Smart Health 32 (2024) 100474

Contents lists available at ScienceDirect

Smart Health

journal homepage: www.elsevier.com/locate/smhl

Semi-Path: An interactive semi-supervised learning framework for gigapixel pathology image analysis

Zhengfeng Lai [a,*], Joohi Chauhan [a], Dongjie Chen [a], Brittany N. Dugger [b], Sen-Ching Cheung [c], Chen-Nee Chuah [a]

[a] Department of Electrical and Computer Engineering, University of California Davis, Davis, CA 95616, USA
[b] Department of Pathology and Laboratory Medicine, University of California Davis, Sacramento, CA 95817, USA
[c] Department of Electrical and Computer Engineering, University of Kentucky, Lexington, KY 40506, USA

ARTICLE INFO

Keywords:
Semi-supervised learning
Pathology image analysis
Active learning

ABSTRACT

The efficacy of supervised deep learning in medical image analyses, particularly in pathology, is hindered by the necessity for extensive manual annotations. Annotating images at the gigapixel level manually proves to be a highly labor-intensive and time-consuming task. Semi-supervised learning (SSL) has emerged as a promising approach that leverages unlabeled data to reduce labeling efforts. In this work, we introduce Semi-Path, a practical SSL framework enhanced with active learning (AL) for gigapixel pathology tasks. Unlike existing methods that treat SSL and AL as independent components where AL incurs significant computational complexity to SSL, we propose a deep fusion of SSL and AL into a unified framework. Our framework introduces Informative Active Annotation (IAA) that employs a SSL-AL iterative structure to effectively extract knowledge from unlabeled pathology data. This structure significantly minimizes labeling efforts and computational complexity. Then, we propose Adaptive Pseudo-Labeling (APL) to address heterogeneity in class distribution, and prediction difficulty that are often observed in real-world pathology tasks. We evaluate Semi-Path on pathology image classification and segmentation tasks over three datasets that include WSIs from breast, colorectal, and brain tissues. The experimental results demonstrate the consistent superiority of Semi-Path over state-of-the-art methods.

1. Introduction

Deep learning frameworks have shown exceptional performance on diverse imaging modalities in pathology (Feng et al., 2020; Graham, Epstein, & Rajpoot, 2020; Pinckaers, Bulten, van der Laak, & Litjens, 2021). For example, ResNet-based (He, Zhang, Ren, & Sun, 2016) networks can achieve superior results in pathology classification problems (Lai et al., 2020) while FCN (Bándi, van de Loo, Intezar, Geijs, Ciompi, van Ginneken, van der Laak, & Litjens, 2017) and U-Net (Oskal, Risdal, Janssen, Undersrud, & Gulsrud, 2019) are two popular deep learning frameworks that produce excellent segmentation results of pathology images. However, the performance of these supervised deep learning methods heavily relies on a large-scale and well-curated labeled dataset (Feng et al., 2020; Graham et al., 2020; Pinckaers et al., 2021). Annotating such a dataset can be a labor-intensive and time-consuming task considering the heterogeneous nature of pathology images and their ultra-high spatial resolutions (often up to 60,000 × 50,000 pixels (Lai et al., 2021)).

* Corresponding author.
E-mail addresses: lzhengfeng@ucdavis.edu (Z. Lai), jhichauhan@ucdavis.edu (J. Chauhan), cdjchen@ucdavis.edu (D. Chen), bndugger@ucdavis.edu (B.N. Dugger), sccheung@ieee.org (S.-C. Cheung), chuah@ucdavis.edu (C.-N. Chuah).

https://doi.org/10.1016/j.smhl.2024.100474
Received 14 March 2024; Accepted 18 March 2024
Available online 26 March 2024

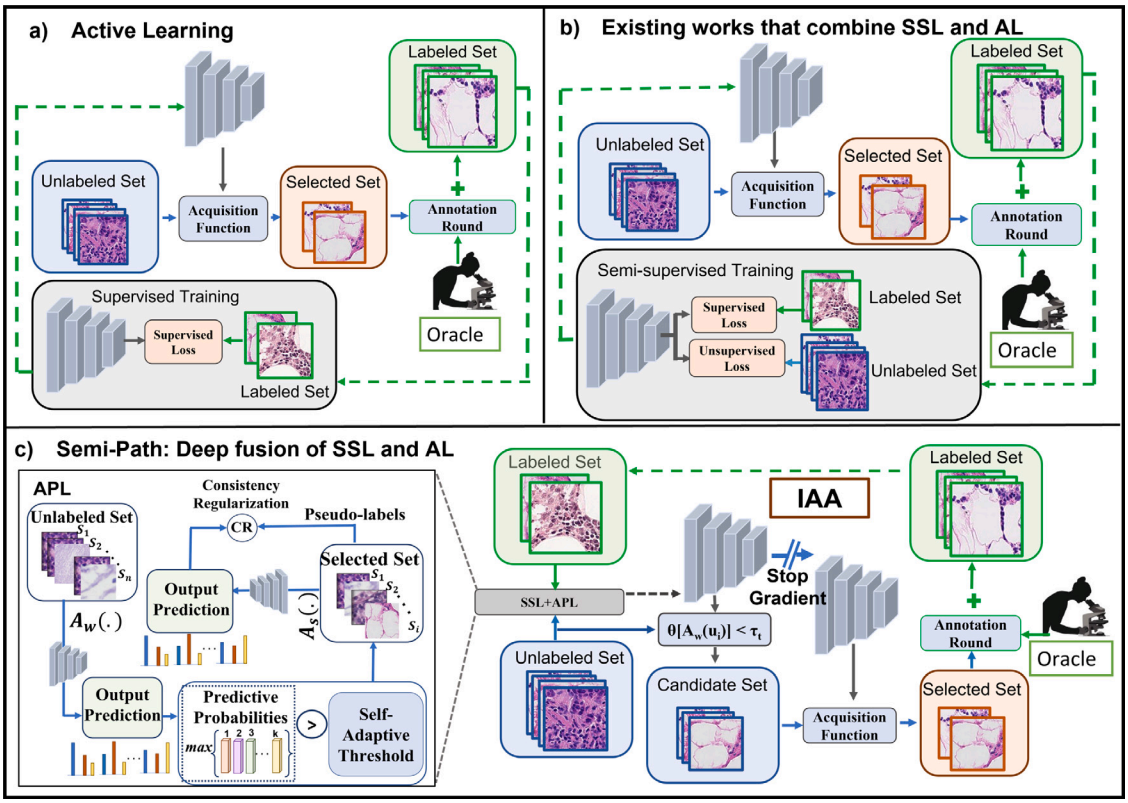

Fig. 1. The difference between Semi-Path and existing methods that combine SSL and AL: instead of treating them as independent components, we propose deep fusing them with an SSL-AL iterative structure. APL is proposed to exploit the unlabeled data during SSL training further.

Semi-supervised learning (SSL) can leverage unlabeled data to reduce the labeling efforts and has shown promising results in computer vision tasks (Chen et al., 2023; Sohn et al., 2020; Wang et al., 2023; Zhang et al., 2021). However, the efforts to deploy them on gigapixel pathology tasks remain limited (Chan, Hosseini, & Plataniotis, 2021; Pulido et al., 2020; You, Dai, Min, Staib, & Duncan, 2023; You, Dai, Min, Staib, Sekhon, & Duncan, 2023). For example, a recent work (Pulido et al., 2020) found the performance of FixMatch (Sohn et al., 2020) on a pathology dataset is unsatisfactory and sensitive to patient diversity. There are two possible reasons why current SSL methods are ineffective. The first reason is that the learning algorithms may not fully utilize unlabeled data: pathology images of ultra-high resolution can be highly similar at the pixel level (Zhang et al., 2022), making them hard to distinguish, even by trained researchers. This challenging characteristic harms the performance of the recent SSL algorithms from computer vision tasks in the early stage of the training process. Second, these SSL algorithms may miss informative samples: the samples not selected for pseudo-labeling (whose prediction confidence is below the pre-defined threshold (Sohn et al., 2020; Xie, Dai, Hovy, Luong, & Le, 2020)) may have extra information useful for learning, e.g., uncertain samples. Based on these two reasons, we propose building an interactive SSL framework to deeply and efficiently extract relevant information from unlabeled data for gigapixel pathology image tasks.

To deeply exploit the information contained in the unlabeled data, we propose Adaptive Pseudo-Labeling (APL) to dynamically select the samples of the unlabeled pool for pseudo-labeling based on the learning status of each class in the SSL training process. In addition, we incorporate active learning (AL) to select the informative samples to construct the labeled set for SSL training. As shown in Fig. 1b, existing works (Gao et al., 2020; Lai et al., 2021) that combine SSL and AL treat SSL and AL as separate and independent steps, which will incur heavy computational complexity from AL on gigapixel pathology images. Different from these methods, we deeply delve into SSL training and propose a deep fusion of SSL and AL to avoid additional computational complexity incurred by AL. To achieve such a deep fusion, we propose an SSL-AL iterative structure to exploit the relationship between SSL and AL. We show that the SSL training can serve a similar function to the inference process in AL so we adapt SSL training as a pruning selection to identify a candidate pool for AL without incurring extra computational cost. Our proposed framework provides a practical deployment of SSL in pathology image analysis. Our main contributions are summarized as follows.

- We explore the applicability of SSL on gigapixel-level pathology images and propose Semi-Path to minimize the labeling efforts.
- To effectively exploit the unlabeled data, we propose Adaptive Pseudo-Labeling (APL) to mitigate the heterogeneity issue.
- We propose Informative Active Annotation (IAA) to enhance SSL with AL by the proposed SSL-AL iterative structure.

- We show that Semi-Path achieves consistently superior performance over state-of-the-art SSL approaches across three datasets and two learning tasks.

2. Related work

Semi-supervised learning (SSL). Recent SSL methods utilize pseudo-labeling, consistency regularization, and/or a combination of them (Koohbanani, Unnikrishnan, Khurram, Krishnaswamy, & Rajpoot, 2021; Lai et al., 2021) to train deep neural networks with both labeled and unlabeled data. Pseudo-labeling methods (Arazo, Ortego, Albert, O'Connor, & McGuinness, 2020; Lee, 2013; Wang, Dong, & Voiculescu, 2022) use the intermediate models to produce pseudo-labels for some of the unlabeled data, which are then combined with the labeled set to train the model further. Consistency-based methods (Miyato, Maeda, Koyama, & Ishii, 2018; Tarvainen & Valpola, 2017) regularize the model training by encouraging consistent predictions on the same sample under various perturbations. Recent works (Chen et al., 2023; Sohn et al., 2020; Wang et al., 2023; Zhang et al., 2021) combine these two approaches and achieve better performance.

Although such methods have also been evaluated on medical domains (Wang, Li, Zheng & Huang, 2022; Wang & Ma, 2023; Wang & Voiculescu, 2023) (such as MRI and CT) and have shown promising results, their applicability on gigapixel pathology tasks has not been sufficiently explored and verified: Pulido et al. (2020) found that the performance of SSL methods can be adversely affected by patient diversity; recent works Koohbanani et al. (2021), You et al. (2022, 2024) also observed that there is still a significant performance gap between fully-supervised learning and SSL. Specifically, there are two challenges for these SSL methods: (1) poor-quality pseudo-labels may result in self-reinforcing errors during the training process (Rizve, Duarte, Rawat, & Shah, 2021); (2) insufficiency in mining the potential of unlabeled data (Huang, Jiang, Aeron, & Hughes, 2024; Lai et al., 2021).

Active learning (AL). Most AL methods choose informative samples by optimizing an acquisition function, which can be defined based on uncertainty, diversity, change in model performance, or a combination of these metrics. The uncertainty-based AL methods use *max entropy* or *max margin* (Joshi, Porikli, & Papanikolopoulos, 2009) as the selection criteria due to their simplicity and effectiveness (Budd, Robinson, & Kainz, 2021). To reduce the number of AL cycles, batch-mode AL methods (Ash, Zhang, Krishnamurthy, Langford, & Agarwal, 2019; Killamsetty, Sivasubramanian, Ramakrishnan, & Iyer, 2021; Kirsch, Van Amersfoort, & Gal, 2019; Sener & Savarese, 2018) were proposed to select a batch of data at once instead of a single data point for labeling. All these AL methods need to perform, for each AL cycle, inference on each unlabeled sample and optimization of the target acquisition function, which can be computationally intensive (Zhao, Zeng, Xu, Chen, & Guan, 2021). The computational challenge can be greatly exacerbated when applied to gigapixel pathological images. For example, it took over one hour with one Nvidia Titan Xp to make the inference on one gigapixel WSI (Lai et al., 2021).

Semi-supervised active learning (SSAL). Although SSL and AL have a similar goal of reducing labeling costs, only a few studies have attempted to combine them into a unified framework. Drugman et al. combined SSL and AL for speech understanding under limited speech data (Drugman, Pylkkonen, & Kneser, 2019). Rhee et al. combined SSL and AL for a pedestrian detection task (Rhee, Erdenee, Kyun, Ahmed, & Jin, 2017). Mahapatra, Schüffler, Tielbeek, Vos, and Buhmann (2013) combined SSL and AL for segmenting MR images. Although these methods attempt to combine SSL and AL, their selection procedures in AL cycles are independent of the model training. Specifically, SSL training and AL selection are two separate steps in the above methods. In addition, none of them have been evaluated on the gigapixel pathology settings.

3. Methodology

Before presenting our framework, we first introduce the basic notions and terminologies. In semi-supervision settings, the training set includes a labeled data pool $\mathcal{X}$ with the corresponding labels $\mathcal{Y}$ and an unlabeled data pool $\mathcal{U}$. The goal is to use both $\mathcal{X}$ and $\mathcal{U}$ to train the model in a semi-supervised manner. We denote the predictive probability from the model as $h(x;\theta)$, where θ refers to the model. We use N_l as the number of images in the labeled pool $\mathcal{X}$. Given labeled data $\mathcal{X} = \{x_i|_{i=1}^{N_l}\}$ and their labels $\mathcal{Y} = \{\mathbf{y}_i|_{i=1}^{N_l}\}$, we apply weak augmentations denoted as $A_w(.)$ (using only flip-and-shift) and minimize the cross-entropy loss through supervised label information as $\mathcal{L}(x,\mathbf{y},\theta) = D_{ce}(\mathbf{y}, h(A_w(x);\theta))$, where $D_{ce}(\cdot,\cdot)$ denotes the cross-entropy between the predictive output and the label. Similar to many recent state-of-the-art SSL algorithms (Chen et al., 2023; Sohn et al., 2020; Wang et al., 2023), we have supervised and unsupervised loss together to train the model θ via the following optimization:

$$\min_{\theta\in\Theta} \Omega(\mathcal{U},\theta) + \frac{\alpha}{N_l}\sum_{\substack{x\in\mathcal{X}\\ \mathbf{y}\in\mathcal{Y}}} \mathcal{L}(x,\mathbf{y},\theta), \tag{1}$$

where α is a hyperparameter that balances the supervised and consistency loss, and Ω refers to the unsupervised loss on the unlabeled data (see in the next section).

3.1. Adaptive Pseudo-Labeling (APL)

The pseudo-label of an unlabeled image is generated based on the model's prediction on the weakly-augmented version of the image when its confidence exceeds a pre-defined threshold (Sohn et al., 2020; Wang & Voiculescu, 2023). The success of the pseudo-labeling with a fixed threshold relies on a necessary assumption that labeled and unlabeled data have similar distributions (Lai, Wang, Gunawan, Cheung, & Chuah, 2022; Xu et al., 2021). This assumption often does not hold in real-life applications, especially in the pathology domain (Pulido et al., 2020), and thus it could hurt the original model's performance (Oliver, Odena, Raffel,

Cubuk, & Goodfellow, 2018). As a result, if the model generates the wrong pseudo-labels due to the domain and distribution shift, and utilizes them as the supervision, the model will be biased and have the prediction error reinforced during the self-training process (Zou & Caragea, 2023; Zou, Zhou, Zhang, & Caragea, 2023).

In this work, we propose an adaptive threshold τ_t based on learning status and learning difficulty to address mismatched distributions in real-life applications. The criteria for selecting an unlabeled sample x for pseudo-labeling is as follows (k refers to the class index):

$$\max_k h(A_w(x);\theta)_k > \tau_t \tag{2}$$

The pseudo-label is then defined as a one-hot vector $\mathbf{p}(x) = \arg\max_k h(A_w(x);\theta)_k$ if x satisfies the above criterion. The adaptive selection is critical in the semi-supervised settings: if τ_t is too high, we may miss some informative samples from the unlabeled data for the supervision; if τ_t is too low, we may incorporate many low-quality pseudo-labels that may hurt the model's performance. On the other hand, the learning status is also different in each training iteration. Hence we propose to design a dynamic threshold for each class based on its learning effect.

A recent SSL work (Zhang et al., 2021) argued the learning status of the SSL model could be reflected by the number of unlabeled samples with predictions above the threshold. Based on this observation, we design τ_t per epoch as:

$$\tau_t = \begin{cases} \tau_{t-1} \cdot \min\{1, N_u^t / N_u^{t-1}\}, & \text{if } 1 < t \le T \\ \tau_1, & \text{otherwise} \end{cases} \tag{3}$$

where t indicates the epoch number and $N_u^t = \sum_{x \in \mathcal{U}} \mathbb{1}(\max_k h(A_w(x);\theta)_k > \tau_{t-1})$, with $\mathbb{1}(\cdot)$ being the indicator function, is to quantify the number of pseudo-labels from last epoch. T is the number of epoch in SSL training. We follow FixMatch (Sohn et al., 2020) to set the initial τ_1 as 0.95: when there are more samples selected out for pseudo-labeling, τ_t will remain at a high level to reserve the most confident samples; otherwise, τ_t can be lower to incorporate more samples for pseudo-labeling to encourage the better utilization of unlabeled data until it reaches the maximum iterations. This dynamic mechanism can adjust the potential bias from the distribution shift: a potential distribution shift may result in the model's over-confidence in the majority class and thus the pseudo-labels on the minority classes can be under-estimated. With the class-aware self-adaptive pseudo-labeling, the model will balance its learning status and generate less biased pseudo-labels.

We now define the unsupervised loss in (1). It is the augmentation-aware consistency loss (Sohn et al., 2020), where the model is trained to enforce the consistency between the predictions on the strongly-augmented unlabeled image $A_s(x)$ and the pseudo-labels $\mathbf{p}(x)$ derived from the weakly-augmented version via a cross-entropy loss:

$$\Omega(\mathcal{U}, \theta) = \frac{1}{N_u^t} \sum_{x \in \mathcal{U}} \mathbb{1}(\max_k h(A_w(x);\theta)_k > \tau_{t-1}) \cdot D_{ce}(\mathbf{p}(x), h(A_s(x);\theta)).$$

3.2. Informative Active Annotation (IAA): Deep fusion of SSL and AL

Existing SSAL methods (Gao et al., 2020; Lai et al., 2021; Zhao et al., 2021) utilized AL as an independent step and incurred extra computational cost. This makes it difficult to be deployed widely in pathology domains for two reasons: first, the computational burden can be extremely heavy when the acquisition function is complicated, e.g., non-linear optimization objective (Choi, Elezi, Lee, Farabet, & Alvarez, 2021; Gudovskiy, Hodgkinson, Yamaguchi, & Tsukizawa, 2020; Kim, Park, Kim, & Chun, 2021); second, even with computationally-efficient acquisition functions, the gigapixel size of WSIs still poses a formidable computational challenge (Jahanifar, Tajeddin, Koohbanani, & Rajpoot, 2021) to make AL inference on them. To deal with these challenges, we propose a novel method called Informative Active Annotation (IAA) to deeply fuse SSL and AL. Fig. 2 illustrates the overall framework of our proposed IAA module. We design a SSL-AL iterative structure specifically to fuse SSL and AL to construct an informative labeled set without incurring additional heavy computational costs.

3.2.1. SSL training as a construction of candidate pool

Since the inference stage for the entire unlabeled set in AL involves extensive computations, especially in the gigapixel-resolution pathology domain, it is essential to explore the potential methods for reducing the dataset size for such inference.

The key objective for AL is to determine samples with high uncertainty. In SSL, we can identify uncertain samples if we are unable to assign pseudo-labels to them. In other words, the samples with a maximum predictive probability below the threshold τ_1, which is self-adapted in the last epoch of this round of SSL training, will remain in the candidate pool $\mathcal{C}_r$. We define $\mathcal{C}_{r-1}$ as the candidate pool from the last round of IAA with $\mathcal{C}_0 = \mathcal{U}$ as the initial candidate pool. For each round of IAA, we can get

$$\mathcal{C}_r = \{x \in \mathcal{C}_{r-1} : \max_k h(A_w(x);\theta)_k < \tau_1\}. \tag{4}$$

Consequently, the pseudo-labeling process employed in SSL training can serve as an acquisition function in AL. The most commonly used acquisition function is the entropy of the predictive probability (Gal, Islam, & Ghahramani, 2017; Myers, 1974). As such, it is important to show that uncertain samples selected with entropy will not be pruned away during SSL training selection using pseudo-labeling. Focusing on binary classification, we prove that minimizing the function used in SSL training is equivalent to maximizing the entropy thus our SSL pruning will not prune uncertain and informative samples.

Theorem 3.1. *In binary classification problems, minimizing confidence score $G = max_k(h(x;\theta))$ is equivalent to maximizing entropy among all the samples.*

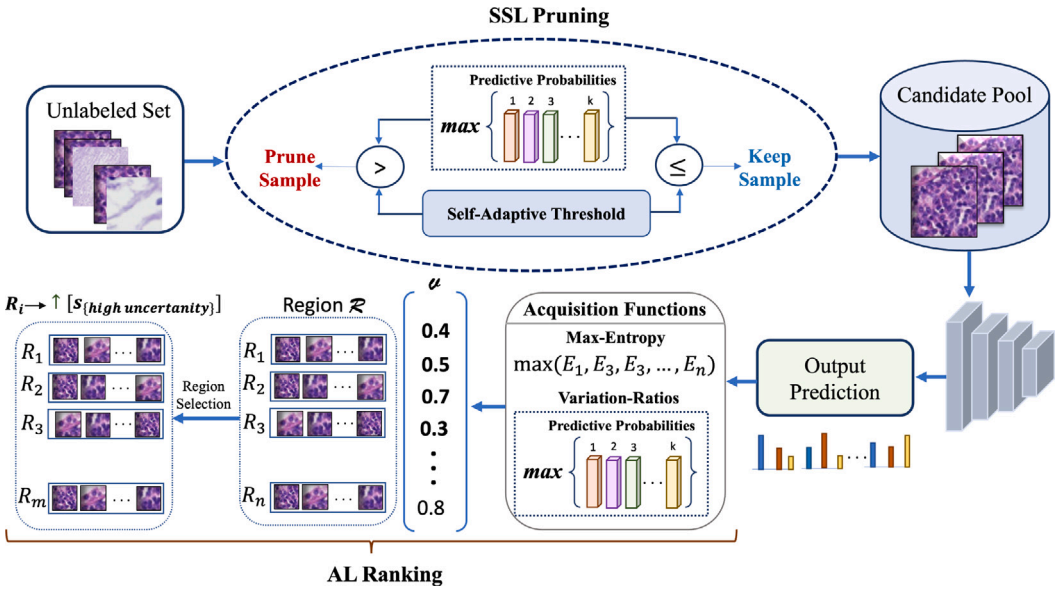

Fig. 2. Architecture of the proposed Informative Active Annotation (IAA) module with a SSL-AL iterative structure. SSL training performs a pruning of the candidate pool for the AL, then AL Ranking utilizes AL's acquisition function to select the samples for labeling.

Proof. Suppose we have two different classes, A and B. The binary entropy function with probability p for class A is given by $H(p) = -p\log(p) - (1-p)\log(1-p)$, where we define $0\log 0 = 0$. Take the derivative of $H(p)$, we get

$$H'(p) = -\log(\frac{p}{1-p}). \tag{5}$$

In addition, the entropy function shows its symmetry, i.e., $H(p) = H(1-p)$. Therefore, $H(p)$ is a monotone decreasing function for $p \in [\frac{1}{2}, 1]$. Moreover, we can obtain a necessary and sufficient condition of the monotony property for $H(p)$: given some samples in the binary classification, the probabilities are given as $\{(p_i, 1-p_i)\}$, the values of entropy functions satisfy $H(p_1) \le H(p_2) \le \cdots \le H(p_n)$ if only if $|\frac{1}{2} - p_1| \le |\frac{1}{2} - p_2| \le \cdots \le |\frac{1}{2} - p_n|$. On the other hand, we try to minimize $G(p) = \max\{p, 1-p\}$. The subgradient of G can be calculated as

$$\nabla G(p) = \begin{cases} -1, & p < \frac{1}{2} \\ 1, & p > \frac{1}{2} \\ \text{Any value between -1 and 1}, & p = \frac{1}{2}. \end{cases} \tag{6}$$

Moreover, $G(p) = G(1-p)$. Hence $H(p)$ and $-G(p)$ have the same monotony and symmetry. That is to say, given some samples in the binary classification, if the probabilities are given as $\{(p_i, 1-p_i)\}$, maximizing $H(p_i)$ among all these samples is equivalent to minimizing $G(p_i), \forall i \in \mathbb{P}$. □

When we have more classes, both H and G are bounded functions so the gap is always finite. Therefore, we set a large enough τ, e.g., 0.95 to make sure the optimal samples selected by G (with $G \ge \tau$) will not be missed. Therefore, we conclude that the SSL pruning will not discard the uncertain and informative samples and serves a similar purpose as the uncertainty-based acquisition functions in binary classification tasks. In fact, it should be pointed out that others have used the maximum of the predictive probability as AL's acquisition function and obtained compatible results with entropy (Gal et al., 2017).

Overall, our proposed SSL module segregates the confident samples and reconstructs the candidate pool of samples not selected for pseudo-labeling. Then the AL inference can be performed solely on this pruned candidate pool, saving computation time and resources by avoiding inference on the entire unlabeled set.

3.2.2. AL ranking for selection

We have shown that the SSL training process can serve as an AL inference. After the pruning of the candidate pool, we step into the AL Ranking of IAA, as shown in Fig. 2. Acquisition functions from AL are introduced in this step to rank samples to fit the labeling budget. If the labeling budget is larger than the size of the candidate pool, we will skip the AL Ranking selection but annotate all samples in the pool with the labeling budget. If the labeling budget is smaller than the candidate pool, we will enter the AL selection: for each cycle, we annotate a subset of the pool until we reach the labeling effort. First, we calculate the score for each sample in the candidate pool and save them in a vector $v \in \mathbb{R}^n$, where n is the size of the candidate pool. Based on the labeling budget, we then select m samples with the lowest confidence among entries of v. However, it is important to define an AL selection

criterion for selecting the most informative samples. One of the widely used selection criteria is based on the degree of uncertainty the current model has on each unlabeled sample.

The gigapixel size of pathology images introduces further challenges in using the traditional AL methods. AL process itself requires high-end computational resources to finish the task in an adequate time frame and when these methods are used for high-resolution pathology images, it may add more complexity to the model and time delays to the pathology image analysis task. In addition, the slow training due to high computational complexity may lead to additional delays in synchronizing the annotator availability after each AL iteration. Considering the gigapixel resolution of pathology images, we evaluate two computationally efficient yet effective methods in IAA: *Max-Entropy* (Gal et al., 2017) and *Variation-Ratios* (Gal et al., 2017; Myers, 1974). For *Max-Entropy* (Gal et al., 2017), samples with high entropy are considered informative and selected for labeling as higher entropy indicates higher uncertainty (Peng, Wang, Liu, & Yang, 2021). For *Variation-Ratios* (Gal et al., 2017; Myers, 1974), we select the highest probability $\max_k\{h(x;\theta)_k\}$ as the acquisition function in each patch's soft label. This value can describe the degree of confidence and uncertainty.

Algorithm 1 Semi-Path with the region-based selection for pathology image segmentation

Input: training dataset $\mathcal{D}$, the number of AL cycles T, labeling budget for each cycle m
Pre-training: obtain f via SimCLR (Chen, Kornblith, Norouzi, & Hinton, 2020) among $\mathcal{D}$
Initialize: randomly select two regions for labeling and tile them into patches to formulate the labeled set $\mathcal{X}_0$
Set an unlabeled data by $\mathcal{U}_0 = \mathcal{D} \setminus \mathcal{X}_0$
Fine-tuning: $\theta_0 = \arg\min_{\theta\in\Theta} L(\mathcal{X}_0, \theta) + \Omega(D, \theta)$
for $t = 0$ to T **do**
  Update $\theta : \theta_{t+1} = \arg\min_{\theta\in\Theta} L(\mathcal{X}_t, \theta) + \Omega(D, \theta)$
  SSL pruning: Record coordinates of patches that are not selected for pseudo labeling
  AL ranking: apply acquisition functions on each patch and save in v, then expand v into region list $\mathcal{R}$
  AL selection: Select the regions $\{\mathcal{R}_i\}_{i=1}^m$
  Split $\{\mathcal{R}_i\}_{i=1}^m$ into a group of patches $\mathcal{W}_s$
  $\mathcal{X}_{t+1} = \mathcal{X}_t \cup (\mathcal{W}_s \times J(\mathcal{W}_s))$
  $\mathcal{U}_{t+1} = \mathcal{U}_t \setminus \mathcal{W}_s$
end for

3.3. Extension to gigapixel-level segmentation tasks

As pathology images are typically at the gigapixel level, annotating the entire WSIs can be more labor-intensive and time-consuming for the segmentation tasks. Here, we show the capability of extending Semi-Path to weakly-supervised pathology image segmentation tasks: instead of requesting pixel-level annotation of an entire WSI, we focus on seeking the most informative regions within the WSI for the annotation. In Lai et al. (2021), they verified a small set of "difficult" regions may be sufficient to improve the model's performance since the visual representations can be similar in gigapixel slides. Therefore, we aim to seek those informative regions to further reduce the labeling efforts by avoiding labeling the whole slide in a weakly-supervised manner.

To enhance the robustness of the uncertainty criterion, we follow Lai et al. (2021) to integrate a region-based selection criterion. Instead of focusing on only one sample at a time, it focuses on a region consisting of a batch of neighboring samples that capture more neighboring information. The regions where most samples have high uncertainty would be selected for annotation queries. Here a region $\mathcal{R} \in \mathbb{R}^{nd\times nd}$ can be tiled into n^2 patches. With the region-based selection, the labeled data set can be quickly expanded as

$$\mathcal{X} = \mathcal{X} \cup (\mathcal{W} \times J(\mathcal{W})), \tag{7}$$

where $\mathcal{W}$ represents a set of patches from $\mathcal{R}$. $J(x)$ is the assigned label for x, $J(\mathcal{W})$ is a set of labels $\{J(x)\}_{x\in\mathcal{W}}$.

Similar to the classification task, the SSL training process itself constructs the candidate pool by collecting the unlabeled patches that fail to be selected for pseudo-labeling. Additionally, we are recording the coordinates of these unlabeled patches. Based on these coordinates, we tile larger regions centered on those patches as the candidate pool. The inference will be processed on the patches contained in these regions. We get a patch-level metric v based on the AL ranking criterion. By taking the mean value of the corresponding entries for each region, v will be transformed into a region-based metric $\hat{v}$. Then, we select the corresponding regions the same way as the one in the classification. The whole process is summarized in Algorithm 1.

4. Experiments

4.1. Datasets and training setup

To show the generalizability of our framework under various pathology image settings, we tested Semi-Path on two commonly-used staining techniques (H&E and Amyloid-β), three tissue regions (breast, colorectal, and brain cortex), collected by three scanners (Pannoramic, NanoZoomer-XR, and Leica Aperio AT2), and three scanning magnifications (10X, 20X, and 40X). The details can be found in Table 1.

Tumor classification task. PCam (Veeling, Linmans, Winkens, Cohen, & Welling, 2018), aiming to classify the existence of a tumor, is a patch-level dataset derived from Camelyon16 (Bejnordi et al., 2017). PCam is collected from patients with breast cancer

Table 1
Summary of the heterogeneity of three datasets.

Dataset	Staining	Tissue type	Scan magnification	Scanner	Slides	Annotators	Task	Disease focus
PCam (Veeling et al., 2018)	H&E	Breast	10X	Pannoramic, NanoZoomer-XR	399	3–4	Classification	Tumor
MHIST (Wei et al., 2021)	H&E	Colorectal	40X	Leica Aperio AT2	328	7	Classification	Polyps
GM/WM (Lai et al., 2021)	Amyloid-β	Brain	20X	Leica Aperio AT2	30	2	Segmentation	Alzheimer's disease

metastasis in the lymph nodes. It contains 327,680 patches at the size of 96 × 96 pixels from 399 Hematoxylin and Eosin (**H&E**) stained Formalin-Fixed Paraffin-Embedded (FFPE) slides and scanned at 10X magnification. We followed the original split to have 262,144 images (75%) in the training set, 32,768 (12.5%) in the validation set, and 32,768 (12.5%) in the hold-out test set (Veeling et al., 2018).

Colorectal polyp classification task. MHIST (Wei et al., 2021) is a clinically-important binary classification between sessile serrated adenomas (SSA) and hyperplastic polyps (HP), which is a challenging problem facing considerable inter-pathologist variability (Abdeljawad et al., 2015). It contains 3152 patches at 224 × 224 pixels extracted from 328 **H&E** stained slides of colorectal polyps. These slides were collected by the Department of Pathology and Laboratory Medicine at Dartmouth-Hitchcock Medical Center (DHMC) and were scanned by Aperio AT2 scanner at 40X magnification. Each image in this dataset has a gold-standard label determined by the majority voting among seven board-certified gastrointestinal pathologists (Wei et al., 2021). We followed the setting in Wei et al. (2021) to split 2175 images into the training phase (training/validation sets) while 977 in the hold-out test set.

GM/WM segmentation task. We also evaluated our framework on a GM/WM segmentation dataset (Lai et al., 2021). It has 30 **Amyloid**-β stained slides (4G8, recognizing residues 17–24, dilution 1:1600, BioLegend catalog number SIG-39200) from FFPE temporal cortex tissue. The slides were digitized by Aperio AT2 scanner at 20X magnification. The average size per slide is around 60,000 × 50,000 pixels. The annotations were provided by a trained researcher and a neuropathologist. In terms of the variety of this dataset, we have 20 WSIs from Alzheimer's disease cases (AD) and 10 from non-Alzheimer's disease cases (NAD). AD slides can contain thousands of plaques in GM (Tang et al., 2019), which differs from the NAD cases. All of these 30 WSIs have been de-identified, lacking personal health information. We followed a previous work (Lai et al., 2021) to split 20 slides into the training phase (training/validation sets) while 10 (6 AD and 4 NAD) in the hold-out test set.

Training details. We use ResNet-18 (He et al., 2016) as the backbone encoder for all experiments. However, our proposed framework also works with Transformer architectures in a plug-and-play setting. We select the current state-of-the-art SSL algorithm, FixMatch (Sohn et al., 2020), as our framework's main baseline and example. We follow the hyper-parameter settings in Sohn et al. (2020) and keep them consistent throughout all of our experiments: the initial confidence threshold τ is set as 0.95, the unlabeled loss weight λ_u as 1, and the ratio of unlabeled data in each mini-batch μ as 2. The hyper-parameters are tuned based on the validation set in the above three learning tasks. We set the batch size as 32 for both labeled and unlabeled data. We use Adam (Kingma & Ba, 2015) with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ as our optimizer in all settings. The learning rate is set as 0.001. For comparison, all computations are conducted on a single GPU (NVIDIA Titan Xp with 12 GB of VRAM).

4.2. Main results on pathology image classification and segmentation tasks

We first evaluate Semi-Path on two pathology classification tasks: tumor classification in breast (PCam (Veeling et al., 2018)) and colorectal polyp classification (MHIST (Wei et al., 2021)). We refer to supervised ResNet-18 (He et al., 2016) as SL. We implement recent popular SSL algorithms as baselines for comparison: Pseudo-Label (Lee, 2013), Mean Teacher (Tarvainen & Valpola, 2017), MixMatch (Berthelot et al., 2019), FixMatch (Sohn et al., 2020), Dash (Xu et al., 2021), and FlexMatch (Zhang et al., 2021). We also include a self-supervised learning algorithm, SimCLR (Chen et al., 2020), to compare with our proposed Semi-Path. SimCLR is pre-trained in a self-supervised manner and then fine-tuned on a small labeled set. We include SimCLR as another label-efficient method for comparison. We select Accuracy, Precision, Recall, F1-score, and AUC as the measuring metrics (TP: True Positive, TN: True Negative, FP: False Positive, FN: False Negative).

Tumor classification. We first set the total labeling budget of PCam (Veeling et al., 2018) as 1000 images (around 0.5% of the entire training set); thus the rest of the images are put into the unlabeled set. To reach the labeling budget, we set the number of IAA cycles as 2. The results are summarized in Table 2. We can find the performance of SL is sensitive to the amount of labeled data: when we use all data for supervised training, it can achieve competitive results; however, the performance degrades significantly when we only use 0.5% labeled data. Then, we fine-tune the self-supervised pre-trained model (Chen et al., 2020) and find the performance is still far from satisfactory due to the large gap between the new pathology task and the pre-trained domain. For SSL, the methods (Sohn et al., 2020; Xu et al., 2021; Zhang et al., 2021) that combine pseudo-labeling and consistency regularization can bring more benefits than SL. Semi-Path can consistently outperform the recent SSL algorithms (Berthelot et al., 2019; Lee, 2013; Sohn et al., 2020; Tarvainen & Valpola, 2017; Xu et al., 2021; Zhang et al., 2021), with 8% improvement over FixMatch (Sohn et al., 2020) and 7.2% over FlexMatch (Zhang et al., 2021) in terms of accuracy. We observed similar results when we set the labeling budget as 1.5%. This shows a better utilization of the unlabeled data by Semi-Path.

Colorectal polyp classification. We have discussed the experiments on a large-scale dataset (PCam (Veeling et al., 2018)) above. To further stress-test the applicability of Semi-Path, we select (MHIST (Wei et al., 2021)) as a more challenging learning task that suffers from inter-rater disagreement. This dataset only consists of 3152 patches but involves 328 slides. Hence the variation of

Table 2
Quantitative comparison on PCam. All scores refer to the macro-averaged values.

Learning manner	Labeled ratio	Algorithm	Accuracy	Precision	Recall	F1-score	AUC
Supervised	100% (Full)	ResNet-18 (He et al., 2016)	92.8 ± 0.30	92.9 ± 0.20	92.6 ± 0.21	92.8 ± 0.20	0.95 ± 0.01
	1.5% (Full)	ResNet-18 (He et al., 2016)	62.8 ± 1.31	69.5 ± 0.67	54.1 ± 2.27	60.8 ± 1.68	0.63 ± 0.02
	0.5% (Full)	ResNet-18 (He et al., 2016)	45.1 ± 2.27	42.0 ± 2.14	33.6 ± 3.30	37.3 ± 2.86	0.51 ± 0.01
Self-supervised	1.5%	SimCLR (Chen et al., 2020)	64.9 ± 1.02	65.2 ± 1.45	60.1 ± 1.70	62.5 ± 1.05	0.64 ± 0.03
	0.5%	SimCLR (Chen et al., 2020)	60.4 ± 1.45	63.3 ± 1.90	58.5 ± 2.01	60.8 ± 1.20	0.61 ± 0.04
Semi-supervised	1.5%	Pseudo-Label (Lee, 2013)	67.9 ± 1.85	70.5 ± 1.90	62.2 ± 1.76	66.1 ± 1.42	0.67 ± 0.04
		Mean Teacher (Tarvainen & Valpola, 2017)	68.8 ± 1.05	69.6 ± 1.21	62.2 ± 1.50	65.7 ± 1.05	0.66 ± 0.02
		MixMatch (Berthelot et al., 2019)	78.9 ± 0.67	78.4 ± 0.99	76.5 ± 0.50	77.4 ± 0.65	0.82 ± 0.03
		FixMatch (Sohn et al., 2020)	83.9 ± 1.29	85.1 ± 0.85	84.1 ± 0.95	84.6 ± 0.90	0.87 ± 0.01
		Dash (Xu et al., 2021)	82.8 ± 1.09	83.0 ± 0.55	81.2 ± 0.60	82.1 ± 0.78	0.85 ± 0.03
		FlexMatch (Zhang et al., 2021)	84.1 ± 1.02	84.9 ± 0.75	84.4 ± 0.80	84.6 ± 0.75	0.86 ± 0.04
		Semi-Path	**89.5** ± 1.05	**89.9** ± 0.70	**89.7** ± 1.02	**89.8** ± 0.78	**0.92** ± 0.03
Semi-supervised	0.5%	Pseudo-Label (Lee, 2013)	55.7 ± 2.05	59.2 ± 2.23	54.5 ± 1.97	56.8 ± 1.55	0.58 ± 0.03
		Mean Teacher (Tarvainen & Valpola, 2017)	62.3 ± 1.87	63.9 ± 2.92	57.2 ± 2.22	60.3 ± 2.49	0.61 ± 0.05
		MixMatch (Berthelot et al., 2019)	65.6 ± 0.34	70.2 ± 1.20	69.0 ± 1.45	69.6 ± 0.95	0.75 ± 0.06
		FixMatch (Sohn et al., 2020)	73.2 ± 0.76	77.5 ± 0.50	73.7 ± 0.25	75.6 ± 0.37	0.84 ± 0.04
		Dash (Xu et al., 2021)	71.0 ± 0.98	75.3 ± 0.78	70.2 ± 0.45	72.7 ± 0.50	0.82 ± 0.02
		FlexMatch (Zhang et al., 2021)	74.0 ± 0.80	78.1 ± 1.15	73.0 ± 0.90	75.5 ± 0.84	0.85 ± 0.03
		Semi-Path	**81.2** ± 0.98	**81.5** ± 0.99	**79.9** ± 0.86	**80.7** ± 0.93	**0.89** ± 0.04

feature representations from each data point in the same class can be amplified (Wang et al., 2021). We follow the same settings in Section 4.2: run three trials for each setting, and report mean value and STD for each metric. We first set 60 labeled patches (around 3% of the training set) as the total labeling budget. We start with 20 labeled data as the initial set and use two IAA cycles to reach the labeling budget. The results are summarized in Table 3. We find that Semi-Path can significantly improve the performance of SSL under this challenging dataset. The performance gain is larger than 4% in all metrics. Then we increase the labeling budget to 200 images (10% of the training set) and achieve similar performance gain. These promising results show that Semi-Path effectively leverages the unlabeled data under this *small-scale* setting.

Pathology image segmentation with weak supervision. In Lai et al. (2020, 2021), they collected an in-house GM/WM segmentation dataset at the gigapixel level and used a weakly semi-supervised way to tackle the segmentation. We follow their setting in this work and show the superiority of our proposed Semi-Path. We select two standard segmentation metrics: IoU score (Rahman & Wang, 2016) and DICE coefficient (Crum, Camara, & Hill, 2006). All results reported in this section are from the **hold-out test** set. We first implement two popular medical image segmentation networks, FCN (Bándi et al., 2017) and U-Net (Oskal et al., 2019), as the SL baselines for comparisons. As FixMatch (Sohn et al., 2020) and SSAL (Lai et al., 2021) have shown their simplicity with the trade-off between the performance and computational complexity, we mainly use these two algorithms for the comparison.

We generate the masks of GM, WM, and background from different methods for each slide. After that, we overlap them on pixel-wise ground truth masks to calculate the IoU score (Lai et al., 2021; Rahman & Wang, 2016). As these WSIs may contain noticeable variability due to the heterogeneous nature of the human brain, we also report standard deviation (STD) to measure the consistency of different methods across the slides in the hold-out test set. Both IoU score (Rahman & Wang, 2016) and STD are summarized in Table 4. As shown in Table 4, if FCN (Bándi et al., 2017) and U-Net (Oskal et al., 2019) are trained in a fully-supervised manner (all 20 slides labeled), they are able to obtain the mean IoU score at 72.91% and 90.19%, respectively. We regard 90.19% as an approximated upper bound performance of SL in this learning task. However, if we reduce the labeling budget to only 2 slides (1 AD and 1 NAD), SL suffers from severe performance degradation. Then, we evaluate FixMatch (Sohn et al., 2020) for the SSL comparison by setting the total labeling budget as 0.1%. As it shows its advantages on very scarce labeled data, e.g. 40 labeled images in CIFAR-10, equivalent to nearly 0.1% of all data in CIFAR-10 (Sohn et al., 2020), we set our total labeling budget as 0.1%. In this case, 0.1% area of 20 WSIs is about 600 patches, equivalent to 24 regions at 1280 × 1280 pixels. We first randomly

Table 3
Quantitative comparison on MHIST. All scores refer to the macro-averaged values.

Learning manner	Labeled ratio	Algorithm	Accuracy	Precision	Recall	F1-score	AUC
Supervised	100% (Full)	ResNet-18 (He et al., 2016)	86.0 ± 0.36	84.8 ± 0.61	83.5 ± 0.55	84.1 ± 0.58	0.89 ± 0.01
Self-supervised	10%	SimCLR (Chen et al., 2020)	63.2 ± 0.95	63.8 ± 1.20	62.2 ± 1.70	63.0 ± 0.85	0.66 ± 0.04
	3%	SimCLR (Chen et al., 2020)	58.6 ± 0.20	60.0 ± 0.55	58.5 ± 0.87	59.2 ± 1.20	0.63 ± 0.03
Semi-supervised	10%	Pseudo-Label (Lee, 2013)	65.4 ± 1.04	66.5 ± 0.98	65.9 ± 0.85	66.2 ± 1.42	0.75 ± 0.03
		Mean Teacher (Tarvainen & Valpola, 2017)	70.1 ± 0.80	71.3 ± 1.05	69.9 ± 0.98	70.6 ± 0.58	0.80 ± 0.02
		MixMatch (Berthelot et al., 2019)	74.1 ± 0.47	75.0 ± 0.75	74.3 ± 0.50	74.6 ± 0.97	0.83 ± 0.03
		FixMatch (Sohn et al., 2020)	70.5 ± 1.55	68.8 ± 2.76	69.5 ± 2.97	69.1 ± 2.86	0.79 ± 0.02
		Dash (Xu et al., 2021)	70.2 ± 1.21	70.0 ± 0.85	71.3 ± 0.69	70.6 ± 1.48	0.81 ± 0.04
		FlexMatch (Zhang et al., 2021)	73.1 ± 0.84	74.6 ± 1.05	75.1 ± 0.90	74.8 ± 0.89	0.82 ± 0.03
		Semi-Path	**81.1** ± 0.89	**81.0** ± 0.78	**81.2** ± 0.90	**81.1** ± 0.62	**0.87** ± 0.04
Semi-supervised	3%	Pseudo-Label (Lee, 2013)	61.2 ± 1.55	61.3 ± 2.00	62.1 ± 1.94	61.7 ± 1.67	0.68 ± 0.02
		Mean Teacher (Tarvainen & Valpola, 2017)	62.3 ± 1.87	63.9 ± 2.92	57.2 ± 2.22	60.3 ± 2.49	0.66 ± 0.05
		MixMatch (Berthelot et al., 2019)	66.8 ± 1.04	68.2 ± 0.99	67.9 ± 1.20	68.7 ± 0.45	0.76 ± 0.03
		FixMatch (Sohn et al., 2020)	62.6 ± 1.35	63.8 ± 1.15	65.6 ± 1.35	64.7 ± 1.25	0.73 ± 0.03
		Dash (Xu et al., 2021)	64.8 ± 0.65	68.3 ± 0.45	67.9 ± 0.67	68.1 ± 0.98	0.76 ± 0.02
		FlexMatch (Zhang et al., 2021)	63.0 ± 1.20	65.8 ± 1.45	64.5 ± 0.96	65.1 ± 0.56	0.73 ± 0.01
		Semi-Path	**76.1** ± 0.90	**75.9** ± 1.23	**75.9** ± 1.09	**75.9** ± 0.99	**0.84** ± 0.02

Table 4
Pixel-wise IoU Scores for AD, NAD, and overall test set.

Method	SL (FCN (Bándi et al., 2017))		SL (U-Net (Oskal et al., 2019))		FixMatch (Sohn et al., 2020)	SSAL (Lai et al., 2021)	Semi-Path
Labeled data	2 WSIs	All WSIs	2 WSIs	All WSIs	0.1%	0.1%	0.1%
AD Back	61.04 ± 5.44	81.13 ± 9.17	59.74 ± 13.9	**96.80** ± 1.48	93.15 ± 2.41	95.01 ± 1.17	95.09 ± 1.21
AD GM	46.98 ± 2.78	76.07 ± 8.91	37.16 ± 9.93	**89.58** ± 5.12	78.57 ± 3.87	88.80 ± 3.92	88.91 ± 4.05
AD WM	27.75 ± 5.50	62.23 ± 14.0	7.57 ± 6.02	**82.53** ± 7.70	56.66 ± 16.4	81.83 ± 5.53	81.95 ± 4.58
NAD Back	66.66 ± 5.17	88.42 ± 1.55	78.46 ± 18.5	**97.36** ± 3.15	97.07 ± 0.31	97.26 ± 0.52	97.33 ± 0.78
NAD GM	50.15 ± 0.49	79.37 ± 2.95	59.59 ± 13.6	**94.42** ± 3.30	83.97 ± 7.76	93.47 ± 1.60	93.59 ± 1.55
NAD WM	19.72 ± 13.6	49.89 ± 12.8	3.02 ± 3.09	**81.25** ± 9.53	22.72 ± 19.0	75.85 ± 11.4	77.95 ± 10.9
Background	63.29 ± 5.81	84.05 ± 9.17	68.28 ± 17.2	**97.02** ± 2.15	94.72 ± 2.71	95.91 ± 1.48	95.99 ± 1.33
GM	48.25 ± 2.66	77.39 ± 7.06	46.13 ± 15.8	**91.52** ± 4.94	80.73 ± 6.01	90.67 ± 3.90	90.78 ± 3.34
WM	24.54 ± 9.80	57.29 ± 14.3	5.75 ± 5.37	**82.02** ± 7.98	43.08 ± 24.0	79.44 ± 8.34	80.35 ± 8.67
Mean	45.36 ± 3.26	72.91 ± 7.56	40.05 ± 10.2	**90.19** ± 3.84	72.84 ± 7.18	88.67 ± 3.12	89.04 ± 2.99

The results are from the hold-out test set. AD refers to Alzheimer's disease cases while NAD refers to Non-Alzheimer's disease cases. 2 WSIs refers to 2 WSIs are labeled, equivalent to 10% regions of all WSIs; all WSIs refers to all WSIs are labeled. 0.1% refers to 0.1% regions of all WSIs are labeled, which can be tiled into 600 patches; so as 0.07% which can be tiled into 400 patches.

select 24 regions from two slides. The remaining regions and the other 18 slides are kept as unlabeled data. It achieves 72.84% of the mean IoU score, which is superior to both SL algorithms trained on two WSIs (limited labeled data). However, its performance on WM of NAD cases remained limited and far from fully-supervised U-Net (Oskal et al., 2019).

Then, we evaluate our deep fusion framework with region-based IAA rounds. We follow Lai et al. (2021) and set the number of rounds as three. We find the performance of Semi-Path can achieve almost 89% of the IoU score, which is comparable to the fully-supervised U-Net (Oskal et al., 2019). The major performance gain comes from the WM regions of NAD cases compared to the original FixMatch (Sohn et al., 2020). It also outperforms a recent semi-supervised active learning framework (Lai et al., 2021) with 2% on the NAD WM regions. The overall performance of Semi-Path is closer to the upper bound performance of this learning task. However, the computation complexity is significantly reduced in Semi-Path compared to SSAL (Lai et al., 2021), where the SSL and AL components are regarded as independent in the framework. Specifically, Semi-Path saves almost 80% of inference time

Table 5
Ablation studies on gigapixel pathology tasks.

SSL	IAA	APL	Accuracy	F1-score
✓			73.2 ± 0.76	75.6 ± 0.37
✓		✓	76.1 ± 0.91	76.1 ± 0.88
✓	✓		78.9 ± 0.43	78.9 ± 0.34
✓	✓	✓	**81.2** ± **0.98**	**80.7** ± **0.93**

(a) Tumor classification task with only 0.5% data labeled.

Methods	IoU score		DICE		AL inference time
	GM	WM	GM	WM	
FixMatch (Sohn et al., 2020)	80.73	43.08	91.02	63.19	–
SSAL (Lai et al., 2021)	90.67	79.44	95.19	88.53	20.7 h
Semi-Path	90.78	80.35	96.04	90.01	3.80 h

(b) GM/WM segmentation with 0.1% areas labeled.

Table 6
Compatibility of IAA with different SSL and its sensitivity to the number of rounds.

SSL Algorithm	Selection	Accuracy	F1-score
Mean Teacher (Tarvainen & Valpola, 2017)	Random	62.3 ± 1.87	60.3 ± 2.49
	IAA	**68.9** ± 1.50	**73.8** ± 1.54
ReMixMatch (Berthelot et al., 2020)	Random	70.1 ± 2.69	72.2 ± 2.05
	IAA	**78.7** ± 0.95	**78.8** ± 0.55
FixMatch (Sohn et al., 2020)	Random	73.2 ± 0.76	75.6 ± 0.37
	IAA	**80.9** ± 0.86	**80.2** ± 0.63

(a) IAA with different SSL

Rounds	Accuracy	Precision	Recall	F1-score
Original	73.2 ± 0.76	77.5 ± 0.50	73.7 ± 0.25	75.6 ± 0.37
1	75.8 ± 1.04	79.1 ± 1.01	75.3 ± 0.94	77.2 ± 0.99
2	80.9 ± 0.86	81.0 ± 0.84	79.5 ± 0.45	80.2 ± 0.63
3	81.2 ± 0.45	81.1 ± 0.21	79.9 ± 0.23	80.5 ± 0.34
4	81.2 ± 0.23	81.2 ± 0.12	80.0 ± 0.21	80.6 ± 0.19
5	81.3 ± 0.27	81.1 ± 0.21	79.8 ± 0.45	80.4 ± 0.37

(b) Effects of IAA rounds

in each IAA round compared to SSAL (Lai et al., 2021). Hence Semi-Path is a computation-efficient framework where SSL and AL are deeply fused and sharing mutual information with each other.

4.3. Empirical analysis on Semi-Path

Ablation studies of APL and IAA. we toggle the APL and IAA modules incrementally in order to show their contributions to Semi-Path's promising results. The 0.5% labeled setting for the tumor classification task is used in this ablation study. As shown in Table 5a, APL can improve the original SSL algorithm (FixMatch (Sohn et al., 2020) used in this study as one example) with 2.9% increase in accuracy. When we use IAA first and then apply APL, it also brings improvement. This shows the effectiveness of APL with/without the proposed IAA. Therefore, APL improves F1-scores and accuracy, evidence of its effectiveness, especially with correctly classifying classes that may have been missed due to fixed threshold. Then we disable APL module and observe the improvement from IAA. We also find that IAA can improve the original SSL algorithm and SSL+APL consistently. Then, we conduct the time analysis on the gigapixel segmentation task. The results are summarized in Table 5b: Semi-Path can achieve superior results while significantly reducing the AL inference time due to the deep fusion design.

Fig. 3 displays the predictive masks from different methods and the ground truth masks. Both of these two cases are from the hold-out test set. The masks of U-Net (Oskal et al., 2019) show supervised learning is sensitive to the amount of labeled data: the results can be close to the ground truth (as shown in the second column) when the variety and volume of the labeled dataset are sufficiently large, but the performance can be greatly degraded (as shown in the third column) if the labeled dataset is limited. On the other side, FixMatch (Sohn et al., 2020) is able to predict the rough boundaries between GM and WM, but the local prediction details are still far away from the ground truth masks. For example, there are considerable quantities of noisy pixels in WM, which means many WM regions are wrongly classified as GM. The predicted masks of Semi-Path are the closest to the ground truth masks.

Compatibility of IAA with different SSL algorithms. To further prove the general benefits of IAA in SSL settings, we also apply IAA to recent popular SSL algorithms to prove its general use and compatibility in SSL settings. The SSL algorithms include Mean

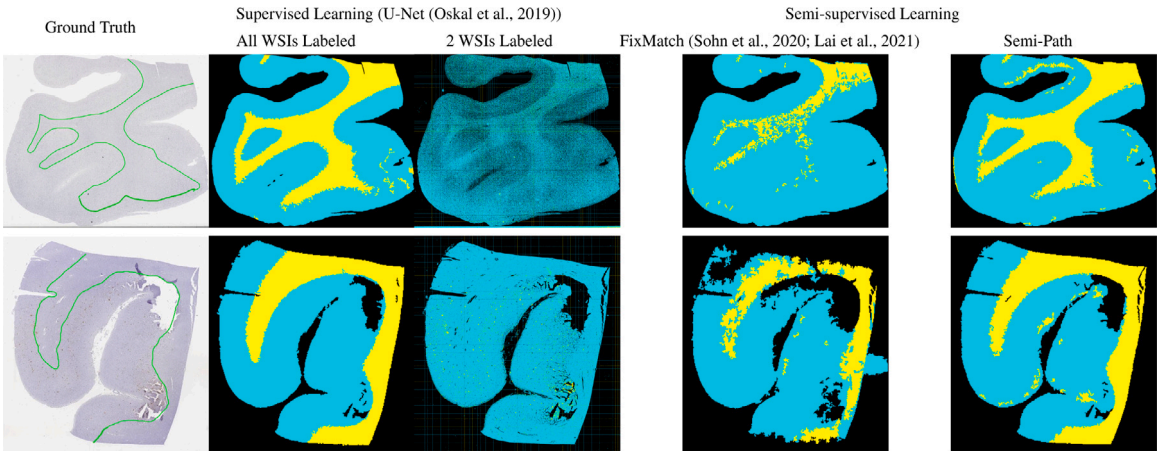

Fig. 3. Segmentation masks visualization: GM, WM, and background are indicated by cyan, yellow, and black, respectively. Both SSL results are using FixMatch (Sohn et al., 2020) as the backbone and only using 0.1% labeled area of 20 WSIs in the training set. All slides are Amyloid-β stained and from brain tissues.

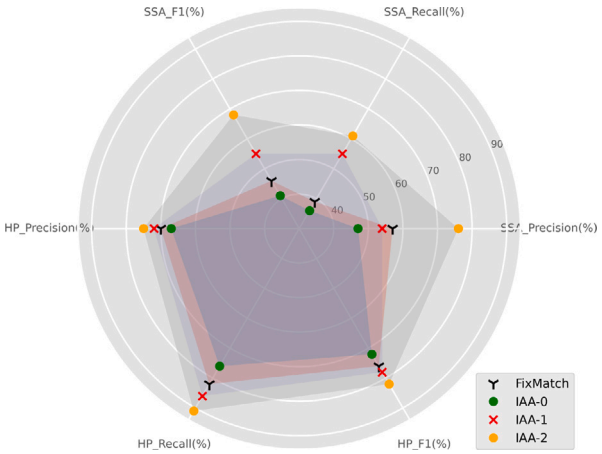

Fig. 4. Comparison on per class's performance in MHIST (Wei et al., 2021) under the labeling budget as only 60 labeled images. IAA-0: the starting stage (20 labeled images), IAA-1: Semi-Path after 1 IAA round (40 labeled dat); IAA-2: Semi-Path after 2 IAA rounds (60 labeled images).

Teacher (Tarvainen & Valpola, 2017), ReMixMatch (Berthelot et al., 2020), and FixMatch. We summarize the results in Table 6a. When we only set 0.5% images as the labeling budget, our proposed IAA module is able to find the informative samples for labeling and boost the performance of these original SSL algorithms consistently.

The sensitivity of IAA on the number of rounds. In Section 4, we use a two-round IAA and show its promising improvements. In this subsection, we further study the effect of additional IAA rounds on eventual performance. For a fair comparison, we set 0.5% of patches in PCam (Veeling et al., 2018) as the total labeling budget and set APL module active. The results are summarized in Table 6b: when we use only one-round IAA to reach the labeling budget, the improvement can be trivial; however when we use more than three rounds to reach the budget, the performance gain is also limited. Considering that each AL round will inevitably incur computation costs and expert interactions, we choose two-round IAA in our study for the trade-off between the performance and clinical interactions.

Data imbalance scenario. To explore the source of performance gain, we look deeper into per class's performance. As we achieve more than 10% of improvement on MHIST (Wei et al., 2021), we report the performance from each IAA round to visualize the benefits for each class. MHIST (Wei et al., 2021) is a challenging dataset due to its limited amount and data imbalance issue (Wang et al., 2021). As shown in Fig. 4, our framework improves the performance (in terms of Recall and Precision) on sessile serrated adenomas (SSA), which demonstrates its effectiveness in classifying the minor class in this dataset. Recent theoretical works (Kim et al., 2020; Lai et al., 2022) found that the generated pseudo-labels from minor classes during SSL training can be significantly underestimated, subsequently biasing the model towards the major classes and degrading the performance eventually. In other words, more minority samples have never been assigned pseudo-labels during the training process. These samples will be split into the candidate pool in our SSL-AL iterative structure of IAA, subsequently, queried for annotations, and ultimately enter the

labeled set. Therefore, our framework has the potential to realign the underlying data imbalance in the unlabeled pool and boost the applicability of SSL in real-world applications.

5. Discussion

We propose Semi-Path, a deployable and active semi-supervised learning framework for relieving labor efforts for gigapixel-resolution pathology applications. We first design class-aware Adaptive Pseudo-Labeling (APL) with dynamic thresholds to select the unlabeled samples for pseudo-labeling based on the learning status of SSL and class difficulty. Then we design and incorporate Informative Active Annotation (IAA) in our SSL framework to further exploit the informativeness in the unlabeled pool. We evaluate Semi-Path on three pathology datasets and show its consistent improvement over other SSL algorithms. Additionally, Semi-Path reduces computational complexity for the AL rounds since SSL and AL are integrated into a unified framework with IAA. We believe Semi-Path has the potential to enhance the applicability of SSL in pathology image applications. The limitation of Semi-Path is that we assume the unlabeled pool has the same classes as the labeled set. In the future work, we aim to build a more generalizable framework that can be applied to other medical domains.

CRediT authorship contribution statement

Zhengfeng Lai: Conceptualization, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Joohi Chauhan:** Validation, Writing – review & editing. **Dongjie Chen:** Visualization, Writing – review & editing. **Brittany N. Dugger:** Data curation, Supervision, Writing – review & editing. **Sen-Ching Cheung:** Supervision, Writing – review & editing, Conceptualization, Methodology. **Chen-Nee Chuah:** Conceptualization, Funding acquisition, Methodology, Project administration, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Chen-Nee Chuah, Brittany N. Dugger reports financial support was provided by National Institutes of Health. Chen-Nee Chuah, Brittany N. Dugger reports financial support was provided by University of California Davis. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that has been used is confidential.

Acknowledgments

This work was supported by the Noyce Initiative UC Partnerships in Computational Transformation Grant and the UC Davis Center for Women's Cardiovascular and Brain Health research program under the HEAL-HER (Heart, BrEast, and BrAin HeaLth Equity Research) award made possible by the Cy Pres funds. This work also received additional partial support from National Institutes of Health grants P30 AG072972 and R01 AG062517. The authors would like to thank the families and participants of the University of California, Davis Alzheimer's Disease Research Center (UCD-ADRC) for their generous donations, as well as all the faculty and staff of the UCD-ADRC. The views and opinions expressed in this manuscript are those of the author and do not necessarily reflect the official policy or position of any public health agency of California or of the United States government.

References

Abdeljawad, K., Vemulapalli, K. C., Kahi, C. J., Cummings, O. W., Snover, D. C., & Rex, D. K. (2015). Sessile serrated polyp prevalence determined by a colonoscopist with a high lesion detection rate and an experienced pathologist. *Gastrointestinal Endoscopy*, *81*(3), 517–524.

Arazo, E., Ortego, D., Albert, P., O'Connor, N. E., & McGuinness, K. (2020). Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *Proc. 2020 int. jt. conf. neural netw.* (pp. 1–8). IEEE.

Ash, J. T., Zhang, C., Krishnamurthy, A., Langford, J., & Agarwal, A. (2019). Deep batch active learning by diverse, uncertain gradient lower bounds. In *ICLR*.

Bándi, P., van de Loo, R., Intezar, M., Geijs, D., Ciompi, F., van Ginneken, B., et al. (2017). Comparison of different methods for tissue segmentation in histopathological whole-slide images. In *Proc. 2017 IEEE int. symp. biomed. imaging* (pp. 591–595).

Bejnordi, B. E., Veta, M., Van Diest, P. J., Van Ginneken, B., Karssemeijer, N., Litjens, G., et al. (2017). Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, *318*(22), 2199–2210.

Berthelot, D., Carlini, N., Cubuk, E. D., Kurakin, A., Sohn, K., Zhang, H., et al. (2020). ReMixMatch: Semi-supervised learning with distribution alignment and augmentation anchoring. In *ICLR*.

Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., & Raffel, C. A. (2019). MixMatch: A holistic approach to semi-supervised learning. *Vol. 32*, In *NeurIPS*.

Budd, S., Robinson, E. C., & Kainz, B. (2021). A survey on active learning and human-in-the-loop deep learning for medical image analysis. *Medical Image Analysis*, Article 102062.

Chan, L., Hosseini, M. S., & Plataniotis, K. N. (2021). A comprehensive analysis of weakly-supervised semantic segmentation in different image domains. *International Journal of Computer Vision*, *129*(2), 361–384.

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *ICML* (pp. 1597–1607).

Chen, H., Tao, R., Fan, Y., Wang, Y., Wang, J., Schiele, B., et al. (2023). SoftMatch: Addressing the quantity-quality trade-off in semi-supervised learning. In *ICLR*.

Choi, J., Elezi, I., Lee, H.-J., Farabet, C., & Alvarez, J. M. (2021). Active learning for deep object detection via probabilistic modeling. In *ICCV* (pp. 10264–10273).

Crum, W. R., Camara, O., & Hill, D. L. (2006). Generalized overlap measures for evaluation and validation in medical image analysis. *IEEE Transactions on Medical Imaging*, *25*(11), 1451–1461.

Drugman, T., Pylkkonen, J., & Kneser, R. (2019). Active and semi-supervised learning in asr: Benefits on the acoustic and language models. arXiv preprint arXiv:1903.02852.

Feng, R., Liu, X., Chen, J., Chen, D. Z., Gao, H., & Wu, J. (2020). A deep learning approach for colonoscopy pathology WSI analysis: accurate segmentation and classification. *IEEE Journal of Biomedical and Health Informatics*, *25*(10), 3700–3708.

Gal, Y., Islam, R., & Ghahramani, Z. (2017). Deep bayesian active learning with image data. In *International conference on machine learning* (pp. 1183–1192). PMLR.

Gao, M., Zhang, Z., Yu, G., Arık, S. Ö., Davis, L. S., & Pfister, T. (2020). Consistency-based semi-supervised active learning: Towards minimizing labeling cost. In *European conference on computer vision* (pp. 510–526). Springer.

Graham, S., Epstein, D., & Rajpoot, N. (2020). Dense steerable filter cnns for exploiting rotational symmetry in histology images. *IEEE Transactions on Medical Imaging*, *39*(12), 4124–4136.

Gudovskiy, D., Hodgkinson, A., Yamaguchi, T., & Tsukizawa, S. (2020). Deep active learning for biased datasets via fisher kernel self-supervision. In *Proc. IEEE/CVF conf. comput. vis. pattern recognit.* (pp. 9041–9049).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proc. IEEE/CVF conf. comput. vis. pattern recognit.* (pp. 770–778).

Huang, Z., Jiang, R., Aeron, S., & Hughes, M. C. (2024). Systematic comparison of semi-supervised and self-supervised learning for medical image classification. arXiv:2307.08919.

Jahanifar, M., Tajeddin, N. Z., Koohbanani, N. A., & Rajpoot, N. M. (2021). Robust interactive semantic segmentation of pathology images with minimal user input. In *ICCV* (pp. 674–683).

Joshi, A. J., Porikli, F., & Papanikolopoulos, N. (2009). Multi-class active learning for image classification. In *Proc. IEEE/CVF conf. comput. vis. pattern recognit.* (pp. 2372–2379). IEEE.

Killamsetty, K., Sivasubramanian, D., Ramakrishnan, G., & Iyer, R. (2021). GLISTER: Generalization based data subset selection for efficient and robust learning. *Vol. 35*, In *AAAI* (pp. 8110–8118).

Kim, J., Hur, Y., Park, S., Yang, E., Hwang, S. J., & Shin, J. (2020). Distribution aligning refinery of pseudo-label for imbalanced semi-supervised learning. *Vol. 33*, In *NeurIPS* (pp. 14567–14579).

Kim, K., Park, D., Kim, K. I., & Chun, S. Y. (2021). Task-aware variational adversarial active learning. In *Proc. IEEE/CVF conf. comput. vis. pattern recognit.* (pp. 8166–8175).

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *ICLR (poster)*.

Kirsch, A., Van Amersfoort, J., & Gal, Y. (2019). Batchbald: Efficient and diverse batch acquisition for deep Bayesian active learning. *Vol. 32*, In *NeurIPS* (pp. 7026–7037).

Koohbanani, N. A., Unnikrishnan, B., Khurram, S. A., Krishnaswamy, P., & Rajpoot, N. (2021). Self-path: Self-supervision for classification of pathology images with limited annotations. *IEEE Transactions on Medical Imaging*, *40*(10), 2845–2856.

Lai, Z., Guo, R., Xu, W., Hu, Z., Mifflin, K., Dugger, B. N., et al. (2020). Automated grey and white matter segmentation in digitized aβ human brain tissue slide images. In *Proc. 2020 IEEE int. conf. multimed. expo workshops* (pp. 1–6). IEEE.

Lai, Z., Wang, C., Gunawan, H., Cheung, S.-C. S., & Chuah, C.-N. (2022). Smoothed adaptive weighting for imbalanced semi-supervised learning: Improve reliability against unknown distribution data. In *International conference on machine learning* (pp. 11828–11843). PMLR.

Lai, Z., Wang, C., Oliveira, L. C., Dugger, B. N., Cheung, S.-C., & Chuah, C.-N. (2021). Joint semi-supervised and active learning for segmentation of gigapixel pathology images with cost-effective labeling. In *ICCV workshop* (pp. 591–600).

Lee, D.-H. (2013). Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning*.

Mahapatra, D., Schüffler, P. J., Tielbeek, J. A., Vos, F. M., & Buhmann, J. M. (2013). Semi-supervised and active learning for automatic segmentation of Crohn's disease. In *Medical image computing and computer-assisted intervention–MICCAI 2013: 16th international conference, Nagoya, Japan, September 22-26, 2013, proceedings, Part II. Vol. 16* (pp. 214–221). Springer.

Miyato, T., Maeda, S.-i., Koyama, M., & Ishii, S. (2018). Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(8), 1979–1993.

Myers, R. H. (1974). *Elementary applied statistics*. Taylor & Francis.

Oliver, A., Odena, A., Raffel, C. A., Cubuk, E. D., & Goodfellow, I. (2018). Realistic evaluation of deep semi-supervised learning algorithms. *Vol. 31*, In *NeurIPS* (pp. 3235–3246).

Oskal, K. R., Risdal, M., Janssen, E. A., Undersrud, E. S., & Gulsrud, T. O. (2019). A U-net based approach to epidermal tissue segmentation in whole slide histopathological images. *SN Applied Sciences*, *1*(7), 1–12.

Peng, F., Wang, C., Liu, J., & Yang, Z. (2021). Active learning for lane detection: A knowledge distillation approach. In *Proc. IEEE/CVF conf. comput. vis. pattern recognit.* (pp. 15152–15161).

Pinckaers, H., Bulten, W., van der Laak, J., & Litjens, G. (2021). Detection of prostate cancer in whole-slide images through end-to-end training with image-level labels. *IEEE Transactions on Medical Imaging*, *40*(7), 1817–1826.

Pulido, J. V., Guleria, S., Ehsan, L., Fasullo, M., Lippman, R., Mutha, P., et al. (2020). Semi-supervised classification of noisy, gigapixel histology images. In *Proc. 2020 IEEE 20th int. conf. bioinform. bioeng.*

Rahman, M. A., & Wang, Y. (2016). Optimizing intersection-over-union in deep neural networks for image segmentation. In *International symposium on visual computing* (pp. 234–244). Springer.

Rhee, P. K., Erdenee, E., Kyun, S. D., Ahmed, M. U., & Jin, S. (2017). Active and semi-supervised learning for object detection with imperfect data. *Cognitive Systems Research*, *45*, 109–123.

Rizve, M. N., Duarte, K., Rawat, Y. S., & Shah, M. (2021). In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. arXiv:2101.06329.

Sener, O., & Savarese, S. (2018). Active learning for convolutional neural networks: A core-set approach. In *ICLR*.

Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C. A., et al. (2020). FixMatch: Simplifying semi-supervised learning with consistency and confidence. *Vol. 33*, In *NeurIPS* (pp. 596–608).

Tang, Z., Chuang, K. V., DeCarli, C., Jin, L.-W., Beckett, L., Keiser, M. J., et al. (2019). Interpretable classification of Alzheimer's disease pathologies with a convolutional neural network pipeline. *Nature Communications*, *10*(1), 1–14.

Tarvainen, A., & Valpola, H. (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*.

Veeling, B. S., Linmans, J., Winkens, J., Cohen, T., & Welling, M. (2018). Rotation equivariant CNNs for digital pathology. In *Int. conf. med. image comput. comput.-assist. intervent.* (pp. 210–218). Springer.

Wang, Y., Chen, H., Heng, Q., Hou, W., Savvides, M., Shinozaki, T., et al. (2023). Freematch: Self-adaptive thresholding for semi-supervised learning. In *ICLR*.

Wang, Z., Dong, N., & Voiculescu, I. (2022). Computationally-efficient vision transformer for medical image semantic segmentation via dual pseudo-label supervision. In *2022 IEEE international conference on image processing* (pp. 1961–1965). IEEE.

Wang, Z., Li, T., Zheng, J.-Q., & Huang, B. (2022). When CNN meet with vit: Towards semi-supervised learning for multi-class medical image semantic segmentation. In *European conference on computer vision* (pp. 424–441). Springer.

Wang, Z., & Ma, C. (2023). Dual-contrastive dual-consistency dual-transformer: A semi-supervised approach to medical image segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 870–879).

Wang, Z., & Voiculescu, I. (2023). Weakly supervised medical image segmentation through dense combinations of dense pseudo-labels. In *MICCAI workshop on data engineering in medical imaging* (pp. 1–10). Springer.

Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Huang, J., et al. (2021). Transpath: Transformer-based self-supervised learning for histopathological image classification. In *Int. conf. med. image comput. comput.-assist. intervent.* (pp. 186–195). Springer.

Wei, J., Suriawinata, A., Ren, B., Liu, X., Lisovsky, M., Vaickus, L., et al. (2021). A petri dish for histopathology image analysis. In *Int. conf. arti. intell. in medi.* (pp. 11–24). Springer.

Xie, Q., Dai, Z., Hovy, E., Luong, T., & Le, Q. (2020). Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems*, *33*, 6256–6268.

Xu, Y., Shang, L., Ye, J., Qian, Q., Li, Y.-F., Sun, B., et al. (2021). Dash: Semi-supervised learning with dynamic thresholding. In *International conference on machine learning* (pp. 11525–11536). PMLR.

You, C., Dai, W., Liu, F., Min, Y., Su, H., Zhang, X., et al. (2022). Mine your own anatomy: Revisiting medical image segmentation with extremely limited labels. arXiv preprint arXiv:2209.13476.

You, C., Dai, W., Min, Y., Liu, F., Clifton, D., Zhou, S. K., et al. (2024). Rethinking semi-supervised medical image segmentation: A variance-reduction perspective. *Advances in Neural Information Processing Systems*, *36*.

You, C., Dai, W., Min, Y., Staib, L., & Duncan, J. S. (2023). Bootstrapping semi-supervised medical image segmentation with anatomical-aware contrastive distillation. In *International conference on information processing in medical imaging* (pp. 641–653). Springer.

You, C., Dai, W., Min, Y., Staib, L., Sekhon, J. S., & Duncan, J. S. (2023). ACTION++: Improving semi-supervised medical image segmentation with adaptive anatomical contrast. arXiv preprint arXiv:2304.02689.

Zhang, B., Wang, Y., Hou, W., Wu, H., Wang, J., Okumura, M., et al. (2021). Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Vol. 34*, In *NeurIPS*.

Zhang, W., Zhu, L., Hallinan, J., Zhang, S., Makmur, A., Cai, Q., et al. (2022). Boostmis: Boosting medical image semi-supervised learning with adaptive pseudo labeling and informative active annotation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 20666–20676).

Zhao, Z., Zeng, Z., Xu, K., Chen, C., & Guan, C. (2021). Dsal: Deeply supervised active learning from strong and weak labelers for biomedical image segmentation. *IEEE Journal of Biomedical and Health Informatics*, *25*(10), 3744–3751.

Zou, H. P., & Caragea, C. (2023). JointMatch: A unified approach for diverse and collaborative pseudo-labeling to semi-supervised text classification. In *The 2023 conference on empirical methods in natural language processing*.

Zou, H., Zhou, Y., Zhang, W., & Caragea, C. (2023). DeCrisisMB: Debiased semi-supervised learning for crisis tweet classification via memory bank. In *Findings of the association for computational linguistics: EMNLP 2023* (pp. 6104–6115).