CO-SEG: AN IMAGE SEGMENTATION FRAMEWORK AGAINST LABEL CORRUPTION

Ziyi Huang[1], *Haofeng Zhang*[2], *Andrew Laine*[3], *Elsa Angelini*[3,4], *Christine Hendon*[1], *Yu Gan*[5]

[1] Department of Electrical Engineering, Columbia University, New York, NY, USA

[2] Department of Industrial Engineering and Operations Research, Columbia University, New York, NY, USA

[3] Department of Biomedical Engineering, Columbia University, New York, NY, USA

[4] NIHR Imperial Biomedical Research Centre, ITMAT Data Science Group, Imperial College London, UK

[5] Department of Electrical and Computer Engineering, The University of Alabama, Tuscaloosa, AL, USA

ABSTRACT

Supervised deep learning performance is heavily tied to the availability of high-quality labels for training. Neural networks can gradually overfit corrupted labels if directly trained on noisy datasets, leading to severe performance degradation at test time. In this paper, we propose a novel deep learning framework, namely Co-Seg, to collaboratively train segmentation networks on datasets which include low-quality noisy labels. Our approach first trains two networks simultaneously to sift through all samples and obtain a subset with reliable labels. Then, an efficient yet easily-implemented label correction strategy is applied to enrich the reliable subset. Finally, using the updated dataset, we retrain the segmentation network to finalize its parameters. Experiments in two noisy labels scenarios demonstrate that our proposed model can achieve results comparable to those obtained from supervised learning trained on the noise-free labels. In addition, our framework can be easily implemented in any segmentation algorithm to increase its robustness to noisy labels.

Index Terms— Deep Learning, Weakly Supervised Learning, Image Segmentation

1. INTRODUCTION

Recent years have witnessed an upsurge of interests in biomedical segmentation. Based on fully convolutional networks, U-Net [1] has been emerging as a classic model which concatenates multi-scale features from the downsampling layers and the upsampling layers. By stacking two U-Net architectures on top of each other, DoubleU-net [2] is an improved version of U-Net aiming to achieve higher performance on specific tasks. CE-Net [3] modified U-Net structure by adopting pretrained ResNet blocks in the feature encoding step to capture high-level spatial information. However, these fully supervised learning algorithms are vulnerable to label noise and their performance may be hugely degenerated by noisy labels. Therefore, under noisy labels, it is important to identify and selectively learn from a clean and reliable subset of samples which mainly include data with clean labels, rather than learning from the whole sample set.

How to improve deep learning performance under noisy labels conditions has caught great attention [4, 5, 6, 7, 8, 9]. One direction is to estimate the mathematical model of a noise distribution [6, 7]. In [7], two procedures were proposed for loss correction and noise transition matrix estimation. Another direction is to directly train on clean samples [8, 9]. Co-teaching [8] trains two networks simultaneously to pick clean samples for each one. However, most current approaches focus on classification tasks, which cannot be applied to the segmentation where labels are spatially arranged in a dense manner. Finally, sample-based reweighting methods [8, 9, 10] just ignore or assign small weights on noisy samples, which can lead to severe overfitting, especially for small datasets.

In this paper, we propose a novel deep learning framework for image segmentation, namely Co-Segmentation (Co-Seg), to handle noisy labels. Our framework integrates the idea of selective training and label correction. In particular, we propose a robust training network to collaboratively learn and select samples with reliable labels. Then a label correction scheme is proposed to enrich the reliable dataset and we retrain a new network on the updated dataset. Experimental results using Co-Seg on noisy labels show performance comparable to supervised learning on noise-free labels. In summary, this paper has the following contributions:
(1) We develop an easily-implemented yet effective framework for image segmentation tasks with noisy labels. It can be easily applied to any deep learning segmentation model to increase learning ability under noisy labels.
(2) We demonstrate that, in multiple noise settings, our model achieves comparable results to supervised training on noise-free datasets.

2. METHODOLOGY

Our proposed framework consists of 3 modules: (1) robust training module under noisy labels, (2) label correction module, and (3) retraining module, as shown in Fig. 1. The robust

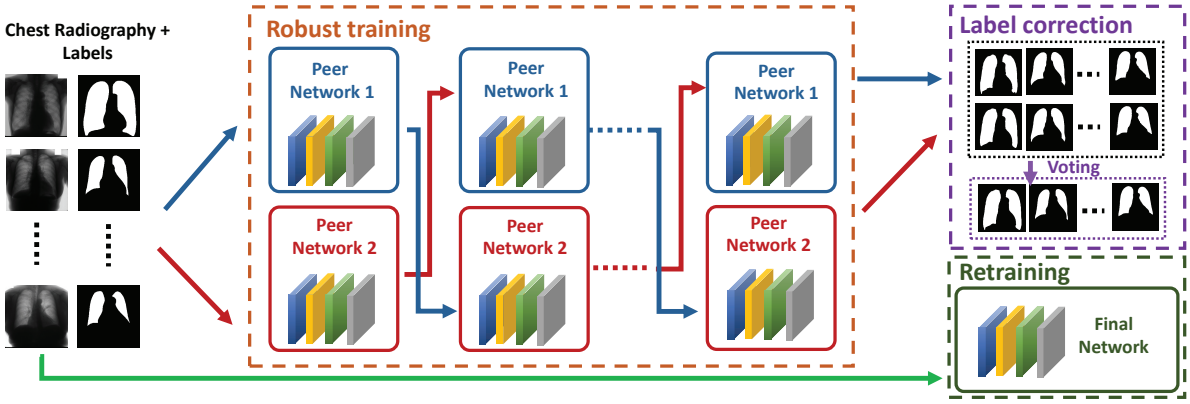

Fig. 1. Algorithm flow of Co-Seg framework.

learning module trains two segmentation networks simultaneously and selects clean samples for their peer networks. Then, the label correction module employs a voting strategy to correct unreliable labels from the noisy samples measured by corruption scores. Lastly, a single segmentation network is trained on the updated dataset to finalize network parameters for future segmentation tasks.

2.1. Robust training against noisy labels

Robust training against noisy labels is very challenging due to memorization effects of deep learning models. Directly training under noisy labels, the networks can gradually overfit noisy samples. Inspired by [8], our robust training module collaboratively trains two networks (peer network 1 & 2) for clean sample selection. Each network picks up a small proportion of high-quality samples in every mini-batch based on sorted corruption scores. Then, such high-quality samples will be fed to its peer network for back propagation. Given a sample, the corruption score S_c is:

$$S_c = -\sum_{x\in\Omega}\sum_{l=1}^{L} g_l(x)\log(p_l(x)) \tag{1}$$

where L is the number of classes, $p_l(x)$ is the estimated probability of class l at pixel position $x \in \Omega$ with Ω the image domain and $g_l(x)$ is the label of the ground truth. The corruption score measures the reliability of the sample. Samples with smaller corruption scores are more likely to be clean. The proportion of samples selected from the corruption score ranking is controlled by α, which is related to the noise level of the dataset.

The loss function of each segmentation network is a combination of a cross entropy loss L_{CE}, a Dice loss L_{Dice}, and a L_2-regularization term on the parameters W_f of the network:

$$L_{total} = L_{CE} + \lambda_1 L_{Dice} + \lambda_2 ||W_f||_2^2 \tag{2}$$

$$L_{Dice} = 1 - \frac{1}{k}\sum_{l=1}^{k}\frac{2\sum_{x\in\Omega}(p_l(x)g_l(x))}{\sum_{x\in\Omega}(p_l(x))^2 + \sum_{x\in\Omega}(g_l(x))^2} \tag{3}$$

where λ_1, λ_2 are tuned parameters. Here, the Dice loss L_{Dice} is used to capture spatial and structural coherence in the segmentation tasks.

2.2. Label correction

We propose to correct noisy labels in biomedical segmentation, rather than ignore or downweight them for two reasons. First, maintaining data size is essential. Training on small-size samples may easily lead to severe overfitting. This is particularly important in biomedical applications where reliable labels at expert level are limited. Second, some noisy samples contain pixels with accurate spatial annotations, which could benefit segmentation. This differs from an image-level classification task where clean and noisy labels are not correlated.

In the label correction module, labels in the training set are corrected based on the voting results from the two collaboratively trained networks obtained from the robust training module. First, we differentiate the noisy samples and the clean samples according to their corruption scores ranking. For each pixel in the noisy samples, we correct their labels if the prediction results from the two networks are the same but different from the input labels. Our updated dataset consists of clean samples with original labels and noisy samples with corrected labels. This process enriches the clean dataset with noise-corrected samples.

2.3. Retraining

Based on the updated dataset, we retrain a final segmentation network, which shares the same network structure as one of the peer networks. Similarly, the retraining uses loss function defined in Eq. 2. Then, this network will be used for future prediction.

3. EXPERIMENTS

We conduct evaluation on segmentation performance in two noisy settings, including both real-world labeling noise from

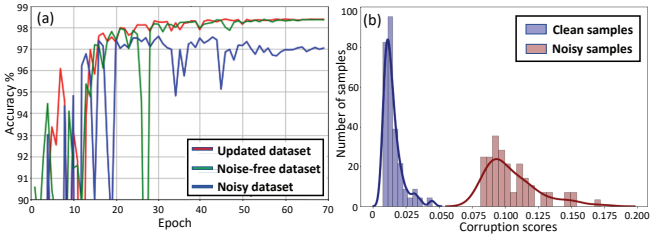

Fig. 2. Benchmark U-Net training behavior with 0.5 noise level and Co-Seg corruption scores. (a) U-Net accuracy over training epochs under Type I noise. (b) Corruption scores under Type II noise: values and probability density functions from the whole training set.

an inexperienced annotator and synthetic noise.

Datatsets. Our evaluation is based on Chest X-ray from the Japanese Society of Radiological Technology (JSRT) dataset [11]. This dataset consists of 247 posterior-anterior (PA) chest radiographs. Ground-truth lung masks were obtained from the annotation of Radiographs (SCR) database [12] at expert level. Following previous work in [13, 14], we resize all images into 256×256 pixels and split the training and testing sets by ID number: the training set contains 124 images with odd ID number and the testing set contains 123 images with even ID.

Label corruption scenarios. We conduct experiments using two scenarios of label corruption. (1) Synthetic boundary (Type I) noise. Manual segmentation variability usually occurs around tissue boundaries due to spatial uncertainty of contrast transition between different tissue types. Following [9], we generate boundary noise by randomly eroding or dilating tissue boundaries by n_i pixels with $1 \leq n_i \leq 8$ in each direction; (2) Inexperienced annotation (Type II) noise. Human annotators tend to have a systematic bias as over and under segmented tissue boundaries. To mimic this real-world noise scenario, an inexperienced annotator who was blind to the algorithm and ground truth, manually labeled the data and generated biased labels.

Experimental setup. As a demonstration of the framework, we choose the classic U-Net [1] as the network architecture to evaluate noise robustness performance and we adopt the same U-Net architecture and hyper-parameters for all segmentation networks in the robust training module (peer network 1 and 2 in Fig. 1) and retraining module (final network). We also use the performance of a single U-Net segmentation network with the same architecture on noise-free training set as a baseline for comparison. The segmentation performance is evaluated by both accuracy (ACC) and Dice coefficient (DIC) in comparison with ground truth. The corrupted training sets are generated by replacing a proportion of clean samples with noisy samples. The noise level (NoL) is defined as the proportion of noisy samples in the training set. Since the noise level in real-world datasets is around $8\% \sim 38\%$ [15], we conduct experiments under 0.1 to 0.5 noise levels for each noise type. Following previous research [8, 9], we assume NoL is a known parameter and $\alpha = 1 - NoL$.

Table 1. Evaluation metrics on lung segmentation with different noise types (Type I and II explained in the text) and noise levels expressed as the proportion of noisy samples.

Network	Noise	Metrics	Noise level for Type I and II				
			0.1	0.2	0.3	0.4	0.5
Co-Seg	Type I	ACC	0.978	0.975	0.978	0.978	0.980
		DIC	0.975	0.973	0.975	0.974	0.976
	Type II	ACC	0.981	0.980	0.980	0.980	0.978
		DIC	0.974	0.974	0.975	0.975	0.973
U-Net	Noise Free	ACC			0.981		
		DIC			0.976		

Results. Table 1 reports evaluation metrics from the JSRT datasets on lung segmentation with different noise types and noise levels, together with the baseline experiment on noise-free dataset using U-Net. The results obtained by our model are comparable with the baseline noise-free U-Net training. Differences are all below 0.6% in both DIC and ACC. Small variations among noise levels are likely caused by model stochasticity. Those results demonstrate that our Co-Seg model can provide robust results under noise levels up to 0.5 with performance competitive with learning using noise-free labels.

We further visualize experimental results to show the effectiveness of the Co-Seg model. Fig. 2(a) compares the training accuracy curves (background and lung segmentation) from the benchmark U-Net trained using the updated dataset, noise-free dataset and the noisy dataset over training epochs. Accuracy with the noisy dataset (blue) decreases after 30 epochs indicating that the network is gradually overfitting the noisy samples. Meanwhile, the accuracy curve with the updated training set (red) is flat and smooth, showing consistency of training quality similar to the curve from noise-free dataset (green). Fig. 2(b) shows the corruption scores and the probability density functions from the entire training set with 0.5 noise level under Type II label noise corruption. The blue/red curves are the probability density functions fitted on the clean/noisy samples. The two probability density functions have (almost) disjoint supports, indicating that corruption scores effectively separate noisy and clean samples.

Figure 3 shows the effect of our label correction module in the two noise scenarios. In Fig. 3 (top row), the Type I noisy labels are well corrected. In Fig. 3 (bottom row), the Type II noisy labels are all well corrected at boundary locations marked by a single arrow. In addition, Co-Seg also fills in the large region of missing pixels marked by a double arrow.

4. CONCLUSIONS

In this paper, we develop a novel collaborative training framework, Co-Seg, to improve segmentation robustness against noisy labels. The robust training module uses two networks to learn the representative features from reliable samples in a noisy dataset. A label correction module employs a voting

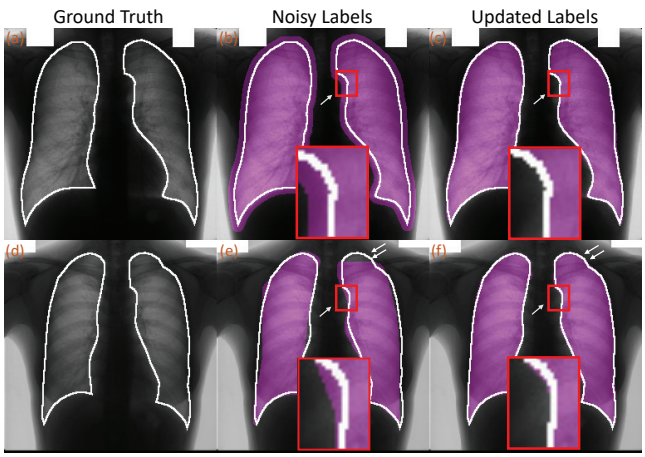

Fig. 3. Label correction results with 0.5 noise level for Type I (top row) and Type II (bottom row) noise types. Ground-truth segmentation lung boundaries are shown in white while the noisy/updated labels are marked in pink.

mechanism to enrich a reliable set prior to final retraining. Experimental results in both synthetic and real-world noise scenarios show that our Co-Seg model is robust to label corruption and achieves comparable results with those trained with noise-free datasets. Importantly, our training scheme is generic and can be easily applied to other deep learning models to increase noise immunity. Future work will focus on extensive validation on more medical segmentation tasks.

5. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted retrospectively using human subject data made available in open access by the Japanese Society of Radiological Technology. Ethical approval was not required as confirmed by the license attached with the open access data.

6. ACKNOWLEDGMENTS

The study was funded in part by the National Institute of Health (4DP2HL127776-02, CPH, subaward of UL1TR003096, YG), National Science Foundation (CRII-1948540, YG).

7. REFERENCES

[1] O. Ronneberger, P.Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention*. 2015, vol. 9351 of *LNCS*, pp. 234–241, Springer.

[2] Debesh Jha, Michael A Riegler, Dag Johansen, Pål Halvorsen, and Håvard D Johansen, "DoubleU-Net: A deep convolutional neural network for medical image segmentation," *arXiv preprint arXiv:2006.04868*, 2020.

[3] Zaiwang Gu, Jun Cheng, Huazhu Fu, Kang Zhou, Huaying Hao, Yitian Zhao, Tianyang Zhang, Shenghua Gao, and Jiang Liu, "Ce-Net: Context encoder network for 2D medical image segmentation," *IEEE Transactions on Medical Imaging*, vol. 38, no. 10, pp. 2281–2292, 2019.

[4] Benoit Frenay and Michel Verleysen, "Classification in the presence of label noise: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 5, pp. 845–869, 2013.

[5] Xingjun Ma, Yisen Wang, Michael E Houle, Shuo Zhou, Sarah M Erfani, Shu-Tao Xia, Sudanthi Wijewickrema, and James Bailey, "Dimensionality-driven learning with noisy labels," *arXiv preprint arXiv:1806.02612*, 2018.

[6] Jacob Goldberger and Ehud Ben-Reuven, "Training deep neural-networks using a noise adaptation layer," in *International Conference on Learning Representations*, 2017.

[7] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu, "Making deep neural networks robust to label noise: A loss correction approach," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1944–1952.

[8] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama, "Co-teaching: Robust training of deep neural networks with extremely noisy labels," in *Advances in Neural Information Processing Systems*, 2018, pp. 8527–8537.

[9] Haidong Zhu, Jialin Shi, and Ji Wu, "Pick-and-learn: Automatic quality evaluation for noisy-labeled image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 576–584.

[10] Zahra Mirikharaji, Yiqi Yan, and Ghassan Hamarneh, "Learning to segment skin lesions from noisy annotations," in *Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data*, pp. 207–215. Springer, 2019.

[11] Junji Shiraishi, Shigehiko Katsuragawa, Junpei Ikezoe, Tsuneo Matsumoto, Takeshi Kobayashi, Ken-ichi Komatsu, Mitate Matsui, Hiroshi Fujita, Yoshie Kodera, and Kunio Doi, "Development of a digital image database for chest radiographs with and without a lung nodule: Receiver operating characteristic analysis of radiologists' detection of pulmonary nodules," *American Journal of Roentgenology*, vol. 174, no. 1, pp. 71–74, 2000.

[12] Bram Van Ginneken, Mikkel B Stegmann, and Marco Loog, "Segmentation of anatomical structures in chest radiographs using supervised methods: A comparative study on a public database," *Medical Image Analysis*, vol. 10, no. 1, pp. 19–40, 2006.

[13] Xiang He, Sibei Yang, Guanbin Li, Haofeng Li, Huiyou Chang, and Yizhou Yu, "Non-local context encoder: Robust biomedical image segmentation against adversarial attacks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33, pp. 8417–8424.

[14] Sangheum Hwang and Sunggyun Park, "Accurate lung segmentation via network-wise training of convolutional networks," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pp. 92–99. Springer, 2017.

[15] Hwanjun Song, Minseok Kim, Dongmin Park, and Jae-Gil Lee, "Learning from noisy labels with deep neural networks: A survey," *arXiv preprint arXiv:2007.08199*, 2020.