ReCo-I2P: An Incomplete Supervised Lymph Node Segmentation Framework Based on Orthogonal Partial-instance Annotation

Litingyu Wang[1], Ping Ye[1], Wenjun Liao[1,2], Shichuan Zhang[1,2], Shaoting Zhang[1,3], and Guotai Wang[1,3](✉)

[1] School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Chengdu, China
[2] Department of Radiation Oncology, Sichuan Cancer Hospital & Institute, Sichuan Cancer Center, School of Medicine, University of Electronic Science and Technology of China, Chengdu, China
[3] Shanghai Artificial Intelligence Laboratory, Shanghai, China
guotai.wang@uestc.edu.cn

Abstract. Quantitative analysis of lymph node volume is instrumental in the diagnosis and treatment of cancer. However, automatic segmentation models for lymph nodes necessitate pixel-level labeling, which is both time-consuming and labor-intensive. The scarcity of pixel-level annotations has thus spurred interest in label-efficient learning as a potential solution. Considering the variance of shapes and locations, and the low-contrast appearance of lymph nodes in computed tomography scans, we propose a new incomplete annotation strategy called orthogonal partial-instance annotation, in which only two orthogonal slices of a small portion of lymph nodes are annotated. To segment as many lymph nodes as possible from such sparse annotations, we propose a prototype-based label-efficient learning framework with a specifically designed loss. Specifically, we extract intra-batch prototypes from the output features of the encoder and store inter-batch prototypes using a momentum-smoothing approach. To re-inject the extracted information from the two kinds of prototypes, we introduce a feature augmentation module that utilizes the extracted prototypes to enhance features. To further complement the predictions generated from enhanced features with those from original features, we design a reliability-based co-teaching strategy based on feature similarity. Experiments demonstrate that our proposed framework outperforms other methods on two mediastinal lymph node datasets. Our implementation is available at https://github.com/HiLab-git/WCODE-PIA.

Keywords: Lymph node segmentation · Incomplete supervised learning · Prototype-based learning

L. Wang and P. Ye contributed equally to this work.

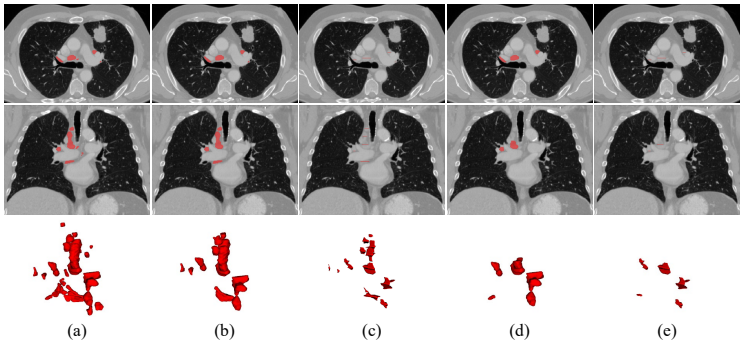

Fig. 1. Illustrations of PIA and oPIA with different annotated ratios. Red represents the pixels annotated by human experts. (a) ground truth - 100%, (b) PIA - 50%, (c) orthogonal PIA - 50%, (d) PIA - 20%, and (e) orthogonal PIA - 20%.

1 Introduction

Lung cancer is characterized by the development of malignant neoplasms, which manifest as uncontrolled cell proliferation within the pulmonary tissues. These neoplasms can also metastasize to adjacent anatomical structures within the lung, such as the lymph nodes, resulting in similar clinical manifestations [1]. As a result, accurate identification and segmentation of mediastinal lymph nodes are critical for evaluating disease progression, facilitating cancer diagnosis, and guiding therapeutic strategies [4,16]. Clinically, contrast-enhanced chest Computed Tomography (CT) scans are the most frequently favored modality in the diagnostic process. Moreover, not only are the enlarged lymph nodes (e.g., those with a short-axis diameter greater than 10 mm) of diagnostic significance, but smaller lymph nodes can also aid in diagnosis, as suggested by [13,17]. Although deep learning methods [3,19] have achieved remarkable success in automatic segmentation, accurately annotating target areas is time-consuming and labor-intensive due to the relatively low contrast between lymph nodes and surrounding tissues in CT scans, as well as the highly variable shapes, sizes, and locations of the lymph nodes.

Label-efficient learning methods, such as using inexact [6,9], incomplete [8,22], and inaccurate [7,20] annotation, have been proposed to reduce the annotation burden. The challenge of inexact and inaccurate learning lies in providing adequate edge information, which is extremely crucial for lymph node segmentation with CT scans. More importantly, current label-efficient learning focuses on the segmentation goal of only a single target region. For tasks with multiple target regions, such as lymph nodes [16] and Crohn's disease [8], many methods will suffer a significant performance degradation. To address these problems, Wang et al. [16] and Ju et al. [8] developed their algorithms for instance-level incomplete annotations at the image- (partial-instance annotation, PIA) and slice-level (target-level incomplete annotation, TIA), respectively. However, TIA

necessitates the labeling of all slices. With the help of the lymph node atlas, PIA (as shown in Fig. 1 (b) and (d)) is a more intuitive choice for lymph node annotation. Considering that labeling lymph nodes on 3D images is still a costly process, we propose a more sparse annotation method called orthogonal PIA (oPIA) (as shown in Fig. 1 (c) and (e)). In this way, a lymph node instance requires labeling only a single slice from the horizontal and coronal plane, which has been proven effective by [2].

We propose several strategies to better leverage oPIA, which is also applicable to PIA. 1) Due to oPIA's extremely sparse annotating information, we design a Prototype-based Feature Enhancement (PFE) strategy. Intra- and inter-batch prototypes are extracted to better aggregate foreground information within and between batches. Then, an attention module is introduced to re-inject reliable foreground information from the prototypes into the features. Finally, two predictions are generated by utilizing the original and enhanced features. 2) To further complement the two predictions, a reliability-based co-teaching strategy is developed. We obtain pseudo labels with high confidence and then evaluate the pixel-level reliability of each prediction through the features before the last layer of the network. Then, complementary information is transferred by learning pseudo labels from the perspective (the reliability map) of another prediction. 3) A combination of losses is advocated to better mine foreground and background information from oPIA, which is the basis for the effectiveness of the previous two strategies. Generally, we propose a **Re**liability-based **Co**-teaching framework enhanced with **I**ntra- and **I**nter-batch **P**rototypes (ReCo-I2P) to address incomplete annotation learning in the extremely sparse and multi-objective scenario. Experiments show that our proposed framework outperforms other state-of-the-art (SOTA) methods on both oPIA and PIA.

2 Method

Fig. 2 illustrates the proposed oPIA learning framework named ReCo-I2P. We introduce VNet [10] with one encoder θ_e and one decoder θ_d as the trainable model. During training, in addition to the prediction P_1 in the normal inference process, the encoder's feature outputs are enhanced by the proposed prototype-based feature enhancement module and then subjected to decoder inference once more to get another prediction P_2. During inference, only the inference of P_1 is needed. For the convenience of the following description, we define X, Y as a training image and the corresponding oPIA.

2.1 Prototype-based Feature Enhancement (PFE)

Prototypes can aggregate category or object information across the entire dataset, providing more informative guidance for the learning of sparse annotations. We denote $P^c \in \mathbb{R}^{B\times 1\times S}$ as the predicted probability map of class c. $F \in \mathbb{R}^{B\times C\times S}$ represents the feature used for prototype extraction. B denotes the batch size

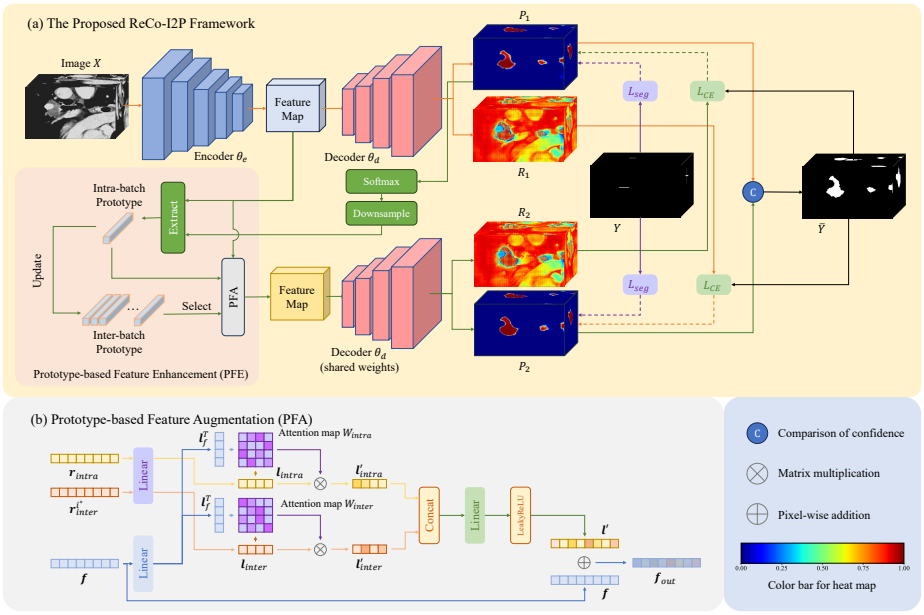

Fig. 2. The overall framework of our approach, which contains a prototype-based feature enhancement (PFE) module assisted with a reliability-based co-teaching strategy.

and S is the spatial size. The extracting process of one prototype $\boldsymbol{r}^c \in \mathbb{R}^{1\times C}$ of class c in batch-level is defined as:

$$\boldsymbol{r}^c = norm\left(\frac{\sum_i^{B\times S} P^c(i)\cdot F(i)}{\sum_i^{B\times S} P^c(i)}\right) \tag{1}$$

where $F(i)$ is a C-channel vector of the i-th pixel of F, while $P^c(i)$ is the predicted probability. $norm(\cdot)$ denotes normalization with 2-norm. In our implementation, the probability map is down-sampled to the same size as the output feature map of the encoder first, and then the prototype is computed.

However, the prototypes generated this way ignore the feature diversity and aggregation of sparse annotation information across batches. Thus, in addition to the intra-batch prototypes, we add a memory bank which contains N inter-batch prototypes. Once an intra-batch prototype $\boldsymbol{r}_{intra}$ is extracted, we calculate cosine similarity with every inter-batch prototype $\boldsymbol{r}_{inter}$ and select the one with the smallest similarity to update:

$$i^* = \underset{i}{\operatorname{argmin}}\left(\left\{SimCos(\boldsymbol{r}_{intra}, \boldsymbol{r}^i_{inter}) | i \in \{1,2,...,N\}\right\}\right) \tag{2}$$

$$\boldsymbol{r}^{i^*}_{inter} = \gamma \times \boldsymbol{r}^{i^*}_{inter} + (1-\gamma)\times \boldsymbol{r}_{intra}, \gamma \in [0,1] \tag{3}$$

$SimCos(\cdot,\cdot)$ denotes the computation of cosine similarity and γ is a hyperparameter of the memoried ratio.

To re-inject useful information of $\boldsymbol{r}_{intra}$ and $\boldsymbol{r}_{inter}$ into the features, we select the inter-batch prototype which is the most similar to the intra-batch prototype to enhance the features using the attention module shown in Fig. 2 (b). Taking the computation between an intra-batch prototype $\boldsymbol{r}_{intra}$ and $\boldsymbol{f} \in \mathbb{R}^{1\times C}$, a pixel from the feature to be enhanced, as an example. They are first linearly projected into a low-dimensional space, obtaining $\boldsymbol{l}_{intra} = linear_r(\boldsymbol{r}_{intra})$ and $\boldsymbol{l}_f = linear_f(\boldsymbol{f})$. Then, the attention map W is calculated through $W = softmax(\boldsymbol{l}_f^T \times \boldsymbol{l}_{intra})$, and apply it to $\boldsymbol{l}_{intra}$ through $\boldsymbol{l}_{intra}^{'} = W \times \boldsymbol{l}_{intra}$. We can obtain $\boldsymbol{l}_{inter}^{'}$ in the same way. Finally, after the concatenation of $\boldsymbol{l}_{intra}^{'}$ and $\boldsymbol{l}_{inter}^{'}$, a linear transformation with LeakyReLU activation is performed to get $\boldsymbol{l}^{'}$. The enhanced feature $\boldsymbol{f}_{out}$ is obtained by adding $\boldsymbol{l}^{'}$ and $\boldsymbol{f}$ together, which is then subjected to the decoder.

2.2 Reliability-based Co-teaching Strategy

Through the enhancement mentioned in Section 2.1, we inject reliable foreground information across batches into the second prediction P_2. Thus, P_1 contains a more reliable background prediction, while P_2 predicts the foreground pixels more confidently. To better utilize the complementary information from each other, firstly, we compute the corresponding confidence maps C_1 and C_2 for the two soft predictions P_1 and P_2 by comparing the predicted probability of each pixel. A high-confident pseudo label $\overline{Y}$ is generated through simple comparison, which provides a better target for our designed co-teaching strategy:

$$\overline{Y}(i) = \begin{cases} H_1(i), & C_1(i) \geq C_2(i) \\ H_2(i), & C_1(i) < C_2(i) \end{cases} \tag{4}$$

where H_1 and H_2 are hard labels generated by P_1 and P_2. However, $\overline{Y}$ inevitably has noise, with most of such occurrences being observed at the edge of the foreground areas. Therefore, we assess the reliability of predictions P_1 and P_2 through the features $F^{'}$ before the final classification layer. The class center $\boldsymbol{f}_c$ is first generated from $F^{'}$ through an average of pixels of class c. Then, the i^{th} pixel of reliability maps R_1 and R_2 is calculated through $R_1(i) = SimCos(\boldsymbol{f}_c, F_1^{'}(i))$ as an example. Finally, we design a co-teaching [7] learning strategy to robustly learn from the pseudo label $\overline{Y}$ from the perspective of another prediction. We utilize the weighted Cross-Entropy loss to complete this design:

$$\mathcal{L}_{wCE}(\hat{Y}, Y; R) = \frac{-\sum_i (R(i) \times \sum_c Y_c(i) log \hat{Y}_c(i))}{\sum_i R(i)} \tag{5}$$

$\hat{Y}$ and Y are the prediction and learning objectives, respectively. R is the weight map. The final loss for reliability-based co-teaching is:

$$\mathcal{L}_{ReCo} = \mathcal{L}_{wCE}(P_1, \overline{Y}; R_2) + \mathcal{L}_{wCE}(P_2, \overline{Y}; R_1) \tag{6}$$

Table 1. Comparison with other methods on LNQ2023 and CT Lymph Node dataset (20%) on DSC(%) and ASSD(mm). The bold represents the best-performed method.

Method	LNQ2023				CT Lymph Node (20%)			
	oPIA		PIA		oPIA		PIA	
	DSC↑	ASSD↓	DSC↑	ASSD↓	DSC↑	ASSD↓	DSC↑	ASSD↓
UpperBound	-	-	-	-	$67.46_{\pm 15.10}$	$5.13_{\pm 3.98}$	$67.46_{\pm 15.10}$	$5.13_{\pm 3.98}$
LowerBound	$11.33_{\pm 7.96}$	$31.78_{\pm 25.04}$	$39.22_{\pm 20.74}$	$24.49_{\pm 16.39}$	$5.95_{\pm 5.08}$	$25.10_{\pm 15.67}$	$44.07_{\pm 20.61}$	$15.43_{\pm 13.63}$
Co-teaching [7]	$41.52_{\pm 19.53}$	$43.20_{\pm 43.84}$	$42.93_{\pm 18.62}$	$34.52_{\pm 40.66}$	$42.20_{\pm 15.18}$	$22.52_{\pm 12.22}$	$40.79_{\pm 15.03}$	$22.51_{\pm 12.19}$
TriNet [20]	$44.08_{\pm 18.99}$	$28.77_{\pm 26.87}$	$43.78_{\pm 19.43}$	$29.40_{\pm 35.76}$	$30.67_{\pm 14.11}$	$20.91_{\pm 11.41}$	$17.58_{\pm 11.36}$	$59.40_{\pm 17.29}$
NRDice [15]	$40.38_{\pm 18.85}$	$57.30_{\pm 53.37}$	$42.05_{\pm 19.93}$	$39.83_{\pm 38.65}$	$22.97_{\pm 12.27}$	$49.04_{\pm 16.41}$	$25.46_{\pm 13.28}$	$44.28_{\pm 17.17}$
GCE [21]	$13.52_{\pm 8.62}$	$26.00_{\pm 15.69}$	$41.45_{\pm 19.09}$	$20.66_{\pm 12.21}$	$6.32_{\pm 4.86}$	$23.53_{\pm 10.77}$	$47.12_{\pm 18.83}$	$14.02_{\pm 10.51}$
RMD [5]	$42.62_{\pm 18.65}$	$35.81_{\pm 41.91}$	$43.21_{\pm 18.98}$	$30.03_{\pm 44.19}$	$30.03_{\pm 14.81}$	$40.70_{\pm 18.02}$	$29.34_{\pm 14.82}$	$39.43_{\pm 16.17}$
DeSCO [2]	$40.46_{\pm 22.73}$	$25.83_{\pm 24.05}$	$37.32_{\pm 19.31}$	$40.42_{\pm 42.40}$	$33.06_{\pm 14.42}$	$27.71_{\pm 11.61}$	$33.90_{\pm 14.39}$	$25.90_{\pm 11.98}$
DBDMP [16]	$50.24_{\pm 14.90}$	$24.18_{\pm 32.77}$	$54.59_{\pm 17.14}$	$17.00_{\pm 21.64}$	$49.94_{\pm 16.06}$	$15.53_{\pm 9.14}$	$51.99_{\pm 16.75}$	$13.69_{\pm 8.72}$
ReCo-I2P(Ours)	$\mathbf{53.94}_{\pm 14.94}$	$\mathbf{16.29}_{\pm 22.32}$	$\mathbf{57.32}_{\pm 17.08}$	$\mathbf{12.63}_{\pm 14.48}$	$\mathbf{51.59}_{\pm 15.50}$	$\mathbf{14.33}_{\pm 8.12}$	$\mathbf{58.26}_{\pm 15.23}$	$\mathbf{8.75}_{\pm 5.50}$

To learn from oPIA, a combination of Cross-Entropy loss $\mathcal{L}_{CE}$, Tversky loss $\mathcal{L}_{Tversky}$ [12] and partial Cross-Entropy loss $\mathcal{L}_{pCE}$ [14] is leveraged:

$$\mathcal{L}_{seg}(\hat{Y}, Y) = \mathcal{L}_{CE}(\hat{Y}, Y) + \mathcal{L}_{Tversky}(\hat{Y}, Y) + \mathcal{L}_{pCE}(\hat{Y}, Y) \tag{7}$$

$\mathcal{L}_{CE}$ and $\mathcal{L}_{Tversky}$ can provide stable supervision signals and allow the prediction of false positives. And different from the $\mathcal{L}_{pCE}$ calculated on all labeled pixels in scribble label [9], including the pixels of background, calculations are only performed on the annotated foreground pixels in our implementation. This design allows correctly labeled pixels to be well-learned while promoting the predictions of more foreground pixels. The overall loss is defined as follows:

$$\mathcal{L}_{total} = 0.5 \times \mathcal{L}_{oPIA} + \epsilon_t \times \mathcal{L}_{ReCo} \tag{8}$$

where $\mathcal{L}_{oPIA} = \mathcal{L}_{seg}(P_1, Y) + \mathcal{L}_{seg}(P_2, Y)$ is the loss calculated between oPIA Y and the predictions P_1, P_2. Following [16], we define ϵ_t based on a ramp-up function: $\epsilon_t = \epsilon \times e^{-5\times(1-t/t_{max})^2}$. In this way, only when the model's features are well learned is the co-teaching strategy implemented to make a more stable and reliable learning process.

3 Experiments

Datasets and Evaluation Metrics. We validated the effectiveness of our method on two publicly available datasets, which are the mediastinal Lymph Node Quantification Challenge (LNQ2023) dataset [4], and a refined version [1] of the CT Lymph Node dataset [11]. LNQ2023 contains 513 contrast-enhanced CT volumes, of which 393 cases are partially annotated for training, while the remaining 120 samples are fully annotated and split into 20 and 100 for validation and test. CT Lymph Node contains 89 contrast-enhanced CT volumes obtained from the National Institutes of Health Clinical Center, which are averagely divided for five-fold cross-validation. All 3D scans are cropped to the lung region by Totalsegmentator [18]. For evaluation, the Dice Similarity Coefficient (DSC) and the Average Symmetric Surface Distance (ASSD) are utilized.

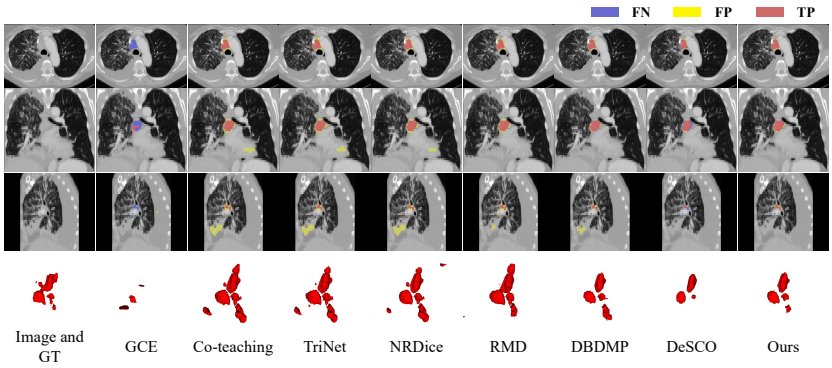

Fig. 3. Qualitative segmentation results on LNQ2023 dataset with oPIA. TP, FP, and FN are generated from the comparison between the ground truth and the prediction.

Implementation Details. We performed experiments on one NVIDIA RTX 2080Ti. The patch size was set to $64 \times 128 \times 224$ for both datasets during our patch-based training. For the LNQ2023 dataset, the best-performing model on the validation set was selected as the final result, while for the CT Lymph Node dataset, the model from the final epoch was selected. For the settings of hyperparameters, the memory ratio γ in eq. 3 was 0.99. The number of inter-batch prototypes N was 3 and 2 for the LNQ2023 and CT Lymph Node datasets. The α of $\mathcal{L}_{Tversky}$ was 0.3 in eq. 7. ϵ and t_{max} was set to 0.1 and 100 in eq. 8, respectively. Unlike other lymph node segmentation methods, there were no post-processing strategies utilized.

Comparison with SOTAs. As pixel-level incomplete annotation shares similar scenario with inaccurate annotation, we evaluated our proposed method against seven state-of-the-art approaches from both inaccurate and incomplete supervision methods: 1) Co-Teaching [7], 2) TriNet [20], 3) NRDice [15], 4) GCE [21], 5) RMD [5], and 6) DBDMP [16], in addition to 7) DeSCO [2], which utilizes the orthogonal annotation. The UpperBound model was trained using full annotations and the LowerBound model was trained solely on oPIA or PIA with $\mathcal{L}_{CE}$ and $\mathcal{L}_{Dice}$. The LowerBound results of the LNQ2023 dataset are not accessible due to the unavailability of the fully annotated training set. Furthermore, as some methods are not specifically designed for orthogonal annotation, we provide performances under oPIA and PIA for a fair and comprehensive evaluation.

The quantitative results are presented in Table 1. Notably, the performances of nearly all methods decline on oPIA, indicating that oPIA presents a more challenging supervision signal than PIA. However, our proposed method demonstrates significant and stable performance improvements on both datasets, regardless of whether oPIA or PIA is leveraged. On the LNQ2023 dataset, our proposed ReCo-I2P achieved an average DSC of 53.94% with oPIA and 57.32% with PIA, outperforming the best state-of-the-art method, DBDMP, by 3.7% and

Table 2. Effectiveness of different modules on validation and test set of LNQ2023 dataset with orthogonal partial-instance annotation. $\mathcal{L}_{seg}$ represents the designed loss for oPIA in eq. 7. Dropout and PEF represent different feature augmentation methods. The impact of $\mathcal{L}_{ReCo}$ with/without the assistance of R is finally explored.

$\mathcal{L}_{seg}$	Augmentation		$\mathcal{L}_{ReCo}$		Validation Set		Test Set	
	Dropout	PFE	*w/o R*	*w R*	DSC↑	ASSD↓	DSC↑	ASSD↓
					$7.99_{\pm 7.30}$	$49.85_{\pm 45.67}$	$11.33_{\pm 7.96}$	$31.78_{\pm 25.04}$
✓					$37.54_{\pm 19.81}$	$38.83_{\pm 27.28}$	$49.40_{\pm 17.63}$	$26.54_{\pm 32.84}$
✓	✓				$39.46_{\pm 19.44}$	$34.16_{\pm 23.92}$	$50.40_{\pm 17.73}$	$24.26_{\pm 41.49}$
✓		✓			$40.79_{\pm 19.83}$	$34.00_{\pm 26.34}$	$51.90_{\pm 17.35}$	$21.09_{\pm 29.42}$
✓		✓	✓		$44.98_{\pm 19.67}$	$26.70_{\pm 25.53}$	$53.29_{\pm 16.27}$	$16.93_{\pm 28.70}$
✓		✓		✓	$\mathbf{45.52}_{\pm 19.00}$	$\mathbf{21.60}_{\pm 25.95}$	$\mathbf{53.94}_{\pm 14.94}$	$\mathbf{16.29}_{\pm 22.32}$

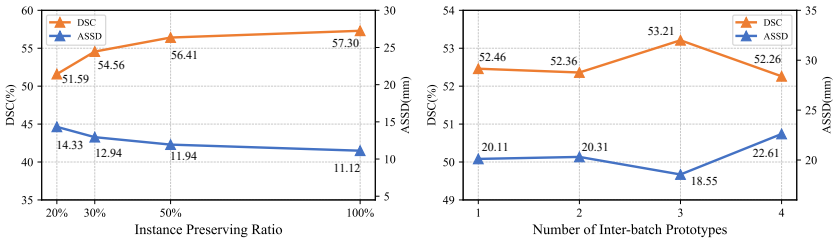

Fig. 4. Sensitivity analysis of the instance preserving ratio on CT Lymph Node dataset (left) and the number of inter-batch prototypes on LNQ2023 dataset (right) with oPIA.

2.7% on DSC, respectively. On the CT Lymph Node dataset, ReCo-I2P trained with PIA achieved an ASSD of 8.75mm, which is notably close to the Upper-Bound's result of 5.13mm. The visualization results, shown in Fig. 3, clearly demonstrate that the segmentation results of our proposed method are closer to the ground truth with fewer false positive regions and more recalls.

Ablation Study. To evaluate the effectiveness of each component, we conducted a comprehensive ablation study on the LNQ2023 dataset. Table 2 presents the quantitative improvements of key modules in our proposed framework. Additionally, we performed a sensitivity analysis to assess the impact of the annotated ratio of lymph nodes and the number of inter-batch prototypes. As shown in Fig. 4, our proposed method maintains stable segmentation performance and does not experience significant performance degradation, even when only a small percentage of the lymph nodes are annotated orthogonally.

4 Conclusions

To summarize, our proposed ReCo-I2P framework, which leverages the information aggregating ability of the prototype and the co-teaching strategy, successfully learns from a novel, cost-friendly, and efficient annotating method of lymph

nodes called orthogonal partial-instance annotation. The reliable foreground information is extracted and re-injected by prototype-based feature enhancement. In addition, complementary information between predictions is transferred through the reliability-based co-teaching strategy. As shown in the results, our method achieves optimal performance on both PIA and oPIA. This indicates that our method is not only applicable to annotate new lymph node datasets under the setting of oPIA but also to improve the annotation quality of existing lymph node datasets (PIA). However, there are not many publicly available datasets that label all lymph nodes in images in the current public community. A proper and comprehensive evaluation of this task remains to be determined.

Acknowledgments. This work was supported by the Natural Science Foundation of Sichuan Province under grant 2025ZNSFSC0644.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bouget, D., Pedersen, A., Vanel, J., Leira, H.O., Langø, T.: Mediastinal lymph nodes segmentation using 3D convolutional neural network ensembles and anatomical priors guiding. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization **11**(1), 44–58 (2023)
2. Cai, H., Li, S., Qi, L., Yu, Q., Shi, Y., Gao, Y.: Orthogonal annotation benefits barely-supervised medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3302–3311 (2023)
3. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation. In: European Conference on Computer Vision. pp. 205–218 (2022)
4. Dorent, R., Khajavi, R., Idris, T., Ziegler, E., Somarouthu, B., Jacene, H., LaCasce, A., Deissler, J., Ehrhardt, J., Engelson, S., et al.: LNQ 2023 challenge: Benchmark of weakly-supervised techniques for mediastinal lymph node quantification. Machine Learning for Biomedical Imaging **3**(MICCAI 2023 LNQ challenge special issue), 1–15 (2025)
5. Fang, C., Wang, Q., Cheng, L., Gao, Z., Pan, C., Cao, Z., Zheng, Z., Zhang, D.: Reliable mutual distillation for medical image segmentation under imperfect annotations. IEEE Transactions on Medical Imaging **42**(6), 1720–1734 (2023)
6. Fu, J., Lu, T., Zhang, S., Wang, G.: UM-CAM: Uncertainty-weighted multi-resolution class activation maps for weakly-supervised fetal brain segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 315–324 (2023)
7. Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I.W., Sugiyama, M.: Co-teaching: Robust training of deep neural networks with extremely noisy labels. In: International Conference on Neural Information Processing Systems. pp. 8536–8546 (2018)
8. Ju, J., Ren, S., Qiu, D., Tu, H., Yin, J., Xu, P., Guan, Z.: A weakly-supervised multi-lesion segmentation framework based on target-level incomplete annotations. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 44–53 (2024)

9. Luo, X., Hu, M., Liao, W., Zhai, S., Song, T., Wang, G., Zhang, S.: Scribble-supervised medical image segmentation via dual-branch network and dynamically mixed pseudo labels supervision. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 528–538 (2022)
10. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 fourth International Conference on 3D Vision (3DV). pp. 565–571 (2016)
11. Roth, H.R., Lu, L., Seff, A., Cherry, K.M., Hoffman, J., Wang, S., Liu, J., Turkbey, E., Summers, R.M.: A new 2.5D representation for lymph node detection using random sets of deep convolutional neural network observations. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 520–527 (2014)
12. Salehi, S.S.M., Erdogmus, D., Gholipour, A.: Tversky loss function for image segmentation using 3D fully convolutional deep networks. In: International Workshop on Machine Learning in Medical Imaging. pp. 379–387 (2017)
13. Stacker, S.A., Achen, M.G., Faries, M.B., Morton, D.L.: The clinical significance of lymph-node metastasis. Lymphangiogenesis in Cancer Metastasis pp. 83–117 (2009)
14. Tang, M., Djelouah, A., Perazzi, F., Boykov, Y., Schroers, C.: Normalized cut loss for weakly-supervised CNN segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1818–1827 (2018)
15. Wang, G., Liu, X., Li, C., Xu, Z., Ruan, J., Zhu, H., Meng, T., Li, K., Huang, N., Zhang, S.: A noise-robust framework for automatic segmentation of COVID-19 pneumonia lesions from CT images. IEEE Transactions on Medical Imaging **39**(8), 2653–2663 (2020)
16. Wang, L., Qu, Y., Luo, X., Liao, W., Zhang, S., Wang, G., et al.: Weakly supervised lymph nodes segmentation based on partial instance annotations with pre-trained dual-branch network and pseudo label learning. Machine Learning for Biomedical Imaging **2**(MICCAI 2023 LNQ challenge special issue), 1030–1047 (2024)
17. Wang, Y., Zhang, S., Zhang, M., Zhang, G., Chen, Z., Wang, X., Yang, Z., Yu, Z., Ma, H., Wang, Z., et al.: Prediction of lateral lymph node metastasis with short diameter less than 8 mm in papillary thyroid carcinoma based on radiomics. Cancer Imaging **24**(1), 155 (2024)
18. Wasserthal, J., Breit, H.C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D.T., Cyriac, J., Yang, S., et al.: TotalSegmentator: Robust segmentation of 104 anatomic structures in CT images. Radiology: Artificial Intelligence **5**(5), e230024 (2023)
19. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: SegFormer: Simple and efficient design for semantic segmentation with transformers. In: International Conference on Neural Information Processing Systems. pp. 12077–12090 (2021)
20. Zhang, T., Yu, L., Hu, N., Lv, S., Gu, S.: Robust medical image segmentation from non-expert annotations with tri-network. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 249–258 (2020)
21. Zhang, Z., Sabuncu, M.R.: Generalized cross entropy loss for training deep neural networks with noisy labels. In: International Conference on Neural Information Processing Systems. pp. 8792–8802 (2018)
22. Zhou, H., Wang, Y., Zhang, B., Zhou, C., Vonsky, M.S., Mitrofanova, L.B., Zou, D., Li, Q.: Unsupervised domain adaptation for histopathology image segmentation with incomplete labels. Computers in Biology and Medicine **171**, 108226 (2024)