Clinical-realistic Annotation for Histopathology Images with Probabilistic Semi-supervision: A Worst-case Study

Anonymous MICCAI 2022 Submission

Paper ID: 431

Abstract. Acquiring pixel-level annotation has been a major challenge for machine learning methods in medical image analysis. Such difficulty mainly comes from two sources: localization requiring high expertise, and delineation requiring tedious and time-consuming work. Existing methods of easing the annotation effort mostly focus on the latter one, the extreme of which is replacing the delineation with a single label for all cases. We postulate that under a clinical-realistic setting, such methods alone may not always be effective in reducing the annotation requirements from conventional classification/detection algorithms, because the major difficulty can come from localization, which is often neglected but can be critical in medical domain, especially for histopathology images. In this work, we performed a worst-case scenario study to identify the information loss from missing detection. To tackle the challenge, we 1) proposed a different annotation strategy to image data with different levels of disease severity, 2) combined semi- and self-supervised representation learning with probabilistic weakly supervision to make use of the proposed annotations, and 3) illustrated its effectiveness in recovering useful information under the worst-case scenario. As a shift from previous convention, it can potentially save significant time for experts' annotation for AI model development.

Keywords: Clinical-realistic Annotation, Histopathology, Probabilistic Semi-supervision, Worst-case study

1 Introduction

Supervised deep learning methods have shown state-of-the-art performances in many applications from natural images to the medical domain [9, 17, 14]. One of the major challenges of supervised methods is their reliance on the quantity and quality of the data and annotations. Using fully delineated segmentation masks for training has long been the common practice. Annotators need to first localize the object of interest from the image, then delineate its boundary. Thus the "annotation difficulty" can come from two sources: localization, and performing a specific type of annotation on the localized region.

For natural image tasks, localization is often less of a concern since our brains are trained to recognize objects from a natural scene. Therefore, the challenge

comes mostly from the process of delineating the complex shapes from the background. To reduce the burden of manual boundary drawing, several "weak" annotation techniques have been proposed, mainly to replace the "strong" full-boundary delineation with easier alternatives, including points [1], bounding boxes [5, 16, 10], and scribbles [12].

In medical domain, there are similar attempts following this strategy [13, 2, 18] to address the issue of tedious boundary delineation. However, performing accurate boundary delineation may not always be the most critical issue. Instead, the localization can pose a bigger challenge for annotators. This is because localization demands much higher expertise comparing to segmentation. In this work, we postulate that "weak" label alone may not be sufficient to make the annotation "easy" since the major difficulty can come from localization. Therefore, to accommodate for clinical-realistic annotation time, we would like to raise this issue to the attention of our society, and propose a shift from convention.

The major contributions of this work are: 1) We provide detailed analysis of the challenges, and simulated a "worst-case" scenario under clinical-realistic time constraint. 2) We proposed a more balanced annotation strategy to tackle the challenges from such time/resource limit, and to better model the difficulties in medical image annotation for AI model development. 3) To utilize the proposed annotations, we designed a semi-supervised learning strategy by using the probabilistic distribution information. With the candidate tasks of tumor detection from histopathology images, we illustrated the potential of the proposed strategies that may significantly ease the annotation burden.

2 Method

In this work, we target at the task of lesion detection from histopathology images. In clinical routine, digital scanners capture the entire sample slide as whole slide image (WSI). Pathologists examine the WSIs carefully under different resolutions, searching for the patterns of cell appearance change indicating diseases. Each individual WSI can have a dimension of $100k \times 100k$ pixels. At this scale, it can be very difficult for both disease region localization and delineation: small lesions can occupy less than 0.1% of the whole image, which is highly possible to be missed; while large lesions can occupy more than 50% of the slide, requiring tens of thousands of clicks if performing full boundary annotation. Hence, the annotation for AI model development often require much more time than what is acceptable in clinical routine [6].

To ease the difficulty of annotation for AI model development so that it can be achieved under a clinical-realistic setting, existing works proposed to use sparse point annotation [8], or diagnosis path [23] to replace full segmentation annotation. As the extreme of "weak" annotation category, a relatively common practice in histopathology domain is to learn from a single label for the entire WSI indicating the clinical findings. Most of time such problem is solved using multi-instance learning [22]. However, although a single label is the most simple from annotation input perspective, we argue that it does not directly relate to

how "easy" the annotation is, and hence may not be sufficiently helpful to ease the annotation burden.

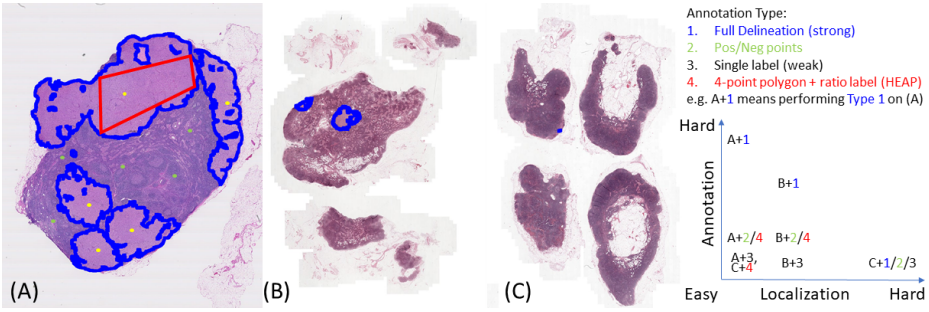

Fig. 1. Different difficulty levels in localization and performing 4 types of annotation: (A-C) give example of three common cases of tumor detection in histopathology: (A) distributed large tumor; (B) focal and medium-size tumor; (C) focal and small tumor. Different annotations illustrated in (A): single 4-point polygon (red), 10-point (5 positive yellow + 5 negative green), and full delineation (blue).

Annotation Strategy As shown in Fig. 1, the tumor region occupies 51%, 3%, and 0.01% of the foreground tissue for Case A, B, and C. For Case A, the annotation challenge mainly comes from delineation. It is easy to capture the tumor site, but takes tedious work to segment it. For Case B, both localization and segmentation are moderately easy. In contrast, for Case C, the challenge mainly comes from localization: it takes pathologist significant effort to very carefully go over the WSI to identify the tiny region that has tumor, but once it is located, it is very easy to do the delineation. Therefore in clinical practice, "weak" label does not necessarily mean less effort. According to clinical studies, diagnosis can have 3.4% false negative, and 5.5% false positive [15]. With second opinion [21], 6.6% can have a change of diagnosis. Hence, "weak" label can be both "hard" and unreliable for cases like C. One aim of this work is to find such an annotation strategy that for different cases, the efforts are always **close to origin** in the 2D coordinate system shown at the right side of Fig. 1, while in the meantime provide as much information as possible for the AI algorithm training.

To tackle the above challenges, we propose a new annotation strategy to account for the difficulty arising from both localization and delineation. Annotators are expected to do two types of annotations: 1) a **single k-point polygon** on the **major** tumor site (if there are multiple tumors) as illustrated by the red polygon in Fig. 1(A), we set $k = 4$ in this study; 2) a rough (stride of 5%) estimation of the **tumor/tissue area ratio**. For 2), annotators will provide a direct judgement for small/medium focal tumors (Fig. 1(B), around 5%), or an estimated multiplier for the polygon area v.s. the whole tumor area(e.g. $\times 4$ for Fig. 1(A)), then the ratio is calculated automatically by dividing the area with tissue area segmented by Otsu thresholding [11]. With our experiments, 5% is a

reasonable stride for this rough estimation. Then for images with different levels of tumor presence, the proposed annotation strategy is:

- Large tumors (e.g. Fig. 1(A)), provide both polygons and ratio
- Medium tumors (e.g. Fig. 1(B)), provide polygon only (if focal), or both if there are multiple tumors and only one is being annotated
- Small tumors (e.g. Fig. 1(C)), provide polygon only, but within clinical-realistic time limit. Note that in the worst-case scenario, most cases in this category will be considered "missed by annotators".

In this way, we provide decent polygon annotations for tumors of all sizes, and with area estimation for further probability modeling on large tumor regions.

Worst-case Scenario As compared with conventional annotation strategy where annotators need to take a lot of time and effort to either delineating the complex boundaries, or identifying the tiny region to provide/reject a positive label, in this work, we set the environment to be clinical realistic in that 1) large tumors are sparsely annotated, and 2) small tumors are located only if identified within time limit. Under this setting, we would like to design a training algorithm by considering both the available information from the above annotation strategy, and the potential uncertainty introduced by the missing tumor regions. To fully test the capability of the proposed method, we did our study under a worst-case scenario. In our experiment dataset of Camelyon 16 [6], out of 111 training cases, 55 are less than 1%, only 25 have tumor region greater than 10%. Given the statistics from [6], under "routine diagnostic setting", pathologists' sensitivity range from 58.9% to 71.9% with mean 62.8%. Hence, it is reasonable to simulate the worst-case scenario by considering all lesions under 1% as missing in annotation.

Semi-supervised Learning with Probabilistic Ratio Supervision We design a semi-supervised method based on the proposed polygon and ratio annotations. We leverage semi- and self-supervised learning techniques to train a base model that can have high false positive rate, then make use of weak supervision from tumor proportional ratio under a probabilistic setting to refine the model.

Fig. 2 illustrated the pipeline of the system. For Stage 1, we follow MixMatch method similar to the one proposed in [8], but with the proposed polygon annotation. We split our dataset into the three categories following the strategy proposed above: WSIs with no annotated tumors, WSIs with both polygon and ratio annotations, and WSIs with polygon annotations only. For Stage 1, we generate positive samples by random sampling from the polygon regions, and negative samples by random sampling from WSIs with no annotations. Although it is possible to get positive patches from the latter, the possibility is low and can be regarded as noise during training.

With trained network from Stage 1, the learnt representation is transferred to Stage 2 and refined by probabilistic ratio supervision. Similar information has been recorded for diagnostic purposes in some clinical protocols. And existing work proposed to utilize such information [19] for subclass identification, where it is modeled as a pseudo label generation process in a deterministic way: the patches are pre-selected and fixed across the entire training process, only the

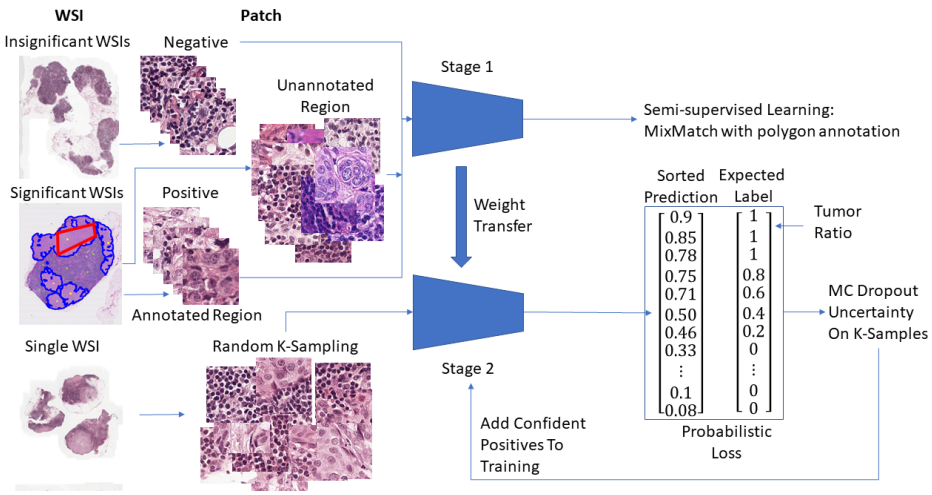

Fig. 2. Proposed semi-supervised pipeline with probabilistic ratio supervision.

pseudo labels will change according to the correct ratio. Hence, it needs a relatively good sampling strategy to begin with, and every step it need to perform inference over the entire dataset. Unfortunately for Camelyon dataset, 55/111 cases have tumor region less than 1%. It is thus neither realistic to correctly sample the potential tumor site, nor to do inference on all locations every step.

Due to this unique class imbalance issue, we propose to model the ratio information in a probabilistic manner. As shown in Fig. 2, at Stage 2, a batch x consists of a fixed number of K patches are randomly sampled from tissue region of a single WSI with rough ratio estimation $r\%$. From a probability perspective, it is expected that there will be around $K \times r\%$ cases of positive finding. Since r is a "rough" estimation, we model the uncertainty with a linear transition label. Specifically, we define a "fuzzy" range r_f around r, indicating that the positive ratio would be at least $r_{\min} = \max(0, r - r_f)$, and at most $r_{\max} = \min(r + r_f, 100)$. To reflect this fuzzy range, we generated a target prediction vector y with length K as:

$$y_i = 1 - \max(\min(i, r_{\max}) - r_{\min}, 0)/(r_{\max} - r_{\min}) \tag{1}$$

such that expected label is 1 for $i <= r_{\min}$, 0 for $i >= r_{\max}$, and linear in between. In this way, the fuzzy sampling is modeled with weaker label supervision. In our experiment for the ease of sampling, we set $K = 100$, and $r_f = 2$, as the ratios are with stride of 5%.

With CNN model f, the probabilistic ratio loss can then be designed as the binary cross entropy loss between the sorted model output $o = \text{sort}(f(x))$ and this "expected target vector" y, the sorted probabilities represent the ordered confidence of a random patch being tumor.

$$\mathcal{L}_{ratio} = BCE(\text{sort}(f(x)), y) \tag{2}$$

Pseudo Labeling with Uncertainty Estimation In order to further make use of the data with ratio annotation, we utilized the Monte Carlo dropout [7] to estimate the uncertainty of the patch predictions from the K patch samples of every WSI. Due to the significant positive/negative imbalance of Camelyon 16 data, we only keep positive samples with high confidence for "feedback training". Specifically, we keep a queue of N samples containing the confident positive patches with their corresponding uncertainty U_N. At every step, out of the sorted predictions sort$(f(x))$, we select the first $K \times r\%$, which are expected to be positive. And according to the estimated uncertainty U_R, we replace k samples from N with the ones from R whose uncertainty is higher $U_R(i) < U_N(j)$, i.e. replacing the more uncertain ones with more confident ones. The loss of these N samples will be

$$\mathcal{L}_{feedback} = BCE(f(x_N), Y_+) \tag{3}$$

where Y_+ is the vector corresponding to positive. The total loss is

$$\mathcal{L} = \mathcal{L}_{ratio} + \lambda \mathcal{L}_{feedback} \tag{4}$$

As $r >= 10$ cases already have polygon annotations being used in Stage 1, we select only $r < 10$ cases in this feedback training. Also in this study, we use $N = 100$, and $\lambda = 1$.

3 Experiments and Results

We test our method using the publicly available Camelyon 16 dataset [6]. The length of the WSIs range from 50k to 200k pixels with two types of microns/pixel: 0.226×0.226 and 0.243×0.243. The training set consists of 111 tumor (with ground truth segmentations) and 159 normal WSIs. Testing set has 129 cases, and organizer provided the ground truth for 48 of them. All results below are tested on these 48 cases as the ground truth for other cases are unavailable. (Note that this can cause some discrepancies from the metric number reported in literature. For example, we used the code and model provided by [11], and on these 48 cases, the FROC score is 0.70, while in the paper [11], the reported FROC is 0.79. Also, the cases without ground truth can not be considered normal, because the result does not match.)

In this study, we compare the proposed annotation and learning strategy to several state-of-the-art alternatives, including fully-supervised, semi-supervised, and weakly-supervised methods with different levels of annotation. The methods include: 1) fully-supervised [11] with ground truth segmentation annotations. This is the expected "upper bound" for the experiments. For this baseline, we experimented with both customized (provided by [11]) and random sampling strategy in patch selection. Also, under our "worst-case" setting where the 55 cases with small tumors are all missed by annotators, we did an experiment with the training data selected from the other 56/111 cases with the customized sampling strategy. 2) SimCLR model [3] trained on both Camelyon and Patch Camelyon dataset in a self-supervised manner, and then fine-tuned using either

56 cases with polygon annotation, or 50% of the patch dataset [20]. 3) MixMatch trained on 10-point or polygon annotations following [8]. 4) Weakly supervised model [4] using 0/1 label only, which is specifically designed and tuned on histopathology and Camelyon 16 data. All algorithms trained on Camelyon 16 data used 224×224 patches at level 0 (40x), while Patch Camelyon network used 96×96 patch at level 2 (10x). Note that these methods are one-stage, i.e. some of them are used as the base model for our Stage 2 finetuning.

We tried our best to reproduce the results by using the original code and model directly if they are available. For 1), we ran inference with the model provided with the customized sampling experiment. Further, we extracted all possible training samples with ground truth annotations at a stride of 128, and trained the model with the training code provided by the authors. For 2), we trained our model following [3], and perform inference under both common settings as other method, and the patch size and level as Patch Camelyon dataset [20]. For 3), we manually generated point annotations, unlabeled and extension sample sets, following the guidelines from [8] and train a model with the code provided by authors. For 4), we directly cite the metric number from the original paper.

Table 1. Quantitative FROC evaluation of different methods in a one-stage manner.

Method	Training Data	Annotation	Amount	Sampling	FROC
Fully-supervised [11]	Camelyon 16	Mask	Full	Customized	0.70
Fully-supervised [11]	Camelyon 16	Mask	Full	Random	0.51
Fully-supervised [11]	Camelyon 16	Mask	56/111	Customized	0.36
SimCLR [3]	Camelyon 16	Polygon	56/111	Random	0.38
SimCLR [3]	Patch Camelyon	0/1 patch	50%	Random	0.39
MixMatch [8]	Camelyon 16	10-point	56/111	Random	0.08
MixMatch [8]	Camelyon 16	Polygon	56/111	Random	0.10
Weakly-supervised [4]	Camelyon 16	0/1 WSI	Full	Random	0.31
Proposed	Camelyon 16	Polygon+ratio	56/111	Random	0.49

For ablation studies, we replaced the two components in the proposed pipeline with other alternatives. In this work, Stage 1 aims to learn a relatively good representation for histopathology images. Therefore, either self-supervised methods, e.g. SimCLR [3] with or without labeled fine-tuning, or semi-supervised method with limited annotation e.g. MixMatch [8], can be applied. We also did an experiment without Stage 1, start Stage 2 training with random initialization. For Stage 2, we choose to disable the uncertainty estimation and feedback strategy, using only the probabilistic ratio supervision. Also, we further relax the stride of ratio estimation from 5% to 10%, which allows for more estimation uncertainty.

As shown in the two tables, the proposed probabilistic ratio supervision significantly promote the performance from Stage 1, (0.38 to 0.44 for SimCLR with labeled finetuning, and 0.08/0.10 to 0.40/0.47 for MixMatch). It is also better

Table 2. Quantitative FROC evaluation of ablation configurations.

Stage 1	Annotation 1	Stage 2	Annotation 2	FROC
Skip	N/A	Probabilistic	Ratio at 5%	0.02
SimCLR [3]	N/A	Probabilistic	Ratio at 5%	0.34
SimCLR [3]	Polygon	Probabilistic	Ratio at 5%	0.44
MixMatch [8]	10-point	Probabilistic	Ratio at 5%	0.40
MixMatch [8]	Polygon	Probabilistic	Ratio at 5%	0.47
MixMatch [8]	Polygon	Probabilistic	Ratio at 10%	0.37
Proposed	Polygon	Probabilistic + feedback	Ratio at 5%	0.49

than other alternative semi- and weakly-supervised methods (0.31). Comparing the supervised method with full segmentation, it performs better than using the 25 annotated cases (0.36), and similar to using random sampling strategy on all annotations (0.51). We noticed that the performance of MixMatch with point annotations [8] seems to work a lot worse than what is presented in the original paper on different dataset. It could be due to that the original paper used customized data, which according to figures, seems to be much more balanced on the tumor/tissue ratio. This class imbalance issue is also raised in the weakly supervised paper [4].

For actual annotation time, the time of polygon + ratio is less than / comparable to the time cost of the 10-point annotations [8] at around 1 minute per case: as shown in Fig. 1(A), the 10-points are preferred to cover all tumor sites (5 points), as well as normal regions (the other 5 points); while the polygon is 4 points, but preferred to cover a large portion of the major tumor. Thus their annotation complexity is comparable. As comparison according to [6], it takes 30 hours to review the 129 testing cases for determining the presence of tumor, with an AUC 0.966. Regarding boundary delineation, although the precise time is not mentioned, there are on average 8800 vertices per WSI, indicating huge annotation effort. We performed an annotation experiment on the case shown in Fig. 1 (A), and it took us 5 hours to finish a decent job.

4 Conclusion

In order to reduce the annotation burden for medical AI model development, in this work, we proposed a shift from conventional annotate strategies where the localization cost is neglected, but can be highly difficult for medical applications. The proposed strategy take both localization and delineation into consideration. With the information annotated, we designed a semi-supervised learning method with probabilistic weak supervision and uncertainty-based feedback. Our results on Camelyon 16 dataset under worst-case clinical realistic setting showed that the proposed annotation and learning strategy achieved better performance than its semi-supervised counterparts, and is comparable to fully supervised method with random sampling strategy.

References

1. Bearman, A., Russakovsky, O., Ferrari, V., Fei-Fei, L.: What's the point: Semantic segmentation with point supervision. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) Computer Vision – ECCV 2016. pp. 549–565 (2016)
2. Cai, J., Tang, Y., Lu, L., Harrison, A.P., Yan, K., Xiao, J., Yang, L., Summers, R.M.: Accurate weakly-supervised deep lesion segmentation using large-scale clinical annotations: Slice-propagated 3d mask generation from 2d recist. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2018. pp. 396–404 (2018)
3. Ciga Ozan, Xu Tony, M.A.: Self supervised contrastive learning for digital histopathology. arXiv preprint arXiv:2011.13971 (2020)
4. Courtiol, P., Tramel, E.W., Sanselme, M., Wainrib, G.: Classification and disease localization in histopathology using only global labels: A weakly-supervised approach. CoRR **abs/1802.02212** (2018)
5. Dai, J., He, K., Sun, J.: Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). p. 1635–1643. ICCV '15, USA (2015)
6. Ehteshami Bejnordi, B., Veta, M., Johannes van Diest, P., van Ginneken, B., Karssemeijer, N., Litjens, G., van der Laak, J.A.W.M., , the CAMELYON16 Consortium: Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. JAMA **318**(22), 2199–2210 (12 2017)
7. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: Proceedings of The 33rd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 48, pp. 1050–1059 (20–22 Jun 2016)
8. Gao, Z., Puttapirat, P., Shi, J., Li, C.: Renal cell carcinoma detection and subtyping with minimal point-based annotation in whole-slide images. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. pp. 439–448 (2020)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. arXiv preprint arXiv:1512.03385 (2015)
10. Khoreva, A., Benenson, R., Hosang, J., Hein, M., Schiele, B.: Simple does it: Weakly supervised instance and semantic segmentation. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1665–1674 (July 2017)
11. Li, Y., Ping, W.: Cancer metastasis detection with neural conditional random field. In: Medical Imaging with Deep Learning (2018)
12. Lin, D., Dai, J., Jia, J., He, K., Sun, J.: Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3159–3167 (June 2016)
13. Maninis, K.., Caelles, S., Pont-Tuset, J., Van Gool, L.: Deep extreme cut: From extreme points to object segmentation. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 616–625 (June 2018)
14. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV). pp. 565–571. IEEE (2016)

15. Ng, J.C., Swain, S., Dowling, J.P., Wolfe, R., Simpson, P., Kelly, J.W.: The Impact of Partial Biopsy on Histopathologic Diagnosis of Cutaneous Melanoma: Experience of an Australian Tertiary Referral Service. Archives of Dermatology **146**(3), 234–239 (03 2010)
16. 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 2015 IEEE International Conference on Computer Vision (ICCV). p. 1742–1750 (2015)
17. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241 (2015)
18. Roth, H.R., Yang, D., Xu, Z., Wang, X., Xu, D.: Going to extremes: Weakly supervised medical image segmentation (2020)
19. Tokunaga, H., Iwana, B.K., Teramoto, Y., Yoshizawa, A., Bise, R.: Negative pseudo labeling using class proportion for semantic segmentation in pathology. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Computer Vision – ECCV 2020. pp. 430–446. Cham (2020)
20. Veeling, B.S., Linmans, J., Winkens, J., Cohen, T., Welling, M.: Rotation equivariant CNNs for digital pathology (Jun 2018)
21. Westra, W.H., Kronz, J.D., Eisele, D.W.: The impact of second opinion surgical pathology on the practice of head and neck surgery: A decade experience at a large referral hospital. Head & Neck **24**(7), 684–693 (2002)
22. Xu, Y., Zhu, J.Y., Chang, E.I.C., Lai, M., Tu, Z.: Weakly supervised histopathology cancer image segmentation and classification. Medical Image Analysis **18**(3), 591–604 (2014)
23. Zheng, Y., Jiang, Z., Zhang, H., Xie, F., Shi, J.: Tracing diagnosis paths on histopathology wsis for diagnostically relevant case recommendation. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. pp. 459–469 (2020)