Joint Semi-supervised and Active Learning for Segmentation of Gigapixel Pathology Images with Cost-Effective Labeling

Zhengfeng Lai*
University of California, Davis
lzhengfeng@ucdavis.edu

Chao Wang*
University of California, Davis
chaowang.hk@gmail.com

Luca Cerny Oliveira
University of California, Davis
lcernyo@ucdavis.edu

Brittany N. Dugger
University of California, Davis
bndugger@ucdavis.edu

Sen-Ching Cheung
University of Kentucky
sccheung@ieee.org

Chen-Nee Chuah
University of California, Davis
chuah@ucdavis.edu

Abstract

The need for manual and detailed annotations limits the applicability of supervised deep learning algorithms in medical image analyses, specifically in the field of pathology. Semi-supervised learning (SSL) provides an effective way for leveraging unlabeled data to relieve the heavy reliance on the amount of labeled samples when training a model. Although SSL has shown good performance, the performance of recent state-of-the-art SSL methods on pathology images is still under study. The problem for selecting the most optimal data to label for SSL is not fully explored. To tackle this challenge, we propose a semi-supervised active learning framework with a region-based selection criterion. This framework iteratively selects regions for annotation query to quickly expand the diversity and volume of the labeled set. We evaluate our framework on a grey-matter/white-matter segmentation problem using gigapixel pathology images from autopsied human brain tissues. With only 0.1% regions labeled, our proposed algorithm can reach a competitive IoU score compared to fully-supervised learning and outperform the current state-of-the-art SSL by more than 10% of IoU score and DICE coefficient.

1. Introduction

Deep learning methods have shown promising performance on pathology images. For example, FCN [4] and U-Net [39] are two popular architectures applied to pathology image segmentation. To tackle the difficulty in training deep neural networks on gigapixel images, Lai *et al.* [31,32] proposed a patch-based method to achieve competitive results compared to FCN and U-Net for segmentation of whole slide images (WSIs) of brain tissues. However, all of the aforementioned methods heavily rely on a well-curated and fully annotated dataset. Dataset annotations are typically conducted by trained experts with sufficient domain knowledge. This can be a time-consuming task and can limit scalability. In Alzheimer's disease (AD) research, neuropathologists normally assess histopathology by detecting and identifying extracellular Amyloid-β plaques [17], one of the hallmarks pathological feature associated with AD [14, 46]. A recent work [48] claimed that more than 30,000 patches from gigapixel WSIs need to be annotated by neuropathology researchers to train a deep learning model to detect the Amyloid-β plaques with acceptable accuracy. Hence the requirement of large amounts of curated data that involves substantial labeling cost may limit the wide-adoption of supervised deep learning methods in real-world medical problems [15, 38].

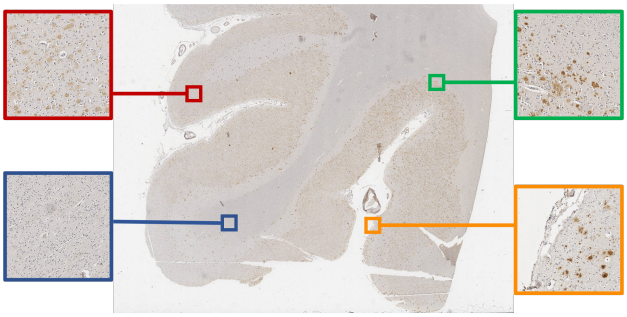


Figure 1. An example of WSI and its GM and WM regions: the red and blue blocks are from GM and WM respectively; the green block contains the boundary between GM/WM; the orange block contains the boundary between tissue and the background.

To circumvent the need of a large labeled dataset, semi-supervised learning (SSL) has been attracting more attention recently as a way to relieve the labeling requirements and build comparable models [6, 19, 24, 47]. Although these methods show good results, efforts on deploying them on pathology images are still limited [8, 40]. Specifically, the

*Equal contributions

selection criterion for labeled data in pathology images remain unclear. There are no well-established SSL guidelines on how to determine the amount of labeled data that is needed.

Active learning (AL), on the other hand, aims at maximizing a model's performance with minimal labeling efforts [3]. One challenge of AL is the *cold start problem*: when the selection of starting set is poor or the size is not sufficiently large, the models trained in subsequent cycles can be high-biased, resulting in poor selection again in the following cycles [23, 29]. Another challenge is that the annotation acquisition during the training process may require substantial rounds of labeling, while at the same time, the annotators cannot guarantee unlimited availability during the whole process. This would notably extend the time needed for the entire project. To reduce the number of rounds needed, batch-mode AL methods are proposed [2,22,27,28,45,52] to select a batch of data to be labeled at once. However, these methods still require many rounds of labeling and have not been evaluated on histology images.

Therefore, to minimize the labeling efforts, a natural direction is to combine AL and SSL as both of them aim to improve a model's performance on limited labeled data. More importantly, SSL and AL are complementary in function: SSL has the potential to leverage the unlabeled data to relieve the *cold start problem* of AL; on the other hand, AL can select the most optimal data to label for SSL. In this paper, we propose a SSL framework with AL-driven region-based labeled data selection. We evaluate our framework on an Amyloid-β stained neuropathology images at gigapixel level to separate grey matter (GM) and white matter (WM), which is an important task for analyzing density and distribution of Amyloid-β plaques in the brain. Figure 1 shows an example of the WSI dataset used to evaluate our approach. Although supervised learning methods [31, 32] have already shown promising results, their performances face a severe degradation when labeled data is limited (as shown in Table 2).

Our contributions can be summarized as follows:

- We combine SSL and AL to construct a label-efficient learning framework with well-defined selection of regions for annotation to construct a small but effectively labeled dataset.
- To tackle the "cold-start" issue of AL, we design a region-based selection criterion to expand the diversity and volume of the labeled set effectively.
- We evaluate the proposed framework on a pathology segmentation task and it outperforms the state-of-the-art SSL (FixMatch [47]) by more than 10% of IoU score and DICE coefficient [12].

2. Related work

2.1. Semi-supervised learning

Semi-supervised learning (SSL) leverages unlabeled data to improve the performance of supervised learning when the amount of labeled data is limited [51]. Classical SSL approaches generate pseudo-labels for each unlabeled sample based on model's prediction, and use them to train the model via pseudo-labeling [1,34], consistency regularization [36, 49], and the combination of them [5, 6, 30, 47]. Recent state-of-the-art SSL algorithms such as FixMatch [47] apply regularization consistency on different augmented views of the same unlabeled data, while generating pseudo labels accordingly during the training process. Although these SSL methods achieve promising results, their applicability on histology images is still under study [40]. There are two challenges for these SSL approaches: 1) pseudo labels, generated based on a poorly estimated class distribution of limited labeled data, increases the difficulty in updating the true distribution of unlabeled data. [44]; 2) another natural question is how to determine the most valuable samples and the size of the samples to be labeled and added to the labeled set.

2.2. Active learning

Active learning (AL) aims to maximize the performance gain of learning by selecting the fewest possible samples and sending them to annotators for labeling [7, 42]. Recent AL approaches include uncertainty-based, diversity-based, and model performance change-based. The uncertainty-based AL approaches use *max entropy* and *max margin* [25] as the selection criteria due to simplicity. Earlier methods [18, 20, 50] worked on the setting that only one sample is selected for labeling in each cycle of AL. Batch-mode AL methods [2, 22, 27, 28, 45, 52] were proposed later to select a batch of samples at once. However, all of these methods rely on a sufficiently large and representative labeled set as the starting set, otherwise the *cold-start* problem would appear and hurt the performance by biasing the model.

2.3. Semi-supervised active learning

Although both SSL and AL share a similar goal for minimizing labeling efforts, only a few works have considered combining them. In [16], they combine SSL and AL for speech understanding to reduce significant errors with limited speech data. Rhee *et al.* built a semi-supervised AL system in the pedestrian detection task [43]. Sener *et al.* redefined AL as core-set selection and considered SSL during AL cycles [45]. However, all of these treated SSL and AL independently without considering their mutual impact on each other [19]. In addition, none of them have been evaluated on the histology images, hence their applicability is still questionable.

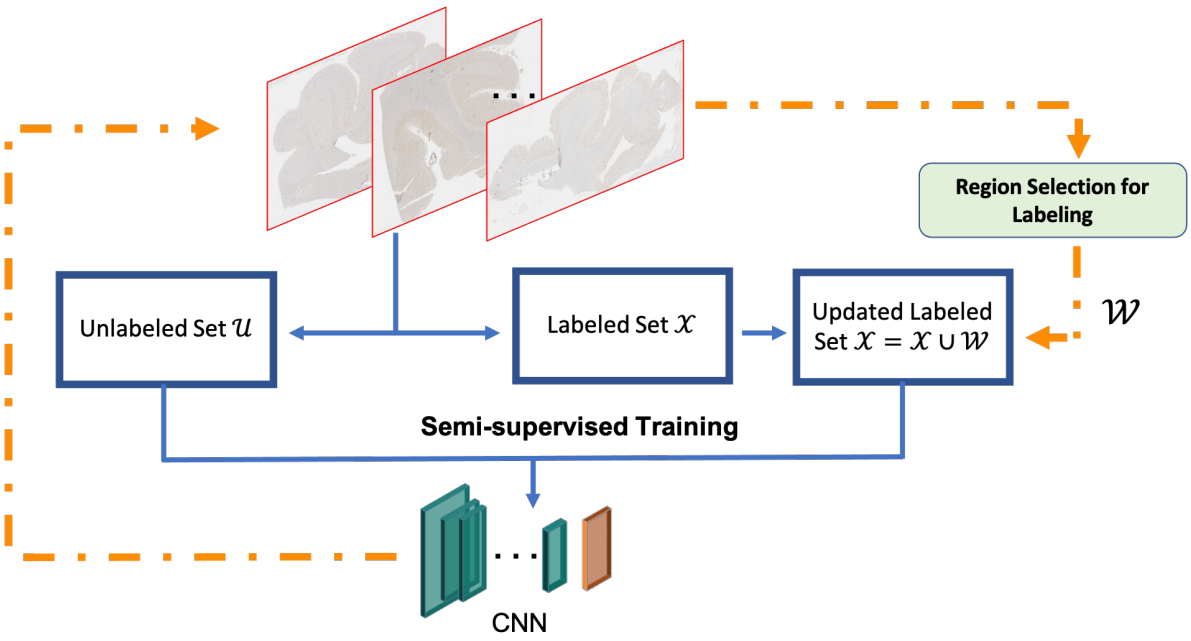


Figure 2. Overview of the proposed semi-supervised active learning framework for gigapixel pathology images.

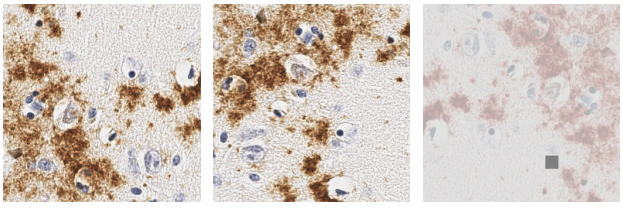
3. Method

3.1. Problem setup

We first set up the GM and WM segmentation problem on gigapixel histopathology images. A recent work [11] revealed that per-pixel classification is not all one needs to tackle a semantic segmentation problem. Lai *et al.* [32] transforms this problem into a patch-based classification problem, where one WSI is tiled into a group of patches. Here we follow their approach and apply semi-supervised learning to classify each patch into three categories, i.e., WM, GM, and background. The training data $\mathcal{D}$ contains a labeled set $\mathcal{X} = \{(x_i, y_i)\}$, and an unlabeled dataset $\mathcal{U} = \{x_j\}$ where $x_i \in \mathbb{R}^{d\times d}$ is the i-th patch and $y_i \in \{0,1\}^3$ is the corresponding one-hot label. Our goal is to train a classifier $h(x,\theta) : \mathbb{R}^{d\times d} \to [0,1]^3$, where θ is the CNN parameter and the k-th component of $h(x,\theta)$ is the predictive probability of the k-th class for an input x. θ is trained by minimizing an objective loss function:

$$\min_{\theta\in\Theta} L(\mathcal{X},\theta) + \Omega(\mathcal{D},\theta), \tag{1}$$

where $L(\mathcal{X},\theta) := \sum_{(x,y)\in\mathcal{X}} l(x,y,\theta)$ and $\Omega(\mathcal{D},\theta) := \sum_{x\in\mathcal{D}} \omega(x,\theta)$. We denote the per-sample supervised loss and regularization as l and ω, respectively. We mainly adopt FixMatch [47], which generates artificial labels using both pseudo-labeling and consistency regularization. Specifically, the pseudo label is generated based on a weakly-augmented unlabeled image (weak), which will be the target to be compared with the output of the model on a strongly-augmented version of the same unlabeled image (strong). One example of weak and strong augmentation operations is shown in Figure 3: weak augmentations include random flip and rotations while strong augmentation include RandAugment [13] and CTAugment [5].



(a) Original patch (b) Weakly augmented (c) Strongly augmented

Figure 3. An example of weak/strong augmentation.

3.2. Semi-supervised active learning

One of the critical issues in applying SSL is the selection of the labeled data $\mathcal{X}$. In this subsection, we aim for an efficient selection of the labeled data by combining SSL and AL. The selection criteria of traditional AL [18,20,50] were designed based on only one sample in each cycle of AL. Here one sample corresponds to one patch in our pathology images. However, one patch may not contain sufficient features for GM/WM and the criterion focusing on only one patch can lead to a haphazard selection for labeling query. Therefore, instead of patch-based selection, we propose a region-based selection criterion to not only reduce the number of labeling queries, but also to provide more "informative" data (details as shown in Session 3.3). Here a region $\mathcal{R} \in \mathbb{R}^{nd\times nd}$ can be tiled into n^2 patches. With the

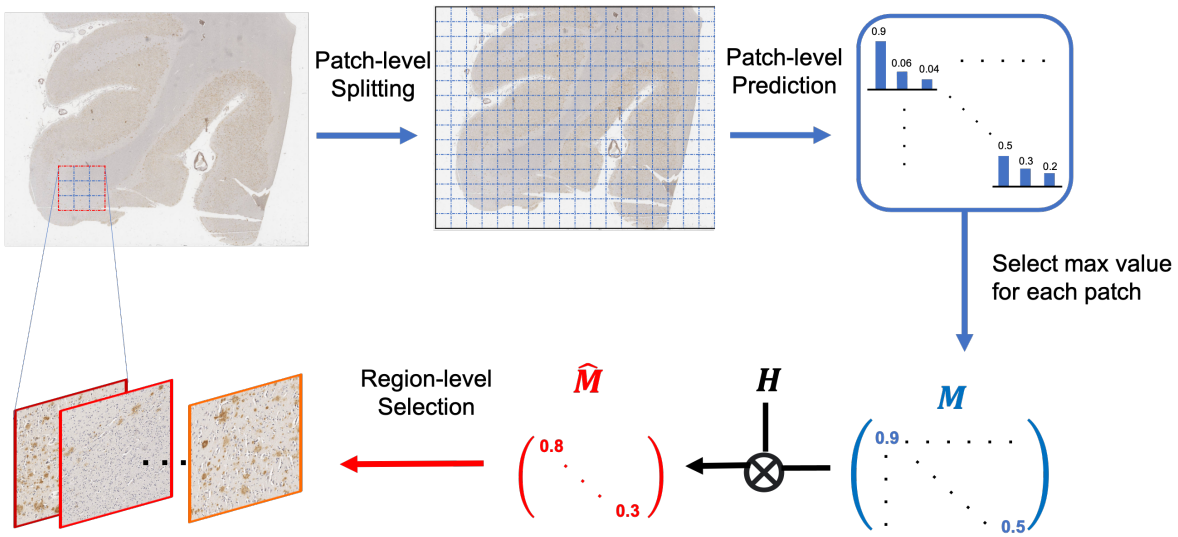


Figure 4. Illustration of the proposed region-based selection criterion: a sliding window is used to traverse the whole slide with semi-supervised learning model to produce patch-level prediction matrix M, which is then convoluted with kernel H to transform it into $\hat{M}$ that contains region-level metrics.

region-based selection, the labeled data set can be quickly expanded as

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

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 $\{J(x)\}_{x\in\mathcal{W}}$. The proposed framework is shown in Figure 2: starting from all WSIs, we deploy our proposed region-based selection criterion (details as shown in Session 3.3) to choose regions for annotation query. These regions would be added into $\mathcal{X}$, subsequently the updated $\mathcal{X}$ and $\mathcal{U}$ are used for semi-supervised training to optimize the encoder. We use m to denote the number of regions to be selected for labeling in one round of active learning. Therefore, m is a hyper-parameter that can be tuned based on the annotation budgets, such as the number of affordable rounds of labeling.

For a well-defined initialization, we follow a recent work [33] that adopts a pre-training process by using a self-supervised module to pretrain the encoder on the unlabeled set $\mathcal{U}$. SimCLR [9] is a simple self-supervised framework that applies contrastive learning to capture visual representations on unlabeled set. We adapt SimCLR [9] to train the encoder f. Then we select several diverse regions among WSIs for annotation query and fine-tune $h = f \circ g$, where g is the last layer for the classifier h as the **initialization** of our AL cycles. For our GM/WM problem, the number of the regions for the fine-tune initialization stage is set as two: one is from the boundary between GM and WM, the other is from the boundary between the tissue and background. These two regions can provide the annotations for all three classes (GM, WM, and background) in this problem.

3.3. Region-based selection criterion

In this subsection, we propose a region-based selection criterion for our AL framework. We hypothesize that labeling samples with high entropy in the prediction should be valuable. However, traditional uncertainty-based methods focus on each individual sample, which may result in haphazard and the "overconfident" issue [19]. This may lead the model to select less valuable samples. Therefore, to enhance the robustness of the uncertainty criterion, we propose a region-based selection criterion to focus on a region consisting of a batch of neighboring samples together instead of on only one sample at once. The regions where the majority of samples have high uncertainty would be selected for annotation query. The details of this process is shown in Figure 4.

Specifically, the proposed metric quantifies the accumulative entropy of predictions over a whole region: for each patch in the region, the value of the most confident class would be selected. Then we can generate a matrix M where the i-th entry is

$$m_i = \max\{h(x_i,\theta)_1, h(x_i,\theta)_2, h(x_i,\theta)_3\}. \quad (2)$$

Here these three entries correspond to WM, GM, and background. As the classifier h is at the patch level, we need to transform the metric from the patch level into the region level. By introducing a kernel H, we process a convolution between M and H, i.e., as $\hat{M} = M * H$. Here $H \in \mathbb{R}^{q\times q}$ could be a mean filter with all the entries being $1/q^2$ to collect the statistic information in each region. The dimension of H is determined based on the original resolution of the

histology images: the size of H should be large enough to contain more areas with sufficient information from each class. In our case, as the boundaries of GM and WM form gradual changes across hundreds of pixels and the size of one patch is set as 256×256, we set the dimension of H as 5×5, which represents a corresponding region of 1280×1280 pixels to include multiple classes in one region. Based on our hypothesis, we will select m regions with m lowest values (regions with most uncertainty) among entries of $\hat{M}$. To avoid redundant selection, the value of entries selected in the previous iterations will be set as infinity. We summarize the whole process in Algorithm 1.

Algorithm 1 Region-based semi-supervised active learning for gigapixel histology images

Input: training dataset $\mathcal{D}$, the total number of steps and cycles S, T, a kernel H, the number of regions m selected in one cycle
Pre-training: obtain f via SimCLR [9] among $\mathcal{D}$
Initialize: randomly select two regions for pixel-wise labeling query and tile them into patches to formulate a labeled datset $\mathcal{X}_0$
Set an unlabeled data by $\mathcal{U}_0 = \mathcal{D}/\mathcal{X}_0$
Fine-tuning: $\theta_0 = \arg\min_{\theta \in \Theta} L(\mathcal{X}_0, \theta) + \Omega(\mathcal{D}, \theta)$
while $t < T$ **do**
  for $s = 0, \ldots, S-1$ **do**
    Patch prediction on $\mathcal{U}_t$
    Generate $\hat{M}_s$ with the i-th entry being $\max\{h(x_i,\theta)_1, h(x_i,\theta)_2, h(x_i,\theta)_3\}$
    Set those entries in M_s corresponding to regions from $\mathcal{X}_t$ as infinity
    $\hat{M}_s = M_s * H$
    Select the regions $\{\mathcal{R}_i\}_{i=1}^{m}$ corresponding to the m lowest value of entries in $\hat{M}_s$
    Split $\{\mathcal{R}_i\}_{i=1}^{r}$ into a group of patches $\mathcal{W}_s$
    $\mathcal{X}_t = \mathcal{X}_t \cup (\mathcal{W}_s, J(\mathcal{W}_s))$
    $\mathcal{U}_t = \mathcal{U}_t / \mathcal{W}_s$
  end for
  Update $\theta : \theta_{t+1} = \arg\min_{\theta \in \Theta} L(\mathcal{X}_t, \theta) + \Omega(\mathcal{D}, \theta)$
  $t = t + 1$
end while

4. Experiments

4.1. Dataset description

This work uses pathology images generated from 5 μm formalin fixed paraffin embedded sections of Superior Middle Temporal Gyri in the temporal cortex. The tissues in these slides were stained with an Amyloid-β antibody 17-24 4G8 with a dilution of 1:1600, from the BioLegend provider (catalog number SIG-39200). This type of staining is routinely used for detection of Amyloid-β [37]. All WSIs were scanned at $20\times$ magnification and digitized by an Aperio AT2 machine, which outputs each WSI scan as a SVS file. Each SVS file has an average resolution of 60,000 $\times$ 50,000 pixels. The area of tissue extracted is in the order of magnitude of one squared inch, which represents ultra-high resolution.

The cohort of this study comprised of 18 cases of deceased patients with pathological diagnosis of Alzheimer's disease (referred as AD cases). The classification of cases as AD followed the method proposed by the Consortium to Establish a Registry for Alzheimer's Disease (CERAD) [35]. The average age at death +/- standard deviation of the AD cases was 84 ± 7 years with 10 males and 8 females. There were also 12 slides that clinicopathological diagnoses of Alzheimer's disease (referred as NAD cases). For racial ethnicity composition, the cohort consisted of: 22 non-Hispanic White (73%), 5 African Americans (17%) and 3 Hispanics (10%). To fully protect data confidentiality, we assign the tag WSI-1 to WSI-18 to AD cases, and WSI-19 to WSI-30 for NAD cases.

As downsampling [10] would lose minute pathological features, we follow a previously published patch based method [32] to tile WSIs into 256×256 patches. 20 WSIs (12 AD cases and 8 NAD cases) are selected for the training process inclusive of the training and validation set. The validation set includes 600 patches from 2 WSIs among these 20 WSIs. The remaining 10 WSIs (6 AD cases and 4 NAD cases) are kept as the **hold-out test** set.

4.2. Training setup

We used the ResNet-18 [21] as the encoder for all experiments. We tuned the following hyper-parameters based on the validation set in the training process. The batch size is 32 for both labeled and unlabeled data. For the training optimizer, we use Adam with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ in our experiments. The learning rate is 0.001. All computations were run on an Intel Xeon W-2102 CPU (4 cores / 4 threads) and a NVIDIA Titan Xp GPU with 12 GB of VRAM. Other hyper-parameters used in FixMatch are summarized in Table 1.

Table 1. FixMatch settings.

Hyper-parameter	Value
Confidence threshold τ	0.95
Unlabeled loss weight λ_u	1
Ratio of unlabeled data in each mini-batch μ	2

4.3. Quantitative comparison

Since our target task is GM/WM segmentation, we choose two standard segmentation metrics — IoU score [41] and DICE coefficient [12]. The results in this and following subsections are from our **hold-out test** set.

Table 2. Pixel-wise IoU Scores for AD, NAD, and overall test set

Method	FCN [47]		U-Net [39]		FixMatch [47]	**Proposed AL-600/AL-400**	
Labeled data	**2** WSIs	**All** WSIs	**2** WSIs	**All** WSIs	**0.1%**	**0.1%**	**0.07%**
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	**96.40** ± 0.95
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	**89.71** ± 3.34
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	**82.19** ± 3.77
AD Mean	45.26 ± 3.55	73.14 ± 9.66	35.40 ± 7.12	**89.64** ± 4.35	76.13 ± 5.89	88.55 ± 3.27	**89.44** ± 2.50
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** ± 1.05
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	92.25 ± 2.36
NAD WM	19.72 ± 13.6	49.89 ± 12.80	3.02 ± 3.09	**81.25** ± 9.53	22.72 ± 19.0	**75.85** ± 11.37	71.90 ± 11.16
NAD Mean	45.51 ± 3.29	72.56 ± 3.97	47.02 ± 10.9	**91.01** ± 3.36	67.92 ± 6.53	**88.86** ± 3.38	87.16 ± 2.86
Test Back	63.29 ± 5.81	84.05 ± 9.17	68.28 ± 17.2	**97.02** ± 2.15	94.72 ± 2.71	95.91 ± 1.48	**96.77** ± 1.05
Test 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	**90.73** ± 3.13
Test WM	24.54 ± 9.80	57.29 ± 14.30	5.75 ± 5.37	**82.02** ± 7.98	43.08 ± 24.0	**79.44** ± 8.34	78.07 ± 8.81
Test 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	88.53 ± 2.75

Rows marked AD contain results on the 6 Alzheimer's disease cases in *hold-out* test set. Rows marked NAD contain results on the 4 non-Alzheimer's disease cases in test set. Rows marked Test contain results on all 10 WSIs. 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.

IoU score. IoU score measures the amount of overlap present in two different measurements of area generated as masks of the WSIs. Although IoU is the ideal metric for evaluating the success of the segmentation, average IoUs cannot account for the variability that occur between different methods. This variability is important to account for as the different Temporal Gyri WSIs contain noticeable differences due to the heterogeneous nature of the human brain. Hence, we choose to also report standard deviation (STD) in Table 2 to measure the stability and consistency of different methods across multiple WSIs in the hold-out test set. We generate the masks of GM, WM, and background per WSI from the trained model and overlap them on pixel-wise ground truth masks. The comparison of IoU scores and STD are summarized in Table 2.

As shown in Table 2, if FCN [4] and U-Net [39] are trained on *all* WSIs with annotations in the training set via fully-supervised learning, they are able to achieve the mean IoU score at nearly 72.91% and 90.19%, respectively. 90.19% would be approximately regarded as an upper bound performance of supervised learning (SL) in this dataset. However, if they are only trained on two annotated WSIs (one AD and one NAD) with SL, the performance has a severe degradation.

After we obtain the benchmark of SL algorithms, we first deploy the current state-of-the-art SSL algorithm, FixMatch [47], on our dataset with the same two WSIs annotated in pixel level. As it shows its advantages on very scarce limited data, e.g. 40 labeled in CIFAR-10, equivalent to nearly 0.1% of all samples in CIFAR-10 [47], we follow their way to set 0.1% of labeled data as the objective. In our case, 0.1% area of 20 WSIs is about 600 patches, equivalent to 24 regions at 1280×1280 pixels. We first adopt a *randomly uniform* selection to select 600 patches from two WSIs (300 patches per WSI, and 100 patches per class). The remaining patches and 18 slides keep as the unlabeled set. It achieves 72.84% of mean IoU score on the hold-out test set, which is superior to FCN/U-Net's performance when trained only on two WSIs (limited labeled data). However, its performance on the WM of NAD cases remains limited and far from SL's.

Then we evaluate our proposed framework with region-based selection cycles. To obtain 0.1% area (or 24 regions tiled into 600 patches), we use AL-400 to refer to our proposed method that selects 16 regions at 1280×1280 pixels in two rounds to generate 400 patches in total for preparing the labeled set of FixMatch [47]. While AL-600 refers to selection of 24 regions in three rounds to generate 600 patches in total for labeling. AL-400 is able to reach almost 89.44% of mean IoU in AD cases, which is very close to the upper

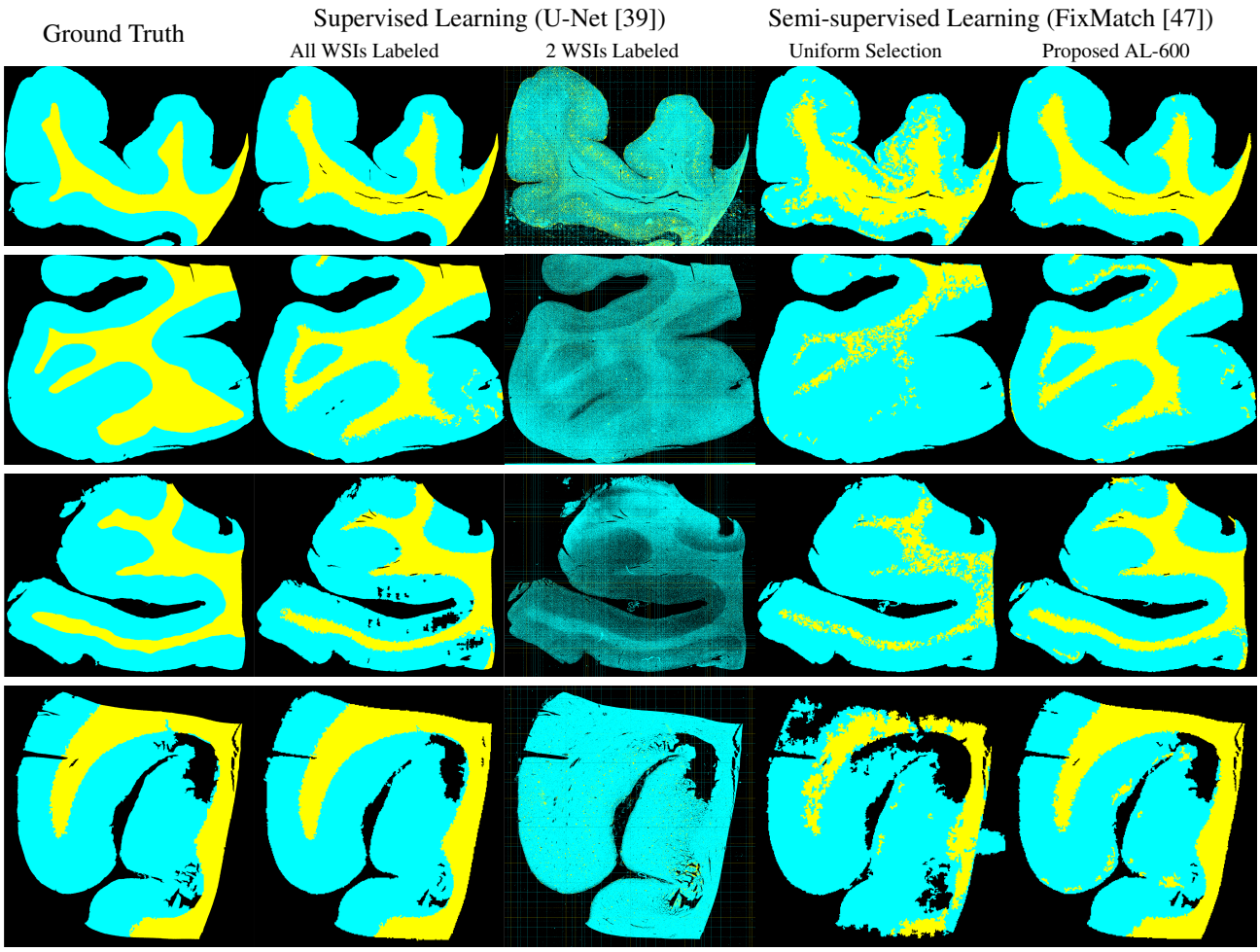


Figure 5. Segmentation masks visualization: GM, WM, and background are indicated by cyan, yellow, and black, respectively. The top two rows are AD cases and the bottom two rows are NAD cases. Both SSL results are only using 0.1% area of 20 WSIs in the training set.

bound of SL when trained on all labeled WSI data. However, there is an notable gap between its IoU score on NAD cases compared to U-Net, especially in WM where the gap is almost 10%. After an addition round for region selection and annotation query, AL-600 improves its performance on NAD cases with nearly 4% of increase on the IoU score of WM. One noticeable issue is that the STD values of WM on NAD cases are relatively larger than the values on other regions, which indicates that the model still has difficulty in predicting WM in NAD slides. This may be caused by the imbalance between WM and GM in NAD cases: the size of WM regions is relatively smaller than that of GM regions. Overall, both AL-400 and AL-600 have potential to get closer to the bound of SL.

DICE coefficient. Besides the IoU score, we also use DICE coefficient [53] to further compare our proposed approach with baseline FixMatch [47]. DICE is one of the most common methods of evaluating image segmentation success in medical imaging [12]. The value is calculated as shown in (3): double the overlapping areas and divide it by the total number of pixels in all areas. As shown in Table 3, the proposed methods outperform significantly on WM regions on DICE coefficient and also slightly outperform on GM regions compared to the baseline FixMatch [47] with random selection for labeling query.

$$\text{DICE} = \frac{2 \times \text{Area of Overlap}}{\text{Total Sum of Pixels in All Areas}} \tag{3}$$

Table 3. DICE coefficient comparison.

Region	FixMatch	AL-400	AL-600
GM	91.02	95.14	95.19
WM	63.19	87.88	88.53

4.4. Segmentation visualization

Figure 5 display the ground truth masks and the predictive masks using different methods. The top two rows are

the masks for AD cases while the bottom two rows are for NAD cases from the hold-out test set. The masks of U-Net(in the second column) indicate that SL is able to get well-defined performance as long as the variety and volume of labeled data are sufficiently large, but the performance could be degraded severely if the variety and volume of labeled data are limited (as shown in the third column). Baseline FixMatch with random selection for the labeled set (in the fourth column) is able to locate the rough boundaries between GM and WM but the shape of WM is far wary from the ground truth masks. There are also considerable amounts of noisy pixels within WM, which means it wrongly classifies some WM regions as GM. Our proposed method with AL selection (in the fifth column) is able to provide more distinguishable boundaries for GM and WM regions. The generated masks are also the closest to our ground truth masks. However, for the second row, the masks of both FixMatch and ours are different from the ground truth masks, which indicates that SSL still requires more labeled data for in order to fully capture the representations of gigapixel histology images.

4.5. A deeper look into AL cycles

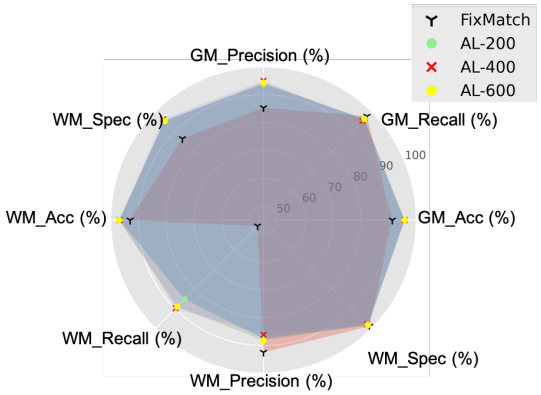


Figure 6. Pixel-wise classification comparison via each AL cycle.

To further analyze the effect of AL cycles on FixMatch [47], the differences of IoU scores in each round are trivial. Therefore, we also select Accuracy, Recall (Sensitivity), Specificity, and Precision to have a deeper look into the AL's effectiveness. The results for each AL cycle on the hold-out test set are summarized in Figure 6. Our proposed framework shows higher precision and specificity in GM regions compared to original FixMatch [47]. It is obvious that WM's recall of FixMatch [47] is the lowest score among all indices. This indicates our proposed framework is able to improve the performance on WM regions.

When we take a deeper look into the performance via AL cycles, from AL-200, AL-400, to AL-600, the model improves its performance on WM regions, such as recall. This fits our design objective: our framework always queries the annotations on difficult regions.

4.6. Analysis on number of selected regions

As stated in Algorithm 1, m is the number of regions selected for annotation query in one round of AL. In Section 4.3 and 4.4, we use three rounds to obtain 24 labeled regions, which are tiled into 600 patches in total. To further reduce the number of cycles, we also try to select the same number of regions based on f in one round. We achieve nearly 87% of mean IoU score, which is only around 2% lower than AL-600 using three rounds of labeling query. This shows a promising direction on further studying the number rounds of labeling query that is required in the histology images.

5. Discussion

In this work, we investigate the applicability of supervised and semi-supervised learning state-of-the-art algorithms in histology images with limited labeled data. We propose a semi-supervised active learning framework with a region-based selection criteria to effectively select regions for labeling query. For our specific dataset, the region-based criterion is more robust than patch-based criterion in seeking the uncertainty regions for labeling and is able to quickly expand the diversity of the labeled set. We evaluate our framework on Amyloid-β stained neuropathology images: our proposed algorithm outperforms the state-of-the-art models with more than 10% of IoU score and DICE coefficient, especially for the WM.

As part of our future work, we plan to test our framework on external datasets, as well as other ultra-resolution histology problems with scarce labeled data and comprehensively measure how much manual annotation efforts can be saved. Since the images used in our study are from a single brain area with a single staining method, our main limitation is the lack of diversity in the data we test. For example, it would be beneficial to the study to expand its evaluation to other differently stained, distinct neuroanatomical areas that will generate significant visual difference from our tested data. If proven successful, the study could be further extended to other gigapixel histology datasets, such as problems dealing with large amounts of annotated WSIs from cancerous biopsies [26] and plaque detection [48].

ACKNOWLEDGMENT

This work was supported by the NSF HDR:TRIPODS grant CCF-1934568. It was also supported by the California Department of Public Health Alzheimer's Disease Program. Funding is provided by the 2019 California Budget Act. 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 the commitments of faculty and staff of the UCD-ADRC.

References

[1] Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2020.

[2] Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. In *International Conference on Learning Representations*, 2019.

[3] Maria-Florina Balcan, Alina Beygelzimer, and John Langford. Agnostic active learning. *Journal of Computer and System Sciences*, 75(1):78–89, 2009.

[4] Péter Bándi, Rob van de Loo, Milad Intezar, Daan Geijs, Francesco Ciompi, Bram van Ginneken, Jeroen van der Laak, and Geert Litjens. Comparison of different methods for tissue segmentation in histopathological whole-slide images. In *IEEE ISBI 2017*, pages 591–595.

[5] David Berthelot, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. In *International Conference on Learning Representations (ICLR)*, 2020.

[6] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. *arXiv:1905.02249*, 2019.

[7] Samuel Budd, Emma C Robinson, and Bernhard Kainz. A survey on active learning and human-in-the-loop deep learning for medical image analysis. *Medical Image Analysis*, page 102062, 2021.

[8] Lyndon Chan, Mahdi S Hosseini, and Konstantinos N Plataniotis. A comprehensive analysis of weakly-supervised semantic segmentation in different image domains. *International Journal of Computer Vision*, 129(2):361–384, 2021.

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607. PMLR, 2020.

[10] Wuyang Chen, Ziyu Jiang, Zhangyang Wang, Kexin Cui, and Xiaoning Qian. Collaborative global-local networks for memory-efficient segmentation of ultra-high resolution images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8924–8933, 2019.

[11] Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation, 2021.

[12] William R Crum, Oscar Camara, and Derek LG Hill. Generalized overlap measures for evaluation and validation in medical image analysis. *IEEE transactions on medical imaging*, 25(11):1451–1461, 2006.

[13] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 702–703, 2020.

[14] Dennis W Dickson. The pathogenesis of senile plaques. *Journal of Neuropathology & Experimental Neurology*, 56(4):321–339, 1997.

[15] Neofytos Dimitriou, Ognjen Arandjelović, and Peter D Caie. Deep learning for whole slide image analysis: an overview. *Frontiers in medicine*, 6:264, 2019.

[16] Thomas Drugman, Janne Pylkkonen, and Reinhard Kneser. Active and semi-supervised learning in asr: Benefits on the acoustic and language models. *arXiv preprint arXiv:1903.02852*, 2019.

[17] Brittany N Dugger and Dennis W Dickson. Pathology of neurodegenerative diseases. *Cold Spring Harbor perspectives in biology*, 9(7):a028035, 2017.

[18] Shai Fine, Ran Gilad-Bachrach, and Eli Shamir. Query by committee, linear separation and random walks. *Theoretical Computer Science*, 284(1):25–51, 2002.

[19] Mingfei Gao, Zizhao Zhang, Guo Yu, Sercan Ö Arık, Larry S Davis, and Tomas Pfister. Consistency-based semi-supervised active learning: Towards minimizing labeling cost. In *European Conference on Computer Vision*, pages 510–526. Springer, 2020.

[20] Thore Graepel and Ralf Herbrich. The kernel gibbs sampler. *Advances in Neural Information Processing Systems*, 13:514–520, 2000.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[22] Steven CH Hoi, Rong Jin, Jianke Zhu, and Michael R Lyu. Batch mode active learning and its application to medical image classification. In *Proceedings of the 23rd international conference on Machine learning*, pages 417–424, 2006.

[23] Neil Houlsby, José Miguel Hernández-Lobato, and Zoubin Ghahramani. Cold-start active learning with robust ordinal matrix factorization. In *International Conference on Machine Learning*, pages 766–774. PMLR, 2014.

[24] Jisoo Jeong, Vikas Verma, Minsung Hyun, Juho Kannala, and Nojun Kwak. Interpolation-based semi-supervised learning for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11602–11611, 2021.

[25] Ajay J Joshi, Fatih Porikli, and Nikolaos Papanikolopoulos. Multi-class active learning for image classification. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2372–2379. IEEE, 2009.

[26] Mahendra Khened, Avinash Kori, Haran Rajkumar, Ganapathy Krishnamurthi, and Balaji Srinivasan. A generalized deep learning framework for whole-slide image segmentation and analysis. *Scientific reports*, 11(1):1–14, 2021.

[27] Krishnateja Killamsetty, Durga Sivasubramanian, Ganesh Ramakrishnan, and Rishabh Iyer. Glister: Generalization based data subset selection for efficient and robust learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8110–8118, 2021.

[28] Andreas Kirsch, Joost Van Amersfoort, and Yarin Gal. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. *Advances in neural information processing systems*, 32:7026–7037, 2019.

[29] Ksenia Konyushkova, Raphael Sznitman, and Pascal Fua. Learning active learning from data. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[30] Chia-Wen Kuo, Chih-Yao Ma, Jia-Bin Huang, and Zsolt Kira. Featmatch: Feature-based augmentation for semi-supervised learning. In *European Conference on Computer Vision*, pages 479–495. Springer, 2020.

[31] Zhengfeng Lai, Runlin Guo, Wenda Xu, Zin Hu, Kelsey Mifflin, Charles DeCarli, Brittany N Dugger, Sen-ching Cheung, and Chen-Nee Chuah. Automated segmentation of amyloid-β stained whole slide images of brain tissue. *bioRxiv 2020.11.13.381871*, 2020.

[32] Zhengfeng Lai, Runlin Guo, Wenda Xu, Zin Hu, Kelsey Mifflin, Brittany N Dugger, Chen-Nee Chuah, and Sen-ching Cheung. Automated grey and white matter segmentation in digitized aβ human brain tissue slide images. In *2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 1–6. IEEE, 2020.

[33] Zhengfeng Lai, Chao Wang, Zin Hu, Brittany N Dugger, Sen-Ching Cheung, and Chen-Nee Chuah. A semi-supervised learning for segmentation of gigapixel histopathology images from brain tissues. In *2021 43nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2021.

[34] Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning (ICML)*, 2013.

[35] Suzanne S Mirra, A Heyman, D McKeel, SM Sumi, Barbara J Crain, LM Brownlee, FS Vogel, JP Hughes, G Van Belle, Leal Berg, et al. The consortium to establish a registry for alzheimer's disease (cerad): Part ii. standardization of the neuropathologic assessment of alzheimer's disease. *Neurology*, 41(4):479–479, 1991.

[36] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993, 2018.

[37] Thomas J Montine, Creighton H Phelps, Thomas G Beach, Eileen H Bigio, Nigel J Cairns, Dennis W Dickson, Charles Duyckaerts, Matthew P Frosch, Eliezer Masliah, Suzanne S Mirra, et al. National institute on aging-alzheimer's association guidelines for the neuropathologic assessment of alzheimer's disease: A practical approach. *Acta Neuropathologica*, 123(1):1–11, 2012.

[38] Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. *Advances in Neural Information Processing Systems*, 31:3235–3246, 2018.

[39] Kay Oskal, Martin Risdal, Emiel Janssen, Erling Undersrud, and Thor Gulsrud. A U-net based approach to epidermal tissue segmentation in whole slide histopathological images. *SN Appl. Sci.*, 1:672, 06 2019.

[40] J Vince Pulido, Shan Guleria, Lubaina Ehsan, Matthew Fasullo, Robert Lippman, Pritesh Mutha, Tilak Shah, Sana Syed, and Donald E Brown. Semi-supervised classification of noisy, gigapixel histology images. In *2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE)*, pages 563–568. IEEE, 2020.

[41] Md Atiqur Rahman and Yang Wang. Optimizing intersection-over-union in deep neural networks for image segmentation. In *International symposium on visual computing*, pages 234–244. Springer, 2016.

[42] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Xiaojiang Chen, and Xin Wang. A survey of deep active learning. *arXiv preprint arXiv:2009.00236*, 2020.

[43] Phill Kyu Rhee, Enkhbayar Erdenee, Shin Dong Kyun, Minhaz Uddin Ahmed, and Songguo Jin. Active and semi-supervised learning for object detection with imperfect data. *Cognitive Systems Research*, 45:109–123, 2017.

[44] Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. *arXiv:2101.06329*, 2021.

[45] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018.

[46] Alberto Serrano-Pozo, Matthew P Frosch, Eliezer Masliah, and Bradley T Hyman. Neuropathological alterations in alzheimer disease. *Cold Spring Harbor perspectives in medicine*, 1(1):a006189, 2011.

[47] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems*, volume 33, pages 596–608, 2020.

[48] Ziqi Tang, Kangway V Chuang, Charles DeCarli, Lee-Way Jin, Laurel Beckett, Michael J Keiser, and Brittany N Dugger. Interpretable classification of alzheimer's disease pathologies with a convolutional neural network pipeline. *Nat. Commun.*, 10(1):1–14, 2019.

[49] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[50] Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. *Journal of machine learning research*, 2(Nov):45–66, 2001.

[51] Jesper E Van Engelen and Holger H Hoos. A survey on semi-supervised learning. *Machine Learning*, 109(2):373–440, 2020.

[52] Kai Wei, Rishabh Iyer, and Jeff Bilmes. Submodularity in data subset selection and active learning. In *International Conference on Machine Learning*, pages 1954–1963. PMLR, 2015.

[53] Kelly H Zou, Simon K Warfield, Aditya Bharatha, Clare MC Tempany, Michael R Kaus, Steven J Haker, William M Wells III, Ferenc A Jolesz, and Ron Kikinis. Statistical validation of image segmentation quality based on a spatial overlap index1: scientific reports. *Academic radiology*, 11(2):178–189, 2004.