A Multitask Framework for Label Refinement and Lesion Segmentation in Clinical Brain Imaging

Yang Yu[1], Jiahao Wang[2], Ashish Jith Sreejith Kumar[1], Bryan Tan[1], Navya Vanjavaka[3], Nurul Hafidzah Rahim[3], Alistair Koh[1], Shaheen Low[1], Yih Yian Sitoh[3], Hanry Yu[2], Pavitra Krishnaswamy[1], and Ivan Ho Mien[1,3](✉)

[1] Institute for Infocomm Research (I^2R), A*STAR, Singapore, Singapore
{yu_yang,Ivan_Ho}@i2r.a-star.edu.sg
[2] National University of Singapore (NUS), Singapore, Singapore
[3] National Neuroscience Institute (NNI), Singapore, Singapore

Abstract. Supervised deep learning methods offer the potential for automating lesion segmentation in routine clinical brain imaging, but performance is dependent on label quality. In practice, obtaining high-quality labels from experienced annotators for large-scale datasets is not always feasible, while noisy labels from less experienced annotators are often available. Prior studies focus on either label refinement methods or on learning to segment with noisy labels, but there has been little work on integrating these approaches within a unified framework. To address this gap, we propose a novel multitask framework for end-to-end noisy-label refinement and lesion segmentation. Our approach minimizes the discrepancy between the refined label and the predicted segmentation mask, and is highly customizable for scenarios with multiple sets of noisy labels, incomplete ground truth coverage and/or 2D/3D scans. In extensive experiments on both proprietary and public clinical brain imaging datasets, we demonstrate that our end-to-end framework offers strong performance improvements over prevailing baselines on both label refinement and lesion segmentation. Our proposed framework maintains performance gains over baselines even when ground truth labels are available for only 25–50% of the dataset. Our approach has implications for effective medical image segmentation in settings that are replete with noisy labels but sparse on ground truth annotation.

Keywords: Multitask learning · Noisy labels · Label refinement · Lesion segmentation

Y. Yu and J. Wang—Contributed equally to this work.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-44917-8_6.

Z. Xue et al. (Eds.): MILLanD 2023, LNCS 14307, pp. 60–70, 2023.
https://doi.org/10.1007/978-3-031-44917-8_6

1 Introduction

The importance of accurate lesion segmentation in clinical brain imaging goes beyond disease diagnosis to treatment decisions and surgical/radiotherapeutic planning [1,13]. Automation saves immense time and manpower resources and to that end, supervised deep learning approaches have achieved state-of-the-art performance in segmenting lesions on brain MRI and CT scans [16–18]. However, these approaches still require large image datasets that must be accurately and laboriously labelled at the pixel level. Compounding this difficulty is the challenge of obtaining high-quality ground truth labels arising from intrinsic (e.g., morphological complexity) to human (e.g., clinical experience) factors.

A practical compromise in the balance between label quality and cost is to achieve full dataset annotation from less experienced annotators while strategically using a limited number of more experienced annotators to review and correct a subset of the acquired lower-quality or noisy labels. To mitigate adverse effects on segmentation accuracy from training models with noisy labels [15], there is a need for an end-to-end capability that automatically refines noisy labels and/or segment lesions, guided by higher-quality ground truth labels.

Some studies develop label refinement methods to map noisy label inputs to ground truth label distributions [11,26] while others design methods to perform segmentation with noisy labels by learning the noise distributions [22,23]. However, the literature on combining label refinement and learning with noisy labels within a unified framework is still in its nascency [4]. Furthermore, there is little work demonstrating these novel methods for complex clinical image segmentation tasks in scenarios with noisy labels from less experienced annotators.

Here, we propose a novel multitask framework that simultaneously refines noisy labels and segments lesions in an end-to-end fashion. We treat the noisy label as an additional input and define a multitask problem to jointly predict the ground truth label distribution and segment lesions of interest. We develop a multitask learning strategy that inherently minimizes the discrepancy between the refined labels and the predicted segmentation mask. Our approach is scalable to multiple sets of noisy labels and/or different levels of specialized ground truth guidance. Our main contributions are:

1. We introduce a novel multitask framework to leverage cross-label discrepancy minimization for end-to-end noisy-label refinement and lesion segmentation.
2. We perform a series of experiments on both proprietary and public brain MRI datasets to demonstrate substantial performance improvements over prevailing baselines. Further, we show that our multitask strategy provides a substantial boost in performance over each of the component networks.
3. We demonstrate that our framework is readily adapted for practical scenarios where specialized ground truth labels can only be obtained for a portion of the dataset, and maintains the improved performance over the baselines even with a 50% drop in ground truth coverage.
4. We also validate the scalability of our modular framework on a 3D lesion segmentation task to show its capability for supporting customized 3D deep learning models.

2 Related Work

Label Refinement on Noisy Labels. To refine noisy labels from multiple annotators, the medical image segmentation community employs techniques such as Simultaneous Truth and Performance Level Estimation (STAPLE) [19]. However, these label fusion approaches require multiple noisy label sets and do not readily incorporate information from associated ground truth labels [8]. Others have developed label refinement methods to map noisy label inputs to ground truth label distributions [11,26] but these have yet to be demonstrated for complex clinical image segmentation tasks. A recent study explored generative adversarial networks (GANs) based on a conditional generative model for ophthalmological conditions [21] but this method is not designed to deal with scenarios with incomplete coverage of associated ground truth labels.

Lesion Segmentation with Noisy Labels. In recent years, various approaches have been proposed to handle noisy or incorrect pixel labels in medical image segmentation tasks [9,24]. Some studies have also proposed and further extended methods that leverage noisy labels for pixel-wise loss correction through meta-guided networks, mask-guided attention, and superpixel-guided learning mechanisms [3,5,12]. However, these methods assume a simple mapping between noisy and ground truth labels, and it is unclear whether such functions can adequately handle the significant real-world variations in distributions of labels (across location, size, and shape for example) from different annotators.

There has also been little work on approaches integrating the refinement of noisy labels with lesion segmentation. To our knowledge, only one study has explored this important combination [4]. This work proposed a GAN-based segmentation framework to integrate label correction and sample reweighting strategies but does not address the inherent discrepancy between the refined labels and the predicted segmentation masks, it would have been helpful as a baseline comparison but their official implementation was not released. Differently, we leverage a cross-label discrepancy minimization strategy [25] and adapt this strategy within a novel multi-task framework for noisy label refinement and lesion segmentation.

3 Methods

Problem Formulation. We denote the n samples as $\mathcal{X} = \{x_1, \ldots, x_n\}$ and $\widetilde{\mathcal{Y}} = \{\widetilde{y}_1, \ldots, \widetilde{y}_n\}$, where x_j and $\widetilde{y}_j$ represent the j-th data input for images and noisy labels respectively. Each data input is also associated with a ground truth label denoted as $\mathcal{Y} = \{y_1, ..., y_n\}$. Given a paired testing input, the task is to both refine the noisy label $\widetilde{y}^T$ and segment lesions in the image x^T.

Multitask Learning Framework. Figure 1 illustrates the proposed framework. Our aim is to learn the network-specific transformation in various tasks.

We accomplish this with two encoder-decoder networks trained in a joint manner for label refinement and lesion segmentation. For both networks, we use the similarity loss to eliminate the cross-label discrepancy, and the supervision loss to leverage the ground truth labels.

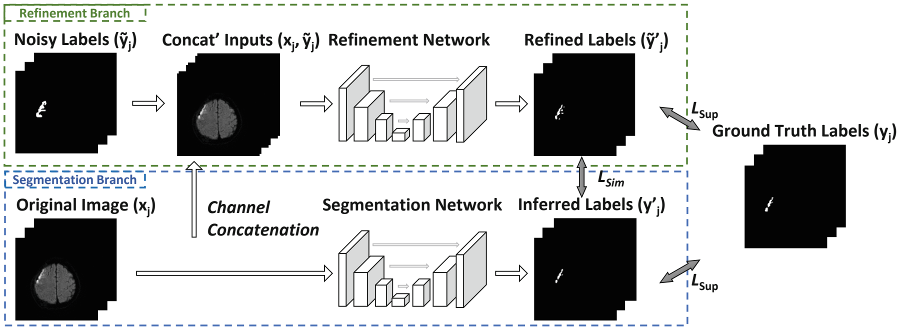

Fig. 1. Proposed end-to-end multitask learning framework for simultaneous label refinement and lesion segmentation.

Specifically, for the j-th input data x_j and $\widetilde{y}_j$, we denote the outputs of the refinement network and segmentation network as $\widetilde{y}'_j$ and y'_j respectively. The similarity loss $\mathcal{L}_{\text{Sim}}$ for both networks is constructed as follows:

$$\mathcal{L}_{\text{Sim}}(x_j, \widetilde{y}_j) = \frac{1}{n}\sum_{j=1}^{n}[(\widetilde{y}'_j - y'_j)^2] \tag{1}$$

The supervision loss $\mathcal{L}_{\text{Sup}}$ for both networks is calculated by:

$$\mathcal{L}_{\text{Sup}}(x_j, \widetilde{y}_j) = \frac{1}{n}\sum_{j=1}^{n}[(1 - \frac{2y_j\widetilde{y}'_j}{\text{Max}((\text{y}_\text{j} + \widetilde{\text{y}}'_\text{j}), \epsilon)}) + (1 - \frac{2y_j y'_j}{\text{Max}((\text{y}_\text{j} + \text{y}'_\text{j}), \epsilon)})] \tag{2}$$

where hyper-parameter ϵ is set to prevent zero-valued denominators. The final objective function across the two networks is then formulated as:

$$\mathcal{L}_{\text{Tot}} = \lambda\mathcal{L}_{\text{Sim}} + \mathcal{L}_{\text{Sup}} \tag{3}$$

where hyper-parameter λ controls the relative importance of the similarity loss. As our proposed framework trains both networks simultaneously in an end-to-end manner, it enables joint refinement of the noisy labels with the prediction of the lesion masks. During testing, the end-to-end refinement and segmentation networks can be used together or separately. The refinement network takes test images with its corresponding noisy labels $\{x^T, \widetilde{y}^T\}$ as inputs and outputs refined labels $\{\widetilde{y}'^T\}$; the segmentation network takes test images $\{x^T\}$ as inputs and outputs inferred labels $\{y'^T\}$.

4 Experiment Setup

Datasets. We evaluated our multitask framework on both proprietary (dwMRI) and public (BraTS 2019, BraTS 2021) MRI datasets (Details in Supplementary Table S1, Figure S1). The dwMRI dataset comprises 298 anonymized diffusion-weighted imaging (DWI) brain studies acquired at a tertiary healthcare institution (National Neuroscience Institute, Singapore) between 2009 to 2019 as part of routine clinical care. Data collection and use were approved by the relevant institutional review boards (SingHealth CIRB Ref 2019/2804). Abnormalities were segmented using ITK-SNAP by two groups of annotators with different experience levels (two radiography students and two neuroradiology fellows) whose annotations yielded one set of noisy labels and ground truth labels respectively. After excluding 7690 images without any DWI abnormalities, 470 images were randomly selected for training/validation (internal splitting ratio 4:1), leaving the remaining 141 images for testing on the 2D task. The BraTS 2019/2021 datasets comprise 335/1251 pre-operative multimodal MRI brain studies of high- and low-grade gliomas with pixel-level annotations categorized into necrotic and non-enhancing tumor (label-1), peritumoral edema (label-2), and enhancing tumor (label-4) [13]. For our experiments, we investigated segmentation of the tumor core (TC) on various sequences by removing label-2 and combining label-4 and label-1 into a single ground truth label. For BraTS 2019, we randomly sampled 500 images for training/validation and 200 images for testing to match with the dwMRI dataset on the 2D task. For BraTS 2021 with more studies provided, we randomly sampled 1000 studies for training/validation and 251 studies for testing on the 3D tasks. To simulate label noise, we applied similar methods used in previous literature [12,23,24,26], namely random shift, random rotation, dilation, and random noise, to the ground truth labels of selected images/studies. Examples of differences between paired ground truth labels and real/simulated noisy labels for both datasets are provided in Supplementary Figure S1.

Implementation Details. The proposed method includes a conventional UNet [18] (2D tasks) and a volumetric UNet [14] (3D tasks) for both refinement and segmentation networks. We demonstrated the results using ResNeXt-50 [20] (2D tasks) and ResNet-110 [6] (3D tasks) as encoder backbones. We implemented all experiments in PyTorch with two Nvidia A5000 GPUs. Hyperparameters are shown in Supplementary Table S2.

Baselines. To compare the performance of our proposed method against other state-of-the-art (SOTA) approaches, we implemented the Conditional GAN algorithm [21], Confident Learning algorithm [24], Attention Module algorithm [9], and volumetric UNet algorithm [14]. We investigate the performance of Conditional GAN for label refinement and the performance of Confident Learning and Attention Module for lesion segmentation on the 2D task. We also examine the performance of Volumetric UNet [14] for lesion segmentation on the 3D task. Hyperparameters are shown in Supplementary Table S3.

Experiments. We performed four experiments to assess the efficacy of our proposed approach for label refinement and image segmentation. In the first experiment, we compared our technique with the aforementioned 2D baselines. In the second experiment, we performed ablation studies to systematically evaluate the effect of each network and similarity loss in our framework. In the third experiment, we simulated scenarios where only a portion (25%, 50%, or 75%) of images have corresponding ground truth labels and assessed performance compared to the fully-paired experiment. In this experiment, we first trained the networks using images with paired noisy and ground truth labels and used the generated pseudo-labels from refined labels as ground truth labels for images with noisy labels only. We further trained the model on both paired and unpaired images using a pseudo-labeling strategy coupled with a deterministic annealing process [10]. In the fourth experiment, we investigated the scalability of our proposed technique to a 3D task with performance compared to the 3D baseline [14].

Table 1. Comparison of refinement and segmentation accuracy for different methods. The means and standard deviations of the DSC (%)/HSD are reported. Best in **bold**.

Task (Methods)	dwMRI		BraTS 2019	
	DSC↑	HSD↓	DSC↑	HSD↓
Refinement (Conditional GAN [21])	49.44 ± 0.45	2.16 ± 0.10	79.84 ± 0.82	3.07 ± 0.03
Refinement (Ours)	$\mathbf{51.46} \pm 0.29$	$\mathbf{2.13} \pm 0.05$	$\mathbf{80.22} \pm 0.07$	$\mathbf{3.02} \pm 0.05$
Segmentation (Confident Learning [24])	34.29 ± 1.62	2.62 ± 0.38	44.13 ± 1.38	3.69 ± 0.11
Segmentation (Attention Module [9])	39.83 ± 1.48	2.47 ± 0.07	56.61 ± 1.38	$\mathbf{3.01} \pm 0.06$
Segmentation (Ours)	$\mathbf{41.96} \pm 2.07$	$\mathbf{2.07} \pm 0.11$	$\mathbf{58.43} \pm 1.29$	3.07 ± 0.03

Evaluation Metrics. We evaluated the Dice Similarity Coefficient (DSC) [2] and Hausdorff Surface Distance (HSD) [7] to quantify the performance of label refinement and lesion segmentation. (Higher DSC and/or lower HSD values indicate better performance.) We employed multiple seeds to compute averaged values and standard deviations for both DSC and HSD metrics.

5 Results

Comparison with SOTA Baselines on Label Refinement. We first benchmarked our results against the recent SOTA baselines for label refinement in Table 1. Our proposed multitask learning approach provides a 2.02%/0.03 boost for the dwMRI dataset and a 0.38%/0.05 improvement for the BraTS 2019 dataset in DSC/HSD over the Conditional GAN refinement baseline. Although simulated noise in the BraTS 2019 dataset is more easily corrected compared with real noise in the dwMRI dataset, the results still display consistently substantial improvements in both actual and simulated noisy labels.

Comparison with SOTA Baselines on Lesion Segmentation. We also benchmarked our results against the recent SOTA baselines for lesion segmentation in Table 1. The Confident Learning approach achieves DSC/HSD of 34.29±1.62 (%)/2.62±0.38 and 44.13±1.38 (%)/3.69±0.11 on dwMRI and BraTS 2019, respectively. When combined with ground truth labels that could better guide the model training process, the Attention Module approach achieves higher DSC of 39.83±1.48 (%)/2.47±0.07 and 56.61±1.38 (%)/3.01±0.06 on dwMRI and BraTS 2019, respectively. Our proposed framework improved over both of these methods with DSC/HSD of 41.96±2.07 (%)/2.07±0.11 and 58.43±1.29 (%)/3.07±0.03 on dwMRI and BraTS 2019, respectively. Our results suggest that the inclusion of noisy labels in our multitask learning approach is able to further enhance semantic consistency among labels with various quality levels.

Table 2. Comparison of refinement and segmentation accuracy for ablation studies. The means and standard deviations of the DSC (%)/HSD are reported. Best in **bold**.

Task (Networks)	dwMRI		BraTS 2019	
	DSC↑	HSD↓	DSC↑	HSD↓
Refinement (Refinement Network)	48.05 ± 0.89	2.23 ± 0.12	78.83 ± 0.25	3.17 ± 0.07
Refinement (Multitask Network w/o $\mathcal{L}_{\text{Sim}}$)	50.37 ± 0.29	2.23 ± 0.02	80.06 ± 0.80	3.08 ± 0.03
Refinement (Multitask Network w/ $\mathcal{L}_{\text{Sim}}$)	$\mathbf{51.46} \pm 0.10$	$\mathbf{2.13} \pm 0.05$	$\mathbf{80.22} \pm 0.07$	$\mathbf{3.02} \pm 0.05$
Segmentation (Segmentation Network)	38.53 ± 2.05	2.05 ± 0.07	55.25 ± 3.16	$\mathbf{3.03} \pm 0.12$
Segmentation (Multitask Network w/o $\mathcal{L}_{\text{Sim}}$)	40.16 ± 2.16	$\mathbf{2.01} \pm 0.26$	55.87 ± 2.58	3.49 ± 0.33
Segmentation (Multitask Network w/ $\mathcal{L}_{\text{Sim}}$)	$\mathbf{41.96} \pm 2.07$	2.07 ± 0.11	$\mathbf{58.43} \pm 1.29$	3.07 ± 0.03

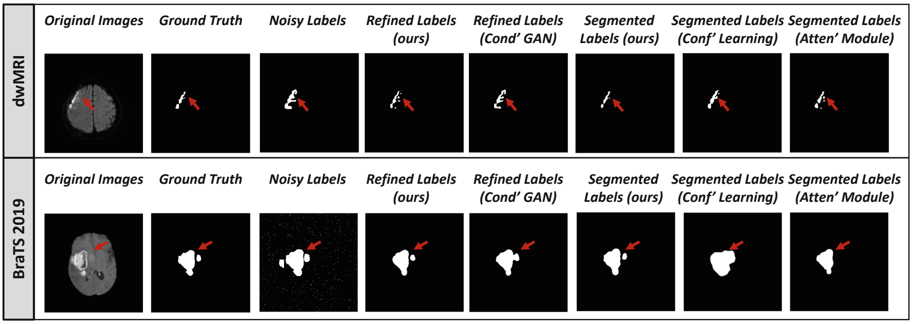

Fig. 2. Visualisation of label refinement and image segmentation on two datasets (dwMRI and BraTS 2019). Subtle differences are indicated by red arrows. (Color figure online)

Visualization of Refined and Segmented Images. For further characterization, we present exemplary results on two images (one from dwMRI and one from BraTS 2019) from the testing sets in Fig. 2. These results demonstrate the potential of our proposed method to correct subtle noises, particularly in connected areas, and also accurately identify semantic lesions, especially when accuracy relies on fine details.

Ablation Studies. Next, we performed ablation studies to systematically assess the contribution of each component in our proposed framework. The results, in Table 2, indicate that injection of the segmentation and refinement paths through a multitask learning framework yields a 3.41% and 3.43% increase in DSC, as well as comparable performance in HSD for the dwMRI dataset; and 1.39% and 3.18% improvement in DSC as well as comparable performance in HSD for the BraTS 2019 dataset on label refinement and lesion segmentation, respectively, compared to using only the refinement or segmentation network alone. The results also show that integration of the similarity loss further improves model performance on lesion detection.

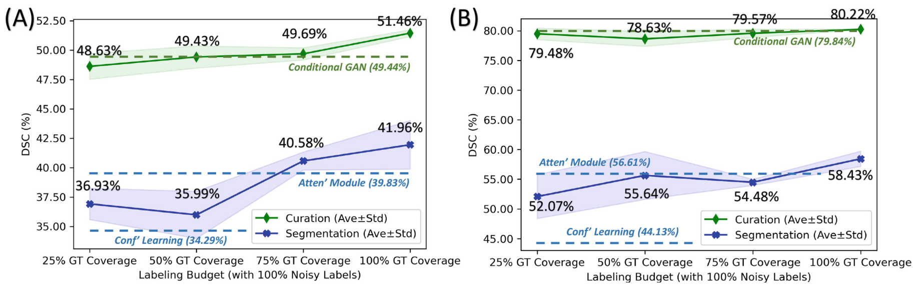

Fig. 3. Performance results in DSC (%) for label refinement and lesion segmentation accuracy on (A) dwMRI (left) and (B) BraTS 2019 (right) datasets in scenarios with incomplete ground truth label coverage, as indicated by the percentage of images paired with ground truth labels. The base DSC (%) for refined labels (green dashed line) and segmented lesions (blue dashed lines) using previous SOTAs with 100% ground truth coverage (%) are also shown for comparison. (Color figure online)

Performance with Incomplete Coverage of Ground Truth Labels. Figure 3 presents the results of the proposed multitask learning framework trained with incomplete labels. In real-world scenarios, it is common to only have a limited percentage of ground truth labels, although ideally, having more ground truth labels would result in better performance. The results show improved performance for the dwMRI and BraTS 2019 datasets with more ground truth labels, due to benefits from pixel information of images providing additional lesion-level information. There are certain fluctuations in DSC values at lower ground truth coverage, but the general trend still remains the same as DSC values increase with greater coverage of paired ground truth labels. Even in the scenario with ground truth labels being available for only 25% of the images, the DSCs for refined labels and segmented lesions still show comparable performance to the baseline DSCs from previous SOTA(s) with 100% ground truth coverage, for the two respective datasets. For lesion segmentation, when only 50–75% of the images have paired ground truth labels, the proposed framework also outperforms the baseline DSCs from previous SOTA(s) with 100% ground truth coverage, by up to 1.70–6.29%/10.35–11.51% for the two respective datasets.

Scalability on 3D Task. We further investigated the scalability of our proposed technique on a 3D task against the recent volumetric UNet baseline for 3D lesion segmentation of the tumor core (TC) in Table 3. When combined with the multitask learning strategy that simultaneously refines noisy labels and segments lesions, the DSC improves from 73.44±0.76 (%) to 76.96±0.80 (%) for lesion segmentation on BraTS 2021. The results also demonstrate that our multitask framework is a general modular framework which can be further adapted to customized 3D deep learning models for enhancement of segmentation accuracy.

Table 3. Comparison of refinement and segmentation accuracy for a 3D task. The means and standard deviations of the DSC (%) are reported. Best in **bold**.

Dataset (Networks)	Refinement (Ours) DSC↑	Segmentation (Ours) DSC↑	Segmentation (Volumetric UNet [14]) DSC↑
BraTS 2021	**88.37** ± 2.14	**76.96** ± 0.80	73.44 ± 0.76

6 Conclusion

In summary, we proposed a novel multitask framework that simultaneously refines noisy labels and segments lesions in an end-to-end manner by leveraging cross-label discrepancy minimization, and conducted extensive experiments on both proprietary and public datasets in 2D and 3D settings to demonstrate its effectiveness. Our work has implications for addressing challenges in brain lesion segmentation that arise from difficulties in acquiring high-quality labels.

Acknowledgements. Research efforts were supported by funding and infrastructure for deep learning and medical imaging research from the Institute for Infocomm Research, Science and Engineering Research Council, A*STAR, Singapore and the National Neuroscience Institute, Singapore.

References

1. Akkus, Z., Galimzianova, A., Hoogi, A., Rubin, D.L., Erickson, B.J.: Deep learning for brain MRI segmentation: state of the art and future directions. J. Digital Imaging **30**, 449–459 (2017)

2. Bertels, J., et al.: Optimizing the dice score and jaccard index for medical image segmentation: theory and practice. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11765, pp. 92–100. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32245-8_11
3. Cai, Z., Xin, J., Shi, P., Zhou, S., Wu, J., Zheng, N.: Meta pixel loss correction for medical image segmentation with noisy labels. In: Medical Image Learning with Limited and Noisy Data: First International Workshop, MILLanD 2022, Held in Conjunction with MICCAI 2022, Singapore, September 22, 2022, Proceedings. pp. 32–41. Springer (2022). https://doi.org/10.1007/978-3-031-16760-7_4
4. Cheng, G., Ji, H., Tian, Y.: Walking on two legs: learning image segmentation with noisy labels. In: Conference on Uncertainty in Artificial Intelligence, pp. 330–339. PMLR (2020)
5. Guo, R., Pagnucco, M., Song, Y.: Learning with noise: mask-guided attention model for weakly supervised nuclei segmentation. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12902, pp. 461–470. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87196-3_43
6. He, K., Zhang, X., Ren, S., Sun, J.: Identity Mappings in Deep Residual Networks. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9908, pp. 630–645. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46493-0_38
7. Huttenlocher, D.P., Klanderman, G.A., Rucklidge, W.J.: Comparing images using the hausdorff distance. IEEE Trans. Pattern Anal. Mach. Intell. **15**(9), 850–863 (1993)
8. Karimi, D., Dou, H., Warfield, S.K., Gholipour, A.: Deep learning with noisy labels: exploring techniques and remedies in medical image analysis. Med. Image Anal. **65**, 101759 (2020)
9. Kim, J., Kim, M., Kang, H., Lee, K.: U-gat-it: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. arXiv preprint arXiv:1907.10830 (2019)
10. Lee, D.H., et al.: Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: Workshop on challenges in representation learning, ICML. p. 896 (2013)
11. Lee, J., et al.: A pixel-level coarse-to-fine image segmentation labelling algorithm. Sci. Reports **12**(1), 8672 (2022)
12. Li, S., Gao, Z., He, X.: Superpixel-Guided Iterative Learning from Noisy Labels for Medical Image Segmentation. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12901, pp. 525–535. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87193-2_50
13. Menze, B.H., et al.: The multimodal brain tumor image segmentation benchmark (brats). IEEE Trans. Med. Imaging **34**(10), 1993–2024 (2014)
14. Myronenko, A.: 3D MRI Brain Tumor Segmentation Using Autoencoder Regularization. In: Crimi, A., Bakas, S., Kuijf, H., Keyvan, F., Reyes, M., van Walsum, T. (eds.) BrainLes 2018. LNCS, vol. 11384, pp. 311–320. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-11726-9_28
15. Obuchowicz, R., Oszust, M., Piorkowski, A.: Interobserver variability in quality assessment of magnetic resonance images. BMC Med. Imaging **20**, 1–10 (2020)
16. Rajpurkar, P., et al.: Deep learning for chest radiograph diagnosis: a retrospective comparison of the chexnext algorithm to practicing radiologists. PLoS Med. **15**(11), e1002686 (2018)

17. Ranjbarzadeh, R., Bagherian Kasgari, A., Jafarzadeh Ghoushchi, S., Anari, S., Naseri, M., Bendechache, M.: Brain tumor segmentation based on deep learning and an attention mechanism using MRI multi-modalities brain images. Sci. Reports **11**(1), 1–17 (2021)
18. 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
19. Warfield, S.K., Zou, K.H., Wells, W.M.: Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation. IEEE Trans. Med. Imaging **23**(7), 903–921 (2004)
20. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1492–1500 (2017)
21. Yang, Y., Wang, Z., Liu, J., Cheng, K.T., Yang, X.: Label refinement with an iterative generative adversarial network for boosting retinal vessel segmentation. arXiv preprint arXiv:1912.02589 (2019)
22. Zhang, L., et al.: Learning to Segment When Experts Disagree. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12261, pp. 179–190. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59710-8_18
23. Zhang, L., et al.: Disentangling human error from the ground truth in segmentation of medical images. In: Proceedings of the 34th International Conference on Neural Information Processing Systems, pp. 15750–15762. ACL (2020)
24. Zhang, M., et al.: Characterizing label errors: confident learning for noisy-labeled image segmentation. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12261, pp. 721–730. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59710-8_70
25. Zhen, L., Hu, P., Wang, X., Peng, D.: Deep supervised cross-modal retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10394–10403 (2019)
26. Zheng, G., Awadallah, A.H., Dumais, S.: Meta label correction for noisy label learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 11053–11061 (2021)