Robust Medical Image Segmentation from Non-expert Annotations with Tri-network

Tianwei Zhang[1], Lequan Yu[2], Na Hu[3], Su Lv[3], and Shi Gu[1(✉)]

[1] Department of Computer and Engineering,
University of Electronic Science and Technology of China, Chengdu, China
gus@uestc.edu.cn
[2] Department of Radiation Oncology, Stanford University, Stanford, CA 94305, USA
[3] Department of Radiology/Huaxi MR Research Center (HMRRC),
West China Hospital of Sichuan University, Chengdu, China

Abstract. Deep convolutional neural networks (CNNs) have achieved commendable results on a variety of medical image segmentation tasks. However, CNNs usually require a large amount of training samples with accurate annotations, which are extremely difficult and expensive to obtain in medical image analysis field. In practice, we notice that the junior trainees after training can label medical images in some medical image segmentation applications. These *non-expert annotations* are more easily accessible and can be regarded as a source of weak annotation to guide network learning. In this paper, we propose a novel Tri-network learning framework to alleviate the problem of insufficient accurate annotations in medical segmentation tasks by utilizing the non-expert annotations. To be specific, we maintain three networks in our framework, and each pair of networks alternatively select informative samples for the third network learning, according to the consensus and difference between their predictions. The three networks are jointly optimized in such a collaborative manner. We evaluated our method on real and simulated non-expert annotated datasets. The experiment results show that our method effectively mines informative information from the non-expert annotations for improved segmentation performance and outperforms other competing methods.

Keywords: Non-expert annotations · Tri-network · Collaborative learning · Segmentation

1 Introduction

Anatomical structure segmentation is one of the key problems in medical image analysis field. In the past years, deep convolutional neural networks (CNNs) have demonstrated promising successes in medical image segmentation tasks [2,8,13]. The high performance of CNNs often relies on a large amount of labeled training data. However, for medical image segmentation tasks, it is time-consuming

T. Zhang and L. Yu—Equal contribution.

A. L. Martel et al. (Eds.): MICCAI 2020, LNCS 12264, pp. 249–258, 2020.
https://doi.org/10.1007/978-3-030-59719-1_25

and expensive to acquire enough reliable annotations, as the annotations are needed to delineate by experienced experts in a slice-by-slice manner [6]. In clinical practice, we noticed that junior trainees can label images after training by a professional doctor. These *non-expert annotations*, which are typically more easily accessible, can be regarded as a source of weak annotation that provides coarsely spatial information but lacks accuracy in detail. Thus a natural question that arises here is whether we can utilize these non-expert annotations to train a segmentation network to alleviate the scarcity of accurate annotation in medical image segmentation tasks. As directly training neural networks with noisy annotations would severely degrade the performance of networks [17], there comes up with the demand of designing special learning strategies that can mitigate the impact of noise labels for deep network training [4,5,9,11,15].

Arpit *et al.* [1] demonstrated that deep networks could benefit by following an 'easy-to-difficult' training procedure under the assumption that the samples with small loss were more likely to be clean. Further, Han *et al.* [4] proposed a framework that co-trained two networks simultaneously and updated each network alternately by the samples with small loss in the other one. Specific for the medical image analysis tasks, Xue *et al.* [16] proposed a sample re-weighting framework for noisy-labeled skin lesion classification, where they removed the high-loss samples during the network training and employed a data re-weighting scheme to weight every reserved sample in one mini-batch. Dgani *et al.* [3] added an additional noise layer to the network for the classification of breast microcalcifications, and Lehtinen *et al.* [7] proposed a learning-from-noisy-sample method and applied it to MR image reconstruction from randomly-sampled data. Although effective on image classification and detection tasks, these methods cannot be straightforwardly applied to the segmentation task, because the noise in image segmentation stands out within the image locally in addition to its label on the global level. As an extension from global label to spatial map, Zhu *et al.* [18] introduced a label quality evaluation strategy to enable neural networks to measure the quality of labels automatically, and Mirikharaji *et al.* [10] proposed to generate a weight map to indicate the more useful pixels and alleviated the influence of the noisy pixels by re-weighting them. Most of previous works provided weight strategy on samples and spatial information partially based on the observation that lower error indicates more informative samples. However, this assumption is arguable for the segmentation task considering that the most informative locations are the boundaries, which may carry a high noise level across different samples. Thus it is necessary to refine the strategy of selecting samples and pixels for the segmentation task by balancing the choices on noise-level and informativity as they are no longer monotonously related due to the nature of segmentation.

In this paper, we aim to develop a learning framework to use noisy non-expert annotations to reduce high-quality annotation effort and combat the inherent noise in real clinical annotations. Following the inspiration of collaborative learning strategy, we propose an efficient framework by extending the Co-teaching [4] to a Tri-teaching network, where two networks jointly select

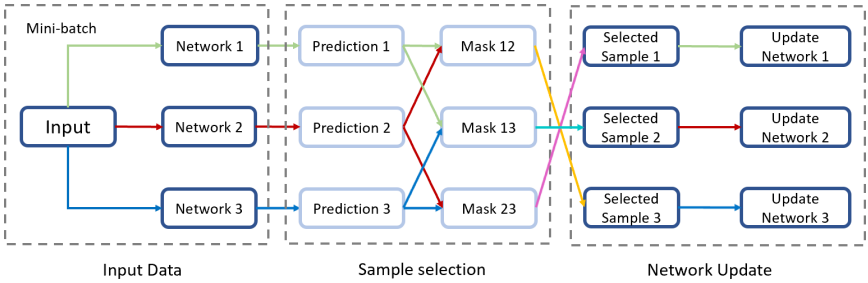

Fig. 1. The pipeline of our Tri-network framework for medical image segmentation from non-expert annotations. We show the procedure in one iteration for illustration.

samples and voxels under the novel strategies designed for two "teachers" rather than one. The introduction of an additional network here not only stabilizes the selecting procedure but also allows for the strategic space to balance noise-level and informativity. To be specific, we train three networks simultaneously and each pair of networks select informative "reliable" samples according to their predictions to guide the third network learning in each iteration. To facilitate the selection of "reliable" samples, we design two feasible strategies according to the consensus and difference between two network outputs. After that, the selected samples are fed into the third network to update its parameters. In this way, three networks jointly learn from the non-expert annotations in a collaborative manner. We evaluate our method on real and simulated non-expert annotation datasets, *i.e.*, stroke lesion segmentation dataset with real noise and public organ segmentation dataset with simulated noise. The results show that our method can effectively use non-expert annotations to improve segmentation performance and outperforms other competing methods.

2 Method

2.1 Overview

We illustrate the training procedure of Tri-network in Fig. 1. The key idea here is to train three networks simultaneously, where each pair of networks guide the third network to mine useful and reliable information from the non-expert annotations. Given a mini-batch of input data, we separately feed them into three networks (*e.g.*, U-Net) at the same time, and acquire three different prediction maps and the corresponding pixel-wise loss maps respecting to the noisy annotations. Next, for each pair of networks (*e.g.*, Network 1 and Network 2), we select those reliable pixels based on the output of two networks with the proposed sample selection strategy and generate one mask (*e.g.*, Mask 12) to represent those selected pixels. After that, we feed the mask to the third network (*e.g.*, Network 3) and guide the network to utilize that useful information for parameter updating. The same procedure is repeated for each network at each training

iteration. In the testing stage, we feed the test data into three trained networks and use the ensemble of the three outputs as the final prediction.

2.2 Sample Selection with Prediction Confidence

As mentioned above, in network training, we generate a mask to represent selected pixel samples from each pair of networks. In other words, those useful samples are which two source networks consider more valuable according to the confidence of their predictions. We then use this mask to guide the training process of the third network so that it can focus more on these valuable pixels, thereby reducing the impact of noisy annotations. Here we propose two sample selection criteria, both of which prove to be effective in dealing with the segmentation problem with non-expert noisy annotations.

Consensus-Based Selection. The first sample selection strategy is based on *consensus* of network predictions. For each mini-batch data, the pair of networks produce two pixel-wise prediction maps and we further get the corresponding confidence map (or loss map) of each network by calculating loss for each pixel between network prediction and the noisy annotations. For each loss map, we sort those pixel-wise loss values and set a loss value threshold to get a binary confidence map B, where one represents T percentage of small-loss pixels (*i.e.*, high confidence) and zero indicates the loss values of corresponding pixels are greater than the specific ratio (*i.e.*, low confidence). Based on the binary confidence map B_1 and B_2 of two networks, we further calculate one binary *consensus map* $M_{cons} = (B_1 == B_2)$. There are two kinds of pixels in the consensus map: pixels with high prediction confidence (*i.e.*, low loss value) in both two networks and pixels with low prediction confidence (*i.e.*, high loss value) in both two networks. The first kind of pixels can be regarded as "clean" pixels and the second kind of pixels can be regarded as "informative" pixels. We feed both two kinds of pixels into the third network and calculate the loss for back-propagation.

Difference-Based Selection. The second sample selection strategy is based on *difference* of network predictions. Similar to the consensus strategy, we first calculate the pixel-wise confidence map (or loss map) for each network prediction. And then we calculate the loss difference map of two networks by subtracting the two loss maps and take the absolute values. In this strategy, we mine useful knowledge by choosing pixels that are greater than a specific proportion, *i.e.*, T percentage of the large-loss-difference, in the above loss difference map and generate a binary *difference map* M_{diff}. Finally, we feed the binary difference map into the third network and update its parameter with these pixels selected by the other two networks which are the same as before.

2.3 Technical Details

The framework was implemented with PyTorch on a TITAN Xp GPU. The three networks in our framework share the same network architecture, *i.e.*, U-Net [12].

The whole framework was optimized with the SGD optimizer with a momentum of 0.9 and a weight decay of 0.0001. The network was trained for 200 epochs with learning rate 0.001 until convergence. To fit the limited GPU memory for training three networks simultaneously, we set the mini-batch size as 8 and resized all the input images to 256×256 pixels. We generate the confidence map (or loss map) by calculating the cross entropy loss for each pixel, and calculate the final loss by averaging the CE loss on selected pixels for back-propagation. For the sample selection strategies, we use all pixels to update all networks at the beginning, and then gradually adjust the specific ratio T during training. Specifically, in the Consensus-based strategy, we set the ratio T as 1 at the beginning, and then linearly decreased it to 50% within 300 iterations and keep it unchanged during the remaining iterations. While in the Difference-based strategy, we set T as 0 at the beginning, and then linearly increased to 50% within the first 300 iterations.

3 Experiment

We evaluated our method on two datasets. We first validated the effectiveness of our proposed method on the stroke lesion dataset with real non-expert annotations and further analyzed the impact of noise level and ratio on a public organ segmentation dataset with simulated non-expert annotations.

3.1 Experiment Setting

Real Clinical Dataset. The clinical stroke lesion segmentation dataset was collected from a multi-modal imaging database of patients with suspected AIS in West China Hospital. This dataset contained all MRI scan sequences of 186 Stroke emergency patients, and the image modality used in our experiments is the FLAIR sequence. We randomly divided the dataset into training set (150 scans) and testing set (36 scans) for our experiments. To acquire the non-expert annotations on the training set, we recruited a junior trainee to annotate all the data after simple training. For the testing data, we invited a six-years-experienced neuroradiologist to annotate the stroke lesion regions as the ground truth.

Simulated Dataset. We employed the public dataset JSRT [14] as a multi-class classification dataset to simulate noisy annotations to further evaluate and analyze the capacity of our method. JSRT has three types of organ annotation information: Lungs, Hearts, and Clavicles. The total of 247 X-ray scans was split into 165 training scans and 82 evaluation scans. Considering that the manual noise of organ segmentation mainly is the inaccurate contours, we simulated the noisy non-expert annotations by randomly conducting morphological changes to the original clean annotations. Specifically, the simulated noise was generated by randomly eroding or dilating the contours of accurate annotations.

Table 1. Comparison with other methods on the clinical stroke dataset.

Method	Dice coefficient[%]
U-Net [12]	62.18
Ensemble U-Net	63.90
Xue *et al.* [16]	64.92
Co-teaching [4]	64.04
Tri-network (Consensus)	**68.12**
Tri-network (D-value)	67.88

Experiment Setting. We trained our framework on training data with non-expert noisy annotations and evaluated the model on testing data by the Dice coefficient score between the predicted segmentation and the accurate ground truth annotations. We compared our Tri-network with multiple recent frameworks including vanilla U-Net [12], ensemble model of U-Net, the loss re-weighting method [16] and Co-teaching [4] under the same setting. For the JSRT dataset, we reported the average results of 5 runs with random data splitting per run.

3.2 Experiments on Real Clinical Stroke Dataset

Quantitative Analysis. We first evaluated the model performance and compared with other related methods on the clinical stroke lesion segmentation dataset (see Table 1). As a baseline, the single U-Net trained with non-expert annotations achieved 62.18% Dice coefficient score on the test data. Among the compared methods, our Tri-network with Consensus-based selection achieved the highest Dice coefficient of 68.12% and Tri-network with Difference-based selection achieved a slightly lower score of 67.88%, both of which outperformed the Ensemble U-net (63.9%), Xue et al. (64.92%), and Co-teaching (64.04%). Compared to the baseline U-Net, all the other methods improved at a certain margin, demonstrating that it was feasible and effective to utilize specialized learning algorithms to train networks with noisy segmentation. For our Tri-network, its improvement over the U-Net (*Ensemble U-Net*) indicated that the selection procedures made a difference besides the ensemble of models. In addition, it also outperformed the sample re-weighting based method [16] and Co-teaching [4], supporting our claim that the training and selection procedure of Tri-network was more robust and adaptive to the segmentation problem.

To further investigate whether the improvement over Co-teaching was solely due to the increased number of networks, we re-trained the Tri-network with degenerated consensus-based sample selection strategy where we only selected small or large loss pixels rather than both. We found that the training process was difficult to converge for the small-loss selection and resulted at 64.27% Dice coefficient score for the large-loss selection. The divergence with small-loss selection was probably caused by the fact that small loss pixels were mainly in

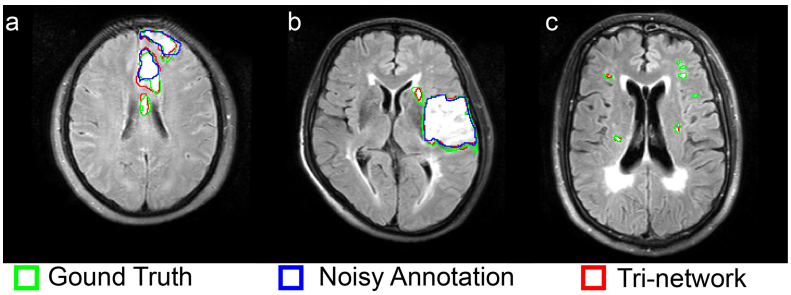

Fig. 2. Visual examples on the clinical dataset. The green, red, and blue colors denote the ground truth annotation, Tri-network, and the noisy annotation, respectively. Our method generates more accurate contours than non-expert annotation (a) and successfully identify small lesions (b & c). (Color figure online)

the center of the lesion, while the most informative samples for segmentation tasks were the pixels near the boundaries. This also supports our early claim in the introduction that the spatial information with a sample distinguishes the segmentation problem from the classification problem. The large-loss selection improved over the baseline and provided a comparable result with Co-teaching approaches, indicating that the efficiency of integrated rules of Tri-networks was beyond the degenerated rule that simply extended Co-teaching to "Tri-teaching" with model ensemble.

Overall, the comparisons we have done prove that both the extension to three network components and integrated sample section strategies and necessary and effective for segmentation with noise annotations.

Qualitative Analysis. In addition to the quantitative comparison, we showed vivid segmentation results in Fig. 2, where the green, red, and blue color denoted the ground truth annotation, Tri-network segmentation, and the non-expert annotation, respectively. Due to the lack of professional medical knowledge, the non-expert annotators often generate annotation noise, including marking areas that are not lesions, missing areas that are lesions, and delineating inaccurate contours of the marked lesions. Especially, in the task of stroke segmentation in FLAIR sequence, where the lesion area was imaged as high signals, we noticed that the non-experts often missed the tiny lesions around the demyelinating area and also delineated inaccurate contours of relative large stroke areas (compare the green v.s. the blue contours in Fig. 2). As we can see from the left panel, the contour of our segmented results was more accurate than the non-expert annotation. And the mid and right panels showed that our framework successfully recognized small lesions near demyelinating area that the non-expert could miss.

3.3 Experiments on Simulated Noisy Dataset

While we have proved the feasibility and effectiveness of our model on the stroke lesion segmentation dataset, the noise-level on this real dataset was fixed. To explore the model generalizability and its capacity to handle different levels of noise, we conducted additional experiments on another public X-ray dataset with simulated noisy label on different levels and ratios. We randomly selected 75% training samples and further randomly eroded/dilated the contour with 5–10 pixels to simulate the non-expert circumstance Fig. 3(c). The U-Net trained with clean labels (fist row in Table 2) performed fairly well and we set it as the upper bound of performance here. Compared with the U-Net trained with clean labels, the performance of U-Net trained with noise label severely decreased on all three organs, especially on the hardest small clavicle. Our method achieved 6.04% and 6.62% average dice improvement with the consensus-based and difference-based sample selection strategy, respectively, while the other learning from noisy label methods produced slightly better performance than the baseline model on this dataset with simulated noise Fig. 3(d). Especially for the clavicle segmentation, our Tri-network achieves about 11% dice score improvement. The average Hausdorff distance for lung, heart, and clavicle is 6.16, 4.83, and 6.63 pixels, respectively. These comparison results demonstrated the effectiveness of our method to utilize noisy annotations on simulated data again.

Table 2. Comparison with other methods on JSRT dataset on Dice metric [%].

Method	Lungs	Heart	Clavicles	Mean
U-Net (no noise)	97.55	94.82	92.35	94.90
U-Net	83.35	86.52	53.29	74.39
Ensemble U-Net	84.52	86.19	58.81	76.51
Xue *et al.* [16]	85.17	86.38	57.11	76.22
Co-teaching [4]	85.06	86.64	57.78	76.49
Tri-network (Consensus)	87.79	88.54	**64.98**	80.43
Tri-network (D-value)	**88.59**	**90.37**	64.07	**81.01**

We also studied the performance of our method under different noisy level and noisy rate. Specifically, we studied two noise levels: low noise level with morphological change within 1 to 5 pixels and high noise level with morphological change within 5 to 10 pixels. For each noisy level, we evaluated our method with noisy rate at 25%, 50%, and 75%, where we conducted random morphological operations for 25%, 50%, and 75% training samples. The clavicle segmentation results are shown in Fig. 4. Compared with U-Net baseline, our method improved the performance under different settings. Overall, our method outperforms other compared methods and the improvement is more obvious at a high noisy rate.

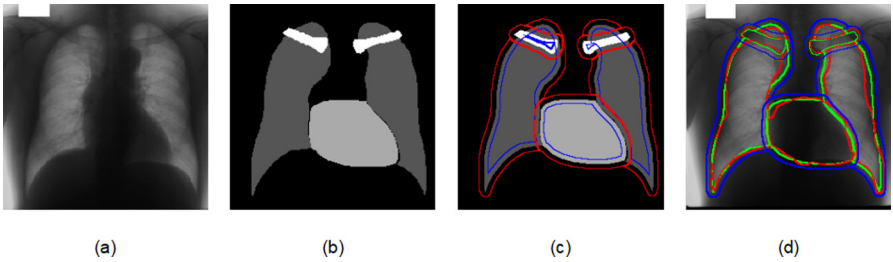

Fig. 3. Simulated noise example and visual result of JSRT Dataset. **(a)** Original X-ray image. **(b)** Ground Truth annotation. **(c)** Two kinds of simulated noise: red contour represents dilation and blue contour represents erosion. **(d)** Visual results of segmentation results: the green, red, and blue colors denote the ground truth annotation, Tri-network, and the noisy annotation, respectively. (Color figure online)

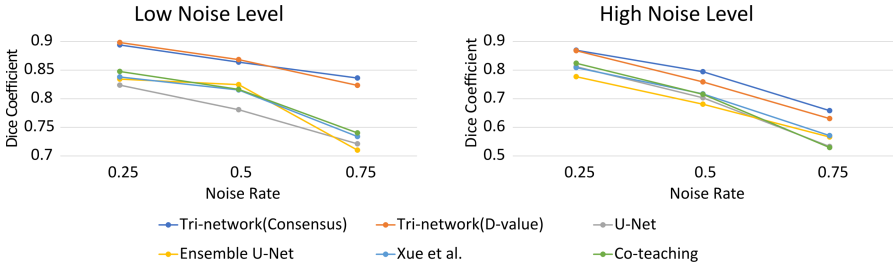

Fig. 4. The performance of clavicle segmentation of different methods on JSRT dataset with different noise settings.

4 Conclusion

In this work, we propose a Tri-network framework with integrated sample selection strategies to tackle the problem of leveraging non-expert annotations for robust medical image segmentation. The Tri-network trains three deep networks simultaneously and employs each pair of networks to guide the third network to mine useful and informative samples during training. The whole framework is optimized in a collaborative manner. As the key part of our framework, we develop two effective sample selection strategies according to the consensus or difference of two network predictions. We verify the effectiveness of our proposed framework on both real non-expert annotated dataset and simulated noisy dataset. The experimental results demonstrate that our method can improve the performance of the network trained with the non-expert annotations and outperform other competing methods.

References

1. Arpit, D., et al.: A closer look at memorization in deep networks 2017. arXiv preprint arXiv:1706.05394 (1938)

2. Ching, T., et al.: Opportunities and obstacles for deep learning in biology and medicine. J. Roy. Soc. Interf. **15**(141), 20170387 (2018)
3. Dgani, Y., Greenspan, H., Goldberger, J.: Training a neural network based on unreliable human annotation of medical images. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pp. 39–42. IEEE (2018)
4. Han, B., et al.: Co-teaching: robust training of deep neural networks with extremely noisy labels. In: Advances in Neural Information Processing Systems, pp. 8527–8537 (2018)
5. Jiang, L., Zhou, Z., Leung, T., Li, L.J., Fei-Fei, L.: MentorNet: learning data-driven curriculum for very deep neural networks on corrupted labels. arXiv preprint arXiv:1712.05055 (2017)
6. Kohli, M.D., Summers, R.M., Geis, J.R.: Medical image data and datasets in the era of machine learning–whitepaper from the 2016 C-MIMI meeting dataset session. J. Digit. Imaging **30**(4), 392–399 (2017)
7. Lehtinen, J., et al.: Noise2noise: learning image restoration without clean data. arXiv preprint arXiv:1803.04189 (2018)
8. Litjens, G., et al.: A survey on deep learning in medical image analysis. Med. Image Anal. **42**, 60–88 (2017)
9. Ma, X., et al.: Dimensionality-driven learning with noisy labels. arXiv preprint arXiv:1806.02612 (2018)
10. Mirikharaji, Z., Yan, Y., Hamarneh, G.: Learning to segment skin lesions from noisy annotations. In: Wang, Q., et al. (eds.) DART/MIL3ID - 2019. LNCS, vol. 11795, pp. 207–215. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-33391-1_24
11. Patrini, G., Rozza, A., Krishna Menon, A., Nock, R., Qu, L.: Making deep neural networks robust to label noise: a loss correction approach. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1944–1952 (2017)
12. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
13. Shen, D., Wu, G., Suk, H.I.: Deep learning in medical image analysis. Ann. Rev. Biomed. Eng. **19**, 221–248 (2017)
14. Shiraishi, J., et al.: Development of a digital image database for chest radiographs with and without a lung nodule. Am. J. Roentgenol. **174**(1), 71–74 (2000)
15. Tanaka, D., Ikami, D., Yamasaki, T., Aizawa, K.: Joint optimization framework for learning with noisy labels. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5552–5560 (2018)
16. Xue, C., Dou, Q., Shi, X., Chen, H., Heng, P.A.: Robust learning at noisy labeled medical images: applied to skin lesion classification. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), pp. 1280–1283. IEEE (2019)
17. Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O.: Understanding deep learning requires rethinking generalization. arXiv preprint arXiv:1611.03530 (2016)
18. 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