Adaptive Region Selection for Active Learning in Whole Slide Image Semantic Segmentation

Jingna Qiu[1(✉)], Frauke Wilm[1,2], Mathias Öttl[2], Maja Schlereth[1], Chang Liu[2], Tobias Heimann[3], Marc Aubreville[4], and Katharina Breininger[1]

[1] Department Artificial Intelligence in Biomedical Engineering,
Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany
`jingna.qiu@fau.de`
[2] Pattern Recognition Lab, Department of Computer Science,
Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany
[3] Digital Technology and Innovation, Siemens Healthineers, Erlangen, Germany
[4] Technische Hochschule Ingolstadt, Ingolstadt, Germany

Abstract. The process of annotating histological gigapixel-sized whole slide images (WSIs) at the pixel level for the purpose of training a supervised segmentation model is time-consuming. Region-based active learning (AL) involves training the model on a limited number of annotated image regions instead of requesting annotations of the entire images. These annotation regions are iteratively selected, with the goal of optimizing model performance while minimizing the annotated area. The standard method for region selection evaluates the informativeness of all square regions of a specified size and then selects a specific quantity of the most informative regions. We find that the efficiency of this method highly depends on the choice of AL step size (i.e., the combination of region size and the number of selected regions per WSI), and a suboptimal AL step size can result in redundant annotation requests or inflated computation costs. This paper introduces a novel technique for selecting annotation regions adaptively, mitigating the reliance on this AL hyperparameter. Specifically, we dynamically determine each region by first identifying an informative area and then detecting its optimal bounding box, as opposed to selecting regions of a uniform predefined shape and size as in the standard method. We evaluate our method using the task of breast cancer metastases segmentation on the public CAMELYON16 dataset and show that it consistently achieves higher sampling efficiency than the standard method across various AL step sizes. With only 2.6% of tissue area annotated, we achieve full annotation performance and thereby substantially reduce the costs of annotating a WSI dataset. The source code is available at https://github.com/DeepMicroscopy/AdaptiveRegionSelection.

Keywords: Active learning · Region selection · Whole slide images

1   Introduction

Semantic segmentation on histological whole slide images (WSIs) allows precise detection of tumor boundaries, thereby facilitating the assessment of metastases [3] and other related analytical procedures [17]. However, pixel-level annotations of gigapixel-sized WSIs (e.g. $100,000 \times 100,000$ pixels) for training a segmentation model are difficult to acquire. For instance, in the CAMELYON16 breast cancer metastases dataset [10], 49.5% of WSIs contain metastases that are smaller than 1% of the tissue, requiring a high level of expertise and long inspection time to ensure exhaustive tumor localization; whereas other WSIs have large tumor lesions and require a substantial amount of annotation time for boundary delineation [18]. Identifying potentially informative image regions (i.e., providing useful information for model training) allows requesting the minimum amount of annotations for model optimization, and a decrease in annotated area reduces both localization and delineation workloads. The challenge is to effectively select annotation regions in order to achieve full annotation performance with the least annotated area, resulting in high sampling efficiency.

We use region-based active learning (AL) [13] to progressively identify annotation regions, based on iteratively updated segmentation models. Each region selection process consists of two steps. First, the prediction of the most recently trained segmentation model is converted to a priority map that reflects informativeness of each pixel. Existing studies on WSIs made extensive use of informativeness measures that quantify model uncertainty (e.g., least confidence [8], maximum entropy [5] and highest disagreement between a set of models [19]). The enhancement of priority maps, such as highlighting easy-to-label pixels [13], edge pixels [6] or pixels with a low estimated segmentation quality [2], is also a popular area of research. Second, on the priority map, regions are selected according to a region selection method. Prior works have rarely looked into region selection methods; the majority followed the standard approach [13] where a sliding window divides the priority map into fixed-sized square regions, the selection priority of each region is calculated as the cumulative informativeness of its constituent pixels, and a number of regions with the highest priorities are then selected. In some other works, only non-overlapping or sparsely overlapped regions were considered to be candidates [8,19]. Following that, some works used additional criteria to filter the selected regions, such as finding a representative subset [5,19]. All of these works selected square regions of a manually predefined size, disregarding the actual shape and size of informative areas.

This work focuses on region selection methods, a topic that has been largely neglected in literature until now, but which we show to have a great impact on AL sampling efficiency (i.e., the annotated area required to reach the full annotation performance). We discover that the sampling efficiency of the aforementioned standard method decreases as the AL step size (i.e., the annotated area at each AL cycle, determined by the multiplication of the region size and the number of selected regions per WSI) increases. To avoid extensive AL step size tuning, we propose an adaptive region selection method with reduced reliance on this AL hyperparameter. Specifically, our method dynamically determines an annotation
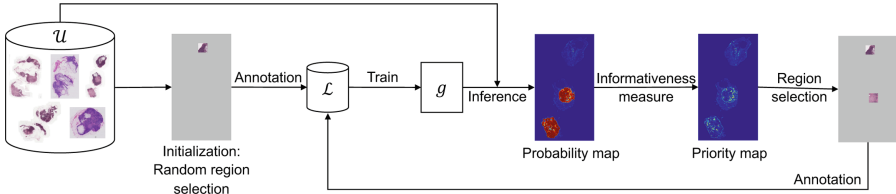
Fig. 1. Region-based AL workflow for selecting annotation regions. The exemplary selected regions are of size 8192×8192 pixels. (Image resolution: $0.25\ \frac{\mu m}{px}$)

region by first identifying an informative area with connected component detection and then detecting its bounding box. We test our method using a breast cancer metastases segmentation task on the public CAMELYON16 dataset and demonstrate that determining the selected regions individually provides greater flexibility and efficiency than selecting regions with a uniform predefined shape and size, given the variability in histological tissue structures. Results show that our method consistently outperforms the standard method by providing a higher sampling efficiency, while also being more robust to AL step size choices. Additionally, our method is especially beneficial for settings where a large AL step size is desirable due to annotator availability or computational restrictions.

2   Method

2.1   Region-Based Active Learning for WSI Annotation

We are given an unlabeled pool $\mathcal{U} = \{X_1 \ldots X_n\}$, where $X_i \in \mathbb{R}^{W_i \times H_i}$ denotes the ith WSI with width W_i and height H_i. Initially, X_i has no annotation; regions are iteratively selected from it and annotated across AL cycles. We denote the jth annotated rectangular region in X_i as $R_{ij} = (c_x^{ij}, c_y^{ij}, w^{ij}, h^{ij})$, where (c_x^{ij}, c_y^{ij}) are the center coordinates of the region and w^{ij}, h^{ij} are the width and height of that region, respectively. In the standard region selection method, where fixed-size square regions are selected, $w^{ij} = h^{ij} = l, \forall i, j$, where l is predefined.

   Figure 1 illustrates the workflow of region-based AL for WSI annotation. The goal is to iteratively select and annotate potentially informative regions from WSIs in $\mathcal{U}$ to enrich the labeled set $\mathcal{L}$ in order to effectively update the model g. To begin, k regions (each containing at least 10% of tissue) per WSI are randomly selected and annotated to generate the initial labeled set $\mathcal{L}$. The model g is then trained on $\mathcal{L}$ and predicts on $\mathcal{U}$ to select k new regions from each WSI for the new round of annotation. The newly annotated regions are added to $\mathcal{L}$ for retraining g in the next AL cycle. The train-select-annotate process is repeated until a certain performance of g or annotation budget is reached.

   The selection of k new regions from X_i is performed in two steps based on the model prediction $P_i = g(X_i)$. First, P_i is converted to a priority map M_i using a per-pixel informativeness measure. Second, k regions are selected based on M_i using a region selection method. The informativeness measure is not the focus

of this study, we therefore adopt the most commonly used one that quantifies model uncertainty (details in Sect. 3.2). Next we describe the four region selection methods evaluated in this work.

2.2   Region Selection Methods

Random. This is the baseline method where k regions of size $l \times l$ are randomly selected. Each region contains at least 10% of tissue and does not overlap with other regions. ***Standard*** [13] M_i is divided into overlapping regions of a fixed size $l \times l$ using a sliding window with a stride of 1 pixel. The selection priority of each region is calculated as the summed priority of the constituent pixels, and k regions with the highest priorities are then selected. Non-maximum suppression is used to avoid selecting overlapping regions. ***Standard (non-square)*** We implement a generalized version of the standard method that allows non-square region selections by including multiple region candidates centered at each pixel with various aspect ratios. To save computation and prevent extreme shapes, such as those with a width or height of only a few pixels, we specify a set of candidates as depicted in Fig. 2. Specifically, we define a variable region width w as spanning from $\frac{1}{2}l$ to l with a stride of 256 pixels and determine the corresponding region height as $h = \frac{l^2}{w}$. ***Adaptive*** (proposed) Our method allows for selecting regions with variable aspect ratios and sizes to accommodate histological tissue variability. The k regions are selected sequentially; when selecting the jth region R_{ij} in X_i, we first set the priorities of all pixels in previously selected regions (if any) to zero. We then find the highest priority pixel (c_x^{ij}, c_y^{ij}) on M_i; a median filter with a kernel size of 3 is applied beforehand to remove outliers. Afterwards, we create a mask on M_i with an intensity threshold of τth percentile of intensities in M_i, detect the connected component containing (c_x^{ij}, c_y^{ij}), and select its bounding box. As depicted in Fig. 3, τ is determined by performing a bisection search over $[98, 100]^{th}$ percentiles, such that the bounding box size is in range $[\frac{1}{2}l \times \frac{1}{2}l, \frac{3}{2}l \times \frac{3}{2}l]$. This size range is chosen to be comparable to the other three methods, which select regions of size l^2. Note that *Standard (non-square)* can be understood as an ablation study of the proposed method *Adaptive* to examine the effect of variable region shape by maintaining constant region size.

2.3   WSI Semantic Segmentation Framework

This section describes the breast cancer metastases segmentation task we use for evaluating the AL region selection methods. The task is performed with patch-wise classification, where the WSI is partitioned into patches, each patch is classified as to whether it contains metastases, and the results are assembled. **Training.** The patch classification model $h(\mathbf{x}, \mathbf{w}) : \mathbb{R}^{d \times d} \to [0, 1]$ takes as input a patch $\mathbf{x}$ and outputs the probability $p(y = 1 | \mathbf{x}, \mathbf{w})$ of containing metastases, where $\mathbf{w}$ denotes model parameters. Patches are extracted from the annotated regions at $40\times$ magnification ($0.25 \frac{\mu m}{px}$) with $d = 256$ pixels. Following [11], a patch is labeled as positive if the center 128×128 pixels area contains at least
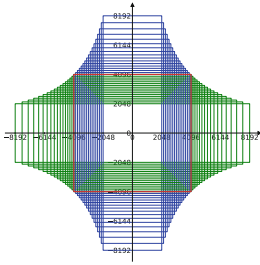
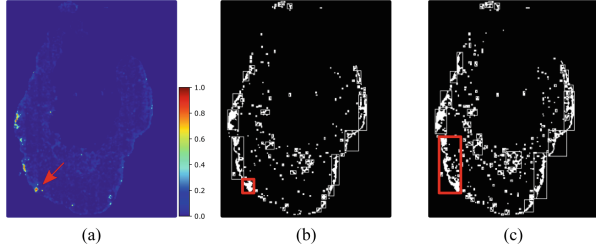**Fig. 2.** *Standard (non-square)*: Region candidates for $l = 8192$ pixels.

Fig. 3. *Adaptive*: (a) Priority map M_i and the highest priority pixel (arrow). (b–c) Bisection search of τ: (b) $\tau = 99^{th}$, (c) $\tau = 98.5^{th}$.

one metastasis pixel and negative otherwise. In each training epoch, 20 patches per WSI are extracted at random positions within the annotated area; for WSIs containing annotated metastases, positive and negative patches are extracted with equal probability. A patch with less than 1% tissue content is discarded. Data augmentation includes random flip, random rotation, and stain augmentation [12]. **Inference.** X_i is divided into a grid of uniformly spaced patches (40× magnification, $d = 256$ pixels) with a stride s. The patches are predicted using the trained patch classification model and the results are stitched to a probability map $P_i \in [0,1]^{W_i' \times H_i'}$, where each pixel represents a patch prediction. The patch extraction stride s determines the size of P_i ($W_i' = \frac{W_i}{s}, H_i' = \frac{H_i}{s}$).

3   Experiments

3.1   Dataset

We used the publicly available CAMELYON16 Challenge dataset [10], licensed under the Creative Commons CC0 license. The collection of the data was approved by the responsible ethics committee (Commissie Mensgebonden Onderzoek regio Arnhem-Nijmegen). The CAMELYON16 dataset consists of 399 Hematoxylin & Eosin (H&E)-stained WSIs of sentinel axillary lymph node sections. The training set contains 111 WSIs with and 159 WSIs without breast cancer metastases, and each WSI with metastases is accompanied by pixel-level contour annotations delineating the boundaries of the metastases. We randomly split a stratified 30% subset of the training set as the validation set for model selection. The test set contains 48 WSIs with and 80 WSIs without metastases [1].

3.2   Implementation Details

Training Schedules. We use MobileNet_v2 [15] initialized with ImageNet [14] weights as the backbone of the patch classification model. It is extended with two

---

[1] Test_114 is excluded due to non-exhaustive annotation, as stated by data provider.

fully-connected layers with sizes of 512 and 2, followed by a softmax activation layer. The model is trained for up to 500 epochs using cross-entropy loss and the Adam optimizer [7], and is stopped early if the validation loss stagnates for 100 consecutive epochs. Model selection is guided by the lowest validation loss. The learning rate is scheduled by the one cycle policy [16] with a maximum of 0.0005. The batch size is 32. We used Fastai v1 [4] for model training and testing. The running time of one AL cycle (select-train-test) on a single NVIDIA Geforce RTX3080 GPU (10GB) is around 7 h.

Active Learning Setups. Since the CAMELYON16 dataset is fully annotated, we perform AL by assuming all WSIs are unannotated and revealing the annotation of a region only after it is selected during the AL procedure. We divide the WSIs in $\mathcal{U}$ randomly into five stratified subsets of equal size and use them sequentially. In particular, regions are selected from WSIs in the first subset at the first AL cycle, from WSIs in the second subset at the second AL cycle, and so on. This is done because WSI inference is computationally expensive due to the large patch amount, reducing the number of predicted WSIs to one fifth helps to speed up AL cycles. We use an informativeness measure that prioritizes pixels with a predicted probability close to 0.5 (i.e., $M_i = 1 - 2|P_i - 0.5|$), following [9]. We annotate validation WSIs in the same way as the training WSIs via AL.

Evaluations. We use the CAMELYON16 challenge metric Free Response Operating Characteristic (FROC) score [1] to validate the segmentation framework. To evaluate the WSI segmentation performance directly, we use mean intersection over union (mIoU). For comparison, we follow [3] to use a threshold of 0.5 to generate the binary segmentation map and report mIoU (Tumor), which is the average mIoU over the 48 test WSIs with metastases. We evaluate the model trained at each AL cycle to track performance change across the AL procedure.

3.3 Results

Full Annotation Performance. To validate our segmentation framework, we first train on the fully-annotated data (average performance of five repetitions reported). With a patch extraction stride $s = 256$ pixels, our framework yields an FROC score of 0.760 that is equivalent to the Challenge top 2, and an mIoU (Tumor) of 0.749, which is higher than the most comparable method in [3] that achieved 0.741 with $s = 128$ pixels. With our framework, reducing s to 128 pixels improves both metastases identification and segmentation (FROC score: 0.779, mIoU (Tumor): 0.758). However, halving s results in a 4-fold increase in inference time. This makes an AL experiment, which involves multiple rounds of WSI inference, extremely costly. Therefore, we use $s = 256$ pixels for all following AL experiments to compromise between performance and computation costs. Because WSIs without metastases do not require pixel-level annotation, we exclude the 159 training and validation WSIs without metastases from all following AL experiments. This reduction leads to a slight decrease of full annotation performance (mIoU (Tumor) from 0.749 to 0.722).
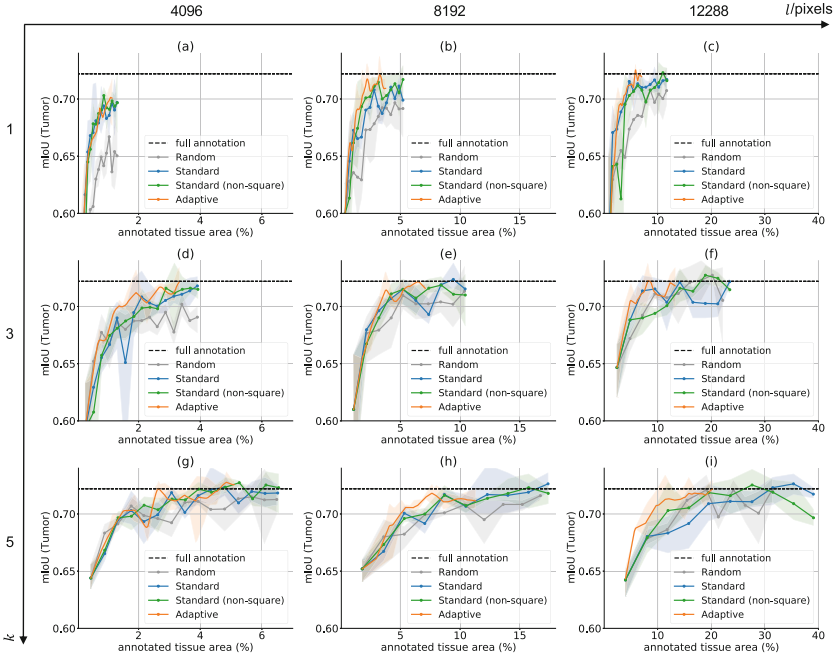
Fig. 4. mIoU (Tumor) as a function of annotated tissue area (%) for four region selection methods across various AL step sizes. Results show average and min/max (shaded) performance over three repetitions with distinct initial labeled sets. The final annotated tissue area of *Random* can be less than *Standard* as it stops sampling a WSI if no region contains more than 10% of tissue. Curves of *Adaptive* are interpolated as the annotated area differs between repetitions.

Comparison of Region Selection Methods. Figure 4 compares the sampling efficiency of the four region selection methods across various AL step sizes (i.e., the combinations of region size $l \in \{4096, 8192, 12288\}$ pixels and the number of selected regions per WSI $k \in \{1, 3, 5\}$). Experiments with large AL step sizes perform 10 AL cycles (Fig. 4 (e), (f), (h) and (i)); others perform 15 AL cycles. All experiments (except for *Random*) use uncertainty sampling.

When using region selection method *Standard*, the sampling efficiency advantage of uncertainty sampling over random sampling decreases as AL step size increases. A small AL step size minimizes the annotated tissue area for a certain high level of model performance, such as an mIoU (Tumor) of 0.7, yet requires a large number of AL cycles to achieve full annotation performance (Fig. 4 (a–d)),

Table 1. Annotated tissue area (%) required to achieve full annotation performance. The symbol "/" indicates that the full annotation performance is not achieved in the corresponding experimental setting in Fig. 4.

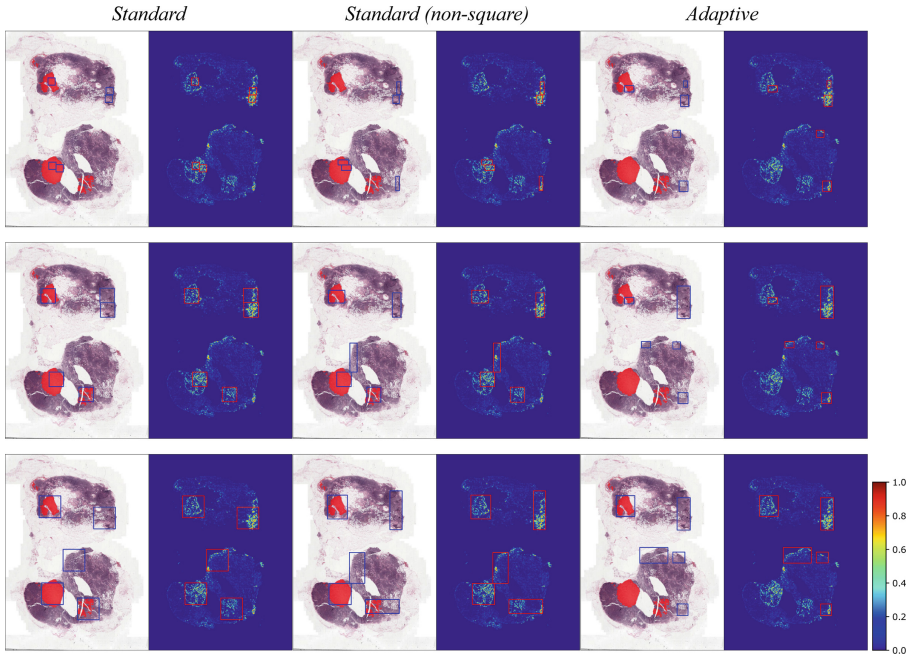| k	1			3			5		
l/pixels	4096	8192	12288	4096	8192	12288	4096	8192	12288
Random	/	/	/	/	/	18.1	/	/	/
Standard	/	/	/	/	9.4	14.2	4.3	17.4	31.5
Standard (non-square)	/	/	11.0	/	/	18.9	3.9	15.7	27.6
Adaptive	/	3.3	5.8	3.3	6.3	8.1	2.6	/	20.0

Fig. 5. Visualization of five regions selected with three region selection methods, applied to an exemplary priority map produced in a second AL cycle (regions were randomly selected in the first AL cycle, $k = 5, l = 4096$ pixels). Region sizes increase from top to bottom: $l \in \{4096, 8192, 12288\}$ pixels. Fully-annotated tumor metastases overlaid with WSI in red. (Color figure online)

resulting in high computation costs. A large AL step size allows for full annotation performance to be achieved in a small number of AL cycles, but at the expense of rapidly expanding the annotated tissue area (Fig. 4(e), (f), (h) and (i)). Enabling selected regions to have variable aspect ratios does not substantially improve the sampling efficiency, with *Standard (non-square)* outperforming *Standard* only when the AL step size is excessively large (Fig. 4(i)). However,

allowing regions to be of variable size consistently improves sampling efficiency. Table 1 shows that *Adaptive* achieves full annotation performance with fewer AL cycles than *Standard* for small AL step sizes and less annotated tissue area for large AL step sizes. As a result, when region selection method *Adaptive* is used, uncertainty sampling consistently outperforms random sampling. Furthermore, Fig. 4(e–i)) shows that *Adaptive* effectively prevents the rapid expansion of annotated tissue area as AL step size increases, demonstrating greater robustness to AL step size choices than *Standard*. This is advantageous because extensive AL step size tuning to balance the annotation and computation costs can be avoided. This behavior can also be desirable in cases where frequent interaction with annotators is not possible or to reduce computation costs, because the proposed method is more tolerant to a large AL step size.

We note in Fig. 4(h) that the full annotation performance is not achieved with *Adaptive* within 15 AL cycles; in Fig. S1 in the supplementary materials we show that allowing for oversampling of previously selected regions can be a solution to this problem. Additionally, we visualize examples of selected regions in Fig. 5 and show that *Adaptive* avoids two region selection issues of *Standard*: small, isolated informative areas are missed, and irrelevant pixels are selected due to the region shape and size restrictions.

4   Discussion and Conclusion

We presented a new AL region selection method to select annotation regions on WSIs. In contrast to the standard method that selects regions with predetermined shape and size, our method takes into account the intrinsic variability of histological tissue and dynamically determines the shape and size for each selected region. Experiments showed that it outperforms the standard method in terms of both sampling efficiency and the robustness to AL hyperparameters. Although the uncertainty map was used to demonstrate the efficacy of our approach, it can be seamlessly applied to any priority maps. A limitation of this study is that the annotation cost is estimated only based on the annotated area, while annotation effort may vary when annotating regions of equal size. Future work will involve the development of a WSI dataset with comprehensive documentation of annotation time to evaluate the proposed method and an investigation of potential combination with self-supervised learning.

References

1. Bejnordi, B.E., et al.: Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA **318**(22), 2199–2210 (2017)
2. Colling, P., Roese-Koerner, L., Gottschalk, H., Rottmann, M.: Metabox+: a new region based active learning method for semantic segmentation using priority maps. arXiv preprint arXiv:2010.01884 (2020)
3. Guo, Z., et al.: A fast and refined cancer regions segmentation framework in whole-slide breast pathological images. Sci. Rep. **9**(1), 1–10 (2019)
4. Howard, J., Gugger, S.: Fastai: a layered API for deep learning. Information **11**(2), 108 (2020)
5. Jin, X., An, H., Wang, J., Wen, K., Wu, Z.: Reducing the annotation cost of whole slide histology images using active learning. In: 2021 3rd International Conference on Image Processing and Machine Vision (IPMV), pp. 47–52 (2021)
6. Kasarla, T., Nagendar, G., Hegde, G.M., Balasubramanian, V., Jawahar, C.: Region-based active learning for efficient labeling in semantic segmentation. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1109–1117. IEEE (2019)
7. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
8. Lai, Z., Wang, C., Oliveira, L.C., Dugger, B.N., Cheung, S.C., Chuah, C.N.: Joint semi-supervised and active learning for segmentation of gigapixel pathology images with cost-effective labeling. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 591–600 (2021)
9. Lewis, D.D., Gale, W.A.: A sequential algorithm for training text classifiers. In: Croft, B.W., van Rijsbergen, C.J. (eds.) SIGIR 1994, pp. 3–12. Springer, Heidelberg (1994). https://doi.org/10.1007/978-1-4471-2099-5_1
10. Litjens, G., Bandi, P., Ehteshami Bejnordi, B., Geessink, O., Balkenhol, M., Bult, P., et al.: 1399 H&E-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset. GigaScience **7**(6), giy065 (2018). http://gigadb.org/dataset/100439
11. Liu, Y., et al.: Detecting cancer metastases on gigapixel pathology images. arXiv preprint arXiv:1703.02442 (2017)
12. Macenko, M., Niethammer, M., Marron, J.S., et al.: A method for normalizing histology slides for quantitative analysis. In: 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, pp. 1107–1110. IEEE (2009)
13. Mackowiak, R., Lenz, P., Ghori, O., et al.: Cereals-cost-effective region-based active learning for semantic segmentation. arXiv preprint arXiv:1810.09726 (2018)
14. Russakovsky, O., et al.: Imagenet large scale visual recognition challenge. Int. J. Comput. Vision **115**(3), 211–252 (2015)
15. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv 2: inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4510–4520 (2018)
16. Smith, L.N.: A disciplined approach to neural network hyper-parameters: part 1-learning rate, batch size, momentum, and weight decay. arXiv preprint arXiv:1803.09820 (2018)
17. Wilm, F., et al.: Pan-tumor canine cutaneous cancer histology (CATCH) dataset. Sci. Data **9**(1), 1–13 (2022)

18. Xu, Z., et al.: Clinical-realistic annotation for histopathology images with proba-
    bilistic semi-supervision: a worst-case study. In: Wang, L., Dou, Q., Fletcher, P.T.,
    Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13432, pp. 77–87. Springer,
    Cham (2022). https://doi.org/10.1007/978-3-031-16434-7_8
19. Yang, L., Zhang, Y., Chen, J., Zhang, S., Chen, D.Z.: Suggestive annotation: a deep
    active learning framework for biomedical image segmentation. In: Descoteaux, M.,
    Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI
    2017. LNCS, vol. 10435, pp. 399–407. Springer, Cham (2017). https://doi.org/10.
    1007/978-3-319-66179-7_46