE-MRIs) with expert labels. The images have the isotropic resolution of $0.625 \times 0.625 \times 0.625$ mm^3. Following the same data preprocessing and split in [21], 80 samples are selected for training and the remaining 20 samples for testing. All the images are cropped to the center of the heart region and the intensities are normalized to zero mean and unit variance. We investigate the scenarios of scarce HQ labeled data, where only 4 (5%) or 6 (7.5%) samples are used as Set-HQ and the rest is utilized as non-expert Set-LQ, simulated by the commonly used label corruption scheme [22,28] including random erosion and dilation with 3–15 voxels.

Implementation and Evaluation Metrics. The framework is based on PyTorch using an NVIDIA GeForce RTX 3090 GPU. 3D V-Net [12] is adopted as the backbone, referring to [21]. We randomly crop patches of $112 \times 112 \times 80$ voxels as the input and use sliding window strategy with stride of $18 \times 18 \times 4$ voxels for inference. The batch size is set to 4 including 2 labeled samples and 2 unlabeled samples. $t_{\max}$ is set to 8,000. K, η and β are empirically set to 8, 0.1 and 0.1. The learning rate is initialized as 0.01 and decayed by multiplication with $(1.0 - t/t_{\max})^{0.9}$. Data augmentation, including random flip and rotation, is applied. Four metrics, including Dice, Jaccard, average surface distance (ASD) and 95% Hausdorff distance (95HD), are adopted for comprehensive evaluation. The code will be available at https://github.com/lemoshu/PLIL.

Comparison Study. The quantitative results are presented in Table 1. H-Sup denotes the supervised baseline that only Set-HQ is utilized, while HL-Sup denotes that Set-HQ and Set-LQ are mixed for supervised learning. We also include recent SSL methods (UAMT [21], CPS [3], CPCL [20] and URPC [11]), HQ-agnostic noisy label learning (NLL) methods (TriNet [24] and PNL [28]) and HQ-aware NLL methods (Decoupled [10] and MTCL [18]). All the methods are implemented with the same backbone and training protocols to ensure fairness. As observed, H-Sup performs poorly with scarce Set-HQ, yet, HL-Sup even further degrades, implying that our simulated LQ labels have led to serious confirmation bias. Relying on some model assumptions [16], SSL methods ignore the LQ labels and exploit the image information only from Set-LQ. Despite effectiveness, it is still difficult for SSL to accurately propagate voxel-level label

Table 1. Quantitative comparison study. Cross-subject standard deviations are shown in parentheses. $*$ indicates $p \leq 0.05$ from Wilcoxon signed rank test when comparing ours with the second best under the HQ-aware setting. The best results are in **bold**.

Methods		Settings			Metrics			
		Set-HQ	Set-LQ	HQ-aware?	Dice [%] ↑	Jaccard [%] ↑	95HQ [voxel] ↓	ASD [voxel] ↓
Sup	H-Sup (Upper Bound)	80	0	-	91.25 (1.93)	83.36 (3.24)	6.32 (6.45)	1.50 (0.61)
	H-Sup	4	0	-	77.84 (8.29)	64.02 (9.60)	22.60 (14.57)	5.03 (1.22)
	HL-Sup	4	76	-	77.21 (7.52)	63.47 (9.66)	21.59 (10.72)	6.35 (2.31)
SSL	UAMT [21]	4	76	-	79.88 (9.69)	67.46 (11.93)	24.11 (14.75)	3.21 (1.56)
	CPS [3]	4	76	-	81.54 (6.56)	69.33 (8.89)	24.54 (16.47)	3.86 (1.08)
	CPCL [20]	4	76	-	82.46 (8.15)	70.92 (11.02)	21.59 (14.49)	3.23 (1.32)
	URPC [11]	4	76	-	78.41 (8.53)	65.27 (14.36)	21.74 (15.03)	6.17 (2.33)
NLL	TriNet [24]	4	76	×	80.12 (7.11)	68.11 (8.36)	20.13 (11.74)	4.85 (1.24)
	PNL [28]	4	76	×	78.01 (7.23)	65.01 (8.22)	18.57 (13.05)	4.78 (1.33)
	Decoupled [10]	4	76	✓	80.74 (6.73)	68.23 (7.31)	16.55 (13.54)	4.90 (1.22)
	MTCL [18]	4	76	✓	83.36 (6.12)	71.53 (8.48)	16.81 (14.13)	3.44 (1.51)
	PLIL (ours)	4	76	✓	**84.91** (3.32)*	**73.93** (5.07)*	**15.49** (12.00)	**3.10** (1.31)
Sup	H-Sup	6	0	-	79.41 (4.86)	65.02 (5.11)	24.36 (14.02)	2.78 (1.01)
	HL-Sup	6	74	-	78.09 (4.45)	64.27 (5.78)	13.18 (9.63)	4.14 (0.76)
SSL	UAMT [21]	6	74	-	83.72 (7.15)	73.10 (9.77)	16.44 (14.07)	2.75 (1.10)
	CPS [3]	6	74	-	82.15 (6.89)	70.26 (9.27)	27.61 (15.07)	2.92 (0.90)
	CPCL [20]	6	74	-	83.99 (5.40)	73.55 (7.56)	19.60 (11.80)	2.79 (0.97)
	URPC [11]	6	74	-	80.52 (7.77)	68.14 (9.53)	22.81 (13.67)	6.18 (1.54)
NLL	TriNet [24]	6	74	×	84.82 (3.68)	74.04 (7.29)	15.37 (7.62)	3.01 (1.19)
	PNL [28]	6	74	×	80.05 (4.72)	68.08 (8.53)	17.02 (10.23)	3.58 (0.83)
	Decoupled [10]	6	74	✓	85.01 (3.76)	74.58 (6.43)	12.35 (8.36)	3.37 (1.02)
	MTCL [18]	6	74	✓	86.06 (4.78)	75.73 (7.18)	12.47 (10.06)	2.87 (1.09)
	PLIL (ours)	6	74	✓	**87.66** (2.61)	**78.12** (4.15)*	**10.93** (9.29)*	**2.41** (0.83)

information when the HQ labeled data is scarce. For the HQ-agnostic methods, TriNet and PNL show effectiveness in alleviating the negative effects brought by the agnostic LQ labels, yet, even fall behind some SSL methods given the violent simulated label noises, revealing that the HQ-agnostic setting may be sub-optimal. For the HQ-aware scenario, Decoupled [10] and MTCL [18] perform well under both labeling settings, demonstrating the benefits of HQ-aware strategy. Our PLIL relies on the manipulation in the feature space, more HQ labeled data will help the network learn more discriminative representations towards accurate isolation. As observed in the 6-HQ-sample setting, PLIL achieves the Dice of 87.66%, only 3.59% away from the upper bound trained with all 80 HQ labeled data. Despite less-discriminative features learned under the 4-HQ-sample setting, PLIL can still achieve respectable results, demonstrating its robustness. The impact of varying expert labeling budgets is further illustrated in Fig. 2(c) while Fig. 2(a) presents exemplar results of our PLIL and other approaches under the 6-HQ-sample setting. Consistently, the predicted mask of our PLIL fits more accurately with the ground truth. To better understand our method, we visualize the estimated noisy-label selection mask $\bar{m}$ for a dilated LQ label y_l of LA in Fig. 2(b), where it can be observed that most dilated regions are well-characterized, further demonstrating the efficacy of our label selection strategy.

Ablation Study and Discussions. To further investigate how our method works, we perform an ablation study under the 4-HQ-sample setting (as presented in Table 2) with the following variants: (i) **PLIL (HQ-agnostic)**: mixing Set-HQ and Set-LQ instead of the separate strategy; (ii) **w/o multi-scale voting**: only utilizing the features before the penultimate convolution for autocratic label isolation; (iii) **w/o** $\mathcal{L}_{ls}$: removing the m-masked supervised loss for the identified clean labeled group; (iv) **w/o** $\mathcal{L}_{nsl}$: removing the $\bar{m}$-masked (noise-selective) stability loss for the identified noisy labeled group. First, our PLIL is tailored for the HQ-aware scenario. As shown in Table 2, the HQ-agnostic input has interfered with the network training and led to obvious performance degradation, showing the efficacy of our separate strategy. We also observe that the arbitration-based multi-scale voting mechanism enables more reliable isolation due to the consideration of different perspectives of the images. When removing $\mathcal{L}_{ls}$, considerable performance degradation can be observed, revealing that our strategy effectively finds out the clean labeled voxels. Besides the productive guidance provided by the isolated clean labeled voxels, the noisy group exploited by the stability learning can further provide informative clues to boost the performance. Empirically, the noisy labeled regions often appear in the challenging areas, which are more sensitive to the perturbations and therefore exploring their perturbed stability during training is rewarding and can enhance the generalizability of the model [15,19]. However, as a methodological study, we only evaluated the methods with the commonly used simulated LQ noisy labels. As observed, the violent simulated noises lead to serious confirmation bias. Some existing NLL methods cannot handle such violent noises well, but some SSL methods, which discard LQ labels, achieve appealing performance. Thus, further clinical validation with real-world amateur noises is an important future work. Besides, to facilitate practical expert-amateur collaboration, we should further consider two intertwined problems in the future: (i) how to cost-efficiently edu-

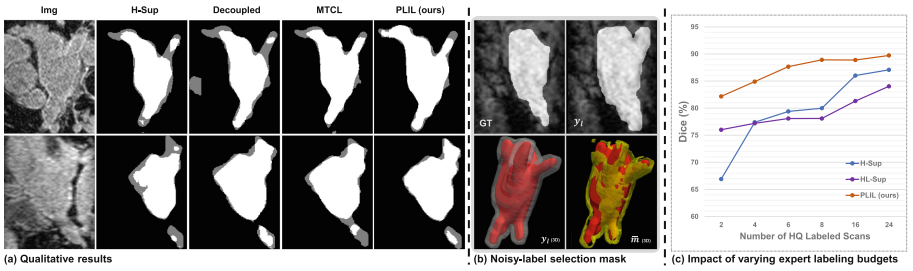

Fig. 2. (a) Examples of LA segmentation results with only 6 HQ labeled data. Grey color represents the inconsistency between the prediction and the ground truth (GT). (b) An example of the dilated LQ label y_l (white in 3D with fused red GT) and the estimated noisy-label selection mask $\bar{m}$ (yellow in 3D with fused red GT). (c) Segmentation performances (indicated by Dice score) with varying expert labeling budgets. (Color figure online)

Table 2. Ablation study with 5% HQ labeled data. The best mean results are in **bold**.

Methods	Metrics			
	Dice [%] ↑	Jaccard [%] ↑	95HQ [voxel] ↓	ASD [voxel] ↓
PLIL (HQ-aware)	**84.91** (3.32)	**73.93** (5.07)	**15.49** (12.00)	**3.10** (1.31)
PLIL (HQ-agnostic)	79.64 (7.46)	66.76 (9.50)	19.48 (11.01)	6.13 (2.55)
w/o multi-scale voting	83.36 (6.83)	72.11 (7.93)	17.36 (11.54)	3.65 (1.77)
w/o $\mathcal{L}_{ls}$	80.65 (9.36)	68.47 (11.53)	23.11 (13.66)	3.12 (1.11)
w/o $\mathcal{L}_{nsl}$	83.22 (7.36)	71.86 (9.59)	19.69 (14.26)	3.26 (1.67)

cate the amateurs on good medical annotation; (ii) how to automatically perform quality controls for the crowdsourced pixel-level labels [8].

4 Conclusion

In this work, we proposed a novel Prototypical Label Isolation Learning (PLIL) framework to robustly learn left atrium segmentation from scarce high-quality labeled data and massive low-quality labeled data. Taking advantage of our multi-scale voting-based prototypical label isolation and adaptive learning scheme for clean and suspected noisy labeled voxels, our approach can robustly exploit the additional low-quality labeled data (e.g., via cheap crowdsourcing), which enables effective expert-amateur collaboration. Comprehensive experiments on the left atrium segmentation benchmark demonstrated the superior performance of our method as well as the effectiveness of each proposed component.

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. Chen, C., Liu, Q., Jin, Y., Dou, Q., Heng, P.-A.: Source-free domain adaptive fundus image segmentation with denoised pseudo-labeling. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12905, pp. 225–235. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87240-3_22
2. Chen, P., Ye, J., Chen, G., Zhao, J., Heng, P.A.: Beyond class-conditional assumption: A primary attempt to combat instance-dependent label noise. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 11442–11450 (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. Cui, W., et al.: Semi-supervised brain lesion segmentation with an adapted mean teacher model. In: Chung, A.C.S., Gee, J.C., Yushkevich, P.A., Bao, S. (eds.) IPMI 2019. LNCS, vol. 11492, pp. 554–565. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-20351-1_43
5. Dolz, J., Desrosiers, C., Ayed, I.B.: Teach me to segment with mixed supervision: confident students become masters. In: Feragen, A., Sommer, S., Schnabel, J., Nielsen, M. (eds.) IPMI 2021. LNCS, vol. 12729, pp. 517–529. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-78191-0_40
6. Guo, X., Yuan, Y.: Joint class-affinity loss correction for robust medical image segmentation with noisy labels. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13434, pp. 588–598. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16440-8_56
7. Kendall, A., Gal, Y.: What uncertainties do we need in Bayesian deep learning for computer vision? arXiv preprint arXiv:1703.04977 (2017)
8. Kentley, J., et al.: Agreement between experts and an untrained crowd for identifying dermoscopic features using a gamified app: reader feasibility study. JMIR Med. Inform. **11**(1), e38412 (2023)
9. Li, S., Gao, Z., He, X.: Superpixel-guided iterative learning from noisy labels for medical image segmentation. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12901, pp. 525–535. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87193-2_50
10. Luo, W., Yang, M.: Semi-supervised semantic segmentation via strong-weak dual-branch network. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12350, pp. 784–800. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58558-7_46
11. 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
12. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: fully convolutional neural networks for volumetric medical image segmentation. In: Fourth International Conference on 3D Vision, pp. 565–571. IEEE (2016)
13. Qu, Y., Mo, S., Niu, J.: DAT: training deep networks robust to label-noise by matching the feature distributions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6821–6829 (2021)
14. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. In: Advances in Neural Information Processing Systems, pp. 4080–4090 (2017)
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. Xiong, Z., et al.: A global benchmark of algorithms for segmenting the left atrium from late gadolinium-enhanced cardiac magnetic resonance imaging. Med. Image Anal. **67**, 101832 (2021)
18. 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
19. Xu, Z., et al.: Ambiguity-selective consistency regularization for mean-teacher semi-supervised medical image segmentation. Med. Image Anal. **88**, 102880 (2023)

20. 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**, 3174–3184 (2022)
21. 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
22. Zhang, M., et al.: Characterizing label errors: confident learning for noisy-labeled image segmentation. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12261, pp. 721–730. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59710-8_70
23. Zhang, P., Zhang, B., Zhang, T., Chen, D., Wang, Y., Wen, F.: Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12414–12424 (2021)
24. Zhang, T., Yu, L., Hu, N., Lv, S., Gu, S.: Robust medical image segmentation from non-expert annotations with tri-network. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12264, pp. 249–258. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59719-1_25
25. Zhang, X., Wei, Y., Yang, Y., Huang, T.S.: SG-One: similarity guidance network for one-shot semantic segmentation. IEEE Trans. Cybern. **50**(9), 3855–3865 (2020)
26. Zhang, Z., Zhang, H., Arik, S.O., Lee, H., Pfister, T.: Distilling effective supervision from severe label noise. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9294–9303 (2020)
27. Zhou, X., Liu, X., Wang, C., Zhai, D., Jiang, J., Ji, X.: Learning with noisy labels via sparse regularization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 72–81 (2021)
28. Zhu, H., Shi, J., Wu, J.: Pick-and-learn: automatic quality evaluation for noisy-labeled image segmentation. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11769, pp. 576–584. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32226-7_64