

Stochastic co-teaching for robust cardiac segmentation in ultrasound with noisy labels

Gino E. Jansen¹

G.E.JANSEN@UVA.NL

Mark J. Schuurin²

Berto J. Bouma³

Ivana Išgum¹

¹ *qurAI group, Informatics Institute, University of Amsterdam, Netherlands*

² *Department of Biomedical Signals and Systems, University of Twente, Netherlands*

³ *Department of Cardiology, Amsterdam UMC UvA, Netherlands*

Editors: Under Review for MIDL 2025

Abstract

In this work, we propose a label noise-robust segmentation framework for left ventricle blood pool segmentation in echocardiography. Based on the stochastic co-teaching approach, our method extends pixel-level filtering of label noise with additional image-level filtering to more effectively prevent noisy labels from backpropagating. We evaluate our framework on the EchoNet-Dynamic dataset, and simulate diverse noisy label scenarios, including over- and undersegmented (i.e., biased) labels. Our results demonstrate that the incorporation of image-based rejection enhances the Dice coefficient by 1.5% points and ejection fraction estimation by 2.3% points with respect to the pixel-based co-teaching framework under heavily biased label noise conditions, and thereby maintains the same performance as on clean data.

Keywords: Echocardiography, Left Ventricle Segmentation, Deep Learning, Label Noise

1. Introduction

Transthoracic echocardiography (TTE) serves as the primary imaging modality for assessing cardiac function. A key quantitative parameter derived from TTE is the left ventricular ejection fraction (LVEF), which quantifies the ratio of blood ejected from the left ventricle relative to its total volume at end-diastole (Lang et al., 2015). Accurate automatic LVEF estimation depends on precisely delineated examples of the left-ventricle blood pool in TTE images, which may be compromised by less experienced annotators, who contribute noisy delineations (Karimi et al., 2020; Shi et al., 2024).

Various strategies have been developed to mitigate the adverse effects of label noise in deep learning. For example, Han et al. (2018) introduced the co-teaching framework, wherein two networks are trained simultaneously on the same mini-batches but cross-select different subsets of instances based on their loss values for optimization. In this approach, each network discards a predetermined fraction of samples with the highest losses to update its peer. Building on this idea, de Vos et al. (2023) proposed stochastic co-teaching (StoCoT), which employs randomly selected rejection thresholds, thereby improving performance in scenarios with unknown noise levels. The authors showed that co-teaching can easily be extended to segmentation tasks, as segmentation is usually framed as a pixel-wise classification problem, and single pixels can be masked out from backpropagation.

In this work, we employ the stochastic co-teaching framework for echocardiography left ventricle segmentation in echocardiography. To prevent a data imbalance between center and border pixels, we perform noisy-label rejection both on the pixel-level and on the image level, and show that this improves performance compared with stochastic co-teaching on the pixel-level only.

2. Method

We propose to adapt the stochastic co-teaching framework, which consists of two parallel networks that identify and filter out noisy labels for *each other's* optimization (cross-updates), thereby mitigating confirmation bias. In addition to the pixel-based rejection scheme proposed by [de Vos et al. \(2023\)](#), we add a frame-based rejection rule where entire images are excluded from training if they are deemed to contain excessive label noise.

The stochastic co-teaching algorithm operates as follows: During training, for each pixel, the posterior probability of the reference class is computed. For a pixel with binary label $y_{\text{true}} \in \{0, 1\}$, the network outputs a logit which is transformed via a sigmoid function to yield y_{pred} . The posterior probability is then defined as $P = y_{\text{true}} \cdot y_{\text{pred}} + (1 - y_{\text{true}}) \cdot (1 - y_{\text{pred}})$ for a binary classification task. A rejection threshold τ is sampled from a Beta distribution ($\alpha = 1, \beta = 8$); a pixel's loss is rejected from backpropagation through the network (i.e., masked out), if $P < \tau$ for the same pixel in the *peer* network. Additionally, an entire image is rejected, if more than 5% of its pixels are discarded through this process.

3. Experiments & Results

We evaluate our method on the EchoNet-Dynamic dataset ([Ouyang et al., 2020](#)), which consists of echocardiogram videos from 10,030 patients with expert delineations of the left-ventricle blood pool in end-systolic and end-diastolic frames. In addition to using the original annotations, we simulate label noise with three settings: **(a)** Positive bias 45%: 5 consecutive morphological dilation operations to 45% of the training images to mimic a consistently over-segmented blood pool by a contributing annotator. **(b)** Negative bias 45%: 5 consecutive erosions to 45% of the training images to simulate consistent under-segmentation. **(c)** Symmetric 80%: 1 to 5 erosions or dilations (uniformly random choice) to 80% of the training images to mimic inconsistent segmentation.

We compare our approach with four training strategies: **(i)** Baseline: A single model trained with binary cross-entropy loss from ([Ouyang et al., 2020](#)); **(ii)** co-teaching: Two networks that each mask out the 5% of pixels with the highest losses per image; **(iii)** stochastic co-teaching (pixel-based); and **(iv)** stochastic co-teaching (pixel- + image-based).

For all experiments, we adopt the settings from ([Ouyang et al., 2020](#)), including a DeepLab-v3 architecture with a ResNet-50 backbone, initialized with random weights, and weight optimization using SGD (learning rate $1e-5$, momentum 0.9). Input images are resized to 112×112 and normalized to zero mean and unit variance. Training runs for 50 epochs, and we report results from the epoch with the best validation loss. Each experiment is repeated 5 times with different random seeds. The co-teaching experiments are performed with an initial warm-up phase of 10 epochs with regular training, and a subsequent 10 epochs of linear stepwise increase of the rejection threshold. Segmentation performance is assessed via the Dice coefficient on annotated frames and clinical importance is assessed using the absolute error of ejection fraction (EF). EF is computed using the end-systolic volume (ESV) and end-diastolic volume (EDV) according to: $EF = \frac{EDV - ESV}{EDV}$, with both volumes estimated as $\text{Volume} = 0.85 \times \frac{A^2}{L}$, where A is the segmentation area, and L the left-ventricle length ([St John Sutton et al., 1998](#)).

Table 1 summarizes the experimental results. The proposed image-based stochastic co-teaching framework outperforms all other methods in the positive bias 45% setting, achieving an average Dice score of 91.7%—comparable to performance on the clean dataset—and a mean absolute EF error of 9.0 percentage points. Under the Symmetric 80% noise condition, the framework yields a modest Dice improvement, though with a slight decline in EF accuracy, while no significant gains are observed for the Negative bias scenario. The experiments on the clean dataset show that the application of stochastic co-teaching preserves baseline performance. Figure 1(b) illustrates

Table 1: Results on the test set, reported as mean $\pm$ standard deviation (computed across 5 random seeds). Dice scores are in % and MAE of EF in percentage points. Best means are in bold; * indicates a significant improvement ($p < 0.05$, Bonferroni-corrected, paired t-tests).

Method	Clean		Pos. bias 45%		Neg. bias 45%		Sym. 80%	
	Dice	MAE	Dice	MAE	Dice	MAE	Dice	MAE
Baseline	91.7 $\pm$ 0.3	8.6 $\pm$ 0.3	89.7 $\pm$ 0.6	12.8 $\pm$ 2.0	90.0 $\pm$ 0.6	10.7 $\pm$ 1.5	90.2 $\pm$ 0.4	9.7 $\pm$ 0.3
Co-teaching	91.3 $\pm$ 0.3	9.5 $\pm$ 0.5	89.4 $\pm$ 0.9	12.5 $\pm$ 1.6	88.8 $\pm$ 1.0	10.7 $\pm$ 2.1	88.4 $\pm$ 4.4	13.1 $\pm$ 5.8
StoCoT (pixel)	91.8 $\pm$ 0.1	8.7 $\pm$ 0.3	90.2 $\pm$ 0.4	11.3 $\pm$ 1.6	89.4 $\pm$ 0.6	11.1 $\pm$ 1.2	90.6 $\pm$ 0.5	10.1 $\pm$ 1.1
StoCoT (pixel+image)	91.8 $\pm$ 0.2	8.6 $\pm$ 0.4	91.7 $\pm$ 0.2*	9.0 $\pm$ 0.4	88.5 $\pm$ 0.3	9.9 $\pm$ 1.1	90.9 $\pm$ 0.3	10.9 $\pm$ 0.8

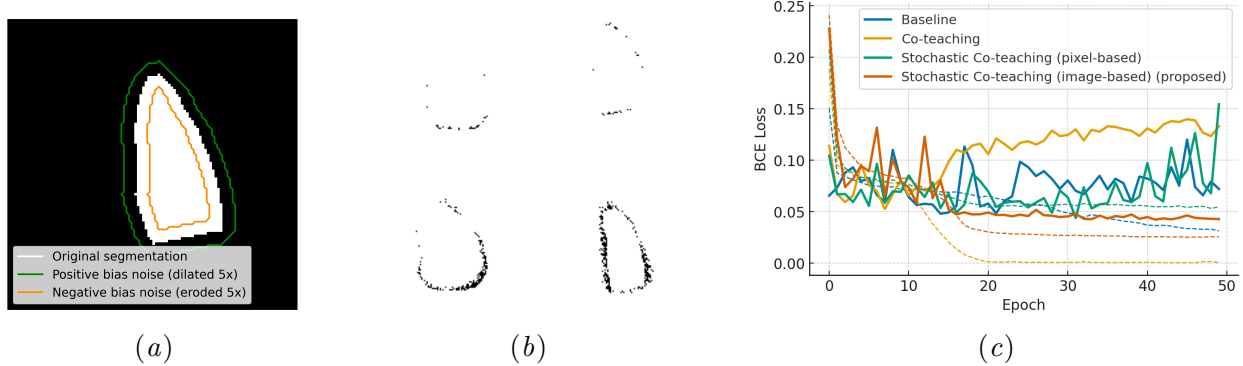


Figure 1: (a) Synthetic label noise examples. (b) Selection masks from stochastic co-teaching (black pixels: rejected; top two are clean, bottom two noisy). (c) Loss curves (solid: validation, dashed: training).

example pixel rejection masks generated during training, where noisier segmentation labels lead to broader rejected regions around the borders. Furthermore, as shown in Figure 1(c), all methods except the proposed image-based approach exhibit signs of overfitting.

4. Discussion and conclusion

We applied and evaluated stochastic co-teaching for training models to segment the left ventricle blood pool in echocardiography. Our method extends the stochastic co-teaching framework by incorporating image-level rejection in addition to pixel-level rejection. This dual rejection strategy effectively filters out mislabeled data, as evidenced by improved performance compared to using pixel-based rejection alone.

We hypothesize that relying solely on pixel-based rejection may lead to an imbalance in training: excessive rejection of border pixels could result in a predominance of central pixels during backpropagation, thereby impairing the learning of boundary details. The image-based rejection helps mitigate this imbalance, ensuring that both border and central regions contribute adequately to the training process. In the negative bias experiment, the proposed method failed because it treated the clean labels as the noisy ones, and vice versa, indicating that a 45% noise rate was too high for this setting. However, in a more realistic noise scenario, the noisy annotator may also be less consistent. Future work should evaluate this method on label noise from expert annotators, and investigate if and why the method underperforms in scenarios with negatively biased noise.

Acknowledgments

This work was supported by Esaote SpA and Pie Medical Imaging BV.

References

- Bob D de Vos, Gino E Jansen, and Ivana Išgum. Stochastic co-teaching for training neural networks with unknown levels of label noise. *Scientific reports*, 13(1):16875, 2023.
- 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. *Advances in neural information processing systems*, 31, 2018.
- Davood Karimi, Haoran Dou, Simon K Warfield, and Ali Gholipour. Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. *Medical image analysis*, 65:101759, 2020.
- Roberto M Lang, Luigi P Badano, Victor Mor-Avi, Jonathan Afilalo, Anderson Armstrong, Laura Ernande, Frank A Flachskampf, Elyse Foster, Steven A Goldstein, Tatiana Kuznetsova, et al. Recommendations for cardiac chamber quantification by echocardiography in adults: an update from the american society of echocardiography and the european association of cardiovascular imaging. *European Heart Journal-Cardiovascular Imaging*, 16(3):233–271, 2015.
- David Ouyang, Bryan He, Amirata Ghorbani, Neal Yuan, Joseph Ebinger, Curtis P Langlotz, Paul A Heidenreich, Robert A Harrington, David H Liang, Euan A Ashley, et al. Video-based ai for beat-to-beat assessment of cardiac function. *Nature*, 580(7802):252–256, 2020.
- Jialin Shi, Kailai Zhang, Chenyi Guo, Youquan Yang, Yali Xu, and Ji Wu. A survey of label-noise deep learning for medical image analysis. *Medical Image Analysis*, 95:103166, 2024.
- M St John Sutton, J-E Otterstat, T Plappert, A Parker, D Sekarski, MG Keane, P Poole-Wilson, and K Lubsen. Quantitation of left ventricular volumes and ejection fraction in post-infarction patients from biplane and single plane two-dimensional echocardiograms. a prospective longitudinal study of 371 patients. *European heart journal*, 19(5):808–816, 1998.