A Weakly-supervised Multi-lesion Segmentation Framework Based on Target-level Incomplete Annotations

Jianguo Ju[1], Shumin Ren[1], Dandan Qiu[1], Huijuan Tu[2(✉)], Juanjuan Yin[1], Pengfei Xu[1], and Ziyu Guan[1(✉)]

[1] School of Information Science and Technology, Northwest University, Xi'an, China
ziyuguan@nwu.edu.cn
[2] Department of Radiology, Kunshan Hospital of Chinese Medicine, Kunshan, China
20214132014@stu.suda.edu.cn

Abstract. Effectively segmenting Crohn's disease (CD) from computed tomography is crucial for clinical use. Given the difficulty of obtaining manual annotations, more and more researchers have begun to pay attention to weakly supervised methods. However, due to the challenges of designing weakly supervised frameworks with limited and complex medical data, most existing frameworks tend to study single-lesion diseases ignoring multi-lesion scenarios. In this paper, we propose a new local-to-global weakly supervised neural framework for effective CD segmentation. Specifically, we develop a novel weak annotation strategy called *Target-level Incomplete Annotation (TIA)*. This strategy only annotates one region on each slice as a labeled sample, which significantly relieves the burden of annotation. We observe that the classification networks can discover target regions with more details when replacing the input images with their local views. Taking this into account, we first design a TIA-based affinity cropping network to crop multiple local views with global anatomical information from the global view. Then, we leverage a local classification branch to extract more detailed features from multiple local views. Our framework utilizes a local views-based class distance loss and cross-entropy loss to optimize local and global classification branches to generate high-quality pseudo-labels that can be directly used as supervisory information for the semantic segmentation network. Experimental results show that our framework achieves an average DSC score of 47.8% on the CD71 dataset. Our code is available at https://github.com/HeyJGJu/CD_cTIA.

Keywords: weakly supervised learning · multi-lesion segmentation · local-to-global · class activation map

1 Introduction

Crohn's disease (CD) is an unexplained inflammatory bowel disease that can occur anywhere along the entire intestinal tract, has a progressive and destructive

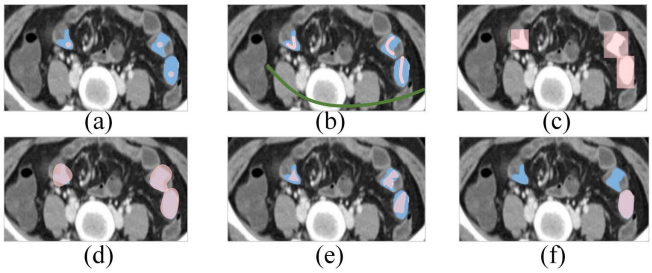

Fig. 1. Illustrations of different annotation forms. Blue and red represent ground truth and different annotation forms, respectively. (a) point annotation, (b) scribble annotation, (c) bounding box annotation, (d) (e) incomplete annotations, and (f) TIA annotation.

course and is increasing in incidence worldwide [21]. Intervening and initiating treatment for CD early can prevent further progression of the disease and improve survival rates. Computed tomography (CT) plays a crucial role in the diagnosis of CD as an auxiliary technology that can characterize the location of the lesion and its relationship to surrounding tissue. Clinically, the segmentation of CD from CT is performed manually by radiologists. Given the escalating annual incidence of CD, this process greatly augments the workload of radiologists. Therefore, many efforts have been made to automatically segment lesions [8, 3]. The current impressive performance is mainly attributed to the availability of large volumes of manually labeled datasets. However, manually labeling these datasets is a time-consuming and labor-intensive task. Recently, to reduce the cost of annotation, researchers have begun to pay attention to weakly supervised semantic segmentation methods. Weakly supervised methods utilize cost-effective weak annotations (as shown in Figure 1) as supervision signals, such as image-level annotations [5, 2], points [20, 10], scribbles [16, 25], bounding boxes [1, 9], and incomplete annotations [23]. These labels enable deep models to learn from large amounts of data with minimal human annotation effort. However, they simply annotate the lesion and cannot provide accurate lesion boundaries. Moreover, labeling each lesion in every medical image can be tedious and inconvenient, especially for multi-lesion applications (such as Crohn's disease). This remains a heavy burden for radiologists.

A natural observation inspires us that in cases of multi-lesion diseases, where a specific disease appears in multiple areas of a single slice, the texture information of each area is often similar to the others to a certain extent. This means that the sampled partial areas in all slices should have consistent texture information with the distribution of texture information in all lesion areas. Based on this observation, we innovatively develop a novel sparse annotation strategy called *Target-level incomplete annotation (TIA)* by considering the similarities between different regions of the same disease. This strategy involves annotating

only one region per slice as a labeled sample (as shown in Figure 1 (f)), which provides accurate boundaries of the target while reducing labeling costs.

We empirically find that local views input into network training can help discover more details about undiscovered semantic regions. Thus, we further propose a novel local-to-global weakly supervised multi-lesion segmentation framework based on TIA to explore the discriminability area in mining strategies by utilizing local views that are cropped from the global view according to TIA. The framework has two parallel branches *i.e.*, the global classification branch and the local classification branch. Specifically, in the global classification branch, we first adopt a multi-scale strategy to obtain global localization information to handle the variable scale of lesions in CD images. We design an affinity cropping network in the local classification branch to crop local views containing global anatomical information using TIA as a reference. When the obtained local views are fed into the classification network for training, accurately locating distinguishable regions can challenge traditional cross-entropy loss supervision. Thus, we propose a local views-based class distance loss that enhances the semantic feature distance between the target and backgrounds. Finally, we propose a new weighting strategy to reweight the results of TIA and two classification branches to generate reliable pseudo-labels. Experimental results show that our framework outperforms other state-of-the-art methods.

2 Method

In Figure 2, we delineate a new local-to-global weakly supervised multi-lesion segmentation framework based on TIA. Specifically, we design global views and local views as the input for the classification branch respectively, and jointly optimize it with local views-based class distance loss and classification loss to obtain more accurate pseudo-labels. Then, we design a joint weight assignment strategy to reweight the results of the two classification branches and TIA to generate the final pseudo-labels. We next detail the framework as follows.

2.1 Global classification branch

A 3D CT image dataset consists of 2D slice sequences for each CD case. Each case is divided into continuous slice sequences $D^{N\times H\times W} = \{D_1, D_2, ..., D_N\}$, where N, H, and W represent the number of slices scanned for a case, height, and width, respectively. These sequences are further divided into two categories, diseased slice sequences $X_o = \{D_1, D_2, ..., D_l\}$ with k ($k \geq 1$) lesions, and healthy slice sequences $X_u = \{D_{l+1}, D_{l+2}, ..., D_N\}$ with k ($k = 0$) lesions. We directly input X_o and X_u into the global classification branch for training. However, the different scales of lesions in CD images can seriously affect the feature extraction ability of the classification network. To address this issue, we incorporate a multi-scale strategy into the classification network. This network can obtain multi-scale information on the global view through simple convolutional transformation. Specifically, we add dilated convolution into the last three layers. Details of our

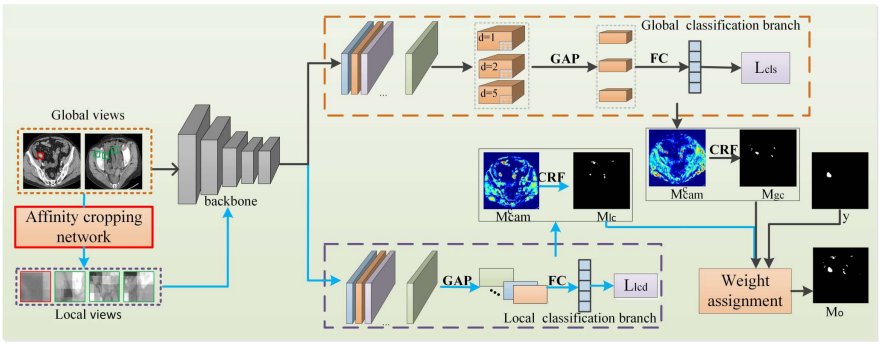

Fig. 2. Illustration of our proposed framework. Our framework contains two parallel branches *i.e.*, the global classification branch and the local classification branch, to generate initial pseudo-labels. Then, we employ a weight assignment strategy to combine TIA and the outputs from the two classification branches to generate final pseudo-labels.

network settings are reported in Section 3. We apply global average pooling (GAP) in the final convolution layer. Then, we classify its output using fully connected layers (FC). Finally, we use the weights of the FC to obtain the class activation map (CAM) for each class. Original CAMs can highlight the most prominent areas in medical images, but they still contain some non-target areas that are mislabeled pixels. Therefore, post-processing [11] is needed after getting the original CAMs to generate more reliable pseudo-labels M_{gc}.

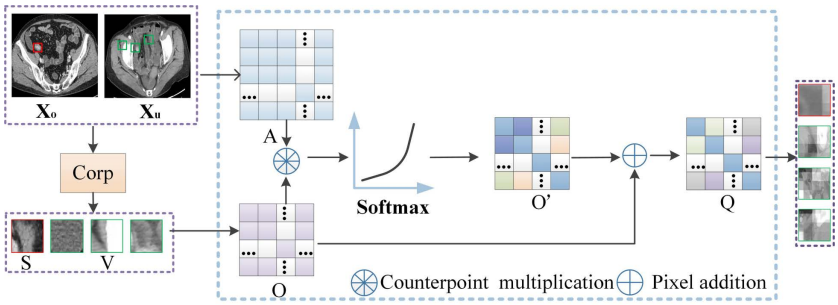

Fig. 3. Illustration of our proposed affinity cropping network.

2.2 Local classification branch

As described in the introduction, a local classification network can discover more discriminative regions by focusing on local views. Based on this observation, we

design a local classification branch to assist the global classification branch in locating more complete target regions. Random cropping [7] on a global view of medical images to obtain local views can destroy anatomical structure information inside the image or even result in no target in the local views, thus misleading the model learning. In our work, we design a novel cropping strategy, termed the affinity cropping network, which uses TIA annotation to accurately locate the target and includes global anatomical information in each local view. As shown in Figure 3, X_o, X_u, and y are the inputs of the affinity cropping network, where X_o and X_u are the global views and y is the TIA annotation of X_o. We crop X_o based on y positioning to obtain foreground local views $S = \{S_1, S_2, ..., S_l\}$. At the same time, we crop X_u based on y positioning and randomly crop multiple times around the cropped area to obtain n background local views $V = \{V_1, V_2, ..., V_n\}$. Next, we multiply each pixel in the local view with all the pixels in the corresponding global view to generate the affinity coefficient P_i of each pixel in the local view. The mathematical expression for calculating P_i is as follows:

$$P_i = -\frac{\sum_{i=1}^{n} A_i O_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} O_i^2}}, \tag{1}$$

where A is the pixel matrix for global view conversion, O is the pixel matrix for local view conversion, and i represents the pixel position index. Then, we normalize the calculated P_i to obtain the final affinity coefficient. Finally, we add the affinity coefficient to the corresponding pixels in the pixel matrix O to obtain the affinity matrix Q. Every pixel of Q contains anatomical information for global view. The mathematical expression for calculating Q is as follows: $Q = O_i + O_i^{'} = O_i + SF(P_i)(i \in 1, 2, ...n)$, where $SF(\cdot)$ represents the normalization operation and i means the pixel position index. We use the local views obtained from the affinity cropping network as input into the local classification branch. Considering the traditional cross-entropy loss l_{cls} cannot accurately distinguish background regions in medical images with unclear background differences. To minimize background interference, we design a local view-based class distance loss during the training process. We use K-means to establish the cluster center s_c, and then we develop a loss function L_{lcd} that increases the distance between foreground and background in local views. This process makes similar classes more alike and discrepancies between different classes more pronounced. The local view-based class distance loss can be expressed mathematically as:

$$L_{lcd} = \sum_{i=1}^{n} d(s_i, s_c) - \sum_{j=1}^{n} d(v_j, s_c), \tag{2}$$

where $d(\cdot, \cdot)$ represents the distance between two pixels, s_i is the foreground pixel, and v_j means the background pixel. This branch still utilizes global average pooling in the last layer, uses the connection layer to calculate probability values, and finally outputs the initial CAM. The same post-processing operation is also used to obtain reliable pixel-level annotation M_{lc}.

2.3 Joint weight assignment strategy

To take full advantage of TIA and the results of both classification branches, we propose a joint weight assignment strategy based on TIA. We believe that what is predicted simultaneously in each branch is more likely to be the target region $M_f = M_{gc} \cap M_{lc}$, so we assign higher weights, while the regions predicted $M_s = (M_{gc} - M_{lc}) + (M_{lc} - M_{gc})$ by only one branch have relatively smaller weights. Moreover, we also apply the TIA, to indicate the target region y, and its weight is set to 1. Finally, we combine these weighted regions to generate the final pseudo-labels M_o. M_o can be expressed as: $M_o = y + W_f \cdot M_f + W_s \cdot M_s$, where W_f represents the weight of M_f, and W_s represents the weight of M_s.

3 Experiments

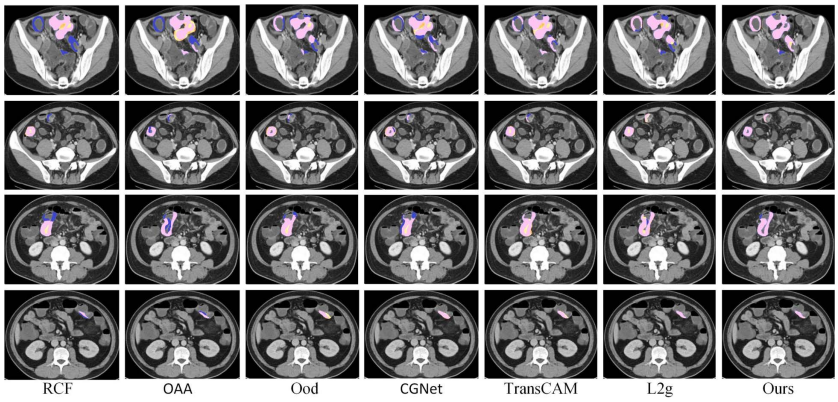

Fig. 4. Qualitative segmentation results on CD71. Blue, yellow, and pink represent the ground truth, predicted error, and the intersection of the predicted and ground truth.

3.1 Dataset and evaluation metric

We conduct experiments on the CD71 dataset to verify the effectiveness of our framework. This dataset is sourced from cooperative hospitals and contains raw CT data from 71 patients with Crohn's disease of the small intestine. We divide the dataset into a training set, a validation set, and a test set containing 51, 10, and 10 cases respectively. The TIA labels of the training set and validation set and the fine labels of the test set are annotated by experts. We evaluate the segmentation performance using the average Dice-Sørensen coefficient (DSC) [22] and mean intersection-over-union (mIoU) [18].

3.2 Implementation details

Our framework is implemented by PyTorch and performed on two NVIDIA GTX 3080 devices. We choose ResNet50 [4] as the backbone for two classification branches. We engage dilated convolution for the last three ResNet blocks in the global classification branch. The dilated rate for the last third layer is 1, for the last second layer is 2, and 5 for the last layer. The stochastic gradient descent optimizer is used to train the models with an initial learning rate of 0.0001. The hyperparameters W_f and W_s are set to 0.8 and 0.5, respectively.

Table 1. Quantitative comparisons of existing state-of-the-art methods on CD71 validation and test sets. (SUP: supervision signal, I: image-level annotation, B: bounding box, S: scribble, U: incomplete annotation, and TIA: our proposed annotation form.)

Method	Backbone	SUP.	val(DSC)	test(DSC)	val(mIOU)	test(mIOU)
RCF [17]	VGG-16	I	31.65	32.12	20.59	21.81
WSSL [19]	VGG-16	B	33.17	34.99	22.75	23.41
Scribble_Saliency [25]	VGG-16	S	36.32	36.01	25.78	25.59
OAA [6]	ResNet-101	I	35.84	36.10	24.66	25.63
Ood [13]	ResNet-101	I	37.20	37.46	27.68	27.83
CGNet [12]	ResNet-101	I	37.42	38.15	27.81	28.47
Mask R-CNN+FL [15]	ResNet-101	U	38.58	38.94	28.93	29.52
TransCAM [14]	ResNet-38	I	39.88	39.40	30.77	30.36
L2g [7]	ResNet-101	I	39.05	39.92	30.07	30.89
VOPC [2]	ResNet-50	I	40.45	41.13	30.91	31.25
DAST [24]	COPLE-Net	U	42.91	42.41	32.90	32.87
Ours	ResNet-50	TIA	**46.79**	**47.80**	**37.86**	**38.41**

3.3 Comparison experiment

To demonstrate the effectiveness and superiority of our framework, we conduct 11 sets of experiments comparing it to the image-level annotation segmentation models (**RCF** [17], **OAA** [6], **Ood** [13], **CGNet** [12], **L2g** [7], **TransCAM** [14], **VOPC** [2]), bounding box annotation segmentation model (**WSSL** [19]), scribble annotation segmentation model (**Scribble_Saliency** [25]), and incomplete annotation segmentation models (**Mask R-CNN+FL** [15], **DAST** [24]). We present our detailed comparison results in Table 1. Our proposed framework consistently outperforms other methods on two evaluation indicators and achieves a score of 47.80% DSC. We find that inaccurate boundary information in weak annotations is the main reason for poor performance in these methods, significantly affecting segmentation performance. L2g [7] also uses a dual-branch classification network to generate pseudo-labels, but the DSC score of our framework increases by 7.88%. The results show that it is necessary to exploit the

global anatomical information of medical images to assist the local classification branch locate the target area. We further visualize the prediction results of our framework and state-of-the-art methods to compare segmentation effects more intuitively. As shown in Figure 4, the single-branch method (*i.e.*, RCF, OAA, Ood, CGNet, and TransCAM) can only provide an approximate location of the target area, whereas the two-branch method (*i.e.*, L2g) locates the target area more completely. The more detailed information provided by the local classification branch has a positive effect on target positioning. Ours achieves more accurate localization and finer results than L2g, thanks to the effectiveness of our multi-scale strategy and affinity cropping network.

3.4 Ablation study

We conduct an ablation study to assess the importance of each proposed component. Table 2 reports the overall accuracies on the test set. In our experiments, G-L-M-ACL provides the best performance, mainly because our framework adds local details while solving the problem of multi-scale and similar lesions and backgrounds to obtain finer segmentation results. These experimental results demonstrate the effectiveness of each component and the rationality of this design for CD segmentation tasks.

Table 2. Ablation study on CD71 dataset. G represents the global classification branch, L represents the local classification branch, M represents the multi-scale strategy, ACL represents the affinity cropping networks and the local view-based class distance loss.

G	L	M	ACL	mIOU	DSC
				4.7	11.3
✓				8.8	17.1
	✓			9.4	19.7
✓	✓			17.5	27.3
✓		✓		17.8	27.5
	✓		✓	20.7	29.7
✓	✓		✓	27.9	37.6
✓	✓	✓	✓	**38.4**	**47.8**

4 Conclusions

In our work, we propose an effect-cost balancing annotation strategy, *i.e.*, *Target-level Incompletely Annotation (TIA)*, and verify its high efficiency. We then apply this annotation form to a weak supervised multi-lesion segmentation setting. We design a novel local-to-global weakly supervised multi-lesion segmentation framework based on TIA. Our framework leverages the knowledge from two views to generate more precise pseudo-labels. We conduct numerous experiments

to validate the effectiveness of our TIA and framework. Although we effectively exploit the observation that the distribution of a certain lesion in a single slice to fit the overall lesions distribution, there is still multi-slice prior information that is not considered, which we will explore further in the future.

Acknowledgments. This work was partially supported by National Natural Science Foundation of China under grant agreements Nos. 62073218, 62273232, 82150301, 62133012. It also partially supported by the Kunshan City Traditional Chinese Medicine (TCM) Science and Technology Development special fund (KZYY202302), the Key research and development projects of Kunshan Ministry of Science and Technology (KS1946), and the Suzhou Medical Association "Imaging Medical Star" general project (2023YX-M04).

Disclosure of Interests. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Chibane, J., Engelmann, F., Anh Tran, T., Pons-Moll, G.: Box2mask: Weakly supervised 3d semantic instance segmentation using bounding boxes. In: European Conference on Computer Vision. pp. 681–699. Springer (2022)
2. Feng, J., Wang, X., Li, T., Ji, S., Liu, W.: Weakly-supervised semantic segmentation via online pseudo-mask correcting. Pattern Recognition Letters **165**, 33–38 (2023)
3. Gao, Y., Dai, Y., Liu, F., Chen, W., Shi, L.: An anatomy-aware framework for automatic segmentation of parotid tumor from multimodal mri. Computers in Biology and Medicine **161**, 107000 (2023)
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
5. Hsieh, Y.H., Chen, G.S., Cai, S.X., Wei, T.Y., Yang, H.F., Chen, C.S.: Class-incremental continual learning for instance segmentation with image-level weak supervision. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1250–1261 (2023)
6. Jiang, P.T., Han, L.H., Hou, Q., Cheng, M.M., Wei, Y.: Online attention accumulation for weakly supervised semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(10), 7062–7077 (2021)
7. Jiang, P.T., Yang, Y., Hou, Q., Wei, Y.: L2g: A simple local-to-global knowledge transfer framework for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16886–16896 (2022)
8. Ju, J., Li, J., Chang, Z., Liang, Y., Guan, Z., Xu, P., Xie, F., Wang, H.: Incorporating multi-stage spatial visual cues and active localization offset for pancreas segmentation. Pattern Recognition Letters **170**, 85–92 (2023)
9. Kervadec, H., Dolz, J., Wang, S., Granger, E., Ayed, I.B.: Bounding boxes for weakly supervised segmentation: Global constraints get close to full supervision. In: Medical imaging with deep learning. pp. 365–381. PMLR (2020)

10. Khalid, N., Froes, T.C., Caroprese, M., Lovell, G., Trygg, J., Dengel, A., Ahmed, S.: Pace: Point annotation-based cell segmentation for efficient microscopic image analysis. In: International Conference on Artificial Neural Networks. pp. 545–557. Springer (2023)
11. Krähenbühl, P., Koltun, V.: Parameter learning and convergent inference for dense random fields. In: International Conference on Machine Learning. pp. 513–521. PMLR (2013)
12. Kweon, H., Yoon, S.H., Kim, H., Park, D., Yoon, K.J.: Unlocking the potential of ordinary classifier: Class-specific adversarial erasing framework for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6994–7003 (2021)
13. Lee, J., Oh, S.J., Yun, S., Choe, J., Kim, E., Yoon, S.: Weakly supervised semantic segmentation using out-of-distribution data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16897–16906 (2022)
14. Li, R., Mai, Z., Zhang, Z., Jang, J., Sanner, S.: Transcam: Transformer attention-based cam refinement for weakly supervised semantic segmentation. Journal of Visual Communication and Image Representation **92**, 103800 (2023)
15. Liu, K., Mokhtari, M., Li, B., Nofallah, S., May, C., Chang, O., Knezevich, S., Elmore, J., Shapiro, L.: Learning melanocytic proliferation segmentation in histopathology images from imperfect annotations. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3766–3775 (2021)
16. Liu, X., Yuan, Q., Gao, Y., He, K., Wang, S., Tang, X., Tang, J., Shen, D.: Weakly supervised segmentation of covid19 infection with scribble annotation on ct images. Pattern recognition **122**, 108341 (2022)
17. Liu, Y., Cheng, M.M., Hu, X., Wang, K., Bai, X.: Richer convolutional features for edge detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3000–3009 (2017)
18. Nowozin, S.: Optimal decisions from probabilistic models: the intersection-over-union case. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 548–555 (2014)
19. Papandreou, G., Chen, L.C., Murphy, K.P., Yuille, A.L.: Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In: Proceedings of the IEEE international conference on computer vision. pp. 1742–1750 (2015)
20. Qu, H., Wu, P., Huang, Q., Yi, J., Yan, Z., Li, K., Riedlinger, G.M., De, S., Zhang, S., Metaxas, D.N.: Weakly supervised deep nuclei segmentation using partial points annotation in histopathology images. IEEE transactions on medical imaging **39**(11), 3655–3666 (2020)
21. Roda, G., Chien Ng, S., Kotze, P.G., Argollo, M., Panaccione, R., Spinelli, A., Kaser, A., Peyrin-Biroulet, L., Danese, S.: Crohn's disease. Nature Reviews Disease Primers **6**(1), 22 (2020)
22. Setiawan, A.W.: Image segmentation metrics in skin lesion: accuracy, sensitivity, specificity, dice coefficient, jaccard index, and matthews correlation coefficient. In: 2020 International Conference on Computer Engineering, Network, and Intelligent Multimedia (CENIM). pp. 97–102. IEEE (2020)
23. Wang, S., Nie, D., Qu, L., Shao, Y., Lian, J., Wang, Q., Shen, D.: Ct male pelvic organ segmentation via hybrid loss network with incomplete annotation. IEEE transactions on medical imaging **39**(6), 2151–2162 (2020)

24. Yang, S., Wang, G., Sun, H., Luo, X., Sun, P., Li, K., Wang, Q., Zhang, S.: Learning covid-19 pneumonia lesion segmentation from imperfect annotations via divergence-aware selective training. IEEE Journal of Biomedical and Health Informatics **26**(8), 3673–3684 (2022)
25. Zhang, J., Yu, X., Li, A., Song, P., Liu, B., Dai, Y.: Weakly-supervised salient object detection via scribble annotations. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12546–12555 (2020)