PLD-AL: Pseudo-Label Divergence-Based Active Learning in Carotid Intima-Media Segmentation for Ultrasound Images

Yucheng Tang[1,2⋆], Yipeng Hu[3], Jing Li[2], Hu Lin[4], Xiang Xu[1], Ke Huang[4⋆⋆], and Hongxiang Lin[2⋆⋆]

[1] School of Mathematical Sciences, Zhejiang University, Hangzhou, China
[2] Zhejiang Lab, Hangzhou, China. hxlin@zhejianglab.edu.cn
[3] Centre for Medical Image Computing and Wellcome/EPSRC Centre for Interventional & Surgical Sciences, University College London, London, UK
[4] Department of Endocrinology, Children's Hospital, Zhejiang University School of Medicine. kehuang@zju.edu.cn

Abstract. Segmentation of the carotid intima-media (CIM) offers more precise morphological evidence for obesity and atherosclerotic disease compared to the method that measures its thickness and roughness during routine ultrasound scans. Although advanced deep learning technology has shown promise in enabling automatic and accurate medical image segmentation, the lack of a large quantity of high-quality CIM labels may hinder the model training process. Active learning (AL) tackles this issue by iteratively annotating the subset whose labels contribute the most to the training performance at each iteration. However, this approach substantially relies on the expert's experience, particularly when addressing ambiguous CIM boundaries that may be present in real-world ultrasound images. Our proposed approach, called pseudo-label divergence-based active learning (PLD-AL), aims to train segmentation models using a gradually enlarged and refined labeled pool. The approach has an outer and an inner loops: The outer loop calculates the Kullback–Leibler (KL) divergence of predictive pseudo-labels related to two consecutive AL iterations. It determines which portion of the unlabeled pool should be annotated by an expert. The inner loop trains two networks: The student network is fully trained on the current labeled pool, while the teacher network is weighted upon itself and the student one, ultimately refining the labeled pool. We evaluated our approach using both the Carotid Ultrasound Boundary Study dataset and an in-house dataset from Children's Hospital, Zhejiang University School of Medicine. Our results demonstrate that our approach outperforms state-of-the-art AL approaches. Furthermore, the visualization results show that our approach less overestimates the CIM area than the rest methods, especially for severely ambiguous ultrasound images at the thickness direction.

⋆ This work was performed when Yucheng Tang was visiting Zhejiang Lab as an intern.
⋆⋆ Corresponding authors: Hongxiang Lin and Ke Huang.

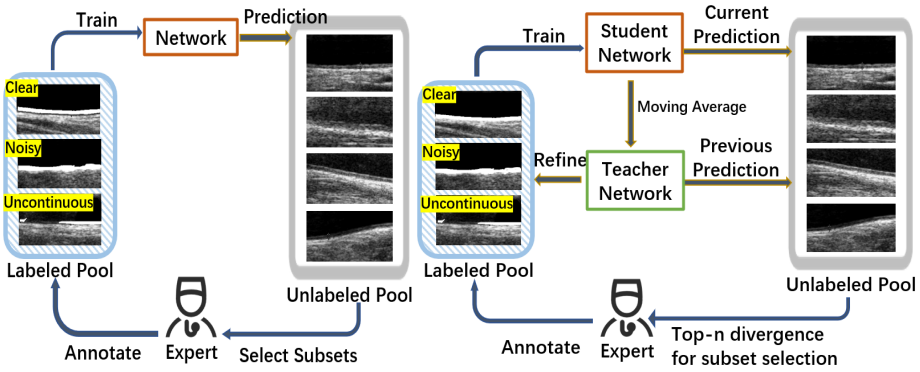
1 Introduction

Carotid intima-media (CIM) segmentation has been widely applied in clinical practice, providing a diagnostic basis for atherosclerotic disease (one of the complications of obesity). To identify the contour of the intima-media, i.e., the structure between the lumen-intima (LI) and the media-adventitia (MA), one of the available solutions is deep learning-based medical image segmentation for CIM. Currently, this CIM segmentation approach faces the challenges of lack of large-quantity images, high-quality labels from ultrasound experts, and a mixture of clear and ambiguous CIM areas in carotid ultrasound images.

Semi-supervised learning recently applies novel frameworks to a general segmentation task [1][2][3][4]. In particular, the combination of consistency regularization and pseudo-labeling utilizes unlabeled data to partially address the lack-of-label issue [5]. A different strategy to efficiently utilize labeling effort is active learning (AL), which can iteratively select a subset of unlabeled data for annotation by experts, but still reach a model performance otherwise requiring a much larger training set. AL has been widely applied to image classification [6][7][8], semantic segmentation [9][10] and medical image segmentation [11][12]. These methods have effectively improved accuracy through experts' involvement. However, carotid ultrasound images are user-end protocol dependent, and with high variability in quality, real-world labels on ultrasound images generally share the same characteristics in high variability. Therefore, after testing several state-of-the-art AL methods, we would like to incorporate methodologies from semi-supervised learning designed to extract predictive information from unlabeled data, and between labeled and unlabeled data, for AL.

In this work, we propose pseudo-label divergence-based active learning (PLD-AL) to obtain accurate CIM segmentation contributing to the clinical diagnosis of obesity and atherosclerotic disease. As shown in Fig. 1, unlike the conventional AL framework that utilizes one machine learning model, PLD-AL is composed by two networks: the student network is fully trained on the current labeled pool, and the teacher network is weighted upon previous itself and the student one. We use divergence, which measures the distance between two model predictions, to select data for annotation. Furthermore, we use the teacher network to refine the labels to reduce the noise of labels and improve the effectiveness of the next network optimization stage.

Our contributions are as follows: we propose PLD-AL, which aims to train segmentation models using a gradually enlarged and refined labeled pool. First, we automatically select and annotate large divergence data between the current and previous AI models, facilitating fast convergence of the AL model to most sound data in the unlabeled pool. Second, we propose a strategy to refine the labels in the labeled pool alternatingly with the proposed label-divergence-based AL algorithm, which improves the robustness compared to the conventional AL approach. We conducted experiments to demonstrate that our method yielded competitive performance gains over other AL methods. Finally, we applied the trained model to a real-world in-house hospital dataset with noisy labels and



(a) Conventional AL (b) Pseudo-label divergence based AL

Fig. 1. (a) Conventional AL that trains a machine learning model to select an unlabeled subset for an expert to annotate. (b) We propose a novel AL framework to progressively annotate data by selecting top-n largest divergence between student and teacher network predictions. Additionally, such a framework can also refine the labeled data assumed to be noisy.

obtain accurate CIM segmentation results. We release our code at `https://github.com/CrystalWei626/PLD_AL`.

2 Method

Section 2.1 establishes mathematical formulation on the main task of CIM segmentation in our AL framework. Our proposed AL approach has two loops: in Section 2.2, the outer loop implements progressive annotation on the automatically selected unlabeled pool; in Section 2.3, the inner loop trains the neural networks on the labeled pool and subsequently refines it through a feedback routine.

2.1 Mathematical notations and formulation

Denote $x \in \mathbb{R}^{I \times J}$ a carotid ultrasound image and $y \in \mathbb{R}^{I \times J}$ the corresponding CIM mask. Let $D_L = X_L \times Y_L$ and X_U be the initial labeled and unlabeled pools, where X_L is the carotid ultrasound image set, and Y_L is the corresponding label set. We aim to improve generalization ability of the AI model by selecting the most informative data in X_U and delivering them to an expert for annotation.

We propose a novel AL framework: PLD-AL for CIM segmentation, as illustrated in Fig. 1 and Alg 1. First, AI models are trained on D_L and used to refine Y_L. Then, the AI models select data from X_U for expert to annotate, forming a new set of labeled data. Finally, we update D_L and X_U and use new D_L to train the same AI models.

Algorithm 1: PLD-AL

Input: Initial labeled pool $D_L = X_L \times Y_L$; Unlabeled pool X_U; Judgment threshold τ; Refining threshold λ;
Initialize θ_S and θ_T;
for $t = 1, \cdots T$ **do**
  $\theta_T^{(0)} \leftarrow \theta_T$; $\theta_S^{(0)} \leftarrow \theta_S$;
  for $k = 1, \cdots, K$ **do**
    $\theta_S^{(k)} := Opt(\theta_S^{(k-1)}; D_L, lr)$; ▷ Optimize $\theta_S^{(k)}$ on D_L
    $\theta_T^{(k)} := \alpha\theta_S^{(k)} + (1-\alpha)\theta_T^{(k-1)}$; ▷ Update $\theta_T^{(k)}$ by EMA
    $\mathcal{M}^{(k)} = \text{mean}_{(x,y)\in D_L}\text{IoU}(y, F(x|\theta_S^{(k)}))$; ▷ Calculate mIoU
    if $k > K_1$ **then**
      $\tilde{\mathcal{M}} = \text{argmin}\sum_{l=1}^{k} \|\tilde{\mathcal{M}} - \mathcal{M}^{(l)}\|_{\ell^2}^2$; ▷ Fit the mIoU curve
      if $\tilde{\mathcal{M}}'(k) - \tilde{\mathcal{M}}'(k-1) < \tau$ **then**
        for $x \in X_L, i \in \{1, 2, ...I\}, j \in \{1, 2, ...J\}$ **do**
          $p_{ij} = F(x(i,j)|\theta_T^{(k)})$; ▷ Predict on teacher network
          $y(i,j) = \text{argmax}\{p_{ij}\}$ if $\max p_{ij} > \lambda$; ▷ Refine Y_L
        $\theta_S \leftarrow \theta_S^{(k)}, \theta_T \leftarrow \theta_T^{(k)}$;
        break;
  $d(x) = \text{mean}_{i,j}\text{Div}_{KL}(x(i,j), \theta_S, \theta_T), x \in X_U$; ▷ Compute KL divergence
  $X_A = \arg_{x\in X_U}\text{TOP}_n d(x)$; ▷ Select unlaleled data
  $Y_A = \{y = \text{Expert}(x) : x \in X_A\}$; ▷ Annotate by expert
  $D_A = X_A \times Y_A; D_L \leftarrow D_L \bigcup D_A$; $X_U \leftarrow X_U \setminus X_A$; ▷ Update D_L, X_U
Output: D_L; θ_T

In each AL iteration, we use a mean-teacher architecture as the backbone of AL. The student and the teacher networks, respectively parameterized by θ_S and θ_T, share the same neural network architecture F, which maps the carotid ultrasound image $x \in \mathbb{R}^{I\times J}$ to the extended three-dimensional CIM mask probability $p \in \mathbb{R}^{I\times J\times 2}$, whose 3rd-dimensional component $p_{ij} \in \mathbb{R}^2$ denotes the softmax probability output for binary classification at the pixel (i, j). We use the divergence between pseudo-labels generated by student and teacher networks to assist in selecting data for the expert to annotate.

2.2 Outer Loop: Divergence based AL

The outer loop is an AL cycle that selects data for the expert to annotate according to the divergence between the predictions of the student and teacher networks. First, we initialize θ_S and θ_T. We complete the inner loop proposed in Section 2.3, and obtain the trained parameters for the student and teacher networks. Then, we select n data from X_U for the expert to annotate. We suggest using the Kullback–Leibler (KL) divergence to assist in selecting data, as shown

in Eq. (1):

$$\mathrm{Div}_{KL}(x(i,j),\theta_S,\theta_T) = \sum_{c=1}^{2} F(x(i,j)|\theta_T)\log\frac{F(x(i,j)|\theta_T)}{F(x(i,j)|\theta_S)}. \tag{1}$$

We consider data prediction uncertainty as a decisive metric for data selection. It is deduced that the KL divergence between the output of the primary and the auxiliary models in a dual-decoder architecture can approximate the prediction uncertainty [13][14].

We compute the KL divergence scores $d(x) = \mathrm{mean}_{i,j}\mathrm{Div}_{KL}(x(i,j),\theta_S,\theta_T)$ of the data in X_U. Let X_A be the subset that contains data x in X_U corresponding to the top-n largest $d(x)$ values (denoted by $\mathrm{TOP}_n d(x)$). With this, we can next obtain the label set Y_A in terms of X_A by means of the expert's annotates and the required post-processing step; see Section 3.1 for details. Lastly, we add the selected dataset with its label set $X_A \times Y_A$ into D_L and delete X_A from X_U. We repeat the above steps until reaching the maximum number of AL iterations.

2.3 Inner Loop: Network optimization and label refinement

The inner loop trains the neural networks by the labeled pool and refines noisy labels through a feedback routine. In the k^{th} epoch of the inner loop, we first use the last labeled pool D_L to optimize the training parameter $\theta_S^{(k)}$ by mini-batch stochastic gradient descent. The loss function consists of a supervised loss L_{sup} between labels and predictions of the student model, and a consistency loss L_{con} between the predictions of the student and the teacher models. These can be implemented using the cross-entropy loss and the mean squared error loss, respectively. Then, we update $\theta_T^{(k)}$ by exponential moving average (EMA) with a decay rate α as Eq. (2):

$$\theta_T^{(k)} = \alpha\theta_S^{(k)} + (1-\alpha)\theta_T^{(k-1)}. \tag{2}$$

We refine noisy labels based on the idea that the fitness soars sharply at first but slows down after the model begins to fit noise[15]. We interrupt the model training before it begins to fit noise, then refine the labels utilizing the current network output. We calculate the model fitness via a series of the intersection over union (mIoU)[16] scores sampled at every training epoch. To estimate the ratio of change of the model fitness, we fit the mIoU curve $\tilde{\mathcal{M}}(k)$ via e.g., the exponential regression formed in Eq. (3) when the length of mIoU series is larger than a designated parameter $K_1 \in \mathbb{N}^+$:

$$\tilde{\mathcal{M}}(k) = a(1-\exp\{-b\cdot k^c\}), \tag{3}$$

where a, b, and c are the fitting parameters to be determined by least squared estimate. Then we calculate the ratio of change of the model fitness γ^k via the derivative of the mIoU curve $\tilde{\mathcal{M}}'(k)$: $\gamma^k = \tilde{\mathcal{M}}'(k) - \tilde{\mathcal{M}}'(k-1)$. When training stops at this epoch k satisfying $\gamma^k < \tau$ (τ is a judgment threshold), we lastly

predict the CIM mask probability $p_{ij} = F(x(i,j)|\theta_T^{(k)})$ via the teacher network for each pixel at (i,j) and update the noisy label $y(i,j)$ in Y_L if $\max p_{ij} > \lambda$ (λ is a refining threshold).

3 Experiments and Results

3.1 Experiment Settings

Implementation Details. We used the PyTorch platform (version 1.13.1) to implement our method. And we adapted the same UNet++ [17] structures as the encoding-decoding structures for the student and the teacher networks. We implemented 1000 training iterations with a total mini-batch size of 14 and initial batch size of labeled data of 2 on an Nvidia GeForce RTX 3070 GPU with 8192 MB of memory (Nvidia, Santa Clara, CA, United States). Since the number of labeled data increases after completing each AL iteration, the batch size of labeled data should increase by 2 synchronously to keep the total epoch num unchanged. We used stochastic gradient descent (SGD) as the optimizer with the parameter settings: momentum (0.9) and weight decay (0.0001). We set EMA decay rate $\alpha = \min\{1 - 1/(iter + 1), 0.99\}$, where $iter$ is the current training iteration number. 2021 regions of interest (ROI) of size 256×128 were cropped from original carotid ultrasound images for model training using template matching technique [18]. We set the number of AL iterations, fixed labeling budget, initial labeled and unlabeled data, and the test data as 5, 200,159, 1857, and 1204, respectively. During each annotation phase, experts manually marked the CIM boundaries with scatters and we subsequently generated the complete CIM masks via the Akima interpolation method [19]. θ_S and θ_T was initialized by the pre-train model[5] on ImageNet [20]. At our best practice, we chose the hyper-parameters $\lambda = 0.8$, $\tau = 0.005$ and $K_1 = 1$.

Dataset. We employed the publicly available Carotid Ultrasound Boundary Study (CUBS) dataset[6][21] and the in-house dataset acquired at Children's Hospital, Zhejiang University School of Medicine. The CUBS dataset contains ultrasound images of the left and right carotid arteries from 1088 patients across two medical centers and three manual annotations of LI and MA boundaries by experts. According to the description of these annotations in the dataset specification, the analytic hierarchy process (AHP) [22] was adapted to weigh the three expert's annotations to obtain accurate labels for testing. We randomly performed morphological transformations (dilation and erosion) by OpenCV [23] on the accurate labels to generate noisy labels for training. The in-house dataset comes from 373 patients aged 6-12, with 2704 carotid ultrasound images. We picked 350 images with visible CIM areas and applied the model trained on CBUS to CIM segmentation. The data acquisition and the experimental protocol have been approved by the institutional review board of Children's Hospital, Zhejiang University School of Medicine.

[5] https://github.com/pprp/timm

[6] https://data.mendeley.com/datasets/fpv535fss7/1

Table 1. Quantitative results of performance comparison, the metrics were calculated over the test dataset and took the mean. Bold font highlights the optimal performance except for the upper limit. The asterisk $*$ denotes $p < 0.001$ compared with the rest methods.

Method	Dice (%) ↑	IoU (%) ↑	ASD (voxel) ↓	95HD (voxel) ↓	Time (S) ↓
Random	70.96 ± 8.26	57.01 ± 8.80	3.79 ± 1.05	15.95 ± 6.41	$\mathbf{118.69 \pm 6.38}$
Entropy[26]	76.62 ± 2.20	63.26 ± 2.75	2.07 ± 0.02	9.21 ± 2.09	192.24 ± 36.13
Confidence[12]	74.86 ± 0.21	61.93 ± 0.34	2.47 ± 0.89	11.15 ± 3.97	166.56 ± 2.93
CoreSet[27]	$79.92 \pm 0.31^*$	67.39 ± 0.43	$1.88 \pm 0.11^*$	$6.33 \pm 0.62^*$	199.78 ± 47.20
CDAL[28]	78.20 ± 1.61	65.15 ± 2.06	2.01 ± 0.10	7.83 ± 1.51	165.15 ± 2.74
Ours	$\mathbf{83.51 \pm 0.28}^*$	$\mathbf{72.33 \pm 0.46}^*$	$\mathbf{1.69 \pm 0.02}^*$	$\mathbf{4.72 \pm 0.11}^*$	139.06 ± 28.66
Upper Limit	84.01	73.03	1.53	4.24	213.77

Evaluation Metrics We utilized dice coefficient (Dice) [24], intersection over union (IoU) [16], average surface distance (ASD), 95% covered Hausdorff distance (95HD) [25], and the average training time of 5 AL iterations as evaluation metrics of the CIM segmentation performance compared to the generated ground truth on the unseen test set.

3.2 Performance Comparison

We evaluated the performance of AL methods on the CIM segmentation task using the CUBS dataset.

Baselines. We compared our method to other AL methods, including AL methods with query strategy based on random selection (Random), entropy increase (Entropy) [26], prediction confidence (Confidence) [12], CoreSet [27] and predicted probability diverse contexts (CDAL) [28]. All of the backbones of these baseline methods are fully supervised models.

Furthermore, we trained a supervised model by the fully labeled pool with accurate labels yielding an upper limit of generalization ability. We compared this upper limit to the performance of all the methods.

Table 1 illustrates the quantitative results of different methods on the test dataset. It shows that our method based on the KL divergence query strategy improves the mean generalization metrics (Dice, IoU, ASD, and 95HD) compared with other AL methods. In particular, it significantly (two-tailed Wilcoxon signed-rank test with $p < 0.001$) outperforms the others in terms of any metric.

3.3 Ablation Study

We conducted ablation study on the CUBS dataset to demonstrate the importance of the label refinement module proposed in Section 2.3. We canceled the label refinement module and substituted the label refinement module with confidence learning (CL) for noise label correction [29].

Table 2 illustrates the results of ablation study experiment. Our method substantially outperforms the method without the label refinement module and

Table 2. Quantitative results of ablation study, the metrics were calculated over the test dataset and took the mean. The abbreviations, Refine and CL, represent the label refinement module and confidence learning[29], respectively. Bold font highlights the optimal performance except for the upper limit. The asterisk $*$ denotes $p < 0.001$ compared with the rest methods.

Method	Dice (%) ↑	IoU (%) ↑	ASD (voxel) ↓	95HD (voxel) ↓	Time (S) ↓
w.o. Refine	80.17 ± 1.37	67.68 ± 1.86	1.97 ± 0.05	6.73 ± 0.99	301.93 ± 27.38
Refine→CL	81.08 ± 1.36	68.9 ± 1.84	1.86 ± 0.10	5.82 ± 0.69	689.09 ± 34.03
w/ Refine	$\mathbf{83.51 \pm 0.28}^*$	$\mathbf{72.33 \pm 0.46}^*$	$\mathbf{1.69 \pm 0.02}^*$	$\mathbf{4.72 \pm 0.11}^*$	$\mathbf{139.06 \pm 28.66}$
Upper Limit	84.01	73.03	1.53	4.24	213.77

slightly outperforms the method with CL. In particular, it significantly (two-tailed Wilcoxon signed-rank test with $p < 0.001$) outperforms the others in terms of all the metrics. Moreover, the training time of our method is significantly reduced compared to CL since CL needs to estimate the uncertainty during training to correct the noisy data smoothly, which leads to more computational cost.

3.4 Application on in-house dataset

We applied the teacher network trained in Section 3.2 to the in-house dataset acquired at a pediatric hospital. Figure 2 visualizes three example images with different CIM area qualities (clear, mildly ambiguous, severely ambiguous). Qualitatively, the generalization ability of the model trained by our method is much better than those trained by other methods, regardless of image quality. Moreover, as shown in Fig. 2, Random over-estimates the CIM area, while CoreSet, CDAL, and our method produces more conservative results but lost continuity in the severely ambiguous image. Quantitatively, the mean Dice, IoU, ASD, and 95HD of our method are $79.20\%, 66.99\%, 1.92$ voxels, and 6.12 voxels, respectively, indicating a small but rational generalization loss on the in-house data.

4 Conclusion

We propose a novel AL framework PLD-AL, by training segmentation models using a gradually enlarged and refined labeled pool to obtain accurate and efficient CIM segmentation. Compared with other AL methods, it achieves competitive performance gains. Furthermore, we applied the trained model to an in-house hospital dataset and obtained accurate CIM segmentation results. In the future, we will extend our approach to subsequently calculate CIM thickness and roughness for clinical evaluation of obesity or atherosclerotic disease. We will also investigate the robustness of the proposed method in terms of inter-expert variations and noisy annotation labels. Our approach merely involves one expert in the loop, which may potentially be sensitive to the expert's experience. Multiple experts may consider minimizing inter-reader differences during human-AI interactive labeling [30].

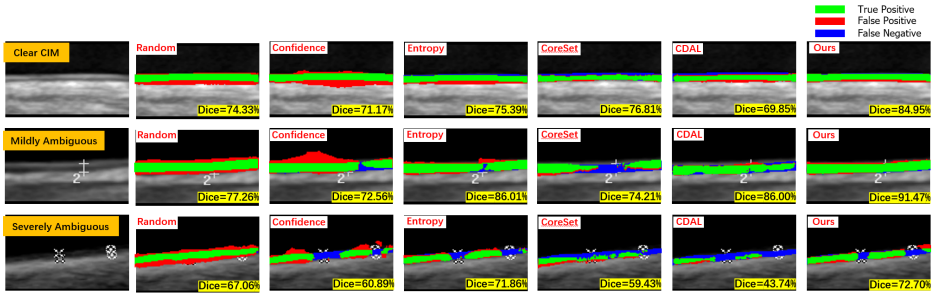


Fig. 2. Qualitative results of application study. It shows the visualization of CIM segmentation on input images with clear, mildly ambiguous, and severely ambiguous CIM areas, respectively. The images are chosen from the in-house dataset. We used the model with the best quantitative results in Section 3.2 to generate the masks. The green, red, and blue masks represent segmented true positive, false positive, and false negative, respectively.

Acknowledgement

This work was supported in part by Research Initiation Project (2021ND0PI02) and Key Research Project (2022KI0AC01) of Zhejiang Lab, National Key Research and Development Programme of China (No. 2021YFC2701902), and National Natural Science Foundation of China (No. 12071430).

References

1. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: NeurIPS, NIPS, Long Beach (2017)
2. Xu, M. C., Zhou, Y., Jin, C., de Groot, M., Alexander, D. C., Oxtoby, N. P., Hu, YP., Jacob, J.: Bayesian Pseudo Labels: Expectation Maximization for Robust and Efficient Semi-supervised Segmentation. In: MICCAI, pp. 580–590. Springer, Singapore (2022)
3. Yao, H., Hu, X., Li, X.: Enhancing pseudo label quality for semi-supervised domain-generalized medical image segmentation. In: AAAI, pp. 3099–3107. AAAI (2022)
4. Liu, F., Tian, Y., Chen, Y., Liu, Y., Belagiannis, V., Carneiro, G.: ACPL: Anti-curriculum pseudo-labelling for semi-supervised medical image classification. In: CVPR, pp. 20697–20706. IEEE Computer Society, New Orleans (2022)
5. Lu, L., Yin, M., Fu, L., Yang, F.: Uncertainty-aware pseudo-label and consistency for semi-supervised medical image segmentation. Biomedical Signal Processing and Control. **79**(2) (2023)
6. Parvaneh, A., Abbasnejad, E., Teney, D., Haffari, G. R., Van Den Hengel, A., Shi, J. Q.: Active learning by feature mixing. In: CVPR, pp. 12237–12246. IEEE Computer Society, New Orleans (2022)
7. Sinha, S., Ebrahimi, S., Darrell, T.: Variational adversarial active learning. In: ICCV, pp. 5972–5981, IEEE, Seoul (2019)

8. Caramalau, R., Bhattarai, B., Kim, T. K.: Sequential graph convolutional network for active learning. In: CVPR, pp. 9583–9592. IEEE Computer Society (2021)
9. Casanova, A., Pinheiro, P. O., Rostamzadeh, N., Pal, C. J.: Reinforced active learning for image segmentation. arXiv preprint arXiv:2002.06583 (2020)
10. Siddiqui, Y., Valentin, J., Nießner, M.: Viewal: Active learning with viewpoint entropy for semantic segmentation. In: CVPR, pp. 9433–9443. IEEE Computer Society (2020)
11. Yang, L., Zhang, Y., Chen, J., Zhang, S., Chen, D. Z.: Suggestive annotation: A deep active learning framework for biomedical image segmentation. In: MICCAI, pp. 399–407. Springer, Quebec City (2017)
12. Xu, Y., Xu, X., Jin, L., Gao, S., Goh, R. S. M., Ting, D. S., Liu, Y.: Partially-supervised learning for vessel segmentation in ocular images. In: MICCAI, pp. 271–281. Springer, Strasbourg (2021)
13. Zheng, Z., Yang, Y.: Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. International Journal of Computer Vision, **129**(4), 1106–1120 (2021)
14. Luo, X., Liao, W., Chen, J., Song, T., Chen, Y., Zhang, S., Zhang, S.: Efficient semi-supervised gross target volume of nasopharyngeal carcinoma segmentation via uncertainty rectified pyramid consistency. In: MICCAI, pp. 318–329. Springer, Strasbourg (2021)
15. Liu, S., Liu, K., Zhu, W., Shen, Y., Fernandez-Granda, C.: Adaptive early-learning correction for segmentation from noisy annotations. In: CVPR, pp. 2606–2616. IEEE Computer Society, New Orleans (2022)
16. Rahman, M. A., Wang, Y., Optimizing intersection-over-union in deep neural networks for image segmentation. In: ISVC, pp. 234–244. Springer, Las Vegas (2016)
17. Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., Liang, J.: Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. IEEE transactions on medical imaging, **39**(6), 1856–1867 (2019)
18. Brunelli, R. Template matching techniques in computer vision: theory and practice. John Wiley & Sons (2009)
19. Akima, H.: A method of bivariate interpolation and smooth surface fitting based on local procedures. Communications of the ACM, **17**(1), 18–20 (1974)
20. He, K., Girshick, R., Dollár, P.: Rethinking imagenet pre-training. In: ICCV, pp. 4918–4927, IEEE, Seoul (2019)
21. Meiburger, K. M., Zahnd, G., Faita, F., Loizou, C., Carvalho, C., Steinman, D. A., Gibello, L., Bruno, R. M., Marzola, F., Clarenbach, R.: DATASET for "Carotid Ultrasound Boundary Study (CUBS): an open multi-center analysis of computerized intima-media thickness measurement systems and their clinical impact". Mendeley Data, V1, doi: 10.17632/fpv535fss7.1 (2021)
22. Sipahi, S., Timor, M.: The analytic hierarchy process and analytic network process: an overview of applications. Management Decision, **48**(5), 775–808 (2010)
23. Bradski, G.: The openCV library. Dr. Dobb's Journal: Software Tools for the Professional Programmer, **25**(11), 120–123 (2000)
24. Bertels, J., Eelbode, T., Berman, M., Vandermeulen, D., Maes, F., Bisschops, R., Blaschko, M. B.: Optimizing the dice score and jaccard index for medical image segmentation: Theory and practice. In: MICCAI, pp. 92–100. Springer, Shenzhen (2019)
25. Aspert, N., Santa-Cruz, D., Ebrahimi, T.: Mesh: Measuring errors between surfaces using the hausdorff distance. In: ICME, pp. 705–708. IEEE, Lausanne (2022)
26. Wang, D., Shang, Y.: A new active labeling method for deep learning. In: IJCNN, pp. 112–119. IEEE, Beijing (2014)

27. Sener, O., Savarese, S.: Active learning for convolutional neural networks: A core-set approach. arXiv preprint arXiv:1708.00489 (2017)
28. Agarwal, S., Arora, H., Anand, S., Arora, C.: Contextual diversity for active learning. In: ECCV, pp. 137–153. Springer, Glasgow (2020)
29. Xu, Z., Lu, D., Wang, Y., Luo, J., Jayender, J., Ma, K., Li, X.: Noisy labels are treasure: mean-teacher-assisted confident learning for hepatic vessel segmentation. In: MICCAI, pp. 3–13. Springer, Strasbourg (2021)
30. Zhang, L., Tanno, R., Xu, M., Huang, Y., Bronik, K., Jin, C., Jacob, J., Zheng, YF., Shao, L., Ciccarelli, O.: Learning from multiple annotators for medical image segmentation. Pattern Recognition, **138** (2023)